Nonparametric Functional Estimation and Related Topics

Edited by

George Roussas

NATO ASI Series

Series C: Mathematical and Physical Sciences - Vol. 335

Nonparametric Functional Estimation and Related Topics

NATO ASI Series

Advanced Science Institutes Series

A Series presenting the results of activities sponsored by the NATO Science Committee, which aims at the dissemination of advanced scientific and technological knowledge, with a view to strengthening links between scientific communities.

The Series is published by an international board of publishers in conjunction with the NATO Scientific Affairs Division

A Life Sciences **B Physics**	Plenum Publishing Corporation London and New York
C Mathematical and Physical Sciences **D Behavioural and Social Sciences** **E Applied Sciences**	Kluwer Academic Publishers Dordrecht, Boston and London
F Computer and Systems Sciences **G Ecological Sciences** **H Cell Biology** **I Global Environmental Change**	Springer-Verlag Berlin, Heidelberg, New York, London, Paris and Tokyo

NATO-PCO-DATA BASE

The electronic index to the NATO ASI Series provides full bibliographical references (with keywords and/or abstracts) to more than 30000 contributions from international scientists published in all sections of the NATO ASI Series.
Access to the NATO-PCO-DATA BASE is possible in two ways:

– via online FILE 128 (NATO-PCO-DATA BASE) hosted by ESRIN, Via Galileo Galilei, I-00044 Frascati, Italy.

– via CD-ROM "NATO-PCO-DATA BASE" with user-friendly retrieval software in English, French and German (© WTV GmbH and DATAWARE Technologies Inc. 1989).

The CD-ROM can be ordered through any member of the Board of Publishers or through NATO-PCO, Overijse, Belgium.

Series C: Mathematical and Physical Sciences - Vol. 335

Nonparametric Functional Estimation and Related Topics

edited by

George Roussas
Intercollege Division of Statistics,
University of California,
Davis, California, U.S.A.

Springer Science+Business Media, B.V.

Proceedings of the NATO Advanced Study Institute on
Nonparametric Functional Estimation and Related Topics
Spetses, Greece
July 29–August 10, 1990

ISBN 978-94-010-5420-1 ISBN 978-94-011-3222-0 (eBook)
DOI 10.1007/978-94-011-3222-0

Printed on acid-free paper

All Rights Reserved
© 1991 Springer Science+Business Media Dordrecht
Originally published by Kluwer Academic Publishers in 1991
Softcover reprint of the hardcover 1st edition 1991

No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without written permission from the copyright owner.

Τῇ Ἱερᾷ Σκιᾷ Τῆς Μητρός Μου

TABLE OF CONTENTS

Preface xi

I. Curve and Functional Estimation

A. BERLINET:
Reproducing Kernels and Finite Order Kernels 3
P. DEHEUVELS:
Laws of the Iterated Logaritm for Density Estimators 19
L. DEVROYE:
Exponential Inequalities in Nonparametric Estimation 31
R. FRAIMAN AND G. PÉREZ IRIBARREN:
Conservative Confidence Bands for Nonparametric Regression 45
T. GASSER AND E. HERRMANN:
Data-Adaptive Kernel Estimation 67
L. GYÖRFI AND E.C. VAN DER MEULEN:
On the Nonparametric Estimation of the Entropy Functional 81

II. Curve and Functional Estimation (Continued)

T. GASSER AND A. KNEIP:
Analysis of Samples of Curves 99
W. HÄRDLE AND E. MAMMEN:
Bootstrap Methods in Nonparametric Regression 111
V.K. KLONIAS:
On the Influence Function of Maximum Penalized Likelihood Density Estimators 125
E. MAMMEN:
Nonparametric Curve Estimation and Simple Curve Characteristics 133
H.-G. MÜLLER AND K. PREWITT:
Applications of Multiparameter Weak Convergence for Adaptive Nonparametric Curve Estimation 141
A. SAMAROV:
On Asymptotic Efficiency of Average Derivative Estimates 167
W. STUTE AND U. WERNER:
Nonparametric Estimation of Elliptically Contoured Densities 173
A.J. VAN ES:
Uniform Deconvolution: Nonparametric Maximum Likelihood and Inverse Estimation 191

III. Parameter Selection, Smoothing

K.P.-S. CHAN AND J.W. KAY:
Smoothing Parameter Selection in Image Restoration 201
M. CSÖRGŐ, P. DEHEUVELS AND L. HORVÁTH:
Estimating the Quantile-Density Function 213
A. CUEVAS AND W. GONZALEZ-MANTEIGA:
Data-Driven Smoothing Based on Convexity Properties 225
M.C. JONES:
Prospects for Automatic Bandwidth Selection in Extensions to Basic Kernel Density Estimation 241
J.S. MARRON:
Root n Bandwidth Selection 251
P. SARDA:
Estimating Smooth Distribution Functions 261
P. VIEU:
Smoothing Techniques in Time Series Analysis 271

IV. Regression Models

M.G. AKRITAS:
Nonparametric Inference in Heteroskedastic Regression 287
M. A. DELGADO:
Bounded Influence Regression in the Presence of Heteroskedasticity of Unknown Form 297
W. GONZALEZ-MANTEIGA AND C.M. CADARSO-SUAREZ:
Linear Regression with Randomly Right-Censored Data Using Prior Nonparametric Estimation 315
L. GYÖRFI:
Universal Consistencies of a Regression Estimate for Unbounded Regression Functions 329
N.E. HECKMAN:
Minimax Bayes Estimation, Penalized Likelihood Methods, and Restricted Minimax Estimation 339
A. KRZYŻAK:
On Exponential Bounds on the Bayes Risk of the Nonparametric Classification Rules 347
Y.P. MACK AND N.J. WULWICK:
Nonparametric Regression Analysis of Some Economic Data 361

V. Dependent Data

A. AZZALINI AND A. BOWMAN:
Nonparametric Regression Methods for Repeated Measurements 377
D. BOSQ:
Nonparametric Prediction for Unbounded Almost Stationary Processes 389
K. HELLAND, K.S. LII AND M. ROSENBLATT:
Monte Carlo and Turbulence 405
T.D. PHAM AND L.T. TRAN:
Kernel Density Estimation Under a Locally Mixing Condition 419
M.B. RAO, G.J. BABU AND C.R. RAO:
Nonparametric Estimation of Survival Functions Under Dependent Competing Risks 431
G.G. ROUSSAS:
Estimation of Transition Distribution Function and Its Quantiles in Markov Processes: Strong Consistency and Asymptotic Normality 443
G.B. SILVEIRA:
L_1 Strong Consistency for Density Estimates in Dependent Samples 463

VI. Time Series Analysis, Signal Detection

R. AHMAD:
Nonparametric Statistical Signal Detection Problems 477
B. AUESTAD AND D. TJØSTHEIM:
Functional Identification in Nonlinear Time Series 493
D. BOSQ:
Modelization, Nonparametric Estimation and Prediction for Continuous Time Processes 509
A. A. GEORGIEV:
Estimation of Chaotic Dynamic Systems with Control Variables 531
M. PAWLAK AND W. GREBLICKI:
Nonparametric Estimation of a Class of Nonlinear Time Series Models 541
J.M. STOYANOV AND P.M. ROBINSON:
Semiparametric and Nonparametric Inference from Irregular Observations on Continuous Time Stochastic Processes 553

VII. Various Topics

A.R. BARRON:
Complexity Regularization with Application to Artificial Neural Networks 561
R. BERAN:
Designing Prediction Bands 577
P.K. BHATTACHARYA:
Analysis of Observational Studies from the Point of View of Nonparametric Regression 587
P. BURMAN AND D. NOLAN:
Some Issues in Cross-Validation 603
J. FAN, Y.K. TRUONG AND Y. WANG:
Nonparametric Function Estimation Involving Errors-in-Variables 613

VIII. Various Topics (Continued)

L. GYÖRFI AND E.C. VAN DER MEULEN:
A Consistent Goodness of Fit Test Based on the Total Variation Distance 631
S.R. JAMMALAMADAKA AND X. WAN:
On a Problem in Semiparametric Estimation 647
B.Ya. LEVIT:
On the Integrable and Approximately Integrable Linear Statistical Models 657
A.B. TSYBAKOV:
Nonparametric Techniques in Image Estimation 669
A.C.M. VAN ROOIJ AND F.H. RUYMGAART:
Regularized Deconvolution on the Circle and the Sphere 679

List of Attendants 691

Contributed Papers 699

Index 701

PREFACE

About three years ago, an idea was discussed among some colleagues in the Division of Statistics at the University of California, Davis, as to the possibility of holding an international conference, focusing exclusively on nonparametric curve estimation. The fruition of this idea came about with the enthusiastic support of this project by Luc Devroye of McGill University, Canada, and Peter Robinson of the London School of Economics, UK. The response of colleagues, contacted to ascertain interest in participation in such a conference, was gratifying and made the effort involved worthwhile. Devroye and Robinson, together with this editor and George Metakides of the University of Patras, Greece and of the European Economic Communities, Brussels, formed the International Organizing Committee for a two week long Advanced Study Institute (ASI) sponsored by the Scientific Affairs Division of the North Atlantic Treaty Organization (NATO). The ASI was held on the Greek Island of Spetses between July 29 and August 10, 1990.

Nonparametric functional estimation is a central topic in statistics, with applications in numerous substantive fields in mathematics, natural and social sciences, engineering and medicine. While there has been interest in nonparametric functional estimation for many years, this has grown of late, owing to increasing availability of large data sets and the ability to process them by means of improved computing facilities, along with the ability to display the results by means of sophisticated graphical procedures. Nonparametric curve estimation has become a subfield of statistics in its own right, and has attracted the interest of many distinguished scholars in probability, theoretical and applied statistics, economics, several areas of engineering, medicine and other fields. Probabilists are attracted by nonparametric curve estimation because of its relation to empirical processes. For theoretical statisticians, the nonparametric approach poses many fascinating and challenging intellectual problems. For applied statisticians, it provides a powerful tool for exploratory data analysis. It is used by control engineers for systems identification, and by modern electrical engineers for pattern recognition. It is used, in information theory, for establishing lower bounds on the performance of estimates and for understanding the limitations of our capabilities. In medicine, it is used for diagnostic purposes, among other things, by way of discriminatory information extracted from huge data bases.

The main objective of the ASI was to provide an opportunity for an extended interaction among research workers from diverse fields, with a view toward bringing the participants up to date on past and current research activity, and to stimulate new research and applications.

The ASI attracted 84 participants, consisting of eminent statisticians, established scientists and younger researchers. Most of the participants came from 13 NATO countries and a few came from 3 non-NATO countries. The scientific activities were organized around a series of two hour lectures delivered by prominent experts. All of the remaining participants had the opportunity to present formal lectures of varying duration. The lectures were presented in a "tutorial" style as advised by the NATO Scientific Affairs Division. Modified versions of 51 of the presented lectures are assembled here and are grouped in six distinct subtitles labeled as: *Curve and Functional Estimation; Parameter Selection, Smoothing; Regression Models; Dependent Data; Time Series Analysis, Signal Detection; and Various Topics.*

Chapters I and II consist of 14 papers in the general area of curve and functional estimation. Chapter III comprises 7 papers dealing with the problem of parameter selection and smoothing procedures. Another 7 papers on regression are grouped together in Chapter IV. Dealing nonparametrically with dependent data has received extensive attention the last few years, and 7 papers on this subject constitute Chapter V. Time series analysis is another area where the application of nonparametric methods has found fertile ground. Various aspects of nonparametric time series analysis methodology are discussed in the 6 papers which constitute Chapter VI. Finally, 10 papers are divided equally into two final chapters, Chapters VII and VIII, under the heading of various topics. These papers deal with a variety of important issues of theoretical and applied interest within the framework on nonparametric functional estimation. Their contents range from complexity regularization to designing prediction bands, to analysis of observational studies, to cross-validation, to nonparametric function estimation involving error-in-variables. Other topics include goodness-of-fit tests, semiparametric estimation, integrable linear statistical models, nonparametric techniques in image estimation, and regularized deconvolution. It is hoped that this volume will be a source of useful information, and will be accessible to a wide spectrum of readers.

The ASI was held at the Anargyreios and Korgialeneios School of Spetses where space was rented to us by the Greek Government. The NATO Scientific Affairs Division was the sole source of financial support for the conference. To this organization, we owe a debt of gratitude for the generous financial support, the explicit and helpful guidelines, the materials we were provided with, and the general assistance and encouragement. All of the above led to a successful meeting for which we wish to register here our deep appreciation. The National Science Foundation of the USA provided the travel expenses for six graduate student participants. On their behalf, our appreciation is expressed herewith. The local arrangements committee, consisting of D. Ioannides, G. Kokolakis, and E. Xekalaki (Chair), helped to ameliorate some local organizational difficulties and deserves our thanks. Mr. P. Paraskevas, Manager of the facilities, helped effectivily with the accommodations and is to be commended. Personal thanks are due to Luc Devroye for his generous advice regarding organizational and editorial work, and to a host of anonymous referees for their utmost support and prompt response to refereeing requests. Finally, my appreciation goes to the staff of the Division of Statistics at the University of California, Davis, for their substantial time and effort, and to Mrs. Nel de Boer, of Kluwer Academic Publishers, whose expert advice provided considerable help in the preparation of this volume.

Davis, California
January 1991

G.G.R.

I. Curve and Functional Estimation

REPRODUCING KERNELS AND FINITE ORDER KERNELS

A. BERLINET
Unité de Biométrie
ENSA.M, INRA, Montpellier II
9, place Pierre Viala
34 060 Montpellier Cedex 1
France

ABSTRACT. Recent literature on density and regression function estimation has shown the interest of kernels of order s, i.e. kernels with (s-1) vanishing moments. We define here a multi-stage procedure to build estimates based on increasing order kernels and leading to a data-driven choice of both the order and the smoothing parameter. Some asymptotic as well as non-asymptotic results are provided for the estimates.

1. Introduction

Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of real-valued independent random variables with common density f and cumulative distribution function F. Let f_n (respectively, F_n) be an estimate of f (respectively, F) built from $X_1,\ldots, X_n$ and to be specified later on, and let $\mathbb{F}_n(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{]-\infty, t]}(X_j)$ be the empirical distribution function associated with the sample.

Following Devroye (1987), let us define $D_r^*(f) = \liminf_{h\downarrow 0} \int |(f * M_h)^{(r)}|$ where $M_h = \frac{1}{h} M\left(\frac{\cdot}{h}\right)$ and M is a mollifier; that is a positive symmetric function integrating to one with compact support [-1, 1] and infinitely many continuous derivatives. Let us suppose that f_n is a standard Parzen-Rosenblatt kernel estimate built from the sample; that is

$$f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

where $(h_n)_{n\in\mathbb{N}}$ is a fixed sequence of positive real numbers tending to zero and K is a measurable function integrating to one. Let L be the Bretagnolle-Huber kernel associated with K; that is

$$\forall x > 0 \qquad L(x) = (-1)^r \int_x^{+\infty} \frac{(y-x)^{r-1}}{(r-1)!} K(y)\, dy\,,\ L(-x) = (-1)^r L(x).$$

Then we have (Devroye, 1987) :

$$\int |Ef_n - f| \le h_n^r \int |L|\ D_r^*(f) \tag{1,1}$$

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 3–18.
© 1991 *Kluwer Academic Publishers.*

provided K is of order at least (r+1) (i.e. K has at least (r+1) finite moments and the first r moments vanish). The same bound is valid for $\sup_{t \in \mathbb{R}} | EF_n(t) - F(t) |$ with $F_n(t) = \int_{-\infty}^{t} f_n(x)\, dx$. This result has been first stated for densities with (r-1) absolutely continuous derivatives. For these densities $D_r^*(f) = \int | f^{(r)} |$ and the proof of the property is based on a Taylor series expansion of f. The introduction of $D_r^*(f)$ and the inequality (1,1) clarified the fact that the interest of kernels of order r was not restricted to densities with r derivatives. Attention will be paid to both global errors and bias terms, since reduction of bias terms entails variance increase, and vice versa. Roughly speaking, if a kernel of order r is used to estimate a density with r derivatives, the bias is of order h_n^r and the optimal MISE (Mean Integrated Squared Error) = $E\left(\int (f_n(x) - f(x))^2\, dx\right)$ is of order $n^{-2r/(2r+1)}$ (Rosenblatt (1956), Bosq and Lecoutre (1987)); a uniform bound for L^1-errors can be found in Devroye (1987).

In part two, we construct finite order kernels from reproducing kernels and show that estimates of the density based on these kernels are actually derivatives of estimates of the cumulative distribution function defined as best L^2-approximations to the empirical distribution function. In part three, we focus on products of polynomials and densities and on the choice of the order and of the smoothing parameter. Computational aspects are also considered. The selection of the order was first considered by Hall and Marron (1987). By performing a mean integrated squared error analysis of the problem, they investigated theoretical properties of kernels with Fourier transform $\exp(-|t|^p)$ and proposed cross-validation as a method for choosing the kernel order and the smoothing parameter. We define a multi-stage procedure to construct estimates based on increasing order kernels and leading to a data-driven choice of both the order and the window width. Finally, we give a strong uniform consistency theorem for multivariate density estimates based on these kernels, extending previous results (Abdous and Berlinet, 1986), and showing how this theorem applies to different kinds of estimates.

2. Construction of finite order kernels

A measurable function K is said to be a kernel of order $p \in \overline{\mathbb{N}}^*$, if

$$\int x^j K(x)\, dx = \begin{cases} 1 & \text{for } j = 0 \\ 0 & \text{for } j = 1, \ldots, p-1 \\ C \neq 0 & \text{for } j = p\,. \end{cases}$$

A common construction of finite order kernels is obtained through piecewise polynomials (Singh (1979), Müller (1984) and Gasser et al (1985)) or by a Fourier transform (Devroye (1987) and Hall and Marron (1987)). We shall mainly be concerned here with products of polynomials and densities. With this type of kernels any choice of support, regularity conditions (such as continuity, differentiability etc) or tail are possible.

Let K_0 be a density function, called the initial kernel. Consider a Reproducing Kernel Hilbert Subspace V of $L^2(K_0)$:

V, endowed with the scalar product $(\varphi, \psi) = \int \varphi(x)\, \psi(x)\, K_0(x)\, dx$, is a Hilbert space of real functions and there is a function $\mathcal{K}(x, y)$ (the reproducing kernel) such that

$$\forall\, x \in \mathbb{R}, \qquad \mathcal{K}(x, .) \in V$$

$$\forall\, \varphi \in V, \forall\, x \in \mathbb{R}, \qquad \int \mathcal{K}(x, u)\, \varphi(u)\, K_0(u)\, du = \varphi(x).$$

Such a framework is encountered for instance in density estimation when using the orthonormal functions method : let μ be a positive σ-finite measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ with density $K_0 = \frac{d\mu}{d\lambda}$. Let us suppose we are given a sample $X_1, \ldots, X_n$ of i.i.d. random variables with common density $\frac{dP_{X_1}}{d\mu}$ lying in $L^2(\mu)$ in which a basis $(\varphi_i)_{i \in \mathbb{N}}$ has been chosen. Then $f = \frac{dP_{X_1}}{d\lambda}$ can be estimated by $f_n(x) = \sum_{i=1}^{q(n)} \frac{1}{n} \sum_{j=1}^{n} \varphi_i(X_j)\, \varphi_i(x)\, K_0(x)$, where q(n) has to be suitably chosen.

The function $\mathcal{K}_{q(n)}(x,y) = \sum_{i=1}^{q(n)} \varphi_i(x)\, \varphi_i(y)$ is the reproducing kernel of the subspace $V_{q(n)}$ of $L^2(\mu)$ generated by $\varphi_1, \varphi_2, \ldots, \varphi_{q(n)}$ and $f_n(x)$ can be rewritten as :

$$f_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathcal{K}_{q(n)}(x, X_j)\, K_0(x).$$

For details see Bosq and Lecoutre (1987), and Györfi (1974) for the sequential case.
When $f \notin L^2(K_0)$, the estimate exhibits an erratic behavior. From the robustness point of view the kernels under consideration below will provide much better estimates.

Theorem 1.

Let $K(x) = \mathcal{K}(x, 0)\, K_0(x)$ and V_r be the space of polynomials of degree at most r.

(a) If the constants belong to V, then $\forall\, y \in \mathbb{R}$, $\mathcal{K}(\,.\,, y)\, K_0(\,.\,)$ integrates to 1.

(b) If the monomial (x^j) $(j \geq 1)$ belongs to V, then $\int x^j\, K(x)\, dx = 0$.

(c) If $\int x^{r+1}\, K(x)\, dx$ is finite and different from zero and if $V_r \subset V$, then K is a kernel of order (r+1).

Proof :

By the reproduction property, for any constant function $C(x) = C$ and any monomial (x^j) $(j \geq 1)$, we have : $\int C\, \mathcal{K}(x, y)\, K_0(x)\, dx = C \Big|_{x=y} = C$

and $$\int x^j \mathcal{K}(x, 0)\, K_0(x)\, dx = (x^j)\Big|_{x=0} = 0.$$

Conversely, under very mild assumptions, satisfied by all reasonable kernel, a kernel of order r can be written in the form $\mathcal{K}(x, 0)\, K_0(x)$ (Berlinet, 1990).

Theorem 2.

Let $f_n(x) = \frac{1}{nh} \sum_{j=1}^{n} \mathcal{K}\left(\frac{x - X_j}{h}, 0\right) K_0\left(\frac{x - X_j}{h}\right)$ be the standard kernel estimate of f defined by means of the kernel $K(x) = \mathcal{K}(x, 0)\, K_0(x)$. Let $F_n(t) = \int_{-\infty}^{t} f_n(x)dx$ be the natural estimate of F obtained by integration. Then $F_n(t) = \Pi(t)$ where

$$\Pi = \arg\min_{\pi \in V} \int \left(\pi\left(\frac{t - u}{h}\right) - \mathbb{F}_n(u)\right)^2 K_0\left(\frac{t - u}{h}\right) du.$$

Proof :

$$F_n(t) = \frac{1}{n} \sum_{j=1}^{n} \int \mathbb{1}_{]-\infty, t]}(X_j + hv)\, \mathcal{K}(v, 0)\, K_0(v)\, dv. \qquad (2,1)$$

The integral in (2,1) is the value at 0 of the projection of $\mathbb{1}_{]-\infty, t]}(X_j + h\,.)$ on the subspace V and $F_n(t)$ is the average of this values. Therefore $F_n(t)$ is the value at 0 of the projection on V of the function $\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{]-\infty, t]}(X_j + h\,.)$; i.e. the solution to the following problem : find π in V minimizing the norm of $\left(\pi(.) - \mathbb{F}_n(t - h\,.)\right)$ and set $F_n(t) = \Pi(0)$.
The conclusion follows.

Corollary.

Under the hypothesis of Theorem 2, if the mapping $\pi(.) \to \pi\left(\frac{t - .}{h}\right)$ is one-to-one in V (as it is the case when $V = V_r$, the space of polynomials of degree at most r), then $F_n(t) = \Pi(t)$ where

$$\Pi = \arg\min_{\pi \in V} \int \left(\pi(u) - \mathbb{F}_n(u)\right)^2 K_0\left(\frac{t - u}{h}\right) du.$$

This last property has been pointed out by Lejeune and Sarda (1989) in particular cases.

Comments : Theorem 2 shows that estimating a density with a kernel like K amounts to simply deriving the best L^2-approximation to the empirical distribution function $\mathbb{F}_n$ by a function $\pi\left(\frac{t - .}{h}\right)$ (π lying in V) with weight $K_0\left(\frac{t - .}{h}\right)$. Such an interpretation tends to indicate that the cumulative distribution function should be more widely used in the determination of the smoothing parameter or of the dimension of models. This is all the more true since the empirical distribution function appears as an intrinsic reference whereas

no equivalent reference exists for the density. Sarda (1990) and Abdous (1990) have done some work in this direction recently.
Many familiar estimates of the density function are actually derivatives of least squares estimates of the cumulative distribution function. As an example, let us consider again the orthonormal functions method mentioned in part 2 and use the same notation.

$$\text{Set } F_n(t) = \int f_n(x)\,dx = \frac{1}{n}\sum_{j=1}^{n}\int \mathbb{1}_{]-\infty,\,t]}(x)\sum_{i=1}^{q(n)}\varphi_i(X_j)\,\varphi_i(x)\,K_0(x)\,dx. \qquad (2,2)$$

The integral in (2,2) is the value at X_j of the projection of $\mathbb{1}_{]-\infty,\,t]}(.)$ on the subspace $V_{q(n)}$ and $F_n(t)$ is the average of these values.

Theorem 3.

If $d(.) = f(x - h\,.\,)$ (respectively, $D(.) = F(t - h\,.\,)$) $\in L^2(K_0)$, then the expectation of $f_n(x)$ (respectively, $F_n(t)$) is the projection of d (respectively, D) on V.

Proof :

$$Ef_n(x) = \left(\frac{1}{h}K\left(\frac{\cdot}{h}\right) * f\right)(x) = \int f(x - hv)\mathcal{K}(v, 0)\,K_0(v)\,dv.$$

The proof is similar for $F_n(t)$.
In part 4, we give a strong uniform consistency result.

3. The polynomial case

In this part, we investigate the case where $V = V_r$, the space of polynomials of degree at most r. No additional work is needed to deal with kernels of order (m,p) used to estimate derivatives of regression or density functions (Singh (1979), Gasser and Müller (1979), Gasser et al (1985), Gasser (1989)) and defined as follows :

a measurable function K is said to be a kernel of order (m, p), $m \in \mathbb{N}$, $p \in \overline{\mathbb{N}}^*$, $m \leq p - 2$, if

$$\int x^j K(x)\,dx = \begin{cases} 0 & \text{for } j \in \{0, \ldots, p-1\} \text{ and } j \neq m \\ (-1)^m\, m! & \text{for } j = m \\ C \neq 0 & \text{for } j = p \end{cases}$$

(the last condition is removed when $p = +\infty$).

A superkernel is a kernel of order $(m, +\infty)$. Müntz's theorem implies that such a kernel cannot have a compact support.

K_0 is a density with finite moments up to order 2s ($s \geq 1$). We have the inclusion : $V_s \subset L^2(K_0)$. All the conditions of orthogonality will refer to $L^2(K_0)$. For simplicity, we identify functions and their classes in $L^2(K_0)$.

Let $P_0 \equiv 1, P_1, P_2, \ldots, P_s$ be the sequence of orthonormal polynomials of V_s and let $\mathcal{K}_s$ be its reproducing kernel :

$$\mathcal{K}_s(x,y) = \sum_{i=0}^{s} P_i(x)P_i(y).$$

Theorem 4.

Let K be a kernel of order (m, p), $p \geq 2$, vanishing with K_0 and belonging to $L^2\left(\frac{1}{K_0}\right)$. Then the function $\frac{K}{K_0} \mathbb{1}_{K_0 \neq 0}$ has, in $L^2(K_0)$, a projection on V_r ($r \leq \min(p-1, s)$) which is independent of K and which represents in V_r the operator $(-1)^m \left.\frac{d}{dx^m}\right|_{x=0}$.

Proof :

The function $\frac{K}{K_0} \mathbb{1}_{K_0 \neq 0}$ is obviously in $L^2(K_0)$. Its projection on V_r is the polynomial Π given by

$$\Pi(x) = \int_{\{K_0(y) \neq 0\}} \mathcal{K}_r(x,y)\, K(y)\, dy = \sum_{i=0}^{r} P_i(x) \int P_i(y)\, K(y)\, dy$$

$$\Pi(x) = \sum_{i=0}^{r} P_i(x) \int \sum_{k=0}^{i} \frac{1}{k!} P_i^{(k)}(0)\, y^k\, K(y)\, dy$$

$$\Pi(x) = \begin{cases} 0 & \text{if } r < m \\ \sum_{i=m}^{r} (-1)^m P_i^{(m)}(0) P_i(x) & \text{if } r \geq m. \end{cases}$$

This proves the first part of the assertion.

Now let $j \in \{0, ..., r\}$ and $\alpha_0^j, \alpha_1^j, ..., \alpha_j^j$ be the coordinates of x^j in the basis $P_0, ..., P_r$.

$$(-1)^m \int x^j \sum_{i=0}^{r} P_i^{(m)}(y) P_i(x) K_0(x)\, dx = (-1)^m \sum_{i=0}^{j} \alpha_i^j P_i^{(m)}(y) = (-1)^m (x^j)^{(m)} \Big|_{x=y}.$$

Theorem 5.

Among the kernels satisfying the assumptions of Theorem 4 with $p = r+1$ ($r \geq 1$) and $m \leq r \leq s$, only one is the product of a polynomial and K_0 :

$$K_r^m(x) = (-1)^m \sum_{i=m}^{p-1} P_i^{(m)}(0) P_i(x)\, K_0(x).$$

Proof :

For $j \in \{0, ..., p-1\}$, let $x^j = \sum_{k=0}^{j} \alpha_k^j P_k(x)$. From Theorem 4

$$\int x^j K_r^m(x)\, dx = \begin{cases} 0 & \text{if } j < m \\ \sum_{i=m}^{j} (-1)^m P_i^{(m)}(0)\, \alpha_i^j & \text{if } j \in \{m, ..., p-1\} \end{cases}$$

and therefore K_r^m is a kernel of order (m, p).

3.1. DETERMINANTAL EXPRESSIONS

To give an explicit formula for K_r^m, we introduce some notation : for $k \geq 1$ and any sequence $\mu = (\mu_i)$ of real numbers, let us denote by M_k^q the Hankel matrix of order k built from $\mu_q, \mu_{q+1}, \ldots, \mu_{q+2k-2}$, and by H_k^q its determinant :

$$M_k^q = \begin{pmatrix} \mu_q & \mu_{q+1} & \cdots & \mu_{q+k-1} \\ \mu_{q+1} & & & \\ \vdots & & & \\ \vdots & & & \\ \mu_{q+k-1} & \cdots & & \mu_{q+2k-2} \end{pmatrix}, \qquad H_k^q = \det(M_k^q).$$

Finally, let $H_{k,m}^q(x)$, $x \in \mathbb{R}$, $m \in \{1, \ldots, k\}$ be the determinant of the matrix obtained from M_k^q by replacing the mth line by $1, x, x^2, \ldots, x^{k-1}$.

Theorem 6.

Let $\mu = (\mu_i)_{0 \leq i \leq 2s}$ be the sequence of (2s + 1) first moments of K_0.

Then $\forall\, x \in \mathbb{R}$, $K_r^m(x) = (-1)^m\, m!\, H_{r+1,m+1}^0(x)\, K_0(x) \,/\, H_{r+1}^0$.

Proof :

Writing $K_r^m(x)$ in the basis $1, x, x^2, \ldots, x^r$ and applying the definition of a kernel of order (m,p), one gets a linear system in the coefficients of $K_r^m(x)$ with matrix M_k^0. Straightforward algebra gives the result.

REMARK : The determinantal form of $K_r^m(x)$ can be used either in practical computations with small values of r, or in theoretical considerations.

In most cases the integration by parts formula shows that kernels of order (m,p) can be obtained as derivatives of kernels of order p. From now on, we will focus on such kernels.

3.2. EXAMPLES

Any choice of K_0, with finite moments up to order 2s ($s \geq 1$), provides a sequence of kernels $K(x) = \mathcal{K}(x, 0)\, K_0(x)$. This choice, possibly made from the observations, has to be further investigated, especially when information is available on the support of f.

(a) $K_0(x) = \frac{1}{2} \mathbb{1}_{[-1,\,1]}(x)$ leads to piecewise polynomial kernels : Legendre kernels.

(b) $K_0(x) = \left(2^{1+1/2k}\,\Gamma(1+1/2k)\right)^{-1} \exp\left(-\frac{x^{2k}}{2}\right)$ gives rise to Gram-Charlier kernels when $k = 1$. The derivatives of the orthonormal polynomials are in this case linear combinations of a bounded number of polynomials of the same system.

(c) $K_0(x) = \left(\frac{2}{\beta}\,\Gamma\left(\frac{\alpha+1}{\beta}\right)\right)^{-1} |x|^{\alpha} \exp\left(-|x|^{\beta}\right)$ gives rise to Laguerre kernels when $\alpha = 0$ and $\beta = 1$.

(d) $K_0(x) = A \exp\left(-\frac{\alpha}{4}(x-\beta)^4 - \frac{\gamma}{2}(x-\beta)^2\right)$ ($\alpha \geq 0$; $\gamma > 0$ if $\alpha = 0$). This family of distributions is characterized by the same property as in (b) with a number of polynomials less than or equal to two.

Some of these kernels have been discussed in detail in the literature (Deheuvels, 1977a). Numerous results concerning orthogonal polynomials with weights, such as those given above and many others, can be found in Freud(1973), Nevai (1973a, 1973b, 1979), Brézinski(1980).

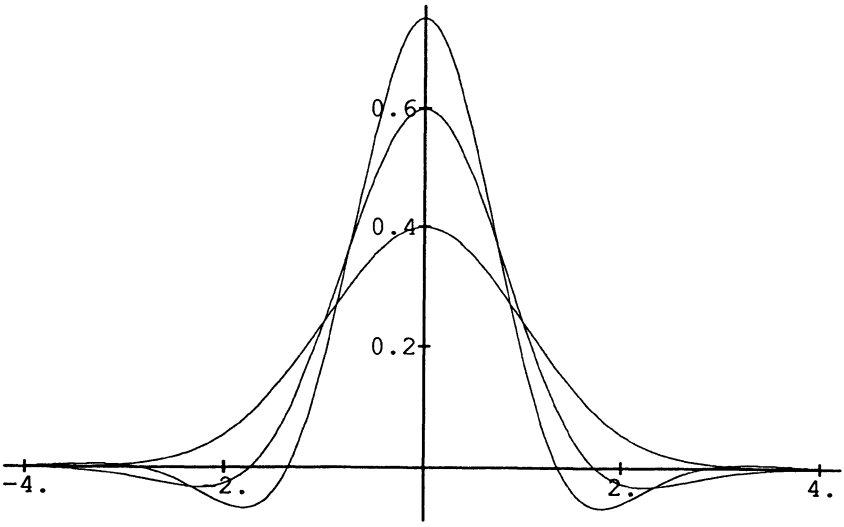


Figure 1. The three first Gram-Charlier kernels represented on [-4, +4].

3.3. COMPUTATIONAL ASPECTS

Only straightforward methods of numerical analysis are needed to calculate our estimates. The orthonormal polynomials can be computed by means of the following relationships :

$$P_n(x) = \frac{Q_n(x)}{\| Q_n \|}, \forall\, n \in \mathbb{N};$$

$$Q_0(x) = 1; \qquad Q_1(x) = x - \int x\, K_0(x)\, dx; \qquad Q_n(x) = (x - \alpha_n)\, Q_{n-1}(x) - \beta_n\, Q_{n-2}(x),\ n \geq 2$$

with $\alpha_n = \dfrac{\int x\, Q_{n-1}^2(x)\, K_0(x)\, dx}{\int Q_{n-1}^2(x)\, K_0(x)\, dx}$ and $\beta_n = \dfrac{\int Q_{n-1}^2(x)\, K_0(x)\, dx}{\int Q_{n-2}^2(x)\, K_0(x)\, dx}$.

The kernel can be computed either iteratively or by means of the Christoffel-Darboux formulas, when the Q_i's are known explicitely :

$$\forall\, x \neq 0, \quad \mathcal{K}_r(x, 0) = \sum_{i=0}^{r} P_i(x)P_i(0) = \frac{1}{\| Q_r \|^2}\left(\frac{Q_{r+1}(x)Q_r(0) - Q_{r+1}(0)Q_r(x)}{x}\right)$$

$$\mathcal{K}_r(0, 0) = \sum_{i=0}^{r} [P_i(0)]^2 = \frac{1}{\| Q_r \|^2}\left[\frac{d}{dx}\left(Q_{r+1}(x)Q_r(0) - Q_{r+1}(0)Q_r(x)\right)\right]\Bigg|_{x=0}.$$

3.4. THE MULTIPLE KERNEL ESTIMATION PROCEDURE FOR THE DENSITY

Above, we argued in favour of the use of the empirical distribution function in the determination of the smoothing parameter, whereas in this subsection, we choose any cross-validation criterion, CV, which preferably would be consistent. For relevant discussion and references, see Berlinet and Devroye (1989). CV is a score function of the sample and of the smoothing parameter h estimating either the likelihood of the sample or some measure of deviation between the estimate and the true unknown density. This includes double (Devroye, 1989) or multiple (Berlinet, 1990) kernel methods.

We start with an estimate $f_{n,0}$ based on K_0 with smoothing parameter h_0 optimizing $CV(K_0, h)$. Then we compute $f_{n,1}$ based on K_1 with smoothing parameter h_1 optimizing $CV(K_1, h)$, and so on. We decide to stop the procedure or not according to the behavior of the sequence $CV(K_i, h_i)$. The order i_0 of the kernel is chosen so as to optimize $CV(K_i, h_i)$. The smoothing parameter corresponding to K_{i_0} is h_{i_0}.

It seems possible to mimick the paper by Hall and Marron (1987) to show that for the minimum $MISE = E\left(\int (f_n(x) - f(x))^2\, dx\right)$:

(1) if f is not too "smooth", then there is a finite optimal order, whereas if f is very "smooth", then, at least asymptotically, the optimal order is infinite;

(2) choosing the smoothing parameter and kernel order by cross-validation is, asymptotically, as good as using the optimal values.

However, in order to implement the procedure on the basis of a fixed sample size, one rather needs theoretical results on the non-asymptotic behavior of $CV(K_i, h_i)$,such as, for instance, convexity properties. Our choice of the order is still empirical : we stop the procedure when no "significant" decrease of CV has occured in the last estimation.

The proposed algorithm is as follows : the quantities n and $X_1, ..., X_n$ are given.
Start out with $f_{0,h_0}(x) = \frac{1}{nh_0} \sum_{j=1}^{n} K_0\left(\frac{x - X_j}{h_0}\right)$, where h_0 optimizes $CV(K_0, h)$.
Then go from step k to step k+1 by means of the following relationship

$$f_{k+1,h}(x) = f_{k,h}(x) + \frac{P_{k+1}(0)}{nh} \sum_{j=1}^{n} P_{k+1}\left(\frac{x - X_j}{h}\right) K_0\left(\frac{x - X_j}{h}\right). \qquad (3,1)$$

Note that the second term of the sum in (3,1) which appears as the correction at step (k+1) integrates to zero.
Much computing time is saved by optimizing $CV(K_i, h_i)$ only over a grid G and storing the values $f_{k,h}(x)$ and $K_0\left(\frac{x - X_j}{h}\right)$ for $h \in G$. In most cases, to evaluate $CV(K_i, h)$, only the values of $f_{k,h}(X_i)$, $1 \leq i \leq n$, will be used. The Fourier Transform of K is easily derived from that of K_0; this is interesting when using the Fast Fourier Transform, especially in case of large sample sizes (Silverman, 1986).
All this is easily transposed to the cumulative distribution function.
Further developments and numerical results will appear in a forthcoming paper.

4. Strong uniform consistency of random kernel-type density estimates

In this section, we study the asymptotic behavior of density estimators based on kernels $K(x) = \mathcal{K}(x, 0)\, K_0(x)$ but with more general dependence with respect to the estimation point x and the observation X_i than through the ratio $\frac{x - X_i}{h}$, where $h = h_n$ is deterministic. Indeed we consider a wide class of multivariate density estimates for which we give sufficient conditions for strong uniform consistency. This class contains all usual kernel estimators and allows non affine random dependence with respect to the observations or the estimation point. Theorem 7 is an extension to nonpositive kernels of the result given in Abdous and Berlinet (1986).
The independent identically distributed random variables $(X_i)_{i \in \mathbb{N}}$ are supposed to be defined on a probability space $(\Omega, \mathcal{A}, P)$ and to take their values in $\mathbb{R}^k$. Their common density f is supposed to be continuous and bounded. For $x \in \mathbb{R}^k$ and $\omega \in \Omega$, f(x) is estimated by

$$f_n(x,\omega) = \frac{1}{n} \sum_{i=1}^{n} K\big(\theta_n(\omega, x, X_i(\omega))\big)\, |\, J_n(\omega, x, X_i(\omega))\, | \qquad (4,1)$$

where

(i) θ_n is a real valued mapping defined on $\Omega \times \mathbb{R}^k \times \mathbb{R}^k$ satisfying the following conditions :

(c1) For almost all ω, $\theta_n(\omega, x, .)$ is, for all x, a diffeomorphism in $\mathbb{R}^k$ with Jacobian $J_n(\omega, x, .)$ and inverse $\theta_n^{-1}(\omega, x, .)$;

(c2) $\forall$ (x, j) $\theta_n(., x, X_j(.))$ and $J_n(., x, X_j(.))$ are measurable.

(ii) K is a linear combination of bounded real-valued functions $(K_j)_{1\le j\le q}$ defined on $\mathbb{R}^k$, such that

(c3) K integrates to one;

(c4) $\forall\, j \in \{1, ..., q\}$ $|K_j|$ is Riemann integrable;

(c5) $\forall\, j \in \{1, ..., q\}$, $\int_0^{+\infty} t^{k-1} L_j(t)dt < +\infty$ where $\forall\, t \ge 0$, $L_j(t)$ is the essential supremum of K_j on $[t, +\infty[$;

The following properties of (θ_n) and (J_n) will be needed in the sequel :

(c6) Almost surely (a.s.), for all (x,y), $\lim_{n\to\infty} \theta_n^{-1}(\omega, x, y) = x$;

(c7) a.s., $\forall\, \delta > 0$, $\sup\left\{ R \mid \lambda^k(B(O, R) \cap E_{n,\delta}(\omega))=0 \right\} \to +\infty$

where

$$E_{n,\delta}(\omega) = \bigcup_{x\in A} E_{n,\delta}(x, \omega),\ E_{n,\delta}(x, \omega) = \left\{ y \mid \| \theta_n^{-1}(\omega, x, y) - x \| \ge \delta \right\}$$

A is a non-empty subset of $\mathbb{R}^k$, λ^k is the Lebesgue measure and B(O, R) is the ball of radius R centered at the origin;

(c8) a.s., $\forall\, R > 0$, $\sup_{\substack{x \in A \\ \|y\| \le R}} \| \theta_n^{-1}(\omega, x, y) - x \| \to 0$;

(c9) There is a sequence of real numbers $(\beta_n)_{n\ge1}$ tending to infinity such that a.s. $\sup_{x, y \in \mathbb{R}^k} | J_n(\omega, x, y) | \le \sqrt{\dfrac{n}{\beta_n \operatorname{Log} n}}$.

REMARKS (i) (c7) is entailed by (c8), which is often easier to check.

(ii) (c5) is satisfied, if $\forall\, \delta > 0$, $\lim_{\|x\|\to\infty} \|x\|^{k+\delta} K(x) = 0$.

Theorem 7 : strong uniform consistency.

Assume (c1) to (c5). If θ_n and J_n satisfy respectively (c7) and (c9), and if f is uniformly continuous on $\mathbb{R}^k$ then

$$\sup_{x \in A} | f_n(x, \omega) - f(x) | \to 0 \text{ a.s.}$$

The proof is based on the inequality :

$$|f_n(x, \omega) - f(x)| \leq |f_n(x, \omega) - g_n(x, \omega)| + |g_n(x, \omega) - f(x)|$$

where $g_n(x, \omega) = \int K\big(\theta_n(\omega, x, y)\big) \, |J_n(\omega, x, y)| \, f(y) \, dy$.
Under our hypothesis, the existence of g_n is a consequence of a change of variable. As a by-product, we get $g_n(x, \omega) = \int K(y) \, f(\theta_n^{-1}(\omega, x, y)) \, dy$ and the property that $\big(K\big(\theta_n(\omega, x, y)\big) \, |J_n(\omega, x, y)|\big)_{n \in \mathbb{N}}$ is a random delta sequence (Prakasa Rao, 1983).

Theorem 8.

(a) Assume (c6). Then $\forall \, x \in \mathbb{R}^k$, $g_n(x,\omega) \to f(x)$ a.s.
(b) Under (c7), if f is uniformly continuous then

$$\sup_{x \in A} |g_n(x, \omega) - f(x)| \to 0 \text{ a.s.}$$

From Theorem 8, it follows that (c1)-(c6) and (c9) entail the a.s. pointwise convergence of the estimate f_n.
We quote a lemma of Devroye and Wagner (1980) in order to fix the notation.

Lemma 1.

Let K be a positive, bounded and Riemann integrable function defined on $\mathbb{R}^k$. For every η, δ, $\rho > 0$, there exists a function

$$K^*(x) = \sum_{i=1}^{N} \alpha_i \, 1_{E_i}(x), \; \forall \, x \in \mathbb{R}^k$$

such that :

(1) $\alpha_1,..,\alpha_N$ are positive real numbers ;

(2) $E_1,\ldots,E_N$ are open disjoint rectangles of $[-\rho,\rho]^k$;

(3) $K^*(x) \leq \sup_{y \in \mathbb{R}^k} K(y) = M_2, \; \forall \, x \in \mathbb{R}^k$;

(4) $|K^*(x) - K(x)| \leq \eta$ on $[-\rho,\rho]^k \cap D^c$;

(5) $D \subset B \subset \bigcup_{i=1}^{M} B_i$ where the B_i's, $i \in \{1, ..., M\}$ are disjoint rectangles of $[-\rho, \rho]^k$ and $\lambda^k(B) < \delta$.

Lemma 2 below sums up the conclusions of the different steps of the proof which follows those in Devroye and Wagner (1980).

Let

$$I_{1,n}(\omega,x) = \int_{\mathbb{R}^k} |J_n(\omega,x,y)| \; |K(\theta_n(\omega,x,y)) - K^*(\theta_n(\omega,x,y))| \, dF(y) \; ,$$

$$I_{2,n}(\omega,x) = \left| \int_{\mathbb{R}^k} |J_n(\omega,x,y)|\, K^*(\theta_n(\omega,x,y))\, dF(y) - \int_{\mathbb{R}^k} |J_n(\omega,x,y)|\, K^*(\theta_n(\omega,x,y))\, dF_n(y) \right| ,$$

$$I_{3,n}(\omega,x) = \int_{\mathbb{R}^k} |\, J_n(\omega,x,y)\, |\ |\, K^*(\theta_n(\omega,x,y)) - K(\theta_n(\omega,x,y))\, |\, dF_n(y) ,$$

$$C_{1,n}(\omega,x) = \left\{ y : \theta_n(\omega,x,y) \in [-\rho,\rho]^k \right\},$$

$$M_1 = \sup_{x \in \mathbb{R}^k} f(x) ,$$

μ_n denotes the empirical measure $\frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ and μ denotes the measure with density f.

We have $$\sup_{x \in A} |\, f_n(x,\omega) - g_n(x,\omega)\, | \le \sum_{i=1}^{3} \sup_{x \in A} I_{i,n}(\omega,x).$$

Lemma 2. Let q = 1 and let K be a probability density. Let η, δ, $\rho > 0$ be fixed quantities as those employed in Lemma 1. Then we have the following bounds :

a) $$\sup_{x \in A} I_{1,n}(\omega,x) \le 2\delta M_1 M_2 + \eta M_1 (2\rho)^k + \sup_{x \in A} \int_{C_{1,n}(\omega,x)} |J_n(\omega,x,y)|\, K(\theta_n(\omega,x,y))\, dF(y) ;$$

b) $$\sup_{x \in A} I_{2,n}(\omega,x) \le N M_2 \sup_{1 \le j \le N} |\mu_n(E_j) - \mu(E_j)| ;$$

c) $$\sup_{x \in A} I_{3,n}(\omega,x) \le \eta M_1 (2\rho)^k + 2 M_1 M_2 \delta + 2 M M_2 \sup_{1 \le i \le M} |(\mu_n - \mu)(B_i)| + \eta\, |(\mu_n - \mu)([-\rho, \rho]^k)| + \sup_{x \in A} \int_{C_{1,n}(\omega,x)} |\, J_n(\omega,x,y)\, |\, K(\theta_n(\omega,x,y))\, d\, F_n(y) ;$$

Under condition (c5), the sum of the integrals in a) and c) is bounded by means of Vapnik-Chervonenkis-Devroye inequality for classes of rectangles. To get the conclusion of Theorem 7 when q = 1 and K is a probability density, it is sufficient to let δ and η tend to 0 and ρ tend to $+\infty$. Theorem 7 follows in a straightforward manner.

5. Applications

The class of estimates described above contains many standard kernel estimates, as is, for example the so-called automatic kernel estimates with smoothing parameter depending on x. Many applications could be given for density, regression and hazard function estimation. We will limit ourselves to five examples concerning the density.

(i) Deheuvels's estimate (1977) is obtained for $\theta_n(\omega, x, y) = A_n(x-y)$ where A_n is an automorphism. In this case, (c8) is satisfied if $\| A_n^{-1} \|^* \to 0$, and (c9) is satisfied if $(n / \mathrm{Log} n) \, |\det A_n|^{-2} \to +\infty$.

(ii) The automatic estimate of Devroye and Wagner (1980) is obtained for $\theta_n(\omega,x,y) = (x-y)/h_n(\omega)$, $\forall\, x, y \in \mathbb{R}^k$, $\forall\, \omega \in \Omega$; $(h_n(\omega))_{n\geq 1}$ is a sequence of positive real numbers depending on the observations X_i, $i \in \{1, ..., n\}$. A sufficient condition for (c8) is that $h_n(\omega) \to 0$ a.s. ; if there exists a sequence $(\beta_n)_{n\geq 1}$ such that $\beta_n \underset{n\to\infty}{\longrightarrow} +\infty$ and $nh_n^{2k}(\omega) / \mathrm{Log}\, n \geq \beta_n$ a.s., then (c9) is satisfied because $| J_n(\omega,x,y) | = 1/h_n^k(\omega)$.

(iii) Chai Genxiang (1984) reexamined this last estimate assuming that K has compact support $[-\rho,\rho]^k$, $\rho > 0$, and that h_n depends both on the observations and on the estimation point x. We have $\theta_n(\omega,x,y) = \alpha\rho.(x-y)/h_n(x,\omega)$, $\forall\, x, y \in \mathbb{R}^k$, $\forall\, \omega \in \Omega$; α is a constant ($\alpha > 1$). Condition (c8) is fullfilled whenever $\sup_{x\in A} h_n(x,\omega) \underset{n\to\infty}{\longrightarrow} 0$ a.s. ; (c9) is satisfied if there exists a sequence $(\beta_n)_{n\geq 1}$ such that

$$\beta_n \underset{n\to\infty}{\longrightarrow} +\infty \text{ and } \inf_{x\in\mathbb{R}^k} (h_n^{2k}(x,\omega))(n/\mathrm{Log}\, n) \geq \beta_n \text{ a.s.}$$

The nearest neighbour estimate is a particular case.

(iv) Bierens 's estimate (1983) is defined by

$$f_n(x,\omega) = \frac{1}{nh_n^k |\mathrm{Det}\, S_n|^{1/2}} \sum_{i=1}^{n} K\left(S_n^{-1/2} \left(\frac{x - \sqrt{1-h_n^2}\, X_i - \left(1 - \sqrt{1-h_n^2}\right) \bar{X}}{h_n} \right) \right)$$

where $\bar{X} = (1/n) \sum_{i=1}^{n} X_i$, $S_n = (1/n) \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^t$ and $(h_n)_n$ is a sequence of real numbers in $]0,1]$; this estimate can be written, up to a multiplicative constant, as the estimate in (1); by Theorem 1,

$$\text{if } h_n \underset{n\to\infty}{\longrightarrow} 0 \text{ and } nh_h^{2k} / \mathrm{Log}\, n \underset{n\to\infty}{\longrightarrow} +\infty \text{, then } \sup_{x\in A} |f_n(x,\omega) - f(x)| \underset{n\to\infty}{\longrightarrow} 0 \text{ a.s.}$$

(v) A transformed kernel estimate, close to the one of Devroye and Györfi (1985), introduced to estimate heavy-tailed densities :

$$f_n(x) = \frac{1}{nh^k} \sum_{i=1}^{n} K\left(\frac{T(x) - T(X_i)}{h} \right) | J_T(X_i) | ,$$

where $T : \mathbb{R}^k \to]-1, 1[^k$ is a diffeomorphism and J_T its Jacobian. Theorem 7 applies with $\theta_n(\omega,x,y) = \frac{T(x) - T(y)}{h}$.

Acknowledgement

Research of the author was sponsored by NSERC Grant A3456 during a stay at McGill University on invitation of Professor Luc Devroye.

References

Abdous, B. (1990) Least squares cross-validation in kernel estimation of distribution function. Preprint.

Abdous, B. and Berlinet, A. (1986) Convergence uniforme presque sûre d'une classe d'estimateurs à noyau, pour la densité. *Comptes Rendus de l'Académie des Sciences de Paris,* 303, pp. 761-764.

Berlinet, A. (1990) The multiple kernel method in nonparametric estimation. Preprint.

Berlinet, A. and Devroye, L. (1989) Estimation d'une densité : un point sur la méthode du noyau. *Statistique et Analyse des Données,* 14, 1, pp. 1-32.

Bierens, H.J. (1983) Sample moments integrating normal kernel estimators of multivariate density and regression functions. *Sankhya. The Indian Journal of Statistics,* 45, series B, pt 2, pp. 160-192.

Bosq, D. and Lecoutre, J.P. (1987) *Théorie de l'Estimation Fonctionnelle.* Economica, Paris.

Brézinski, C. (1980) *Padé-Type Approximation and General Orthogonal Polynomials.* Birkhauser, Basel.

Chai Genxiang, (1984) Consistent estimation of random window-width kernel of density function. *Scientia Sinica,* series A, XXVII, 11, pp. 1155-1163.

Deheuvels, P. (1977a) Estimation non-paramétrique de la densité par histogrammes généralisés. *Revue de Statistique Appliquée,* 25, pp. 5-42.

Deheuvels, P. (1977b) Estimation non-paramétrique de la densité par histogrammes généralisés (II). *Publications de l'Institut de Statistique des Universités de Paris,* 22, pp. 1-23.

Devroye, L. (1987) *A Course in Density Estimation.* Birkhauser, Boston.

Devroye, L. (1989) The double kernel method in density estimation. *Annales de l'Institut Henri Poincaré,* 25, 4, pp. 553-580.

Devroye, L. and Györfi, L.(1985) *Nonparametric Density Estimation : The L_1 View.* John Wiley, New York.

Devroye, L. and Penrod, C.S. (1984) The consistency of automatic kernel density estimates. *Annals of Statistics,* 12, 4, pp. 1231-1249.

Devroye, L. and Wagner, T.J. (1980) The strong uniform consistency of kernel density estimates. *Journal of Multivariate Analysis,* V. North Holland, Amsterdam, pp. 59-77.

Freud, G. (1973) On polynomial approximation with the weight $\exp(- x^{2k}/2)$. *Acta Mathematica Academiae Scientiarum Hungaricae,* 24, pp. 363-371.

Gasser, T. (1989) The analysis of biomedical signals by nonparametric methods. *Statistique et Analyse des Données,* 14, 1, pp. 85-101.

Gasser, T. and Müller, H.G. (1979) Optimal convergence properties of kernel estimates of derivatives of a density. In T. Gasser and M. Rosenblatt (eds.), *Smoothing techniques for curve estimation,* Lecture Notes in Mathematics 757, Springer Verlag, pp. 144-154.

Gasser, T., Müller, H.G. and Mammitzsch, V. (1985) Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society* , series B, 47, pp. 238-252.

Györfi, L. (1974) Estimation of probability density and optimal decision function in RKHS, in *Progress in Statistics*, J. Gani, K. Sarkadi and I. Vincze editors, North-Holland, Amsterdam, pp. 281-301.

Hall, P. and Marron, J.S. (1987) Choice of kernel order in density estimation. *Annals of Statistics,* 16,pp. 161-173.

Lejeune, M. and Sarda, P. (1989) Smooth estimators of distribution and density functions. Preprint.

Müller, H.G. (1984) Smooth optimum kernel estimators of densities, regression curves and modes. *Annals of Statistics,* 12,pp. 766-774.

Nevai, P. (1973a) Some properties of orthogonal polynomials corresponding to the weight $(1 + x^{2k})^{\alpha} \exp(- x^{2k})$ and their application in approximation theory. *Soviet Math. Dokl.,* 14, pp. 1116-1119.

Nevai, P. (1973b) Orthogonal polynomials on the real line associated with the weight $|x|^{\alpha} \exp(- |x|^{\beta})$, I. *Acta Mathematica Academiae Scientiarum Hungaricae,* 24, pp. 335-342.

Nevai, P. (1979) *Orthogonal polynomials.* Memoirs of the American Mathematical Society, 219.

Prakasa Rao, B.L.S. (1983) *Nonparametric Functional Estimation.* Academic Press.

Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics,* 27, pp. 832-837.

Sarda, P. (1990) Smoothing parameter selection for smooth distribution functions. Preprint.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Singh,R.S. (1979)Mean squared errors of estimates of a density and its derivatives. *Biometrika*, 66, pp. 177-180.

LAWS OF THE ITERATED LOGARITHM FOR DENSITY ESTIMATORS

Paul DEHEUVELS
L.S.T.A., Université Paris VI
4 Place Jussieu, T. 45-55, E3
75252 Paris Cedex 05
FRANCE

ABSTRACT. We obtain laws of the iterated logarithm for density estimates. Our results are based on functional laws of the iterated logarithm for the increments of empirical processes due to Deheuvels and Mason (1990).

1. INTRODUCTION

Let $U_1, U_2, \ldots$ be a sequence of independent random variables with a uniform distribution on $(0,1)$, and let $U_n(t) = n^{-1}\#\{U_i \le t: 1 \le i \le n\}$ denote the empirical distribution function based on $U_1, \ldots, U_n$. Consider the uniform empirical process $\alpha_n(t) = n^{1/2}(U_n(t)-t)$, and, for $0<a<1$, $0 \le t \le 1-a$, and $0 \le s \le 1$, set $\xi_n(a,t;s) = \alpha_n(t+as)-\alpha_n(t)$.

A sequence of constants $\{a_n, n \ge 1\}$ will be said to satisfy the M. Csörgő-Révész-Stute ([CRS]) conditions if

$$0<a_n<1 \text{ for } n \ge 1,\ a_n \downarrow 0 \quad \text{and} \quad na_n \uparrow \infty \quad \text{as} \quad n \to \infty; \tag{S.1}$$

$$(\log(1/a_n))/\log\log n \to \infty \quad \text{as} \quad n \to \infty; \tag{S.2}$$

$$na_n/\log n \to \infty \quad \text{as} \quad n \to \infty. \tag{S.3}$$

Recently, Deheuvels and Mason (1990) obtained the following functional law of the iterated logarithm [LIL] for the sets of functions $\{\xi_n(a_n,t;\cdot),\ 0 \le t \le 1-a_n\}$. In order to describe this law, we need to introduce some notation. To this end, denote by $B(0,1)$ the set of all bounded functions on $[0,1]$, and, for any $f \in B(0,1)$, set $\|f\| = \sup_{0 \le s \le 1}|f(s)|$. Endow $B(0,1)$ with the topology defined by the sup-norm $\|\cdot\|$, and, for any $\varepsilon>0$ and $A \subset B(0,1)$, set $A^{\varepsilon} = \{f \in B(0,1), \|f-g\|<\varepsilon \text{ for some } g \in A\}$. Furthermore, denote by S the set of all absolutely continuous functions f on $[0,1]$ having Lebesgue derivative $\dot{f}$ such that $\int_0^1 \dot{f}^2(s)ds \le 1$. The definition of S is due to Strassen (1964). The aforementioned result of Deheuvels and Mason (1990) may now be stated in the following Theorem A.

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 19–29.
© 1991 *Kluwer Academic Publishers.*

Theorem A. Let $\{a_n, n\geq 1\}$ be a sequence satisfying the [CRS] conditions. Then, for any $\varepsilon>0$, there exists a.s. an $n_\varepsilon<\infty$ such that for all $n\geq n_\varepsilon$, $\{\xi_n(a_n,t;\cdot)/(2a_n\log(1/a_n))^{1/2} : 0\leq t\leq 1-a_n\}\subset S^\varepsilon$. Moreover, for any $\varepsilon>0$ and $h\in S$, there exists a.s. an $N_\varepsilon<\infty$ such that there exists for all $n\geq N_\varepsilon$ a $0\leq t_n = t_{n,\varepsilon,h}\leq 1-a_n$ such that $\|\xi_n(a_n,t_n;\cdot)/(2a_n\log(1/a_n))^{1/2}-h\|<\varepsilon$.

Proof. See Deheuvels and Mason (1990), §3. ∎

As a consequence of Theorem A, Deheuvels and Mason (1990) gave new proofs and extensions of the laws of the iterated logarithm due to Stute (1982) and Mack (1983). The corresponding results are stated in Theorems B and C below. Before stating these theorems, we will introduce some further notation. In the sequel, $X_1, X_2, \ldots$ will be a sequence of independent and identically distributed random variables on $(-\infty,\infty)$, having common distribution function $F(x) = P(X_1\leq x)$ and density $f(x) = F'(x)$ assumed to be continuous and positive on a given bounded non-void interval $[A,B]$. Let $F_n(x) = (1/n)\#\{X_i\leq x : 1\leq i\leq n\}$, and consider the following estimators of f. Let

$$f_n(x) = (n\lambda_n)^{-1}\sum_{i=1}^{n} K((x-X_i)/\lambda_n) = \int_{-\infty}^{\infty} \lambda_n^{-1}K(\lambda_n^{-1}(x-t))dF_n(t) \tag{1.1}$$

be the kernel estimate of f (Rosenblatt (1956), Parzen (1962)) based on a sequence $\{\lambda_n, n\geq 1\}$ of positive smoothing parameters (bandwidth) satisfying suitable assumptions specified below. The kernel $K(\cdot)$ satisfies the following conditions:

$$K(\cdot) \text{ is of bounded variation on } (-\infty.\infty); \tag{K.1}$$

$$\int_{-\infty}^{\infty} K(u)du = 1; \tag{K.2}$$

$$\text{For some } 0<M<\infty,\ K(u) = 0 \text{ for all } |u|\geq M. \tag{K.3}$$

Another estimate of f is as follows. Let

$$\hat{f}_n(x) = \mu_n/\inf\{h>0 : F_n(x+h/2)-F_n(x-h/2)\geq\mu_n\} \tag{1.2}$$

be the nearest neighbor density estimator (Fix and Hodges (1951), Loftsgaarden and Quesenberry (1965)), where $\{\mu_n, n\geq 1\}$ is another sequence of positive smoothing parameters. Furthermore, set

$$Ef_n(x) = \int_{-\infty}^{\infty} \lambda_n^{-1}K(\lambda_n^{-1}(x-t))dF_n(t), \tag{1.3}$$

and

$$M\hat{f}_n(x) = \mu_n/\inf\{h>0 : F(x+h/2)-F(x-h/2)\geq\mu_n\}. \tag{1.4}$$

In the sequel, we will assume at times that λ_n is random and depends on $X_1,\ldots,X_n$. In this case, it is noteworthy that $Ef_n(x)$ is not equal to the expectation $E(f_n(x))$ of $f_n(x)$. On the other hand, if λ_n is non-random, we have $Ef_n(x) = E(f_n(x))$.

We have the following LIL's for f_n and $\hat{f}_n$.

Theorem B. Let $\{\lambda_n, n\geq 1\}$ be a non-random sequence satisfying the [CRS] conditions. Then, under (K.1)-(K.2), and for any A<C<D<B, we have

$$\lim_{n\to\infty} \sup_{C\leq x\leq D} \pm \left(\frac{f_n(x)-Ef_n(x)}{\sqrt{f(x)}}\right)\left(\frac{n\lambda_n}{2\log(1/\lambda_n)}\right)^{1/2} = \left(\int_{-\infty}^{\infty} K^2(u)du\right)^{1/2} \text{ a.s.} \quad (1.5)$$

Theorem C. Let $\{\mu_n, n\geq 1\}$ be a non-random sequence satisfying the [CRS] conditions. Then, for any A<C<D<B, we have

$$\lim_{n\to\infty} \sup_{C\leq x\leq D} \pm \left(\frac{\hat{f}_n(x)-M\hat{f}_n(x)}{f(x)}\right)\left(\frac{n\mu_n}{2\log(1/\mu_n)}\right)^{1/2} = 1 \quad \text{a.s.} \quad (1.6)$$

Proofs. See §4 in Deheuvels and Mason (1990). See also Stute (1983), Mack (1983) and Hall (1990). ∎

The motivation for this paper is that, in the recent literature, there has been a considerable interest in kernel estimators where the bandwidth λ_n depends on the data $X_1,\ldots,X_n$, and eventually, also on x. A typical example of such data-driven bandwidth selections is as follows. Consider the Mean Inegrated Square Error defined by

$$\text{MISE}(\lambda_n) = E\int_I (f_n(x)-f(x))^2 dx, \quad (1.7)$$

where I is a specified sub-interval of $(-\infty,\infty)$. Assuming some further regularity conditions (such as the existence of a square integrable second derivative f" of f), it may be proved (see, e.g., Devroye and Györfi (1984)) that any choice of λ_n which minimizes $\text{MISE}(\lambda_n)$ satisfies

$$\tilde{\lambda}_n = (1+o(1))En^{-1/5} \quad \text{as} \quad n\to\infty, \quad (1.8)$$

where E is a constant depending on I and f. Many authors have proposed techniques enabling to choose $\lambda_n = \lambda(n;X_1,\ldots,X_n)$ in such a way that

$$\lambda_n/(En^{-1/5}) \to 1 \quad \text{as} \quad n\to\infty \quad \text{(in probability or almost surely).} \quad (1.9)$$

We do not offer here a discussion of such methods (for such a discussion, see, for instance, Park and Marron (1988) and the references therein), and limit ourselves to the observation that theorems such as Theorem B or Theorem C cannot be applied in the case of such data-driven bandwidths. In the sequel, we will adapt the methodology of Deheuvels and Mason (1990) to show that versions of Theorems B and C still hold in this case.

In Section 2, we will present LIL's of this type. In Section 3, we will review further functional LIL's for increments which are actually available.

2. LAWS OF THE ITERATED LOGARITHM FOR DATA DRIVEN KERNEL ESTIMATORS

In this section, we will assume that $f_n(\cdot)$ is as in (1.1) with a bandwidth $\lambda_n = \lambda(n;x;X_1,\ldots,X_n)$ satisfying the following general conditions. Below, assume that [C,D] is an arbitrary (but fixed) non-void sub-interval of [A,B]. We will make the following assumptions on λ_n.

There exists a function $H(\cdot)$ which is positive and continuous on [A,B], and a sequence $\{a_n, n\geq 1\}$ of (possibly dependent upon $X_1,\ldots,X_n$) positive random variables such that (L.1)

$$\sup_{C\leq x\leq D} |a_n^{-1}H(x)^{-1}\lambda(n;x;X_1,\ldots,X_n)-1| \to 0 \quad \text{a.s.} \tag{2.1}$$

Let $\{a_n, n\geq 1\}$ be as in (2.1). There exists a sequence $\{b_n, n\geq 1\}$ of positive constants such that the complement of (L.2)

$$\{1/\rho \leq a_n/b_n \leq \rho\} \text{ holds a.s. finitely often for all } \rho>1 \tag{2.2}$$

sufficiently large.

We will denote likewise by (L.3) and (L.4) the assumptions obtained by replacing formally in (L.1), (2.1) and (L.2), (2.2), respectively, almost sure convergence by convergence in probability (for (2.2), this amounts to assume that both a_n/b_n and b_n/a_n are $O_p(1)$ as $n\to\infty$).

The following examples illustrate (L.1)-(L.2) and (L.3)-(L.4).

Example 2.1. Moore and Yackel (1977) and Mack and Rosenblatt (1979) consider the choice of $\lambda(n;x;X_1,\ldots,X_n)$ taken equal to the distance of x to its k-th nearest neighbor in the sample $X_1,\ldots,X_n$. By choosing $\mu_n = k/n$, we see that this choice amounts to taking

$$\lambda_n = \lambda(n;x;X_1,\ldots,X_n) = (2\mu_n)/\hat{f}_n(x), \tag{2.3}$$

where $\hat{f}_n(x)$ is as in (1.2). Assume that $k=k_n$ satisfies the conditions:

$$k_n/\log n \to \infty \text{ and } k_n/n \to 0 \text{ as } n\to\infty. \tag{2.4}$$

Then, Devroye and Wagner (1977) have proved that

$$\sup_{C\leq x\leq D} |\hat{f}_n(x)-f(x)| \to 0 \text{ a.s.} \tag{2.5}$$

It follows that for λ_n as given in (2.3), assumptions (L.1) and (L.2) hold with $a_n = b_n = \mu_n = nk_n$ and $H(x) = 2/f(x)$ for $C\leq x\leq D$.

Example 2.2. Let I = [A,B], and set $R(g) = \int_I g^2(x)dx$. For $j=1,\ldots,n$, denote by $f_{j:n}$ the leave-one-out kernel estimator constructed from the

data $X_1,\ldots,X_n$ with X_j deleted, as in (1.1). Next, choose $0<E_1<E<E_2<\infty$, where E is as in (1.8), and let $\hat{\lambda}_n = \lambda(n;X_1,\ldots,X_n)$ to be the largest minimizer, over the interval $[E_1 n^{-1/5}, E_2 n^{-1/5}]$, of the cross-validated score function

$$CV(\lambda_n) = R(f_n) - 2n^{-1}\sum_{j=1}^{n} f_{j:n}(X_j). \tag{2.6}$$

Assuming some appropriate smoothness conditions on $K(\cdot)$ and $f(\cdot)$, Hall and Marron (1987) showed that $\hat{\lambda}_n$, as defined above, satisfies

$$n^{1/10}(\hat{\lambda}_n/\tilde{\lambda}_n - 1) \xrightarrow[d]{} N(0,\sigma^2), \tag{2.7}$$

where "$\xrightarrow[d]{}$" denotes convergence in distribution, $\tilde{\lambda}_n$ is as in (1.8), and σ^2 is a constant depending on f.

From, (1.14), it follows that $\hat{\lambda}_n = \lambda(n;X_1,\ldots,X_n)$ satisfies (L.3)-(L.4) with $a_n = \hat{\lambda}_n$, $b_n = \tilde{\lambda}_n$ and $H(\cdot) = 1$.

The main results of this section are stated in the following two theorems.

Theorem 2.1. Assume that (K.1)-(K.3) hold, together with (L.1)-(L.2). Furthermore, assume that $\{b_n, n\geq 1\}$, as specified in (L.2), satisfies the CRS conditions. Then, we have

$$\lim_{n\to\infty} \sup_{C\leq x\leq D} \pm \left(\frac{f_n(x)-Ef_n(x)}{\sqrt{f(x)}}\right)\left(\frac{n\lambda_n}{2\log(1/\lambda_n)}\right)^{1/2} = \left(\int_{-\infty}^{\infty} K^2(u)\,du\right)^{1/2} \text{ a.s.} \tag{2.8}$$

Theorem 2.2. Assume that (K.1)-(K.3) hold, together with (L.3)-(L.4). Furthermore, assume that $\{b_n, n\geq 1\}$ as specified in (L.2), satisfies the [CRS] conditions. Then, we have

$$\lim_{n\to\infty} \sup_{C\leq x\leq D} \pm \left(\frac{f_n(x)-Ef_n(x)}{\sqrt{f(x)}}\right)\left(\frac{n\lambda_n}{2\log(1/\lambda_n)}\right)^{1/2} = \left(\int_{-\infty}^{\infty} K^2(u)\,du\right)^{1/2} \tag{2.9}$$

in probability.

Proof. We follow the argument of the proof of Theorem 4.1 of Deheuvels and Mason (1990) with simple modifications. In the first place, assume, without loss of generality, that $M = 1/2$ in (K.3), and set

$$\tilde{f}_n(x) = f_n\left(x + \frac{1}{2}\lambda_n\right) = (n\lambda_n)^{-1}\sum_{i=1}^{n} \tilde{K}((X_i-x)/\lambda_n), \tag{2.10}$$

where $\tilde{K}(u) := K\left(\frac{1}{2} - u\right)$. Since f is continuous and $\lambda_n \to 0$ a.s., all we need for (2.8) to hold is to show that, for any $A<C<D<B$, we have

$$\lim_{n\to\infty} \sup_{C\leq x\leq D} \pm \left(\frac{\tilde{f}_n(x)-E\tilde{f}_n(x)}{\sqrt{f(x)}}\right)\left(\frac{n\lambda_n}{2\log(1/\lambda_n)}\right)^{1/2} = \left(\int_0^1 \tilde{K}^2(u)\,du\right)^{1/2} \text{ a.s.} \tag{2.11}$$

For later use, note that (K.3) implies that $\tilde{K}(u) = 0$ for $u \notin [0,1]$. Without loss of generality, set $F_n(x) = U_n(F(x))$, where $U_n(\cdot)$ is as in Section 1. Following Deheuvels and Mason (1990) (see their §4), it can be verified that

$$n^{1/2}\lambda_n(\tilde{f}_n(x)-E\tilde{f}_n(x)) = -\int_0^1(\alpha_n(F(x+\lambda_n u))-\alpha_n(F(x)))d\tilde{K}(u). \tag{2.12}$$

Next, we assume that x varies in a small sub-interval (c,d) of [C,D]. Here, we choose d-c>0 so small that $\sup_{0\le u\le 1}|F(x+\lambda_n u)-F(x)-\gamma_n u|\le\gamma_n\delta$, where $\delta>0$ is an arbitrary constant and $\gamma_n := a_n\sup\{f(x)H(x) : c\le x\le d\}$. Now consider $\xi_n(\gamma_n,F(x);u) = \alpha_n(F(x)+\gamma_n u)-\alpha_n(F(x))$ for $x\in(c,d)$ and $0\le u\le 1$. Under the assumptions of Theorem 2.1, there exists almost surely a $\rho>1$ such that, for all n sufficiently large, $b_n/\rho\le\gamma_n\le\rho b_n$. Since $\{b_n, n\ge 1\}$ satisfies the [CRS] conditions, so does $\{\rho b_n, n\ge 1\}$. Applying Theorem A to this case, we see that, for any given $\varepsilon>0$, we have, almost surely for all large n, $\xi_n(\rho b_n,F(x);\cdot)/(2\rho b_n\log(1/\rho b_n))^{1/2} \in S^{\varepsilon/4\rho}$. Now we make the obvious remark that if $h \in S$, then the function defined by $h_\nu(s) = \nu^{-1/2}h(\nu s)$ for $0\le s\le 1$ and $0<\nu\le 1$ satisfies $h_\nu \in S$. Thus, by choosing $\nu=\gamma_n/(\rho b_n) \in [1/\rho^2,1]$, it follows that, a.s., for all large n and $x\in(c,d)$, we have $\xi_n(\gamma_n,F(x);\cdot)/(2\gamma_n\log(1/\rho b_n))^{1/2} \in S^{\varepsilon/4}$. Finally, a simple argument based on the fact that $(\log(1/\lambda_n))/\log(1/\rho b_n) \to 1$ shows that $\xi_n(\gamma_n,F(x);\cdot)/(2\gamma_n\log(1/\lambda_n))^{1/2} \in S^{\varepsilon/2}$ a.s. for all large n and $x\in(c,d)$. In other words, there exists a function $h \in S$ such that for all $0\le v\le 1$, $|(\alpha_n(F(x)+\gamma_n v)-\alpha_n(F(x)))/(2\gamma_n\log(1/\lambda_n))^{1/2}-h(v)|<\varepsilon/2$. This, in turn, implies, by choosing $\delta \le \varepsilon^2/4\rho$, that for all $0\le u\le 1$

$$|(\alpha_n(F(x+\lambda_n u)) - \alpha_n(F(x)))/(2\gamma_n\log(1/\lambda_n))^{1/2} - h(u)|<\varepsilon/2$$
$$+ \sup\{|g(u)-g(v)| : 0\le u,\ v\le 1,\ |u-v| \le \rho\delta,\ g \in S\} \le \varepsilon/2 + \sqrt{\rho\delta} \le \varepsilon.$$

This last inequality, combined with (2.12), shows that

$$\limsup_{n\to\infty} n^{1/2}\lambda_n(\tilde{f}_n(x)-E\tilde{f}_n(x))/(2\gamma_n\log(1/\lambda_n))^{1/2}$$
$$\le \sup\{-\int_0^1 f(u)d\tilde{K}(u) : f\in S\} + \varepsilon\int_0^1|d\tilde{K}(u)|, \tag{2.13}$$

where $\int_0^1|d\tilde{K}(u)|$ denotes the total variation of $\tilde{K}$. By Schwarz's inequality, we see that $|\int_0^1 f(u)d\tilde{K}(u)| = |\int_0^1 \dot{f}(u)\tilde{K}(u)du|\le\left(\int_0^1|\tilde{K}^2(u)du\right)^{1/2}$. Recalling the definition of γ_n and using the fact that $\varepsilon>0$ can be

chosen as small as desired, we see that, given any $\theta>0$, we may choose an $\eta>0$ such that, whenever $0<d-c<\eta$, we have

$$\limsup_{n\to\infty}\left\{\sup_{c\le x\le d} \pm \left(\frac{\tilde{f}_n(x)-E\tilde{f}_n(x)}{\sqrt{f(x)}}\right)\left(\frac{n\lambda_n}{2\log(1/\lambda_n)}\right)^{1/2}\right\}$$

$$\le \theta + \left(\int_0^1 \tilde{K}^2(u)du\right)^{1/2} \text{ a.s.} \tag{2.14}$$

By (2.14), it follows readily that, almost surely,

$$\limsup_{n\to\infty}\left\{\sup_{c\le x\le d} \pm\left(\frac{\tilde{f}_n(x)-E\tilde{f}_n(x)}{\sqrt{f(x)}}\right)\left(\frac{n\lambda_n}{2\log(1/\lambda_n)}\right)^{1/2}\right\}\le\left(\int_0^1 \tilde{K}^2(u)du\right)^{1/2}. \tag{2.15}$$

For the second half of (2.8), we will make use of the observation that, for any $\varepsilon>0$ and $h\in S$, there exists a finite sequence $h^{(1)},\ldots,h^{(N)}$ of functions belonging to S such that, for any $1/\rho^2\le\nu\le1$, there exists a $j=j(\nu)$ such that $\|h-h^{(j)}\| < \varepsilon/4$ (recall that $g_\nu(u) = g(\nu u)/\sqrt{\nu}$ for $0\le u\le1$). Following the same steps as those used in the proof of (2.13), we see from Theorem A that there exists an $N_\varepsilon < \infty$ such that for any $j=1,\ldots,N$, there exists an $x_{n:j}\in(c,d)$ such that

$$\|\xi_n(\rho b_n,F(x_{n:j});\cdot)/(2\rho b_n\log(1/\rho b_n))^{1/2}-h^{(j)}\|\le\varepsilon/4\rho.$$

By choosing $\nu = \gamma_n/(\rho b_n)\in[1/\rho^2,1]$ and $j=j(\nu)$, it follows that

$$\|\xi_n(\gamma_n,F(x_{n:j});\cdot)/(2\gamma_n\log(1/\rho b_n))^{1/2}-h\|\le\varepsilon/2. \tag{2.16}$$

Finally, by imposing the condition that $0<d-c<\delta$ sufficiently small, we see that the same arguments as those used in (2.13), in combination with (2.16), show that

$$\liminf_{n\to\infty} n^{1/2}\lambda_n\left(\tilde{f}_n(x_{j:n})\right)-E\tilde{f}_n(x_{j:n})/(2\gamma_n\log(1/\lambda_n))^{1/2}$$

$$\ge -\int_0^1 h(u)d\tilde{K}(u) - \varepsilon\int_0^1|d\tilde{K}(u)| \quad \text{a.s.} \tag{2.17}$$

Recalling that our assumptions imply that $\sup\{|\gamma_n/(\lambda_n f(x))-1| : c\le x\le d\}$ may be rendered arbitrarily small, by imposing the condition that $0\le d-c\le\delta$ is small enough, we see that (2.17) implies that

$$\liminf_{n\to\infty}\left\{\sup_{C\le x\le D} \pm\left(\frac{f_n(x)-Ef_n(x)}{\sqrt{f(x)}}\right)\left(\frac{n\lambda_n}{2\log(1/\lambda_n)}\right)^{1/2}\right\}$$

$$\ge -\int_0^1 h(u)dK(u) = -\int_0^1 \dot{h}(u)dK(u) \text{ a.s.} \tag{2.18}$$

Since the choice of $h \in S$ is arbitrary, we may choose

$$h(u) = K(u)/\left(\int_0^1 K^2(s)ds\right)^{1/2}.$$

For such a choice of h, the right-hand side of (2.18) is equal to $\left(\int_0^1 K^2(u)du\right)^{1/2}$. This, in combination with (2.15), completes the proof of Theorem 2.1. The proof of Theorem 2.2 is very similar and will therefore be omitted. ∎

Remark 2.1. A close examination of the proof of Theorems 2.1-2.2 just given shows that Assumptions (L.1)-(L.2) may be replaced by the assumption that $\lambda_n(n;x;X_1,\ldots,X_n)/b_n$ is almost surely bounded away from zero and infinity, uniformly over the interval $C\le x\le D$, and satisfies a uniform Lipschitz condition in this interval.

Corollary 2.1. Assume that $\lambda_n = (2\mu_n)/\hat{f}_n(x)$ is as in (2.3), where $\hat{f}_n$ is as in (1.2), and $\{\mu_n, n\ge 1\}$ satisfies the [CRS] conditions. Then

$$\lim_{n\to\infty} \sup_{C\le x\le D} \pm \left(\frac{f_n(x)-Ef_n(x)}{\sqrt{f(x)}}\right)\left(\frac{n\mu_n}{2\log(1/\mu_n)}\right)^{1/2} = \left(\int_{-\infty}^{\infty}\tilde{K}^2(u)du\right)^{1/2} \text{ a.s.} \quad (2.19)$$

Proof. It follows readily from Theorem 2.1 combined with the results of Example 2.1. ∎

Remark 2.2. Corollary 2.1 describes the limiting oscillation behavior of the plug-in estimator discussed by Moore and Yackel (1977) and Mack and Rosenblatt (1979). Similar results could be obtained by choosing for λ_n a function of a preliminary estimate of the density f and of n.

Corollary 2.2. Assume that $\lambda_n = \hat{\lambda}_n$ is as in Example 2.2, and satisfies (2.7). Then

$$\lim_{n\to\infty} \sup_{C\le x\le D} \pm \left(\frac{f_n(x)-Ef_n(x)}{\sqrt{f(x)}}\right)\left(\frac{n\lambda_n}{2\log(1/\lambda_n)}\right)^{1/2} = \left(\int_{-\infty}^{\infty}K^2(u)du\right)^{1/2} \quad (2.20)$$

in probability. Moreover, in (2.20) we may replace λ_n by $\tilde{\lambda}_n$.

Proof. It follows readily from Theorem 2.2 combined with the results of Example 2.2. Recall that $\tilde{\lambda}_n$ is defined in (1.8). ∎

Remark 2.3. Corollary 2.2 describes the limiting oscillation behavior of the classical MISE cross-validated data-driven estimator. Interestingly enough, this behavior is (up to the order given in (2.20)) asymptotically identical to that of the exact optimum MISE estimator obtained when $\lambda_n = \tilde{\lambda}_n$, and $\tilde{\lambda}_n$ is as in (1.8).

Remark 2.4. We may use the argument of Hall (1990) to show that the assumption (K.3) is not necessary for the validity of the conclusion of Theorems 2.1 and 2.2. We omit the details of this argument.

3. CONCLUSION

The preceding examples show the power of the methods based on functional LIL's for increments of empirical processes for obtaining LIL's for density estimators. In this section, we review some of the results due to Deheuvels and Mason (1990) and Deheuvels (1990) for ranges not covered by the [CRS] conditions. In the first place, we consider the case where the sequence of constants $\{a_n, n\geq 1\}$ satisfies

$$na_n/\log n \to v \in (0,\infty). \tag{S.4}$$

The following Theorem D is due to Deheuvels and Mason (1990).

Theorem D. Let $\{a_n, n\geq 1\}$ be a sequence satisfying (S.4). Denote by Δ_v the set of all absolutely continuous non-decreasing functions g on [0,1], having Lebesgue derivative g such that

$$\int_0^1 (\dot{g}(u)\log(\dot{g}(u)/v)-\dot{g}(u)+v)du \leq 1, \tag{3.1}$$

with the convention that $0\log 0 = 0$. Then, for any $\varepsilon>0$, there exists almost surely an $n_\varepsilon < \infty$ such that for all $n\geq n_\varepsilon$,

$$\{(n/\log n)(F_n(t+a_n\cdot)-F_n(t)) \ :0\leq t\leq 1-a_n\} \subset \Delta_v^\varepsilon \ . \tag{3.2}$$

Moreover, for any $g \in \Delta_v$, there exists, almost surely an $N_\varepsilon < \infty$ such that for all $n \geq N_\varepsilon$, there exists, a $0 \leq t=t_{n,\varepsilon,g} \leq 1-a_n$ such that

$$\|(n/\log n)(F_n(t+a_n\cdot) - F_n(t))-g\| < \varepsilon. \tag{3.3}$$

Proof. See Deheuvels and Mason (1990), §2. ∎

Theorem E below is due to Deheuvels (1990) and covers the case where

$$(\log(1/a_n))/\log\log n \to w\in(0,\infty). \tag{S.5}$$

Theorem E. Let $\{a_n, n\geq 1\}$ be a sequence satisfying (S.1)-(S.5). Then, for any $\varepsilon>0$, there exists almost surely an $n_\varepsilon < \infty$ such that, for all $n \geq n_\varepsilon$,

$$\{\xi_n(a_n,t;\cdot)/(2(w+1)a_n\log\log n)^{1/2} \ :0\leq t\leq 1-a_n\} \subset S^\varepsilon \ . \tag{3.4}$$

Moreover, we have infinitely often in n with probability one

$$S \subset \{\xi_n(a_n,t;\cdot)/(2wa_n\log\log n)^{1/2} \ :0\leq t\leq 1-a_n\}^\varepsilon \ . \tag{3.5}$$

Proof. See Deheuvels (1990). ∎

The range covered by Theorem D corresponds typically to bandwidths for which the density estimators are not uniformly consistent. On the other hand, Theorem E corresponds to very large bandwidths (of the order of a negative power of logn), which are not likely to be

used in statistical practice. Therefore, both theorems correspond to boundary situations outside the range of interest which is covered by the [CRS] conditions. For this reason, we will not state here the versions of Theorems 2.1 and 2.2 which correspond to these theorems. They have, however, the advantage of giving a complete description of the increment of the uniform empirical process when combined with Theorem A.

Similar results, to be published in the near future, enable us to treat the almost sure pointwise behavior of density estimators. Also, extensions of these results to higher dimensions will be provided in forthcoming publications.

4. REFERENCES

Deheuvels, P. (1990). 'Functional laws of the iterated logarithm for large increments of empirical and quantile processes', submitted.

Deheuvels, P. and Mason, D.M. (1990). 'Functional laws of the iterated logarithm for the increments of empirical and quantile processes', submitted.

Devroye, L. and Györfi, L. (1984). 'Nonparametric Density Estimation: The L_1 View', Wiley, New York.

Devroye, L. and Wagner, T.J. (1977). 'The strong uniform consistency of nearest neighbor density esimates.' Annals of Statistics 5, 536-540.

Fix, E. and Hodges, J.L. (1951). 'Discriminatory analysis. Non-parametric discrimination consistency properties.' Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

Hall, P. (1990). 'On the law of the iterated logarithm for density estimators.' Statistics and Probability Letters, 9, 237-240.

Hall, P. and Marron, J.S. (1987). 'Extent to which least-squares cross-validation minimizes integrated square error in nonparametric density estimation'. Probability Theory and Related Fields 74, 567-581.

Loftsgaarden, D.O. and Quesenberry, C.P. (1965). 'A nonparametric estimate of a multivariate density function.' Annals of Mathematical Statistics 36, 1049-1051.

Mack, Y.P. (1983). "Rate of strong uniform convergence of k-nn density estimates.' Journal of Statistical Planning and Inference 8, 185-192.

Mack, Y.P. and Rosenblatt, M. (1979). 'Multivariate k-nearest neighbor density estimates.' Journal of Multivariate Analysis 9, 1-15.

Moore, D.S. and Yackel, J.W. (1977). 'Consistency properties of nearest neighbor density estimates.' Annals of Statistics 5, 143-154.

Park, B. and Marron, J.S. (1988). 'Comparison of data-driven bandwidth selectors.' Technical Report # 1759, University of North Carolina, Chapel Hill.

Parzen, E. (1962). 'On the estimation of a probability density function and mode,' Annals of Mathematical Statistics 33, 1065-1076.

Rosenblatt, M. (1956). "Remarks on some nonparametric estimates of a density function.' Annals of Mathematical Statistics 27, 832-837.

Strassen, V. (1964). 'An invariance principle for the law of the iterated logarithm.' Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 3, 211-226.

Stute, W. (1982). 'A law of the iterated logarithm for kernel density estimators.' Annals of Probability 10, 414-422.

EXPONENTIAL INEQUALITIES IN NONPARAMETRIC ESTIMATION

Luc DEVROYE
Division of Statistics
University of California at Davis
Davis, CA. 95616 USA

ABSTRACT. We derive exponential inequalities for the oscillation of functions of random variables about their mean. This is illustrated on the Kolmogorov-Smirnov statistic, the total variation distance for empirical measures, the Vapnik-Chervonenkis distance, and various performance criteria in nonparametric density estimation. We also derive bounds for the variances of these quantities.

1. Introduction.

Hoeffding (1963) showed that for independent random variables $X_1, X_2, \ldots, X_n$ with $a_i \leq X_i \leq b_i$,

$$\mathbf{P}\left\{\left|\sum_{i=1}^{n}(X_i - \mathbf{E}X_i)\right| > t\right\} \leq 2e^{-2t^2/\sum_{i=1}^{n}(b_i-a_i)^2}\ ,\ t > 0.$$

Perhaps the best known form of this inequality is obtained when the X_i's are i.i.d. Bernoulli (p) random variables. In that case, we obtain Chernoff's bound (1952) for the binomial distribution: if X is binomial (n, p), then

$$\mathbf{P}\left\{|X - np| > t\right\} \leq 2e^{-2t^2/n}\ ,\ t > 0.$$

Various extensions of these inequalities have been developed over the years. The generalization to martingales due to Hoeffding (1963) and Azuma (1967) has led to interesting applications in combinatorics and the theory of random graphs (for a survey, see McDiarmid, 1989). We have used it in density estimation (Devroye, 1988, 1989).

In this paper, we collect various extensions of Hoeffding's inequality and highlight their applications in the nonparametric estimation of densities and distribution functions. For completeness, the proofs of the inequalities are sketched as well. In the last section, we present new bounds for the variance of functions of independent random variables. The inequalities are illustrated on examples in nonparametric estimation, and are shown to be sharper than those obtained from the Efron-Stein inequality.

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 31–44.
© 1991 *Kluwer Academic Publishers.*

2. Inequalities for martingale difference sequences.

Hoeffding (1963) mentioned that his inequalities would also be valid when applied to martingale difference sequences. To make things a bit more precise, let us consider a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and a nested sequence $\mathcal{F}_0 = \{\emptyset, \Omega\} \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}$ of sub-σ-fields of $\mathcal{F}$. A sequence of integrable random variables $X_0, X_1, X_2, \ldots$ is a *martingale* if

$$\mathbf{E}\{X_{n+1} \mid \mathcal{F}_n\} = X_n \quad \text{a.s.}, \quad \text{each } n \geq 0.$$

A sequence of integrable random variables $Y_1, Y_2, \ldots$ is a *martingale difference sequence* if for every $n \geq 0$,

$$\mathbf{E}\{Y_{n+1} \mid \mathcal{F}_n\} = 0 \quad \text{a.s.}.$$

Note that any martingale $X_0, X_1, X_2, \ldots$ leads to a natural martingale difference sequence by defining

$$Y_n = X_n - X_{n-1}, \quad n \geq 1.$$

And any martingale difference sequence $Y_1, Y_2, \ldots$ in turn yields a natural martingale by defining X_0 in an arbitrary fashion and setting

$$X_n = \sum_{i=1}^{n} Y_i + X_0.$$

For any nested sequence of sub-σ-fileds $\mathcal{F}_0 = \{\emptyset, \Omega\} \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \cdots \subseteq \mathcal{F}$ and any integrable random variable X, we can define *Doob martingale* by setting

$$X_n = \mathbf{E}\{X \mid \mathcal{F}_n\}, \quad n \geq 0.$$

Thus, $X_0 = \mathbf{E}X$, and if X is $\mathcal{F}_n$-measurable, then $X_n = X$, and

$$X - \mathbf{E}X = \sum_{i=1}^{n} (X_i - X_{i-1}).$$

We begin with an inequality along the lines suggested by Hoeffding (1963) and Azuma (1967) (see McDiarmid, 1989):

THEOREM 1. *Let $\mathcal{F}_0 = \{\emptyset, \Omega\} \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \cdots \subseteq \mathcal{F}_n$ be a nested sequence of σ-fields. Let the integrable random variable X be $\mathcal{F}_n$-measurable, and define the Doob martingale $X_k = \mathbf{E}\{X \mid \mathcal{F}_k\}$. Assume that for $k = 1, \ldots n$, there exist random variables Z_k, $\mathcal{F}_{k-1}$-measurable, and constants c_k such that*

$$Z_k \leq X_k \leq Z_k + c_k.$$

Then for $t > 0$,

$$\mathbf{P}\{X - \mathbf{E}X \geq t\} \leq 2e^{-2t^2/\sum_{i=1}^{n} c_i^2}, \quad t > 0,$$
$$\mathbf{P}\{X - \mathbf{E}X \leq -t\} \leq 2e^{-2t^2/\sum_{i=1}^{n} c_i^2}, \quad t > 0.$$

This Theorem uses a simple Lemma due to Hoeffding (1963):

LEMMA 1. *Let X be a random variable with $\mathbf{E}X = 0$, $a \le X \le b$. Then for $\lambda > 0$,*

$$\mathbf{E}\left\{e^{\lambda X}\right\} \le e^{\lambda^2(b-a)^2/8}.$$

PROOF. Note that by convexity,

$$\begin{aligned} e^{\lambda x} &\le \frac{x-a}{b-a}e^{\lambda b} + \frac{b-x}{b-a}e^{\lambda a}\ , \quad a \le x \le b\ , \\ \mathbf{E}e^{\lambda X} &\le \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} \\ &= (1-p+pe^{\lambda(b-a)})e^{-p\lambda(b-a)}\ , \text{ where } p = \frac{a}{a-b} \in [0,1] \\ &\stackrel{\text{def}}{=} e^{\varphi(u)}\ , \end{aligned}$$

where $u = \lambda(b-a)$, $\varphi(u) = -pu + \log(1-p+pe^u)$. But it is easy to see that

$$\begin{aligned} \varphi'(u) &= -p + \frac{p}{p+(1-p)e^{-u}}\ , \\ \varphi''(u) &= \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \le \frac{1}{4}\ , \\ \varphi(0) &= \varphi'(0) = 0\ , \end{aligned}$$

and by Taylor's series expansion with remainder,

$$\varphi(u) \le \frac{u^2}{8} = \frac{\lambda^2(b-a)^2}{8}\ . \blacksquare$$

PROOF. Set $Y_k = X_k - X_{k-1}$, $S_k = \sum_{i=1}^k Y_i = X_k - X_0$. Note that $S_n = X_n - \mathbf{E}X = X - \mathbf{E}X$. Also,

$$\begin{aligned} \mathbf{P}\left\{X - \mathbf{E}X \ge t\right\} &= \mathbf{P}\left\{S_n \ge t\right\} \\ &\le e^{-\lambda t}\mathbf{E}\left\{e^{\lambda S_n}\right\} \quad \text{for} \lambda > 0 \quad \text{(by Chernoff's bounding method)} \\ &= e^{-\lambda t}\mathbf{E}\left\{e^{\lambda S_{n-1}}\mathbf{E}\left\{e^{\lambda Y_n}|\mathcal{F}_{n-1}\right\}\right\} \\ &\le e^{-\lambda t}\mathbf{E}\left\{e^{\lambda S_{n-1}}\right\}e^{\lambda^2 c_n^2/8} \quad \text{(Lemma 1)} \\ &\le e^{-\lambda t}e^{(\lambda^2/8)\sum_{i=1}^n c_i^2} \quad \text{(iterate previous argument)} \\ &= e^{-2t^2/\sum_{i=1}^n c_i^2} \quad \text{(take } \lambda = 4t/\sum_{i=1}^n c_i^2\text{)}\ . \end{aligned}$$

The second inequality in Theorem 1 is obtained when we replace X by $-X$. $\blacksquare$

3. McDiarmid's extension of Hoeffding's inequality.

The following extension of Hoeffding's inequality is useful for random variables that are complicated functions of independent random variables, and that are relatively robust to individual changes in the values of the random variables.

THEOREM 2. *(McDiarmid, 1989) Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to R$ satisfies*

$$\sup_{\substack{x_1,\ldots,x_n \\ x'_1,\ldots,x'_n \in A}} |f(x_1,\ldots,x_n) - f(x_1,\ldots,x_{i-1},x'_i,x_{i+1},\ldots,x_n)| \le c_i \ , \ 1 \le i \le n \ .$$

Then

$$\mathbf{P}\left\{|f(X_1,\ldots,X_n) - \mathbf{E}f(X_1,\ldots,X_n)| \ge t\right\} \le 2e^{-2t^2/\sum_{i=1}^n c_i^2} \ .$$

PROOF. Define $Y = Y_n = f(X_1,\ldots,X_n)$, and let $\mathcal{F}_n$ be the σ-field generated by $X_1,\ldots,X_n$. Define

$$Y_k = \mathbf{E}\left\{Y|\mathcal{F}_k\right\} \ , \ Z_k = \operatorname{ess\,inf}\left\{Y_k|\mathcal{F}_{k-1}\right\} \ , \ W_k = \operatorname{ess\,sup}\left\{Y_k|\mathcal{F}_{k-1}\right\} \ ,$$

so that $Z_k \le Y_k \le W_k$. We can apply Theorem 1 directly to Y_n if we can show that $W_k - Z_k \le c_k$. But this follows from

$$\begin{aligned}
W_k &= \operatorname{ess\,sup}\left\{\mathbf{E}\left\{f(X_1,\ldots,X_n)|\mathcal{F}_k\right\}|\mathcal{F}_{k-1}\right\} \\
&\le \sup_{x\in A} \mathbf{E}\left\{f(X_1,\ldots,X_{k-1},x,X_{k+1},\ldots,X_n)|\mathcal{F}_k\right\} \\
&= \sup_{x\in A} \mathbf{E}\left\{f(X_1,\ldots,X_{k-1},x,X_{k+1},\ldots,X_n)|\mathcal{F}_{k-1}\right\} \\
&\le \inf_{x\in A} \mathbf{E}\left\{f(X_1,\ldots,X_{k-1},x,X_{k+1},\ldots,X_n)|\mathcal{F}_{k-1}\right\} + c_k \\
&\le \operatorname{ess\,inf}\left\{\mathbf{E}\left\{f(X_1,\ldots,X_n)|\mathcal{F}_k\right\}|\mathcal{F}_{k-1}\right\} + c_k \\
&= Z_k + c_k \ . \ \blacksquare
\end{aligned}$$

4. Applications.

4.1. CHERNOFF'S BOUND.

Let $X_1,\ldots,X_n$ be Bernoulli (p_1) ,..., Bernoulli (p_n) random variables, and consider $\sum_{i=1}^n X_i$. Clearly, the conditions of Theorem 2 are fulfilled with $c_i \equiv 1$. Thus,

$$\mathbf{P}\left\{\left|\sum_{i=1}^n (X_i - p_i)\right| \ge t\right\} \le 2e^{-2t^2/n} \ .$$

As a special case, $p_i \equiv p$ for all i, we obtain Chernoff's bound (1952) for a binomial (n,p) random variable X:

$$\mathbf{P}\left\{|X - \mathbf{E}X| \ge t\right\} \le 2e^{-2t^2/n} \ .$$

4.2. The Kolmogorov-Smirnov statistic.

Let F_n be the standard empirical distribution function based upon an i.i.d. sample $X_1, \ldots, X_n$ drawn from a distribution function F on the real line. The Kolmogorov-Smirnov distance is

$$\sup_x |F_n(x) - F(x)| \ .$$

Note that changing one point X_i can increase or decrease F_n by at most $1/n$ over a certain interval. Thus the condition of Theorem 2 is fulfilled with $c_i \equiv 1/n$. We have

$$\mathbf{P}\left\{\left|\sup_x |F_n(x) - F(x)| - \mathbf{E}\sup_x |F_n(x) - F(x)|\right| \geq \frac{t}{\sqrt{n}}\right\} \leq 2e^{-2t^2} \ .$$

In this respect, we note that Dvoretzky, Kiefer and Wolfowitz (1956) showed that

$$\mathbf{P}\left\{\sup_x |F_n(x) - F(x)| \geq \frac{t}{\sqrt{n}}\right\} \leq Ce^{-2t^2}$$

for some $C > 0$, while Massart (1990) proved that one can take $C = 2$. Massart's bound and the inequality derived from Theorem 2 do not imply each other.

4.3. Nonparametric density estimation: the L_1 norm.

Let f_n be the Parzen-Rosenblatt kernel estimate of a density f based upon an i.i.d. sample $X_1, \ldots, X_n$ drawn from f (Rosenblatt, 1956; Parzen, 1962):

$$f_n(x) = f_n(x; X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) \ .$$

Here K is a given function (the kernel) integrating to one, $K_h(u) = \frac{1}{h}K(\frac{u}{h})$, and $h > 0$ is a smoothing factor. An important criterion for evaluating the performance of a density estimate is $\int |f_n - f|$. This random variable satisfies the conditions of Theorem 2 with $c_i \equiv 2\int |K|/n$ as we will now see. Take any numbers $x_1, \ldots, x_n$ and $x'_1, \ldots, x'_n$ with $x'_i = x_i$ except for $i = j$. Then,

$$\begin{aligned}
\left|\int |f_n(x; x_1, \ldots, x_n) - f(x)|\, dx - \int |f_n(x; x'_1, \ldots, x'_n) - f(x)|\, dx\right| & \\
&\leq \int |f_n(x; x_1, \ldots, x_n) - f_n(x; x'_1, \ldots, x'_n)|\, dx \\
&\leq \frac{1}{n}\int \left|K_h(x - x_j) - K_h(x - x'_j)\right|\, dx \\
&\leq \frac{2\int |K|}{n} \ .
\end{aligned}$$

Thus, dropping the $(.)$ and dx, we have

$$\mathbf{P}\left\{\left|\int |f_n - f| - \mathbf{E}\int |f_n - f|\right| > t\right\} \leq 2e^{-nt^2/2\int^2 |K|} \ .$$

This inequality improves over one first published in Devroye (1988), where the exponent had the constant 32 instead of 2. The present improvement was independently pointed out to me by Pinelis (1990).

REMARK 1. We recall that $\sqrt{n}\mathbf{E}\int|f_n - f| \to \infty$ for the kernel estimate when one of these conditions holds:

(1) $\lim_{n\to\infty} h = 0$;
(2) The characteristic function for f has unbounded support;
(3) $\int\sqrt{f} = \infty$.

See e.g. Devroye and Györfi (1985) or Devroye (1988). When this is the case, a simple application of Chebyshev's inequality shows that

$$\frac{\int|f_n - f|}{\mathbf{E}\int|f_n - f|} \to 1 \quad \text{in probability.}$$

In other words, the L_1 eror behaves asymptotically like a deterministic sequence, just as averages do when the weak law of large numbers applies.

REMARK 2. For the standard histogram estimate, regardless of the bin width, we have

$$\mathbf{P}\left\{\left|\int|f_n - f| - \mathbf{E}\int|f_n - f|\right| > t\right\} \le 2e^{-nt^2/2} .$$

Thus, just as for the kernel estimate, we have an inequality that is valid for all f and n, and for all choices of the bin widths and the smoothing factors. The non-asymptotic character of the inequalities will undoubtedly make them useful tools for further applications.

REMARK 3. By the boundedness of $\int|f_n - f|$, we note that $\int|f_n - f| \to 0$ in probability if and only if $\mathbf{E}\int|f_n - f| \to 0$. But if these quantities tend to zero in the indicated senses, by the results of this section, for every $\epsilon > 0$ and $t > 0$, it is possible to find n_0 such that for $n > n_0$,

$$\mathbf{P}\left\{\int|f_n - f| > t\right\} \le 2e^{-n(1-\epsilon)t^2/2\int^2|K|} .$$

Thus, weak convergence of the L_1 error implies complete convergence. This observation is at the basis of the equivalence results of Devroye (1983), but the present proof is much shorter. We note that a sufficient condition for the weak (and thus complete) convergence for the kernel estimate is that $h \to 0$ and $nh \to \infty$ as $n \to \infty$ (see Devroye, 1987, where a proof is given based upon results of Scheffé (1947) and Glick (1974)). When the kernel K has at least one non-vanishing moment, then these conditions are necessary as well.

REMARK 4. The condition in Theorem 2 demands that each individual sample point have a limited influence on $\int|f_n - f|$. This may not be the case for some data-based methods of choosing h.

4.4. L_p NORMS.

Define the L_p norm of g by $\|g\|_p = (\int |g|^p)^{1/p}$, where $p \geq 1$ is fixed. If f_n is shorthand for $f_n(x; x_1, \ldots, x_n)$, $g_n \equiv f_n(x; x'_1, \ldots, x'_n)$ and if $x_1, \ldots, x_n$ and $x'_1, \ldots, x'_n$ are sequences of numbers with $x_i = x'_i$ except for $i = j$, then

$$\begin{aligned}\left| \|f_n - f\|_p - \|g_n - f\|_p \right| &\leq \|f_n - g_n\|_p \\ &\leq \left\| \frac{1}{nh} K_h(x - x_j) - \frac{1}{nh} K_h(x - x'_j) \right\|_p \\ &\leq \frac{2}{nh^{1-1/p} \|K\|_p}\end{aligned}$$

by Minkowski's inequality. From Theorem 2, we then deduce the following inequality:

$$\mathbf{P}\left\{ \left| \|f_n - f\|_p - \mathbf{E}\,\|f_n - f\|_p \right| \geq t \right\} \leq 2e^{-nt^2 h^{2-2/p}/(2\|K\|_p^2)} .$$

The inequality remains formally valid even if $p = \infty$, in which case we obtain the supremum norm.

Assume for the sake of simplicity that K is a bona fide density. We claim that the relative stability result, i.e.,

$$\frac{\|f_n - f\|_p}{\mathbf{E}\,\|f_n - f\|_p} \to 1 \quad \text{in probability,}$$

holds whenever $h \to 0$, $nh \to \infty$ and $1 \leq p < 2$. Of course, for the norms to make sense, we have to additionally assume that $K, f \in L_p$. Assume for simplicity that $K \geq 0$. To prove the claim, we first havve to establish that for any density f, there exists a constant $c > 0$ such that

$$\mathbf{E}\,\|f_n - f\|_p \geq c \max\left(h^2, 1/\sqrt{nh}\right) .$$

Under smoothness and tail conditions on f, this result is rather standard. The generalization to all f requires some work. Back to the statement. It clearly suffices to show that the variance of $\|f_n - f\|_p$ is $o(\mathbf{E}^2\,\|f_n - f\|_p)$ by Chebyshev's inequality. The variance is $O(1/(\sqrt{n}h^{1-1/p}))$ — this follows from the exponential bound shown above. If $h \leq n^{-1/5}$, then the statement is easily verified since $h^{\frac{1}{2}-\frac{1}{p}} \to \infty$. If $h \geq n^{-1/5}$, then we need only verify that $\sqrt{n}h^{3-1/p} \to \infty$.

Interestingly, the first p for which the relative stability result fails is $p = 2$. We can only obtain it from the inequality shown above when $nh^5 \to \infty$, a condition that is known to yield suboptimal values for h for all densities (not just all smooth ones!). However, this does not mean that the relative stability result is not valid in L_2 for $h \sim n^{-1/5}$. Indeed, Hall (1982) proved that

$$\frac{\|f_n - f\|_2^2}{\mathbf{E}\,\|f_n - f\|_2^2} \to 1 \quad \text{in probability,}$$

under certain conditions on f, h and K.

4.5. Uniform deviation of empirical measures.

An i.i.d. sample$X_1, \ldots, X_n$ with common probability measure μ on the Borel sets $\mathcal{B}$ of R^d induces an empirical probability measure μ_n by

$$\mu_n(B) = \frac{1}{n} \sum_{i=1}^{n} I_B(X_i) \ ,$$

where I is the indicator function, and $B \in \mathcal{B}$. The total variation distance between μ_n and μ is

$$T_n \stackrel{\text{def}}{=} \sup_{B \in \mathcal{B}} |\mu_n(B) - \mu(B)| \ .$$

Clearly, $T_n \equiv 1$ if μ is nonatomic, so the total variation distance is rather restrictive. Vapnik and Chervonenkis (1971) considered instead

$$V_n \stackrel{\text{def}}{=} \sup_{A \in \mathcal{A}} |\mu_n(B) - \mu(B)|$$

where $\mathcal{A}$ is a suitable subclass of the Borel sets. For example, if $\mathcal{A} = \{(-\infty, x] : x \in R\}$ and $d = 1$, then V_n is the standard Kolmogorov-Smirnov distance discussed above. They showed in particular that

$$\mathbf{P}\{V_n \geq t\} \leq 4s(\mathcal{A}, 2n)e^{-nt^2/8} \ , \ nt^2 \geq 1,$$

where

$$s(\mathcal{A}, n) = \max_{x_1, \ldots, x_n) \in R^{dn}} N_{\mathcal{A}}(x_1, \ldots, x_n)$$

and $N_{\mathcal{A}}(x_1, \ldots, x_n)$ is the number of different sets in

$$\{\{x_1, \ldots, x_n\} \cap A \mid A \in \mathcal{A}\} \ .$$

For many families $\mathcal{A}$ that are not too rich, such as all halfspaces, all intersections of a finite number of halfspaces, all balls, etc., $s(\mathcal{A}, n) \leq n^D$ for a finite D (the "Vapnik-Chervonenkis dimension"), so that the Vapnik-Chervonenkis bound decreases exponentially with n. For extensions and discussions, see Devroye (1982), Gaenssler (1983), Pollard (1984) and Alexander (1984).

If we replace X_j in the sample by X_j' while holding all the other elements fixed, V_n changes by at most $1/n$, so that from Theorem 2,

$$\mathbf{P}\{|V_n - \mathbf{E}V_n| \geq t\} \leq 2e^{-2nt^2} \ .$$

This implies that $\sqrt{n}(V_n - \mathbf{E}V_n) = O(1)$ in probability. For the limit law theory of $\sqrt{n}V_n$, we refer to Dudley (1978).

4.6. Variable kernel estimates.

Breiman, Meisel and Purcell (1977) introduced the variable kernel estimate

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{H_i}(x - X_i) ,$$

where H_i is a function of X_i and the data, $X_1, \ldots, X_n$. If H_i is a function of X_i and n only, then the inequality of section 4.3 applies unaltered.

A more interesting choice of H_i is that in which it becomes a function of the distance between X_i and its k-th nearest neighbor among the data. Replacing X_i by X_i' affects at most ck of the H_j's, where c is some universal constant depending upon the dimension only. This is seen by noting that X_i can be among the k nearest neighbors of at most $c'k$ of the X_j's, where c' depends upon d only (Devroye and Penrod, 1986). Thus, $\int |f_n - f|$ changes by at most ck/n. Hence,

$$\mathbf{P}\left\{\left|\int |f_n - f| - \mathbf{E}\int |f_n - f|\right| > t\right\} \leq 2e^{-2nt^2/(c^2k^2)} .$$

Thus, $\mathrm{Var}\{\int |f_n - f|\} = O(k^2/n)$. Depending upon the choice of k, this can be used to establish the relative stability of the estimate.

4.7. Recursive kernel density estimates.

The bound of section 4.2 remains valid for density estimates of the form

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h_i}(x - X_i) ,$$

where h_i depends upon i only; this is an estimate attributed to Wolverton and Wagner (1969). Consider estimates of the form

$$f_n(x) = \sum_{i=1}^{n} p_i K_{h_i}(x - X_i) ,$$

where $(p_1, \ldots, p_n)$ is a probability weight vector, and both p_i and h_i depend upon i and possibly n. This general form of estimate goes back to Deheuvels (1973). The condition of Theorem 2 holds with $c_i \equiv 2p_i \int |K|$. Clearly,

$$\sum_{i=1}^{n} c_i^2 = 4\sum_{i=1}^{n} p_i^2 \left(\int |K|\right)^2 \leq 4 \max_{1\leq i\leq n} p_i \left(\int |K|\right)^2 .$$

This in turn can be used in the upper bound of Theorem 2.

Deheuvels (1973) proposed the latter estimate, based upon a fixed sequence $h_1, h_2, \ldots$, with $p_j = h_j / \sum_{i=1}^{n} h_i$. Assume furthermore that h_n oscillates so slightly that $\sum_{i=1}^{n} h_i^2 \leq Anh_n^2$ and $\sum_{i=1}^{n} h_i \geq Bnh_n$ for some positive constants A, B. Then we see that upon proper substitution of the various bounds,

$$\mathbf{P}\left\{\left|\int |f_n - f| - \mathbf{E}\int |f_n - f|\right| > t\right\} \leq 2e^{-Bnt^2/(2A\int^2 |K|)} .$$

Deheuvels' estimate is relatively stable for all f, whenever $h \to 0$, $nh \to \infty$, K has at least one non-vanishing moment, and h satisfies the regularity condition mentioned above.

5. Inequalities for the variance.

In many applications, one would like to obtain information about the variance, or the oscillation, of a random variable of the form $f(X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are i.i.d. random vectors. Often, $f(\cdot)$ is a rather complicated function of the data (see the examples in the previous section). One of the first general tools in this respect is the Efron-Stein inequality (Efron and Stein, 1981; see also Vitale, 1984).

THE EFRON-STEIN INEQUALITY. *Let f be a symmetric function of its n arguments, and let $X_1, \ldots, X_{n+1}$ be i.i.d. random vectors. Define*

$$S_i = f(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{n+1}) \ ,$$
$$S = S_{n+1} = f(X_1, \ldots, X_n) \ ,$$

and

$$\overline{S} = \frac{1}{n+1} \sum_{i=1}^{n+1} S_i \ .$$

Then

$$\mathrm{Var}\{S\} \le \sum_{i=1}^{n+1} \mathbf{E}\left\{\left(S_i - \overline{S}\right)^2\right\} = (n+1)\mathbf{E}\left\{\left(S - \overline{S}\right)^2\right\} \ .$$

When the right-hand-side of the inequality is worked out, and some further bounding is used, we obtain the following result:

$$\begin{aligned}
\mathrm{Var}\{S\} &\le (n+1)\mathbf{E}\left\{\left(\frac{1}{n+1}\right)^2 \left(\sum_{i=1}^{n+1}(S - S_i)\right)^2\right\} \\
&\le \frac{1}{n+1}\mathbf{E}\left\{(n+1)\sum_{i=1}^{n+1}(S - S_i)^2\right\} \\
&= \sum_{i=1}^{n+1}\mathbf{E}\left\{(S - S_i)^2\right\} \\
&= \sum_{i=1}^{n}\mathbf{E}\left\{(S - S_i)^2\right\} \\
&= n\mathbf{E}\left\{(S_2 - S_1)^2\right\} \ .
\end{aligned}$$

Assume next that a condition similar to that of Theorem 2 holds:

$$\sup_{\substack{x_1,\ldots,x_n \\ x'_1,\ldots,x'_n \in A}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n)| \le c \ , \ 1 \le i \le n \ . \qquad (*)$$

Then $|S_2 - S_1| \le c$, and thus, $\mathrm{Var}\{S\} \le nc^2$.

The derivation given here is that of Devroye (1987), where it was used to show that for the kernel estimate, regardless of the choice of h or the nature of f,

$$\mathrm{Var}\left\{\int |f_n - f|\right\} \leq \frac{4\int^2 |K|}{n} .$$

In 1986, Steele obtained a related inequality:

STEELE'S INEQUALITY. *Let f be an arbitrary function of its n arguments, and let $X_1, \ldots, X_n, X_1', \ldots, X_n'$ be i.i.d. random vectors. Define*

$$S_i = f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n) ,$$

and

$$S = f(X_1, \ldots, X_n) .$$

Then

$$\mathrm{Var}\{S\} \leq \frac{1}{2}\sum_{i=1}^{n} \mathbf{E}\left\{(S - S_i)^2\right\} .$$

Note that the symmetry of f is no longer a requirement. Also, under condition (*), $\mathrm{Var}\{S\} \leq nc^2/2$. This yields an improvement by a factor of 2 over the Efron-Stein based bound. It is possible to improve these results even further, as in Theorem 3 below.

THEOREM 3. *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to R$ satisfies*

$$\sup_{\substack{x_1,\ldots,x_n \\ x_1',\ldots,x_n' \in A}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i , \ 1 \leq i \leq n .$$

Then

$$\mathrm{Var}\{f(X_1, \ldots, X_n)\} \leq \frac{1}{4}\sum_{i=1}^{n} c_i^2 .$$

PROOF. Let $\mathcal{F}_i$ be the σ-algebra generated by $X_1, \ldots, X_i$. Let $Y = Y_n = f(X_1, \ldots, X_n)$. Then $Y_i = \mathbf{E}\{Y \mid \mathcal{F}_i\}$ forms a Doob martingale process. We formally set, as usual, $\mathcal{F}_0 = \{\emptyset, \Omega\}$, so that $Y_0 = \mathbf{E}Y$. Thus,

$$\begin{aligned}
\mathrm{Var}\{Y\} &= \mathbf{E}\left\{(Y - Y_0)^2\right\} \\
&= \mathbf{E}\left\{\left(\sum_{i=1}^{n}(Y_i - Y_{i-1})\right)^2\right\} \\
&= \mathbf{E}\left\{\sum_{i=1}^{n}(Y_i - Y_{i-1})^2\right\} + 2\sum_{1\leq i<j\leq n} \mathbf{E}\left\{(Y_i - Y_{i-1})(Y_j - Y_{j-1})\right\} \\
&= \mathbf{E}\left\{\sum_{i=1}^{n}(Y_i - Y_{i-1})^2\right\} ,
\end{aligned}$$

where we used the martingale property to show that the cross product terms are zero: for $i < j$, we have

$$\begin{aligned}&\mathbf{E}\{(Y_i - Y_{i-1})(Y_j - Y_{j-1}) \mid \mathcal{F}_{j-1}\}\\ &\quad = (Y_i - Y_{i-1})\left(\mathbf{E}\{Y_j|\mathcal{F}_{j-1}\} - Y_{j-1}\right) = 0 \quad \text{almost surely.}\end{aligned}$$

Theorem 3 follows from the above result if we can show that

$$\mathbf{E}\{(Y_i - Y_{i-1})^2|\mathcal{F}_{i-1}\} \le c_i^2/4 \ .$$

To see this, we observe that if

$$Z_i = \operatorname{ess\,inf}\{Y_i - Y_{i-1}|\mathcal{F}_{i-1}\} \ , \ W_i = \operatorname{ess\,sup}\{Y_i - Y_{i-1}|\mathcal{F}_{i-1}\} \ ,$$

then, as shown in the proof of Theorem 2, $W_i \le Z_i + c_i$, and thus, given $\mathcal{F}_{i-1}$, $Y_i - Y_{i-1}$ is a zero mean random variable taking values in the set $[Z_i, Z_i + c_i]$. But an arbitary random variable X taking values in a set $[a, b]$ has variance not exceeding $\mathbf{E}(X - (a+b)/2)^2 \le (b-a)^2/4$, so that

$$\mathbf{E}\{(Y_i - Y_{i-1})^2|\mathcal{F}_{i-1}\} \le c_i^2/4 \ .$$

This concludes the proof of Theorem 3. ∎

REMARK 1. For the kernel estimate, we obtain

$$\operatorname{Var}\left\{\int |f_n - f|\right\} \le \frac{\int^2 |K|}{n} \ ,$$

which is an improvement by a factor of 4 over the inequality shown in Devroye (1987), which was based upon the Efron-Stein bound. This improvement was suggested to me by Pinelis (1990), who mentions a range of inequalities in a much more general framework.

REMARK 2. For the histogram estimate, we obtain

$$\operatorname{Var}\left\{\int |f_n - f|\right\} \le \frac{1}{n} \ ,$$

REMARK 3. Without further work, we also have

$$\operatorname{Var}\left\{\sup_x |F_n(x) - F(x)|\right\} \le \frac{1}{4n}$$

for the Kolmogorov-Smirnov distance. Similarly, borrowing the notation of section 4.5, we have

$$\operatorname{Var}\left\{\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right\} \le \frac{1}{4n} \ .$$

6. Acknowledgement.

The author's research was supported by NSERC grant A3456.

7. References.

K. S. Alexander, "Probability inequalities for empirical processes and a law of the iterated logarithm," *Annals of Probability*, vol. 12, pp. 1041–1067, 1984.

K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal*, vol. 37, pp. 357–367, 1967.

L. Breiman, W. Meisel, and E. Purcell, "Variable kernel estimates of multivariate densities," *Technometrics*, vol. 19, pp. 135–144, 1977.

H. Chernoff, "A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations," *Annals of Mathematical Statistics*, vol. 23, pp. 493–507, 1952.

P. Deheuvels, "Sur une famille d'estimateurs de la densité d'une variable aléatoire," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 276, pp. 1013–1015, 1973.

P. Deheuvels, "Sur l'estimation séquentielle de la densité," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 276, pp. 1119–1121, 1973.

L. Devroye, "Bounds for the uniform deviation of empirical measures," *Journal of Multivariate Analysis*, vol. 12, pp. 72–79, 1982.

L. Devroye, "The equivalence of weak, strong and complete convergence in L1 for kernel density estimates," *Annals of Statistics*, vol. 11, pp. 896–904, 1983.

L. Devroye, *A Course in Density Estimation*, Birkhäuser, Boston, 1987.

L. Devroye, "An application of the Efron-Stein inequality in density estimation," *Annals of Statistics*, vol. 15, pp. 1317–1320, 1987.

L. Devroye, "Asymptotic performance bounds for the kernel estimate," *Annals of Statistics*, vol. 16, pp. 1162–1179, 1988.

L. Devroye, "The kernel estimate is relatively stable," *Probability Theory and Related Fields*, vol. 77, pp. 521–536, 1988.

L. Devroye, "The double kernel method in density estimation," *Annales de l'Institut Henri Poincaré*, vol. 25, pp. 533–580, 1989.

L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L1 View*, John Wiley, New York, 1985.

L. Devroye and C. S. Penrod, "The strong uniform convergence of multivariate variable kernel estimates," *Canadian Journal of Statistics*, vol. 14, pp. 211–219, 1986.

R. M. Dudley, "Central limit theorems for empirical measures," *Annals of Probability*, vol. 6, pp. 899–929, 1978.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of a sample distribution function and of the classical multinomial estimator," *Annals of Mathematical Statistics*, vol. 33, pp. 642–669, 1956.

B. Efron and C. Stein, "The jackknife estimate of variance," *Annals of Statistics*, vol. 9, pp. 586–596, 1981.

P. Gaenssler, *Empirical Processes*, Lecture Notes-Monograph Series, vol. 3, Institute of Mathematical Statistics, Hayward, CA., 1983.

N. Glick, "Sample-based classification procedures related to empiric distributions," *IEEE Transactions on Information Theory*, vol. IT-22, pp. 454–461, 1976.

P. Hall, "Limit theorems for stochastic measures of the accuracy of density estimators," *Stochastic Processes and Applications*, vol. 13, pp. 11–25, 1982.

W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

C. McDiarmid, "On the method of bounded differences," Technical Report, Institute of Economics and Statistics, Oxford University, 1989.

E. Parzen, "On the estimation of a probability density function and the mode," *Annals of Mathematical Statistics*, vol. 33, pp. 1065–1076, 1962.

I. F. Pinelis, "To the Devroye's estimates for distributions of density estimators," Technical Report, 1990.

D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, New York, 1984.

M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Annals of Mathematical Statistics*, vol. 27, pp. 832–837, 1956.

H. Scheffé, "A useful convergence theorem for probability distributions," *Annals of Mathematical Statistics*, vol. 18, pp. 434–458, 1947.

J. M. Steele, "An Efron-Stein inequality for nonsymmetric statistics," *Annals of Statistics*, vol. 14, pp. 753–758, 1986.

V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, vol. 16, pp. 264–280, 1971.

R. A. Vitale, "An expansion for symmetric statistics and the Efron-Stein inequality," in: *Inequalities in Statistics and Probability*, pp. 112–114, ed. Y. L. Tong, IMS, Hayward, CA., 1984.

C. T. Wolverton and T. J. Wagner, "Recursive estimates of probability densities," *IEEE Transactions on Systems. Science and Cybernetics*, vol. 5, p. 307, 1969.

CONSERVATIVE CONFIDENCE BANDS FOR NONPARAMETRIC REGRESSION

Ricardo FRAIMAN
and
Gonzalo PEREZ IRIBARREN

Centro de Matemática. Universidad de la República Oriental de Uruguay.
Eduardo Acevedo 1139, Montevideo, Uruguay.

ABSTRACT

In this paper conservative confidence bands for the regression function , closely related to those proposed by Hall and Titterington (1988) are considered. However we allow to appear dependence between the local averages which will lead to more efficent confidence bands. Confidence bands based on local medians are also considered and for both cases we find optimal choices of the number of observations to use at each cell and the number of subintervals to be considered. A comparison between both methods is given and some examples are considered. The proposal is very simple and gives confidence bands for which calculation of widths is very easy.

1. Introduction.

In a very nice paper Hall and Titterington (1988) propose conservative confidence bands for the regression function related to those of Knafl, Sacks and Ylvisaker (1988), in that they are based on linear (in the data) estimates of the regression function at any given point. However their proposal is much simpler and gives confidence bands for which calculation of widths are very easy. Some related work have also be done by Wahba (1983) and by Silverman (1985) from a Bayesian point of view, that leads to splines methods to construct confidence bands. See also Härdle and Bowman (1988) for an asymptotic approach based on bootstrapping techniques.

Basically the idea is to divide the interval where we are working, let say the interval [0,1] into r subintervals $[(i-1)h, ih]$, $1\leq i \leq r$, $h=m\delta$ for m an integer that satisfy $n=mr$, where n is the sample size and δ is the distance between points at which observations are taken. At each of these cells using a local average of the response variables in the subinterval, $\overline{Y}_i = m^{-1}(Y_{(i-1)m} + \dots + Y_{im-1})$ construct a confidence interval for the average of the regression function g on the cell, and obtain from it a confidence band for the function $g(x)$, $x \in [(i-1)h, ih]$ using some constraint in the local behaviour of the function g. Since the observations that are used at each cell are independent,

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 45–66.
© 1991 *Kluwer Academic Publishers.*

the confidence band at the whole interval is just obtained by working with the product of those probabilities.

Following the same idea we will consider two extensions:
(i) allowing to appear dependence between the local averages $\overline{Y}_i$ by using more observations than those who are in the corresponding cell to construct the confidence band at each cell.
(ii) using local medians instead of local means in order to obtain robust confidence bands. In this last case we will also allow to appear the kind of dependence described on (i).

In both cases we will not require to the distribution of the errors to be gaussian. However for (i) we will need to calculate percentiles of averages of variables with the error's distribution.

We will show that for a fixed m, conservative confidence bands using k observations, k>m, at each local average can be obtained that, in general, lead to more efficient confidence bands. Optimal k values can be found for a fixed m, and also an optimal pair (k,m) can be obtained under mild assumptions. Since n=mr we will refer in what follows to the pairs (k, r).

Some examples are given for (i) and (ii) together with tables that compares the lenght of different confidence bands. Also optimal k and (k, r) values are given. They can be easily computed at each problem since the bandwidth lenght does not depend on the data except on the error's distribution, the distance between points and a bound for the derivative of the regression function. The case of unknown variance is also considered for local means under normality assumptions.

2. Conservative Confidence Bands Based on Local Averages.

Let us consider the following regression model: $Y(x) = g(x) + e(x)$, where g is a continuous function and e(x) a white noise. We will give some conservative confidence bands for g(x) when observations are avalaible at the points $x_j = j\delta$, i.e. , when we have

$$Y_j = g(j\delta) + e_j \qquad j=0,1,\dots,n \tag{2.1}$$

where δ is a positive number, $\{e_j : j\geq 0\}$ is a sequence of i.i.d. random variables with $E(e_j)=0$ or symmetric around 0. We may assume without loss of generality that we are dealing with the interval [0,1]. In this case $\delta n=1$.

As in Hall and Titterington (1988) we divide the interval [0,1] into r subintervals $[(i-1)h,ih]$ $1\leq i\leq r$, $h=m\delta$ for m an integer that satisfies n=mr. Each $j\delta$ will belong to some cell $[(i-1)h,ih]$. At each cell we consider a local average of k data points

$$\overline{Y}_i = k^{-1}(Y_{p_i}+\dots+Y_{p_i+k-1}) \tag{2.2}$$

where $p_i = (i-1)m - [(k-m)/2]$, and [x] denotes the greatest integer less or equal than x.

Note that if k=m these are the local averages considered by Hall and Titterington (1988), while if k>m the variables $\overline{Y}_i$, $\overline{Y}_{i+1}$ will be correlated. For instance if k=2m the sequence $\{\overline{Y}_i, i\geq 1\}$ is 3-dependent.

Throughout this section we will only assume:

H1. g(x) is a continuous function that verifies a Lipschitz condition on each subinterval $I_i=[p_i\delta,(p_i+k)\delta]$ with constant c_i i=1, ... , r.

We will consider the case when k<m, k=m and k>m. For the first two cases the variables $\{\overline{Y}_i : i=1,\dots, r\}$ are independent while in the last case, as noted above, they are dependent.

However we will show that for a fixed m, conservative confidence bands using k>m can be obtained that, in general, leads to more efficient confidence bands (shorter ones) depending on the relationship between $c=\max\{c_i : i=1,\dots,n\}$ and F. For this case optimal k values can be found relative to some bounds given in Lemma 2.1. On the other hand also an optimal pair (k, r) can be found when $c_i=c$ for $i=1,\dots, r$. In both cases optimal values can be obtained from the distribution F and the values of n, c and the confidence level. (In the normal case it will depend on σ^2, n,c and the confidence level).

The results on this section are based on the following two Lemmas.

Lemma 2.1. Assume H1 and let $x_i=p_i\delta$, $z=x_i+yk\delta$ for $[(k-m)/2]/k\le y\le[(k+m)/2]/k$ and

$$\bar{g}_i=k^{-1}\sum_{h=0}^{k-1} g(x_i+h\delta)\ .\ \text{Then we have that}$$

$$|g(z)-\bar{g}_i|\le\Delta_{k,i}$$

where $\Delta_{k,i}=c_i\delta(k^2+m^2+2m)/4k$ if $k\ge m$ and k-m is even; $\Delta_{k,i}=c_i\delta(k^2+m^2-1)/4k$ if $k\ge m$ and k-m is odd; $\Delta_{k,i}=(m+1)/2$ if k<m and m-k is even and $\Delta_{k,i}=m/2$ if k<m and m-k is odd.
The proof can be found in the Appendix.

Lemma 2.2. a) Let $U_1,U_2,\dots,U_{(r-1)p+k}$ be i.i.d. random variables with common continuous distribution function F, $p\ge1$, $p+1\le k<2p+1$ and consider the following sums of random variables:

$$\mathcal{U}_s = U_{(s-1)p+1} + U_{(s-1)p+2} +\dots+ U_{(s-1)p+k} \quad \text{for } s=1,\dots,r$$

each of which have distribution function F_k^*, the k-convolution of F with itself.

Let $\mathcal{V}_1, \mathcal{V}_2,\dots,\mathcal{V}_r$ be independent random variables with the same distribution function F_k^* as $\mathcal{U}_1$ (i.e. independent copies of $\mathcal{U}_1, \mathcal{U}_2,\dots,\mathcal{U}_r$). Then we have that

$$P(\mathcal{U}_1\le t_1, \mathcal{U}_2\le t_2, \dots ,\mathcal{U}_r\le t_r) \ge P(\mathcal{V}_1\le t_1, \mathcal{V}_2\le t_2, \dots ,\mathcal{V}_r\le t_r)$$

and

$$P(\mathcal{U}_1>t_1, \mathcal{U}_2>t_2, \dots ,\mathcal{U}_r>t_r) \ge P(\mathcal{V}_1>t_1, \mathcal{V}_2>t_2, \dots ,\mathcal{V}_r>t_r)\ ,\ \text{for all point } (t_1,\dots t_r)\in R^r.$$

b) Moreover, the same result holds for $lp+1\le k\le(l+1)p+1$ and $l\ge1$.
The proof can be found in the Appendix.

Remark 2.1. Note that in (a), the sequence $\{\mathcal{U}_k\ k\ge1\}$ is 2-dependent, and in the general case considered in (b) the same sequence is (l+1)-dependent.

Remark 2.2. For the gaussian case, Lemma 2.2 can be derived directly from Slepian Lemma (see for instance Leadbetter, Lindgren and Rootzén (1983), Normal comparison Lemma on page 81). Conversely Slepian inequality can be derived from Lemma 2.2.

Remark 2.3. Lemma 2.2 will allow us to obtain conservative confidence bands when we use k>m. In particular if F is symmetric we have that

$$P(|\mathcal{U}_1|\le u,\dots,|\mathcal{U}_r|\le u)\ge 2P(\mathcal{V}_1\le u,\dots,\mathcal{V}_r\le u) - 1$$

as can be easily seen from the well known inequality $P(A\cap B)\ge P(A)+P(B) - 1$ with $A=\{\mathcal{V}_1\le u,\dots,\mathcal{V}_r\le u\}$ and $B=\{\mathcal{V}_1\ge -u,\dots,\mathcal{V}_r\ge -u\}$.

Corollary 2.1. Let $\mathcal{U}_1,\dots,\mathcal{U}_r$ and $\mathcal{V}_1,\dots,\mathcal{V}_r$ be as in Lemma 2.2 and normally distributed, and θ a positive random variable independent of $\{\mathcal{U}_1,\dots,\mathcal{U}_r,\mathcal{V}_1,\dots,\mathcal{V}_r\}$. Then we have that

$$P(\frac{\mathcal{U}_1}{\theta}\le u,\dots,\frac{\mathcal{U}_r}{\theta}\le u) \ge P(\frac{\mathcal{V}_1}{\theta}\le u,\dots,\frac{\mathcal{V}_r}{\theta}\le u)\,. \tag{2.3}$$

In particular we can apply this result to the case $\theta=(\nu/s)^{1/2}$ where ν has a chi-square distribution with s degrees of freedom.

The conclusion follows inmediately from

$$E[\,E(\mathbf{1}_{\{\max_{h=1\dots,r}\mathcal{U}_h\le\theta u\}}/\theta)] \ge E[\,E(\mathbf{1}_{\{\max_{h=1\dots,r}\mathcal{V}_h\le\theta u\}}/\theta)]$$

where $\mathbf{1}_A$ is the indicator function of the set A.
Moreover when $\theta=(\nu/s)^{1/2}$ and ν has a chi-square distribution with s degrees of freedom (2. 3) is just

$$\int_0^\infty P(\max_{h=1\dots,r}\mathcal{U}_h\le u(\nu/s)^{1/2}/\nu=\zeta)f_\nu(\zeta)d\zeta \ \ge\ \int_0^\infty P(\max_{h=1\dots,r}\mathcal{V}_h\le u(\nu/s)^{1/2}/\nu=\zeta)f_\nu(\zeta)d\zeta$$

where $f_\nu(.)$ is the chi-square density for s degrees of freedom.

2.1 CONFIDENCE BANDS FOR THE REGRESSION FUNCTION WHEN σ^2 IS KNOWN.

For the lack of simplicity let us consider the case when $\{e_j : j>1\}$ is a gaussian white noise. From (2.1) we have that

$$\bar{Y}_i-\bar{g}_i = k^{-1}\sum_{j=p_i}^{p_i+k-1} e_j = \bar{e}_i$$

where Y_i is defined in (2.2), $\bar{g}_i$ is defined in Lemma 2.1 and $\bar{e}_i$ is a gaussian random variable with zero mean and variance σ^2/k. Therefore by Lemma 2.2 and Remark 2.3, given $0<\alpha<1$ we can find $a>0$ such that

$$1-\alpha \le P(-a<\bar{e}_i<a \text{ for } i=1,\dots,r) = P(\bar{Y}_i - a < \bar{g}_i < \bar{Y}_i + a \text{ for } i=1,\dots,r). \tag{2.4}$$

More precisely the value a in (2.4) is obtained from the relationship

$$P(\max_{i=1\dots,r}|\bar{e}_i|\le a) = P(k^{1/2}\sigma^{-1}\max_{i=1\dots,r}|\bar{e}_i|\le ak^{1/2}\sigma^{-1}) \ge 2\Phi^r(ak^{1/2}\sigma^{-1})-1$$

where Φ stands for the standard normal distribution function, choosing $a=\sigma k^{-1/2}\Phi^{-1}[(1-\alpha/2)^{1/r}]$.

On the other hand from the proof of Lemma 2.1 it is easy to see that

$$|g(z)-\bar{g}_i| \le c_i\delta\,[k+6y-1+2yk(y-1)]/2 = \Delta_{k,i}^{+}(y) \quad \text{for } y\ge 0$$

and

$$|g(z)-\bar{g}_i| \le c_i\delta(k-1-2yk)/2 = \Delta_{k,i}^-(y) \quad \text{for } y<0. \tag{2.5}$$

(We recall that $k \ge m$ implies $y \ge 0$.)

Then the confidence bands are given by

$$1-\alpha \le P(Y_i-a-\Delta_{k,i}(y) \le g(z) \le Y_i+a+\Delta_{k,i}(y) \quad \forall\ z \in I_i,\ i=1,\ldots,r)$$

where $\Delta_{k,i}(y)$ is $\Delta_{k,i}^+(y)$ or $\Delta_{k,i}^-(y)$ respectively. However in practice if $m\delta$ is small enough we will use the uniform bounds $\Delta_{k,i}$ given in Lemma 2.1 instead of $\Delta_{k,i}(y)$, (which corresponds to a minimax situation) to obtain the confidence bands.

With straightforward modifications we can obtain an analogous result for the non-gaussian case in terms of the percentiles of the distribution function of the error's average.

2.2 WIDTHS OF CONFIDENCE BANDS FOR FIXED R AND DIFFERENT VALUES OF K.

In this Section we will compare the efficiency of the conservative confidence bands for different values of r and k at the gaussian case using the uniform bounds $\Delta_{k,i}=\Delta_k$. From the last Section we have that the width of a conservative confidence band with level $1-\alpha$ at each subinterval I_i is given by

$$L_{k,i} = L_k = 2\,\sigma\, k^{-1/2}\Phi^{-1}[(1-\alpha/2)^{1/r}] + 2\Delta_k.$$

The difference between L_m and L_k is given by

$$L_m-L_k = 2\,\sigma\Phi^{-1}[(1-\alpha/2)^{1/r}]\,(m^{-1/2} - k^{-1/2}) + 2[\Delta_m - \Delta_k]$$

In this case it is easy to see that if $k<m$, then $\Delta_k \ge \Delta_{m-1}$ and Δ_{m-1} is essentially equal to Δ_m (the difference is $c_i\delta/2$). Therefore the case $k<m$ will lead to widther confidence bands. However if we use $\Delta_k(y)$ instead of Δ_k and $k<m$, it will happens that $\Delta_k(y) > \Delta_{m-1}(y)$ for some y values and $\Delta_k(y) < \Delta_{m-1}(y)$ for other y values. For instance if $\sigma^2=0$, $k=1$ will be an optimal selection for some values of y but not for all the possible values at the subinterval. If $k \ge m$ we have that

$$2[\Delta_m - \Delta_k] = c_i\delta(m+1-(k^2+m^2+2m)/2k) = -c_i\delta(k-m)(k-m-2)/2k \tag{2.6}$$

if $k-m+1$ is even, and

$$2[\Delta_m - \Delta_k] = c_i\delta(m+1-(k^2+m^2-1)/2k) = -c_i\delta[(k-m)^2-2k-1]/2k \tag{2.7}$$

otherwise.

Then

$$0 \le L_m-L_k \iff \sigma/c_i\delta \ge p(k)/2\ \Phi^{-1}[(1-\alpha/2)^{1/r}]\,(m^{-1/2} - k^{-1/2})$$

with p(k) being the right hand member of (2.6) or (2.7) respectively divided by $-c_i\delta$.

This relationship does not depend on the data except on the values of σ, c_i and δ.

Then, we can also find an optimal pair (k,r). In Table I we give the bandwiths lenght for different values of k and r and the optimal selection for a real example of radiocarbon age and tree-ring age already considered by Hall and Titterington (1988, Example 5.2 pag. 245) using the estimated variance given there, δ=2400/180 and c_i=c=1. We recall that the bandwidth lenght does not depend on the data except on those values,so for each problem an optimal solution , in this sense, can be found easily. The best value for m=k is obtained at m=6. However the pair (m,k)=(6,11) gives a smaller band and at the pair (4,9) we obtain the minimal value, with a relative efficency increased on 22.7% $((L_m-L_k)/L_k)$).

In Table II we give the bandwidth lenght for different values of α, n, δ, σ, c_i, r and k for normal errors, which shows the optimal k value for each r (or m) and its relative efficiency.

RADIOCARBON AGE AND TREE-RING AGE EXAMPLE
CONFIDENCE LEVEL=0.95, n=180, C=1

m	6	5	4	3	9	15
r	30	36	45	60	20	12
optimal k value	11	10	9	8	14	20
Lenght for Lk	196.3	191.4	188.32	188.33	218.6	275.6
Lenght for Lm	231.2	233.9	242.5	265.7	242	292.95

Table I

CONFIDENCE LEVEL=.9,r=10, δ=.01,C=1					CONFIDENCE LEVEL=.95,r=50, δ=.0005,C=1					
σ	1	.8	.4	.01	σ	1	.8	.4	.2	.0001
optimal k value	65	57	37	11	optimal k value	559	481	305	195	41
Lk length	.969	.837	.536	.115	Lk length	.417	.36	.227	.144	.020
Lm length	1.73	1.41	.759	.12	Lm length	1.06	.85	.434	.227	.0206

TABLE II

In Table III we find the optimal pair (k,r)=(66,4) for α=0.1, n=100 δ=0.01 σ=c=1.

CONFIDENCE LEVEL=0.9, n=100, δ=0.01, σ=C=1

r	1	2	4	5	10	20	25	50	100
k*	100	77	66	65	65	68	69	73	75
Lk*	1.338	.992	.927	.93	.969	1.02	1.037	1.086	1.133
Lm	1.338	1.06	1.153	1.216	1.734	2.563	2.92	4.389	6.586

TABLE III

In Figure 2.1 a scatterplot of generated data points (Yj, jδ) j=0,...,2000 for normal errors with :ero mean and variance 0.04, δ=0.0005 verifying the model

$$Y_j=0.5 - |j\delta-0.5| + e_j$$

; given. In Figure 2.2 a graph of the step function which takes values B^+_i and B^-_i, i=1,...,r, where B^+_i and B^-_i denotes the upper and lower bounds of a 95% confidence band, which are plotted by ›ining succesive values by straight lines together with the graph of the true function is given. We consider additional 77 data points at each extreme of the [0,1] interval). We use for r=50 the ptimal k value of 195 which gives a bandwidth lenght of 0.1446. In this case the confidence band orresponding to k=m=40 has a length of 0.2276.

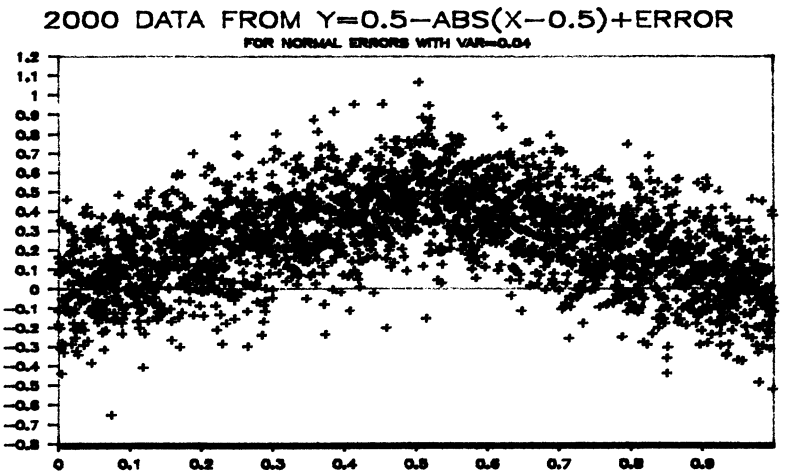


Figure 2.1

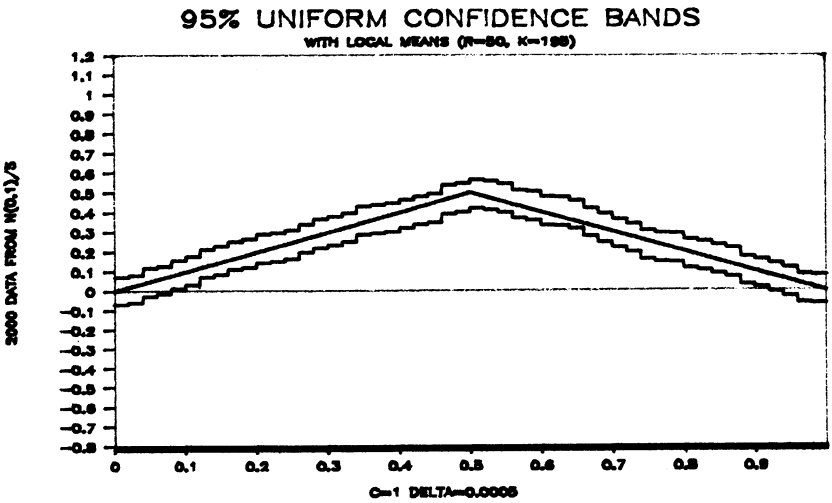


Figure 2.2

Figure 2.3 shows a similar scatterplot of 400 double-exponential generated data points with variance 1/16, verifying the model

$$Y_j=0.5 - |j\delta-0.5| + e_j$$

($c_j=c_0=1$). A conservative 90% uniform confidence band based on local means is plotted (r=10, k=131) together with the true function. A conservative method based on a Berry-Esséen type bound was considered to obtain the critical values. Details are given in the Appendix.

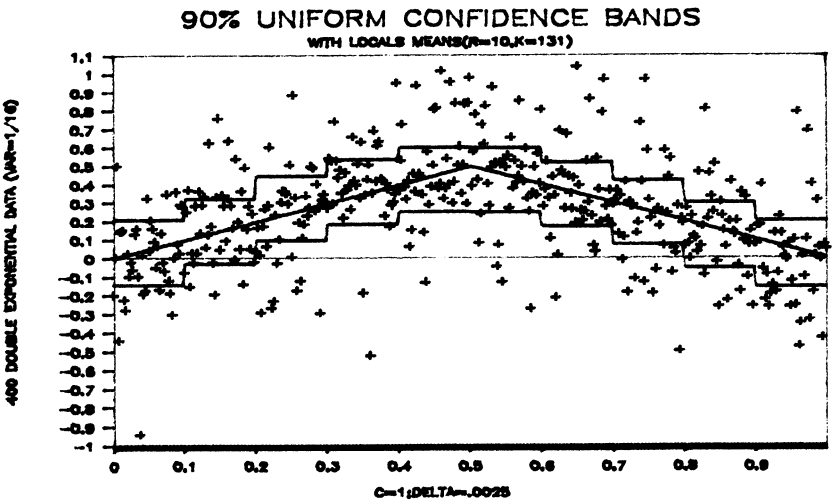


Figure 2.3

2.3 CONFIDENCE BANDS FOR THE REGRESSION FUNCTION WHEN σ^2 IS UNKNOWN AT THE GAUSSIAN CASE.

If σ^2 is unknown we may construct a slight overestimate of it,as in Hall and Titterington (1988). Since the difference of two i.i.d. normal variables is independent of their sum the set of differences $Y_{2j}-Y_{2j-1}$ will be independent of the estimates used to construct the confidence bands if we consider only those differences for which the corresponding sums appears in the error's average. If k=m it is just to exclude those differences for which 2j or 2j-1 is of the form im for some integer i. We denote this set by $\mathcal{S}$. Then if

$$\hat{\sigma}^2=(s-1)^{-1}\Big\{ \sum_{j\in\mathcal{S}} (Y_{2j}-Y_{2j-1})^2 \; - s^{-1} [\sum_{j\in\mathcal{S}} (Y_{2j}-Y_{2j-1})]^2 \Big\}$$

where s is the number of elements in $\mathcal{S}$, we have that $(s-1)\hat{\sigma}^2/\sigma^2$ has a chi-square distribution with s-1 degrees of freedom and a noncentrality parameter which depends on the function g. Therefore using now Corollary 2.1 we obtain conservative confidence bands now related to the Student's distribution function with s-1 degrees of freedom instead of the normal distribution.

<u>A remark on the asymptotic behaviour.</u> We have seen that for normal data

$$L_{m,i}\leq 2\,\sigma m^{-1/2}\Phi^{-1}[(1-\alpha/2)^{1/r}] + c_i\delta(m+1). \tag{2.8}$$

Let us suppose that $r\to\infty$, $m\to\infty$ (therefore $n=mr\to\infty$) and $c_i=c_0$. Then since $(2\log r)^{-1/2}\Phi^{-1}[(1-\alpha/2)^{1/r}]\to 1$ (see for instance, Leadbetter, Lindgren and Rootzén (1983)) we will have that $L_{m,i}\to 0$ provided that $\log r/m\to 0$. Actually, the first term in the right hand of (2.8) converges to zero, as well as $c_0(m+1)\delta=c_0(m+1)/mr$ converges to zero.

If we choose $k\geq m$ in such a way that $k/n\to 0$, and $k/\log r\to\infty$, for $r\to\infty$, we arrive to the same conclusion for L_k.

Similar results can be obtained in the non gaussian case, which will depend on the tail behaviour of the distribution F.

3. Conservative confidence bands based on local medians.

In this Section we will obtain conservative confidence bands based on local medians, that will have natural robustness properties with high breakdown point. These results can be applied to nonparametric regression models without any requirement to the distribution of the errors F (even without any moment condition or symmetry assumptions). Moreover, the confidence bands based on local medians will also be conservative in the sense that their level, obtained for a given F, will be larger for any other distribution funcion G which is stochastically smaller than F. From a robust point of view the confidence bands obtained for a given F can be applied when the true distribution G lies in a whole neigborhood that is controlled by F in the sense that F is more heavy tailed than G.

As before, we will also consider the case when k>m when the bands are constructed from local medians of groups of observations with some overlapping, which will introduce dependence between the local medians, in order to attain more efficiency.

When the regression function g is monotonic, we will also be able to reduce the bandwidth lenght of the conservative band.

The results are based on two Lemmas, similar to those on the preceeding Section, one who deals with the "interpolation" error and the other one who deals with the stochastic error, that has some interest by itself.

Let $Y(x) = g(x) + e(x)$ where g is a continuous function and $e(x)$ is a white noise with distribution function F, median(F) = 0. Observations are avalaible as before at the points $x_j=j\delta$ i.e.

$$Y_j=g(j\delta)+e_j \qquad j=0,\dots,n \tag{3.0}$$

where δ is a positive number and $\{e_j:\ j\geq 0\}$ is a sequence of i.i.d. random variables with distribution F.

As before we divide the interval [0,1] into r subintevals $[(i-1)h,ih]$ $1\leq i\leq r$, $h=m\delta$ for m an integer , $n=mr$. Each $j\delta$ will belong to some cell $[(i-1)h,ih]$. At each cell we consider a local median of lenght k $\tilde{Y}_i=\text{median}(Y_{p_i},\dots,Y_{p_i+k-1})$ for $p_i=(i-1)m-[(k-m)/2]$ and

$$\tilde{G}_i=\text{median}(g(x_i), g(x_i+\delta),\dots,g(x_i+(k-1)\delta)) = \text{median}(g(x_i+s\delta), s=0,\dots,k-1)$$

for $x_i=p_i\delta$. Then we have that

$$\tilde{Y}_i=\text{median}((g(x_i+s\delta) + e_{p_i+s}) , s=0,\dots,k-1) =$$
$$\tilde{G}_i + \text{median}((g(x_i+s\delta)-\tilde{G}_i + e_{p_i+s}) , s=0,\dots,k-1) . \tag{3.1}$$

<u>Lemma 3.1</u>. Let $z\in[(i-1)h,ih]$ and assume that g verifies a Lipschitz condition of order one with constant c_i in the subinterval $[p_i\delta,\ (p_i+k-1)\delta]$. Then

a) $|g(z)-\tilde{G}_i| \leq c_i\delta \max((k+1)/4 ; (m+1)/2)$.

b) If in addition g is monotonic, then $|g(z)-\tilde{G}_i| \leq c_i\delta\ (m+1)/2$.
The proof can be found in the Appendix.

<u>Lemma 3.2</u>. a) Let $U_1,U_2,\dots,U_{(r-1)p+k}$ be i.i.d. random variables with common continuous distribution function F, $p\geq 1$, $p+1\leq k<2p+1$ and consider the following random vectors:
$\mathcal{U}_s = (U_{(s-1)p+1}, U_{(s-1)p+2},\dots,U_{(s-1)p+k})$ for $s=1,\dots,r$.

Let $\mathcal{V}_1, \mathcal{V}_2,\dots,\mathcal{V}_r$ be independent copies of $\mathcal{U}_1, \mathcal{U}_2,\dots,\mathcal{U}_r$ (i.e. $\{ \mathcal{V}_1, \mathcal{V}_2,\dots,\mathcal{V}_r \}$ is a independent set of random vectors, each of them with the same distribution function as $\mathcal{U}_1$).

Denote by $\mathcal{U}_s^{(h)}$ the h-order statistics of the vector $\mathcal{U}_s$, (h=1,2,...,k) and $\mathcal{V}_s^{(h)}$ the h-order statistics of the vector $\mathcal{V}_s$, s=1,2,...,r. Then we have that

$$P(\mathcal{U}_1^{(h)}\leq t, \mathcal{U}_2^{(h)}\leq t, \dots ,\mathcal{U}_r^{(h)}\leq t) \geq P(\mathcal{V}_1^{(h)}\leq t, \mathcal{V}_2^{(h)}\leq t, \dots ,\mathcal{V}_r^{(h)}\leq t)$$

and

$$P(\mathcal{U}_1^{(h)}>t, \mathcal{U}_2^{(h)}>t, \dots ,\mathcal{U}_r^{(h)}>t) \geq P(\mathcal{V}_1^{(h)}>t, \mathcal{V}_2^{(h)}>t, \dots ,\mathcal{V}_r^{(h)}>t)$$

for all real t.

b) Moreover, the same result holds for $lp+1\leq k\leq (l+1)p+1$ and $l\geq 1$.
The proof can be found in the Appendix.

3.1 CONFIDENCE BANDS BASED ON LOCAL MEDIANS.

We will now construct conservative confidence bands based on local medians by using Lemma 3.1 and Lemma 3.2. Let $x \in [(i-1)h, ih]$ and suppose that g verifies the condition given in Lemma 3.1. Then we have that

$$|g(x_i+s\delta)-\tilde{G}_i| \le c_i\delta \ (k+1)/2.$$

Therefore

$$-c_i\delta \ (k+1)/2 + \text{median}(U_{p_i+s}, s=0,\dots,k-1) \le \text{median}(g(x_i+s\delta)-\tilde{G}_i + U_{p_i+s}, s=0,\dots,k-1) \le$$

$$c_i\delta \ (k+1)/2 + \text{median}(U_{p_i+s}, s=0,\dots,k-1).$$

On the other hand by Lemma 3.1

$$|g(x)-\tilde{G}_i| \le c_i\delta \max((k+1)/4 ; (m+1)/2).$$

Putting things together we have

$$\tilde{Y}_i - \text{median}(U_{p_i+s}, s=0,\dots,k-1) - c_i\delta \ (k+1)/2 - c_i\delta \max((k+1)/4 ; (m+1)/2) \le g(x) \le$$

$$\tilde{Y}_i - \text{median}(U_{p_i+s}, s=0,\dots,k-1) + c_i\delta \ (k+1)/2 + c_i\delta \max((k+1)/4 ; (m+1)/2)$$

and so if A and B are such that $A \le \text{median}(U_{p_i+s}, s=0,\dots,k-1) \le B$ we have that

$$\tilde{Y}_i - B - c_i\delta [(k+1)/2 + \max((k+1)/4 ; (m+1)/2)] \le g(x) \le$$

$$\tilde{Y}_i - A - c_i\delta [(k+1)/2 + \max((k+1)/4 ; (m+1)/2)]$$

for $x \in [(i-1)h, ih]$.

Finally if a and b are such that

$$(1-\alpha)^{1/r} = P(a \le \text{median}(U_{p_i+s}, s=0,\dots,k-1) \le b) \tag{3.2}$$

we get our robust uniform confidence bands by using Lemma 3.2. Effectively,

$$P\Big\{\tilde{Y}_i - b - c_i\delta [(k+1)/2 + \max((k+1)/4 ; (m+1)/2)] \le g(x) \le$$

$$\tilde{Y}_i - a + c_i\delta [(k+1)/2 + \max((k+1)/4 ; (m+1)/2)]\Big\} \ge$$

$$P\left\{a\le \text{median}(U_{p_i+s}\ ,\ s=0,\dots,k-1)\le b, i=1,\dots,r\right\}\ge P^{\,r}\left\{a\le \text{median}(U_{p_i+s}\ ,\ s=0,\dots,k-1)\le b\right\}=(1-\alpha) \tag{3.3}$$

We summarize this result in the following Theorem.

Theorem 3.1. Let $Y_1, \dots, Y_n$ be observations verifying (3.0) and assume that for a fixed $k\ge m$ at each subinterval $[p_i\delta, (p_i+k)\delta]$ $i=1,\dots,r$ the function g satisfy a Lipchitz condition of order one with constant c_i. Then a $1-\alpha$ conservative confidence band is given by the left hand side of (3.3) with a and b verifying (3.2).

In the symmetric case we take a=-b.

Corollary 3.1. In the monotone case, $\max((k+1)/4;(m+1)/2)$ can be replaced by $(m+1)/2$ in (3.3). Moreover, if we know that the function g is nondecreasing, then each cell can be divided into two subintervals. At the left one we just need to add the interpolation term $(m+1)/2$ to the lower bound, while in the right one we only need to add it to the upper bound. If g is nonincreasing the same result holds changing left by right subintervals.

Remark 3.1. As in the case of local means it can be shown that if $r/k \to 0$ as $r\to\infty$ and $k/n \to 0$, the bandwidth lenght L_k converges to zero for all F symmetric around zero with density f(x). Efectively, under these assumptions it can be proved that

$$F(b_k)=1/2+1/2[(\log r)/k]^{1/2} + o([(\log r)/k]^{1/2})$$

where b_k is such that

$$P(\text{median}(\xi_1,\dots,\xi_{2k+1}) \in (-b_k,b_k))=(1-\alpha)^{1/r}$$

and $\{\xi_i,\ i=1,\dots,2k+1\}$ are i.i.d. random variables with common symmetric distribution F. On the other hand the condition $k/n \to 0$, ensures that the "interpolation error" converges also to zero, which entails the desired result.

3.2 SOME EXAMPLES OF CONSERVATIVE BANDS.

Some artificial examples are considered in what follows. In Figure 3.1 a 90% confidence band based on local means is constructed from 200 data verifying the model

$$Y_i=0.25 - |x_i-0.5| + e_i \qquad \text{for } i=1,\dots,200$$

for e_i with distribution function $F_1(./4)$, where F_1 is the standard Cauchy distribution, while using percentiles from a normal distribution with variance 1/16. Figure 3.1 shows clearly how bad can be the bands constructed from local means when the error's distribution have heavy tails. Note that not only the real function has nothing to do with the constructed interval but also there is no possible continuous function in this band. The bandwidth length of 90% conservative bands constructed from local means with Cauchy percentiles is also given in Figure 3.9.

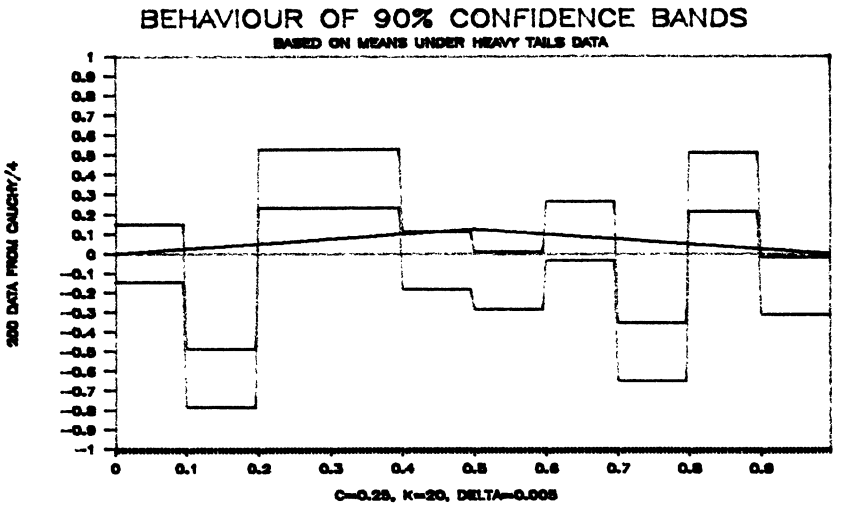


Figure 3.1

Figures 3.2 and 3.3 corresponds to 200 data from a linear model $Y_i=0.5\,x_i + e_i$ $i=1,...200$ (monotone case), when the error's distribution is $F_1(./10)$ and 95% confidence bands are constructed from local means and medians respectively, for r=10, k=41, δ=0.005 and c=0.5. For local means we have now used percentiles from $\Phi(x/10)$. For the same model, in Figure 3.4 a 95% confidence band is constructed based on 200 data when the error's distribution is $F_2(./10)$ and F_2 is a Student distribution with 3 degrees of freedom.

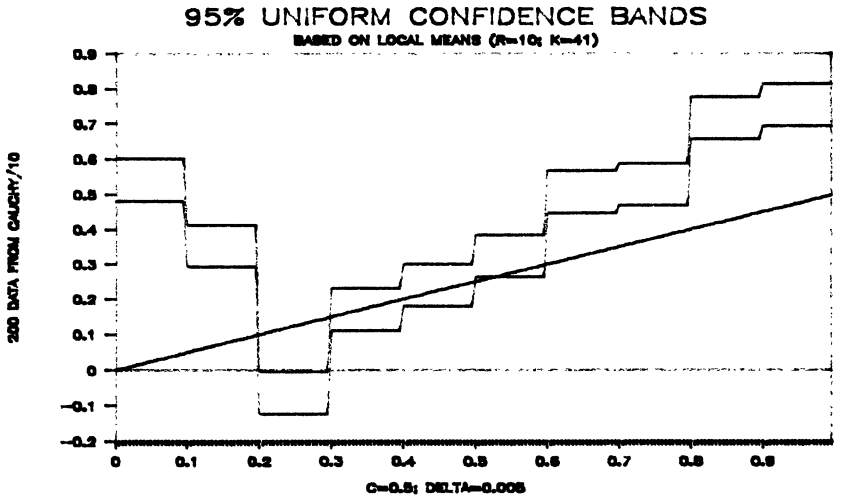


Figure 3.2

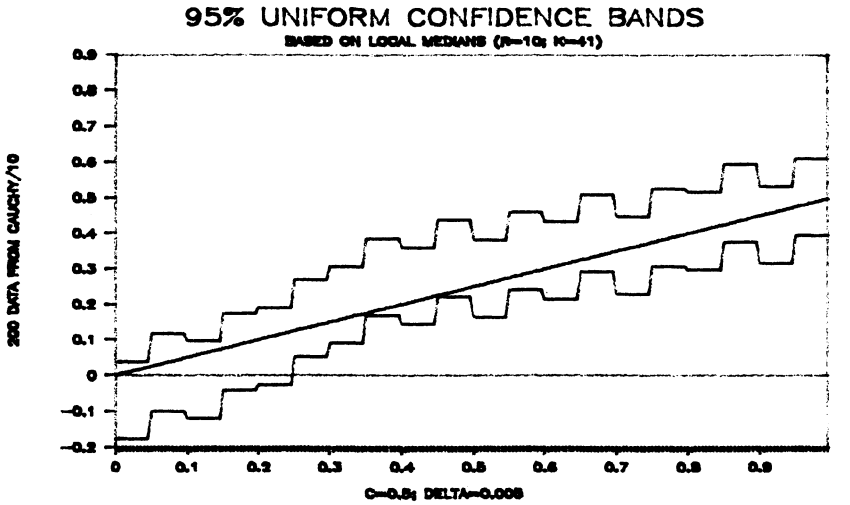


Figure 3.3

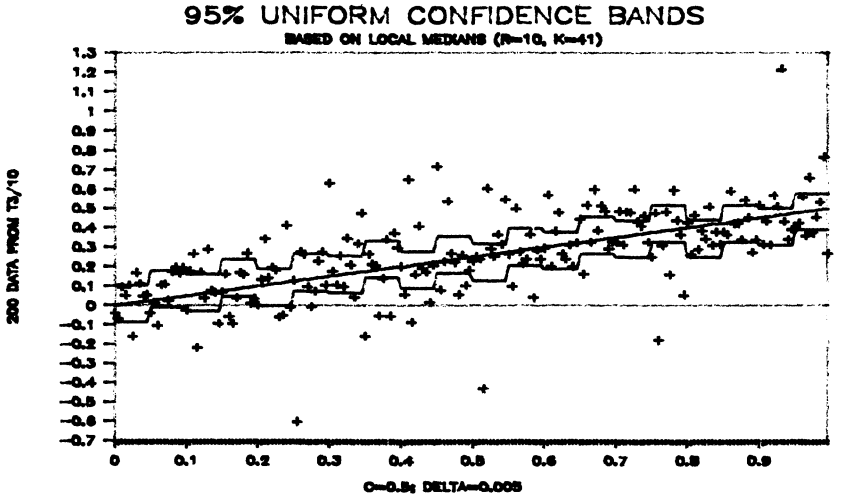


Figure 3.4

In Figures 3.5 and 3.6 the model

$Y_i=0.5 - |x_i-0.5| + e_i$ for $i=1,\ldots,2000$

is consider, for c=1, $0 \le x_i \le 1$, δ=0.0005 while e_i are random variables with a Student distribution with 3 degrees of freedom and variance 0.04 ; and a 0.95 level uniform conservative band is constructed based on local medians. The value k=81 is the optimal one for r=50.

We recall that these bands can be used not only for the distribution function F from which they have been constructed but also for any other distribution function G whose tails are controlled by F.

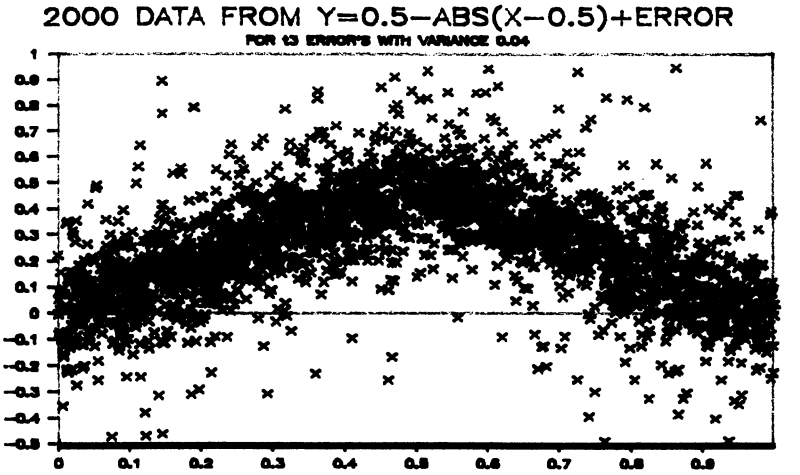


Figure 3.5

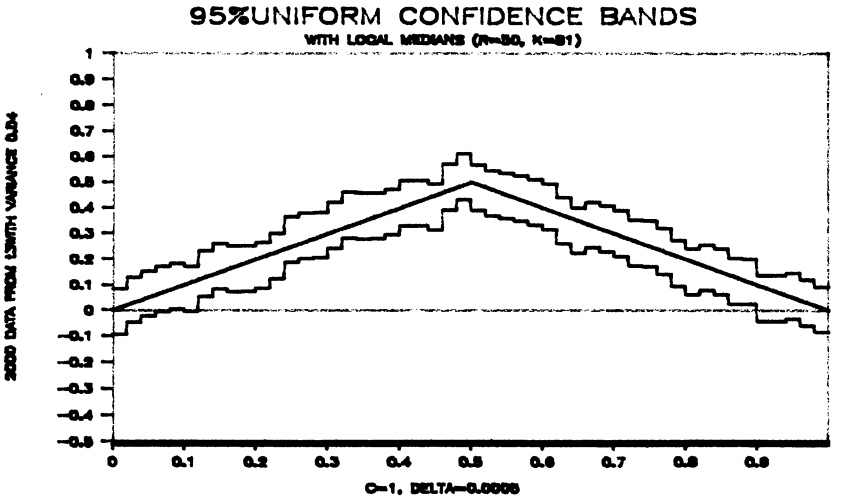


Figure 3.6

The statistical component of the bandwidth lenght, i.e., the median percentiles can also be easily computed from percentiles of the binomial distribution function as is well known. Fixed the confidence level 1-α and the number of subintervals r the percentiles for a symmetric F can be calculated as follows:

a) Find the value $p=p(a)=P(e_1 \geq a)$ for which

$$P\left(\sum_{i=1}^{2s+1} Z_i \leq s\right) = (1-\alpha/2)^{1/r}$$

where Z_i are i.i.d. Bernoulli random variables with parameter p , and k=2s+1, for different values of k.

b) For $q=q(a)=1-p(a) = P(e_1 \leq a)=F(a)$ find a according to the distribution function F for different values of k.

Note that once the q(.) values are calculated they can be used for different distribution functions. Adding the deterministic component we obtain the bandwidth length. Also optimal (k,r) values can be obtained varing those parameters.

In Figure 3.7 a comparison of the bandwidth lenght of 90% conservative confidence bands constructed from local means and medians is given for normal data with variance 1/16, c=0.25, r=10 and δ=1/200. Figures 3.8 and 3.9 gives the same information for data from a Cauchy distribution. Finally some percentiles for the conservative bands are given for the standard normal distribution, Student with 3 degrees of freedom and Cauchy distribution in Figure 3.10.

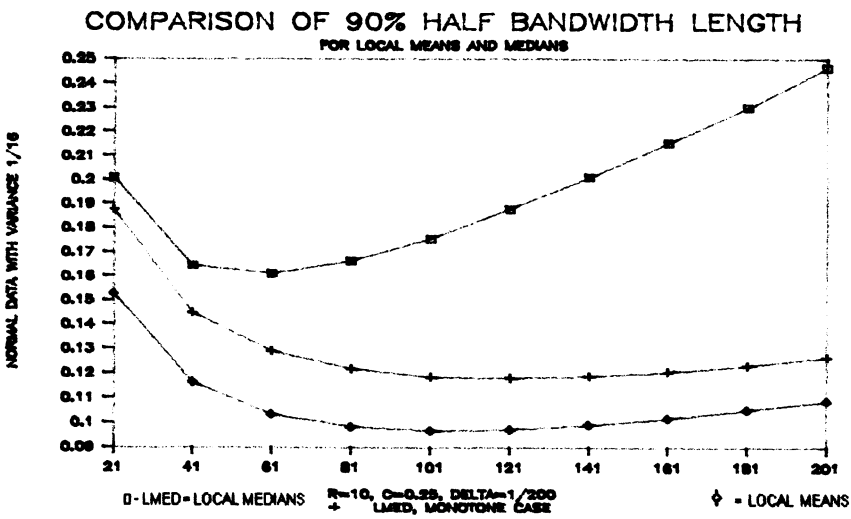


Figure 3.7

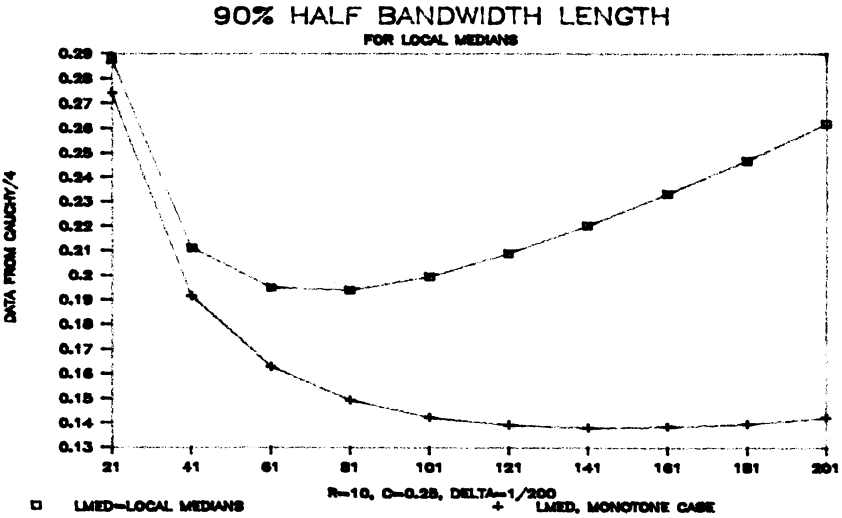


Figure 3.8

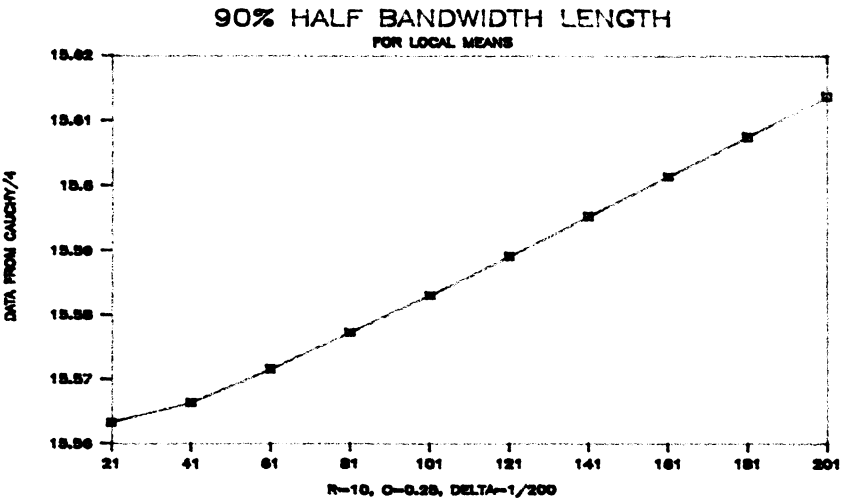


Figure 3.9

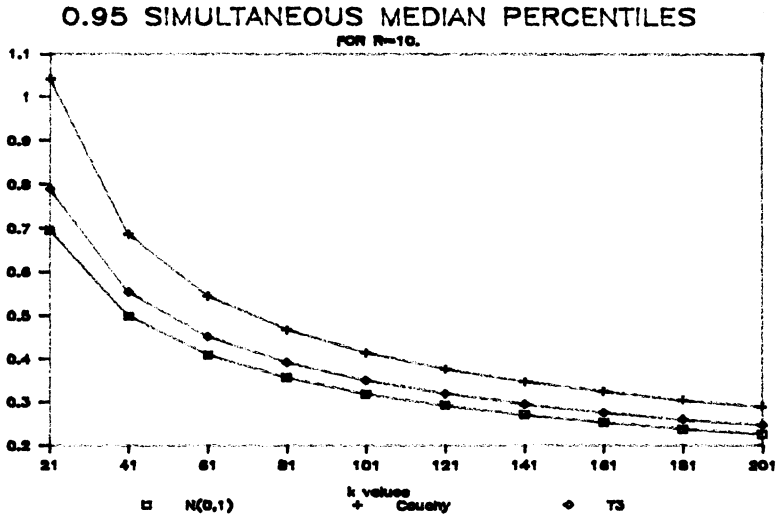


Figure 3.10

Appendix.

Proof of Lemma 2.1. We have that

$$|g(z) - \bar{g}_i| = k^{-1} \left| \sum_{h=0}^{k-1} [g(z) - g(x_i + h\delta)] \right| \le c_i\, k^{-1} \sum_{h=0}^{k-1} |h - yk| \tag{4.1}$$

If $k \ge m$, we have that $y \ge 0$, and the right hand member of (4.1) is equal to

$$c_i\, k^{-1} \delta [yk + (yk-1) + \ldots + (yk-q) + (q+1-yk) + \ldots + (k-1-yk)] =$$

$$c_i\, k^{-1} \delta [yk(2q+2-k) + k(k-1)/2 - q(q+1)]$$

where $q=[yk]$, and $[(k-m)/2]/k \le y \le [(k+m)/2]/k$. If k-m is even, we have that the worst selection for y is $y=(k+m)/2k$. In this case we have that

$$|g(z) - \bar{g}_i| \le c_i\, k^{-1} \delta [(k+m)(k+m+2-k)/2 + k(k-1)/2 - (k+m)((k+m)/2+1)/2] =$$

$$c_i\, \delta (k^2 + m^2 + 2m)/4k.$$

If k-m is odd, it is easy to see that

$$|g(z) - \bar{g}_i| \le c_i \delta (k^2 + m^2 - 1)/4k$$

choosing now $yk=(k+m-1)/2$ (or $(k-m-1)/2$).

Analogously we obtain the bounds for the case $k \le m$.

Proof of Lemma 2.2. a) Let

$$\mathcal{U}_1 = \sum_{i=1}^{p} U_i + \sum_{i=p+1}^{k} U_i = W_1+W_2$$

$$\mathcal{U}_2 = \sum_{i=p+1}^{k} U_i + \sum_{i=k+1}^{2p} U_i + \sum_{i=2p+1}^{p+k} U_i = W_2+W_3+W_4$$

$$\mathcal{U}_3 = \sum_{i=2p+1}^{p+k} U_i + \sum_{i=p+k+1}^{3p} U_i + \sum_{i=3p+1}^{2p+k} U_i = W_4+W_5+W_6$$

.....

$$\mathcal{U}_r = \sum_{i=(r-1)p+1}^{(r-1)p+k} U_i + \sum_{i=(r-1)p+k+1}^{rp} U_i + \sum_{i=rp+1}^{(r-1)p+k} U_i = W_{2r-2}+W_{2r-1}+W_{2r}.$$

Note that in particular if r=2 we have only $\mathcal{U}_1=W_1+W_2$ and $\mathcal{U}_2=W_2+(W_3+W_4)$ with W_1, W_2 and (W_3+W_4) being independent random variables; and that if k=2p, then $W_3, W_5, \ldots$, are lacking of sense).

It is clear that we will have proved the Lemma if we prove the following inequality:

$$P(W_1+W_2 \le t_1, W_2+W_3+W_4 \le t_2, W_4+W_5+W_6 \le t_3, \ldots, W_{2r-2}+(W_{2r-1}+W_{2r}) \le t_r) \ge$$

$$P(W_1+W_2 \le t_1, W'_2+W_3+W_4 \le t_2, W_4+W_5+W_6 \le t_3, \ldots, W_{2r-2}+(W_{2r-1}+W_{2r}) \le t_r) \quad (4.2)$$

with W'_2 an independent copy of W_2 i.e. a new random variable with the same distribution function that W_2 but independent of the set $\{W_1, W_2, \ldots, W_{2r}\}$.

We rewrite (4.2) for r>1 as:

$$E\left(1_B E\left(F^*_{k-p}((t_1 - W_1)) \wedge (t_2 - W_3 - W_4)\right) / W_1, W_3, W_4, W_5, \ldots, W_{2r}\right) \ge$$

$$E\left(1_B E\left(F^*_{k-p}((t_1 - W_1))\, F^*_{k-p}((t_2 - W_3 - W_4)\right) / W_1, W_3, W_4, W_5, \ldots, W_{2r}\right)$$

where B is a event in the sample space of $\{W_1, W_3, W_4, W_5, \ldots, W_{2r}\}$. The case r=2 is easily reformulated without conditional expectation.

Finally, since $0 \le F^*_{k-p}(.) \le 1$, we have that $F^*_{k-p}(u \wedge v) \ge F^*_{k-p}(u) \cdot F^*_{k-p}(v)$ which concludes the proof.

Changing F by 1-F and $\wedge$ by $\vee$ with the same argument we get the inequality

$$P(\mathcal{U}_1>t, \mathcal{U}_2>t, \ldots, \mathcal{U}_r>t) \ge P(\mathcal{V}_1>t, \mathcal{V}_2>t, \ldots, \mathcal{V}_r>t).$$

(b) follows in an analogous way to (a).

On critical values for exponential data. A conservative method based on a Berry-Esséen type bound was considered to obtain the critical values on example 2.2. ; namely that for the normalized convolution of k double-exponential distribution functions with variance σ, we have that

$$|F_{k,\sigma}(x) - \Phi(x/\sigma)| \leq (\pi k)^{-1} \tag{4.3}$$

where $F_{k,\sigma}(x)$ stands for the distribution function of the sum of k independent double-exponential random variables with variance σ, and Φ denotes the standard normal distribution. To prove (4.3) we begin with the following inequality, which can be found, for instance, in Feller (1966), Lemma 2, p.512)

$$|G_k(x) - \Phi(x/\sigma)| \leq \pi^{-1} \int_{-\infty}^{+\infty} |t|^{-1} |\psi^k(t) - e^{-t^2/2}| dt.$$

where G_k denotes the k-convolution of G with itself, and ψ is the characteristic function of G.

Therefore if F is the double-exponential distribution we have that

$$\sup_x |F_{k,\sigma}(x) - \Phi(x/\sigma)| = \sup_x |F_{k,1}(x) - \Phi(x)| \leq$$

$$\pi^{-1} \int_{-\infty}^{+\infty} |t|^{-1} |(1+t^2/2k)^{-k} - e^{-t^2/2}| dt \leq \pi^{-1} \int_{-\infty}^{+\infty} |t|^{-1} |(1+t^2/2k)^{-k} - (1+t^2/2k)^{-(k+1)}| dt$$

since

$$(1+t^2/2k)^{-k} \geq e^{-t^2/2} \geq (1+t^2/2k)^{-(k+1)} .$$

Finally we have that

$$\pi^{-1} \int_{-\infty}^{+\infty} |t|^{-1} |(1+t^2/2k)^{-k} (1-(1+t^2/2k)^{-1})| dt = \pi^{-1} \int_{-\infty}^{+\infty} |t| (1+t^2/2k)^{-(k+1)}/2k dt$$

$$\leq (2k\pi)^{-1} \int_{-\infty}^{+\infty} |t| e^{-t^2/2} dt = (k\pi)^{-1}$$

which concludes the proof.

For k=131, we obtain then that $|F_{131,1}(x) - \Phi(x)| \leq 0.002429$. Then for example considered in Figure 2.2 we have that $F(a) \geq \Phi(4a(131)^{1/2}) - 0.00243 = (0.90)^{1/10} = 0.994883$, i.e. $4a(131)^{1/2} = 2.7836 = \Phi^{-1}(0.994883 + 0.00243)$ and $a = 0.0608$.

Proof of Lemma 3.1. a) Let $\mathcal{A}_i$ and $\mathcal{B}_i$ be the sets $\mathcal{A}_i = \{g(x_i), g(x_i+\delta), \ldots, g(x_i+(k-1)\delta)\}$ and $\mathcal{B}_i = \{g(x_i+[(k-m)/2]\delta), \ldots, g(x_i+[(k+m)/2]\delta)\}$ respectively, and consider the two complementary situations:

i_1) $\tilde{G}_i \in \mathcal{B}_i$

i_2) $\tilde{G}_i \in \mathcal{A}_i \setminus \mathcal{B}_i$

It can be easily be seen that in the first case we have that $|g(z) - \tilde{G}_i| \le c_i\delta(m+1)/2$, while if i_2) holds $|g(z) - \tilde{G}_i| \le c_i\delta(k+1)/4$, if we take into account that g(.) is a continuous function and the subintervals are of the form $[x_i+l\delta, x_i+(l+1)\delta]$, all of the same lenght. Note that the first property implies the connection of the graph of g(.).

b) follows easily from the fact that the monotonicity of g ensures that $\tilde{G}_i \in \mathcal{B}_i$.

<u>Proof of Lemma 3.2.</u> a) Let Z_i be the random variable defined by $Z_i = \mathbf{1}_{\{U_i \le t\}}$. Then,

$$P(\mathcal{U}_1^{(h)} \le t, \mathcal{U}_2^{(h)} \le t, \dots, \mathcal{U}_r^{(h)} \le t) =$$

$$P\Big(Z_1+\dots+Z_k \ge h,\ Z_{p+1}+\dots+Z_k+\dots+Z_{p+k} \ge h,\ Z_{2p+1}+\dots+Z_{2p+k} \ge h, \dots$$

$$, Z_{(r-1)p+1}+ \dots +Z_{(r-1)p+k} \ge h\Big)$$

and

$$P(\mathcal{U}_1^{(h)} \ge t, \mathcal{U}_2^{(h)} \ge t, \dots, \mathcal{U}_r^{(h)} \ge t) =$$

$$P\Big(Z_1+ \dots +Z_k \le h-1,\ Z_{p+1}+\dots+Z_{p+k} \le h-1,\ \dots, Z_{(r-1)p+1}+ \dots +Z_{(r-1)p+k} \le h-1\Big).$$

Now consider as in the proof of Lemma 2.2 the descomposition given by

$$\sum_{i=1}^{p} Z_i + \sum_{i=p+1}^{k} Z_i = W_1+W_2$$

$$\sum_{i=p+1}^{k} Z_i + \sum_{i=k+1}^{2p} Z_i + \sum_{i=2p+1}^{p+k} Z_i = W_2+W_3+ W_4$$

$$\sum_{i=2p+1}^{p+k} Z_i + \sum_{i=p+k+1}^{3p} Z_i + \sum_{i=3p+1}^{2p+k} Z_i = W_4+W_5+ W_6$$

.....

$$\sum_{i=(r-1)p+1}^{(r-1)p+k} Z_i + \sum_{i=(r-1)p+k+1}^{rp} Z_i + \sum_{i=rp+1}^{(r-1)p+k} Z_i = W_{2r-2}+W_{2r-1}+ W_{2r}$$

and the proof can by obtained following line by line the argument already used in the proof of Lemma 2.2.

<u>Remark 4.2</u>. If we substitute t in Lemma 3.2 t by $t_1, \dots, t_r$ and h by $h_1,\dots,h_r$ we get a similar inequality to that of the statement of Lemma 3.2.

References.

Feller, W. (1966). An Introduction to Probability Theory and its Applications. J. Wiley, vol. 2.

Hall, P. and Titterington, D.M. (1988). "On Confidence Bands in Nonparametric Density Estimation and Regression". J. Multivariate Anal., 27, 228-254.

Härdle, W. and Bowman, A.W. (1988). "Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands". J. Amer. Statist. Assoc. 83, 102-110.

Knafl, G., Sacks, J. and Ylvisaker, D. (1985). "Confidence Bands for Regression Functions". J. Amer. Statist. Assoc. 80, 683-691.

Leadbetter, M.R., Lindgren, G. and Rootzen, H. (1983). Extremes and related properties of random sequences and processes. Springer-Verlag, N.Y. , Heidelberg, Berlin.

Silverman, B. W. (1985). "Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion)". J. Roy. Statist. Soc. Ser. B, 47, 1-52.

Wahba, G. (1983). "Bayesian confidence intervals for the cross-validated smoothing spline". J. Roy. Statist. Soc. Ser. B. 45, 133-150.

DATA-ADAPTIVE KERNEL ESTIMATION

THEO GASSER
Zentralinstitut für seelische Gesundheit
I5
6800 Mannheim 1
Germany

EVA HERRMANN
Sonderforschungsbereich 123
Universität Heidelberg
Im Neuenheimer Feld 294
6900 Heidelberg
Germany

ABSTRACT. Kernel estimators are by now well established for curve estimation in a broad variety of problems. Here the regression problem is studied. Basic properties are summarized and a comparison with other popular estimators is then made. The choice of bandwidth or smoothing parameter is decisive in many ways; a data-adaptive optimization of this choice offers many advantages. An iterative plug-in rule based on the asymptotic formula is introduced and contrasted with cross-validation type selectors. The latter prove to be inferior asymptotically and in simulations mainly due to their large variability. A difficult problem is then choosing the bandwidth appropriately when residuals are correlated. Introducing a second tuning parameter to adapt to the unknown correlation structure leads to a solution which shows good properties in theory and practice. It is important that this may be achieved without postulating some parametric model for the residual time series.

1. Introduction

Let us assume that data $Y_1, \ldots, Y_n$ have been collected on some fixed design $0 \leq t_1 \leq \ldots \leq t_n \leq 1$ to determine some response function m according to the following model:

$$Y_i = Y(t_i) = m(t_i) + \varepsilon_i \qquad (i = 1, \ldots, n) \tag{1}$$

where the ε_i are independent zero mean residuals. A constant variance $\sigma^2 = E(\varepsilon_i^2)$ is assumed for the sake of simplicity only. The random design model with $m(t_i) = E(Y_i|t_i)$ is at some place also considered, when (Y_i, t_i) is a bivariate sample with ordered random variables $\{t_i\}$ from a sample of random variables with marginal density f. The fixed design may be distributed "regularly" according

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 67–79.
© 1991 *Kluwer Academic Publishers.*

to some density f in the following sense: $t_i = F^{-1}\big((i-0.5)/n\big)$ where F is some distribution on $[0,1]$ with density $f = F'$. Whenever m is nonlinear in t, it is difficult - often close to arbitrary - which of the infinity of nonlinear parametric models to choose. The choice of a wrong model inevitably leads to bias, which is often substantial (Gasser et al., 1984). Once such a model has been established in some field of application, it is difficult to change. Thus bias effects may have a long life, since they cannot be eliminated by repeating experiments, in contrast to variability. Thus, there is good reason for investigating nonparametric regression estimators.

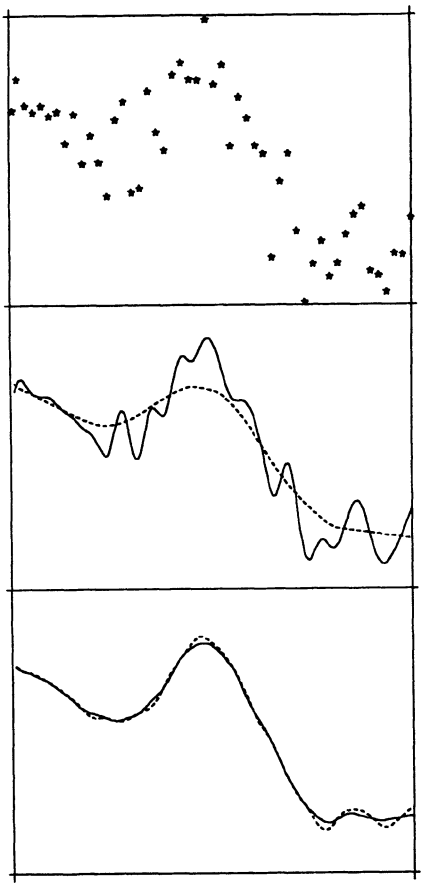


Figure 1. Simulated example with the data in the upper part, over- and under-smoothed kernel estimators in the middle part, and plug-in estimator and optimal kernel estimator in the lower part.

In this paper we want to draw a line between some of our own developments,

rather than focus on details of derivation. After introducing kernel estimators of the convolution type (Gasser and Müller, 1979, 1984; Gasser et al., 1985), we will concentrate on two topics:

I. Why choose these kernel estimators and not any of the other nonparametric methods such as smoothing splines or Nadaraya-Watson-type kernel estimators?

II. How should we select the bandwidth from the data when residuals are independent - when they are dependent? (The latter problem is obviously much more difficult)

Problem II is illustrated in fig. 1 for simulated data (n=50), showing the effect of over- and undersmoothing by choosing some fixed bandwidth and also the data-adaptive plug-in estimator to be described late in comparison with the optimal kernel fit. Obviously, there is not much to choose between these two.

2. Definitions, Basics and Comparisons with other Estimators

In many applications, it is important to obtain derivatives of the underlying true curves instead or in addition to the regression function itself. The kernel estimator $\hat{m}_\nu$ $(\nu = 0, 1, 2)$ of $m^{(\nu)}$ is defined as follows:

$$\hat{m}_\nu(t; h) = \sum_{i=1}^{n} g_i(t; h; \nu) Y_i \tag{2}$$

where: $$g_i(t; b; \nu) = \frac{1}{h^{\nu+1}} \int_{s_{i-1}}^{s_i} K_\nu\left(\frac{t-u}{h}\right) du$$

$$s_i = (t_i + t_{i+1})/2$$

$$K_\nu = \text{kernel}$$

$$h = \text{bandwidth}$$

In order to apply estimator (2), we have to specify K_ν and h. Kernels satisfying various optimality criteria can be derived analytically (Gasser et al., 1985), and they are valid independent of m and σ^2 and apply also when estimating derivatives. By allowing for certain asymmetric weighting schemes, it is possible to estimate till the boundary of the data, while keeping the rates of the interior. Inevitably, a price in perfomance has to be paid for small sample size when estimating at the boundary especially when estimating derivatives. Optimal kernels can be represented as Legendre polynomials and this in turn leads to a fast algorithm for kernel estimation, working in $O(n)$ operations. A sketch of this algorithm has been given in Gasser and Kneip (1989).

The optimal bandwidth depends essentially on the problem at hand (for an asymptotic formula, see below). For a better understanding of the properties of

the estimators, bias and variance are essential (Gasser and Müller, 1984):

$$\mathrm{Bias}\big(\hat{m}_\nu(t;h)\big) = (h^{k-\nu}/k!) \cdot \int_{-1}^{1} K_\nu(x) x^k dx \cdot m^{(k)}(t) + o(h^{k-\nu}) \tag{3}$$

$$\mathrm{var}\big(\hat{m}_\nu(t;h)\big) = \frac{\sigma^2}{nh^{2\nu+1} f(t)} \int_{-1}^{1} K_\nu(x)^2 dx + o\Big(\frac{1}{nh^{2\nu+1}}\Big).$$

Here, k is the order of the kernel, i.e. the order of the first non-vanishing moment. For random design, variance is different and has to be multiplied by some factor: it is 2 when putting $s_i = t_i$ (done for technical convenience in Jennen-Steinmetz and Gasser, 1988), but it is 1.5 for $s_i = (t_i + t_{i+1})/2$ (Chu and Marron in an unpublished manuscript). Some simple convergence properties may be obtained from (3), and more refined ones may be found in the literature. Relations (4) lead to the asymptotic mean itegrated squared error AMISE and minimization of AMISE with respect of h to the asymptotically optimal bandwidth h_{asy}:

$$h_{\mathrm{asy}} = \Big(C \cdot \frac{\sigma^2 \int f(t)^{-1} dt}{\int m^{(k)}(t)^2 dt} \cdot \frac{1}{n}\Big)^{1/(2k+1)} \tag{4}$$

where, of course σ^2 and $\int m^{(k)}(t)^2 dt$ depend crucially on the data at hand. There is a rich literature on other nonparametric curve estimators, in particular:

1. Other types of kernel estimators (Priestley-Chao; Nadaraya-Watson type),
2. smoothing splines,
3. k-nearest-neighbour (k − NN) estimators.

The question then comes up, whether any of these is best or superior for some class of problems. For equally spaced design,they are equivalent, and the estimator suggested by Priestley and Chao (1972) is asymtotically equivalent to ours in general. The kernel estimator $\tilde{m}$ considered by Nadaraya (1964) and Watson (1964) for $\nu = 0$ is popular and often thought to be 'natural' for random design.

$$\tilde{m}(t;h) = \frac{\sum_{i=1}^{n} K\big((t-t_i)/h\big) Y_i}{\sum_{i=1}^{n} K\big((t-t_i)/h\big)} \tag{5}$$

Its variance is in fact the one given in (3) both for fixed and random design, but its bias is rather awkward ($k = 2$):

$$\mathrm{Bias}\big(\tilde{m}(t;h)\big) = h^2 \int_{-1}^{1} K(x) x^2 dx \big[m''(t)/2 + m'(t) f'(t)/f(t)\big] + o(h^2). \tag{6}$$

It has been argued in Gasser and Engel (1990) that the dependence of the bias

on m' and f'/f may jeopardize a clear interpretation of the estimated curves. Furthermore, the Nadaraya-Watson estimator is inferior to ours in a minimax-sense, and may in fact become arbitrarily bad in MISE. Regarding the other estimators, it may be advisable to introduce a design-variable bandwidth $b(t) = b_0 \cdot f(t)^{-\alpha}$ $(0 \leq \alpha \leq 1)$ (Jennen-Steinmetz and Gasser, 1988). The ordinary kernel estimator corresponds then to $\alpha = 0$, the smoothing spline to $\alpha = 1/4$ (Silverman, 1984) and the k – NN estimator to $\alpha = 1$. The properties of estimators may then be studied via the real parameter α. As it turned out none of the estimators is uniformly optimal, but the kernel estimator ($\alpha = 0$) is minimax optimal. Beyond minimax optimality, one may say that k – NN estimators perform well only when the design f and the structure of the regression problem (in terms of m'') are well-matched.

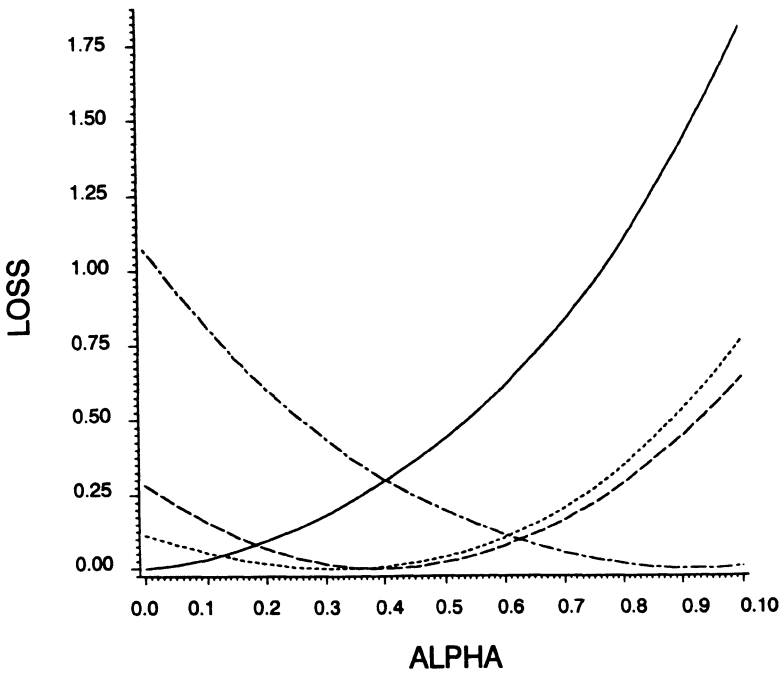


Figure 2. Loss in MISE for variable bandwidth kernel estimator for a Beta(8, 8) design distribution and different regression functions consisting of Gaussion peak at 0.2 (joined line), 0.3 (dotted line), 0.4 (dashed line) and 0.5 (dash-dotted line) respectively.

This is illustrated in fig. 2, depicting

$$\mathrm{Loss}(\alpha) = \big(\mathrm{MISE}(\alpha) - \mathrm{MISE}(\alpha_{\mathrm{opt}})\big)/\mathrm{MISE}(\alpha_{\mathrm{opt}})$$

for a Beta(8, 8) distributed design and a regression function consisting of a Gaussian peak centerd either at 0.2, 0.3, 0.4 or 0.5. The more fundamental problem of specifying the best smoothers - in terms of the asymptotic constant - has been dealt with in Nussbaum (1985) and Speckman (1985).

3. Bandwidth Selection for Independent Residuals

The choice of the smoothing parameter h is decisive for the statistical quality of the resulting estimate. A data-adaptive optimaization of the bandwidth could thus be statistically beneficial, but could also make the method more objective and more economical to use. A subjective choice is in many fields of application not well tolerated for general scientific reasons. The effect of under- and oversmoothing was illustraged in figure 1 for a set of simulated data. Two different criteria of perfomance appeared in the literature.

$$\mathrm{ISE}(h) = \int_0^1 v(t)\big(\hat{m}_\nu(t;h) - m(t)\big)^2 dt \tag{6}$$

$$\mathrm{MISE}(h) = \mathrm{E}\big(\mathrm{ISE}(h)\big). \tag{7}$$

Here v is a weight function introduced to avoid boundary effects. The prominent method for bandwidth selection for density and regression problems was cross-validation (see Härdle et al., 1988, and the literature cited there). Cross-validation consists in computing the sum-of-squares of one-leave-out residuals and minimizing it with respect to the bandwidth. A number of methods based on ordinary residuals (generalized cross-validation, minimum unbiased risk estimation etc.) is asymptotically equivalent (Härdle et al., 1988) to cross-validation. They share quite large a variability in the bandwidth selected, which makes them somewhat unreliable. There are also some algorithmic problems with cross-validation. The variability problem led us to start investigating an estimator based on h_{asy} (4). For the unknown quantities σ^2 and $\int v(t)m^{(k)}(t)^2 dt$ estimators have to be plugged-in, and the resulting bandwidth selector is thus called a "plug-in estimator $\hat{h}_{\mathrm{opt}}$". The algorithm works as follows:

1. Estimate σ^2 nonparametrically as $\hat{\sigma}^2$ following Gasser et al. (1986).

2. Estimate the optimal bandwidth h_{opt} with estimation of $\int v(t)m^{(k)}(t)^2 dt$ by the following iterative procedure (described for $k = 2$ for simplicity):

   (i) Set $\hat{h}_0 = 1/n$.
   (ii) Iterate $i = 1, 2, \ldots$

$$\hat{h}_i = \left(C \cdot \frac{\hat{\sigma}^2}{\int_0^1 v(t)\hat{m}_2(t;\hat{h}_{i-1}\cdot n^{1/10})^2 dt} \cdot \frac{1}{n}\right)^{1/5} \tag{8}$$

(iii) Stop after 11 iterations and set $\hat{h}_{\text{opt}} = \hat{h}_{11}$.

For a detailed treatment of this estimator the interested reader may consult Gasser et al. (1990). It seems rather natural to use kernel estimators to determine the second derivative in the denominator; the real problem is the choice of the pilot bandwidth for estimating the respective functional. To perform iterations is essential, since an arbitrarily prespecified pilot bandwidth may be grossly wrong. It is clear that this pilot bandwidth has to be inflated with respect to the bandwidth for estimating m itself, and $n^{1/10}$ is the asymptotically correct rate to optimize on variability (see the Theorem below). The 11 iterations are also in accordance with the theory; the first 7 iterations are necessary to achieve the right order and thereafter the variance gets improved. Figure 3 illustrates for simulated data how the iterations approach the unknown optimal value, and compares this with cross-validated bandwidth (two good, two bad, and two medium realizations habe been choosen out of 400 replicates).

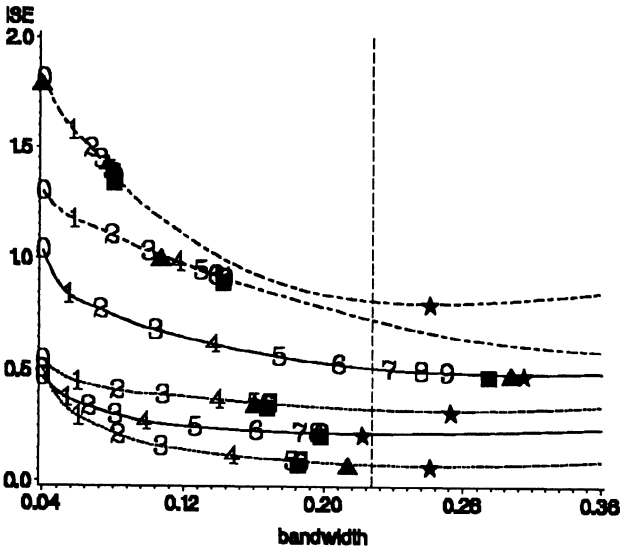


Figure 3. Six realizations in ISE for a sinefunction with $n = 25$, $\sigma^2 = 1$, two at 0.1 quantile $\text{ISE}(\hat{h}_{\text{opt}}/\min \text{ISE}(h)$ (solid line), two at 0.5 quantile (dashed), two at 0.9 quantile (dotted); quantiles with respect to plug-in. Numbers indicate iterations, ■ = plug-in selector, ▲ = cross-validation selector, ★ = true optimum. Curves are displaced by 0.1 each to improve visibility.

Let S_p denote a class of smooth functions, i.e. of functions whose ℓ-th derivative belongth to the Lipschitz-class $\mathrm{Lip}(p-\ell)$, with $\ell < p \leq \ell + 1$ and $p \geq 0$, and assume some further regularity conditions to be satisfied.

Then the following asumptotic results hold for the plug-in estimator $\hat{h}_{\mathrm{opt}}$ of the MISE-optimal h_{opt} $(\nu = 0,\ k = 2)$:

THEOREM. Assume $m \in S_p$ $(p > 2)$ and $m''(t) \neq 0$ for some $t \in [\delta, 1-\delta]$. Set $q = \min\{2, p-2\}$. After $(8 + 5/q)$ iterations the following holds:

$$\hat{h}_{\mathrm{opt}} = h_{\mathrm{opt}}\big(1 + O(n^{-\beta}) + O_p(n^{-(3+q)/10})\big)$$

with a β satisfying

(a) $\beta \geq q/10$ in general.

(b) $\beta \geq \big(2 + \min\{2, p-4\}\big)/10$ if $p > 4$ and if a 6-th order kernel is used for the second derivative.

(c) $\beta = 1/2$ if $p > 7$ and if a very special 8-th order kernel is used for the second derivative.

The first term represent the deterministic part of the error, the second term the variability. For the latter, a relative rate $O_p(n^{-1/2})$ can be achieved contrasted to a relative rate $O_p(n^{-1/10})$ for cross-validation type estimators (Härdle et al., 1988). Hall and Marron (1989) have shown in the context of density estimation, that the latter rate represents an upper bound. This result is suprising in view of the fact that minimum unbiased risk estimation is well established in parametric model selection (Mallows, 1973), but shares in our problem the poor rate of convergence with cross-validation. The suspicion of a break down of the plug-in estimator in case of non-smooth functions cannot be confirmed: it does neither break down theoretically (Gasser et al., 1990) nor in the simulations, but its rate is then inferior to cross-validation.

It is more difficult to choose the vector of bandwidth appropriately in high dimensional regression problems by an eyeball analysis. The data-adaptive plug-in estimator just presented has been successfully generalized to two dimensions in Wand et al. (1990); for various reasons the two-dimensional problem is of particular interest, and it carries the flavour of higher-dimensional problems. Figure 4 illustrates this for a simulated 2-dimensional regression problem, comparing data with the plug-in fit, the true optimal data fit and the true surface. The one realization with median perfomance is shown which is obviously good. Estimating derivatives from noisy data is also more intricate than the regression itself and of substantial practical interest. The plug-in bandwidth selector can also be generalized to this class of problems (Herrmann et al., 1990).

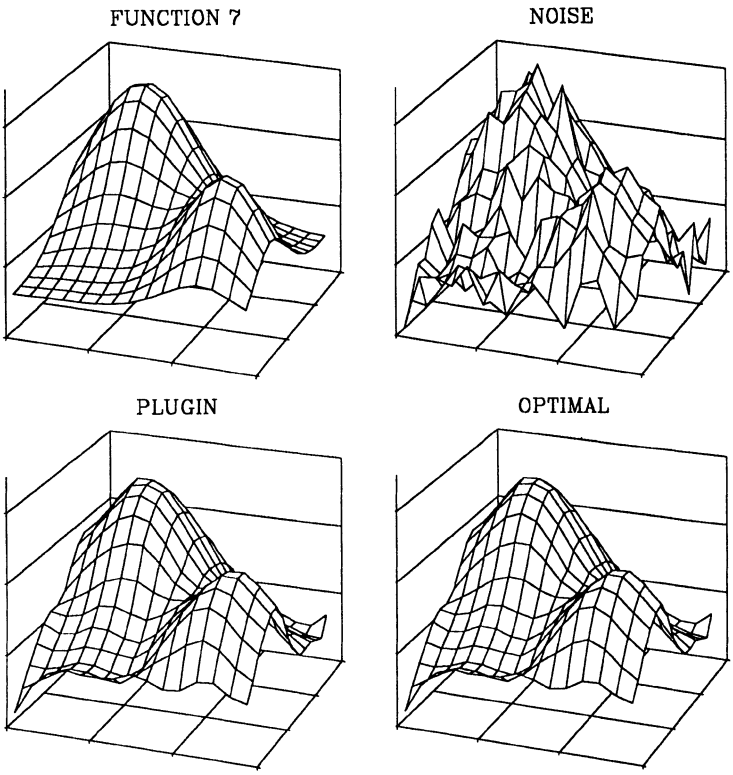


Figure 4. Simulation example with $15 \cdot 15$ points on a rectangular grid. Realization with median performance in ISE of the two-dimensional plug-in estimator. The upper left part shows the true regression function, the upper right part the noisy data, the lower left part the two-dimension plug-in estimator and the lower right part the true optimal kernel estimator for this data.

4. Bandwidth Selection for Regression with Dependent Residuals

Independence is a crucial assumption in many statistical procedures, and this is also true for nonparametric regression. Some aspects have been covered in Hart and Wehrly (1988). Here, only bandwidth selection is considered. When naively applying bandwidth selection valid for independent residuals, very bad regression fits are usually obtained for correlated residuals. This becomes obvious in applications and can be explained theoretically in a straight-forward way (Hart, 1988, see also below). The common pattern is a much too small bandwidth, close to interpolation, since the correlated noise is mistaken for signal under assumption of independence. Cross-validation has been modified by leaving out $k \geq 1$ points to suppress some dependence structure (Vieu and Hart, 1989). This introduces

a further so far unspecified parameter in a technique with major drawbacks (sec. 3). Another possibility is to assume some parametric model for the residual time series (Chiu, 1989). Apart from some technical problems with this approach, it also contracdicts the nonparamtric handling of the regression problem.

When developping our method based on the results of sec. 3 (Herrmann et al., 1989), we wanted a bandwidth selector that would work for a broad variety of correlation patterns and would perform well in case of independence. Let us assume equally spaced data and $k = 2$ for convenience. The following asymptotically MISE-optimal bandwidth h_{asy} can easily be obtained:

$$h_{\text{asy}} = \left(C \cdot \frac{S}{\int m''(t)^2 dt} \cdot \frac{1}{n}\right)^{1/5} \tag{9}$$

$$\text{where} \qquad S = \gamma_0 + 2\sum_{\nu=1}^{\infty} \gamma_\nu \tag{10}$$

$$\gamma_\nu = \mathrm{E}\varepsilon_i \varepsilon_{i+\nu}.$$

In the frequent case where $S > \sigma^2$, too small a bandwidth results when falsely assuming independent residuals. The plug-in estimator $\hat{h}_{\text{opt}}$ (8) of sec. 3 for estimating the MISE-optimal bandwidth h_{opt} is again used. The estimator $\hat{\sigma}^2$ has to be replaced by indiciously chosen estimator of the functional S. For m-dependent residuals Müller and Stadtmüller (1988) proposed an estimator of S based on residuals $\hat{\varepsilon}_{i,\nu} = Y_i - Y_{i-\nu}$. In order to improve on bias, we consider the generalization of the residuals introduced in Gasser et al. (1986):

$$\hat{\varepsilon}_{i,\nu,\mu} = Y_i - \tfrac{\nu}{\nu+\mu} Y_{i+\mu} - \tfrac{\mu}{\nu_\mu} Y_{i-\nu}. \tag{11}$$

With the help of the residuals, we can obtain an estimator $\hat{\gamma}_{nu}$ for γ_ν after suitable algebraic manipulations, leading to an estimator $\hat{S}_m$ of S (Herrmann et al., 1990):

$$\hat{S}_m = \sum_{\nu=-m}^{m} \hat{\gamma}_\nu \tag{12}$$

For m-dependent residuals and linear functions m, $\hat{S}_m$ is an unbiased estimator of S. Assuming some mixing condition for the ε_i and some smoothness for m, $\hat{S}_m$ is consistent when $m \to \infty$, $m^2/n \to 0$ as $n \to \infty$. Thus, we have constructed a further approximation problem for the time series which results in choosing the integer value m "well". Two rules for choosing m adaptively from the data have been derived on intuitive grounds and some asumptotic arguments. Both chowed a rather good performance in extensive simulations. An example is given in fig. 5, showing boxplots for the bandwidth, in comparison the true optimal ones, the optimal ones estimated from the data using the plug-in estimator plus one of the two rules for choosing m and the naive selector assuming independent residuals.

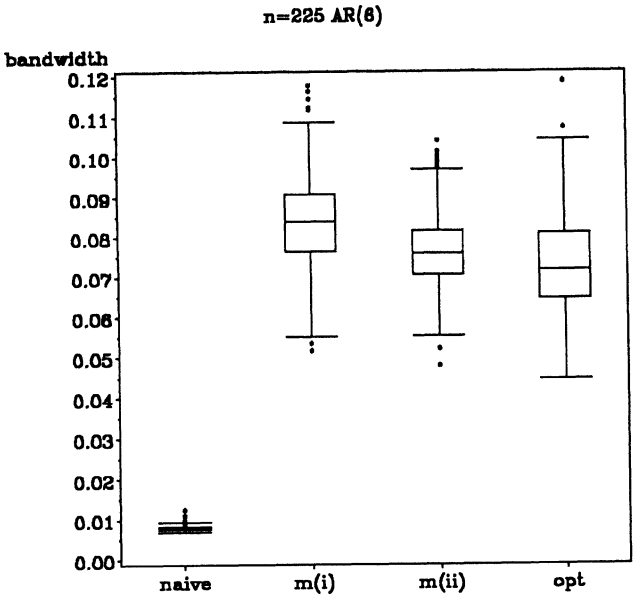


Figure 5. Boxplots for estimated bandwidths for 400 simulations with $n = 225$ and residuals following an AR(6) process with spectral density with a decreasing trend superposed by two peaks. Comparison of both modified plug-in estimators with data-driven choice of m, the naive plug-in estimator and the true ISE-optimal bandwidth.

Under assumptions which are not too restrictive, an asymptotic result similar to the one of sec. 3 can be derived:

$$\hat{h}_{\text{opt}} = h_{\text{opt}}\big(1 + O(n^{-q/10} + |\mathrm{E}(\hat{S}) - S|) + O_p(n^{-(3+q)/10} + |\hat{S} - \mathrm{E}(S)|)\big)$$

It is a nice feature that the effects of estimating the functional $\int m''(t)^2 dt$ and the functional S seperate into different additive terms.

References

Chiu, S.T. (1989) 'Bandwidth selection for kernel estimate with correlated noise', Statist. Probab. Lett. **8**, 347–354.

Chu, C.K. and Marron, J.S. (1989) 'Choosing a kernel regression estimator', Report, University of North Carolina, Chapel Hill.

Gasser, Th. and Engel, J. (1990) 'The choice of weights in kernel regression estimation', Biometrika **77**, 377–381.

Gasser, Th. and Kneip, A. (1989) Discussion of 'Linear smoothers and additive models' by A. Buja, T. Hastie and R. Tibshirani. Ann. Statist. **17**, 532–535.

Gasser, Th.; Kneip, A. and Köhler, W. (1990) 'A flexible and fast method for automatic smoothing', J. Amer. Statist. Assoc., to appear.

Gasser, Th. and Müller, H.G. (1979) 'Kernel estimation of regression functions', in Th. Gasser and M. Rosenblatt (eds.). Smoothing techniques for curve estimation, Springer, Heidelberg, pp. 23–68.

Gasser, Th. and Müller, H.G. (1984) 'Estimating regression functions and their derivatives by the kernel method', Scand. J. Statist. **11**, 171–185.

Gasser, Th.; Müller, H.G.; Köhler, W.; Molinari, L. and Prader, A. (1984) 'Nonparametric regression analysis of growth curves,' Ann. Statist. **12**, 210–229.

Gasser, Th.; Müller, H.G. and Mammitzsch, V. (1985) 'Kernels for nonparamtric curve estimation', J. Roy. Statist. Soc. Ser. B **47**, 238–252.

Gasser, Th.; Sroka, L. and Jennen-Steinmetz, Ch. (1986) 'Residual variance and residual pattern in nonlinear regression', Biometrika **73**, 625–633.

Härdle, W.; Hall, P. and Marron, J.S. (1988) 'How far are automatically chosen regression smoothing parameters from their optimum?' J. Amer. Statist. Assoc. **83**, 86–95.

Hart, J.D. (1988) 'Kernel regression estimation with time series errors', Report.

Hart, J.D. and Wehrly, T. E. (1986) 'Kernel regression estimation using repeated measurements data', J. Amer. Statist. Assoc. **81**, 1080–1088.

Herrmann, E.; Gasser, Th. and Kneip, A. (1990) 'Choice of bandwidth for kernel regression when residuals are correlated', Report, Universität Heidelberg.

Herrmann, E.; Kneip, A. and Gasser, Th. (1990) 'Differentiation with kernel estimators: choice of bandwidth', Report, Universität Heidelberg.

Jennen-Steinmetz, Ch. and Gasser, Th. (1988) 'A unifying approach to nonparametric regression estimation', J. Amer. Statist. Assoc. **83**, 1084–1089.

Mallows, C. (1973) 'Some comments on C_p', Technometrics **15**, 86–101.

Müller,H.G. and Stadtmüller, U. (1988) 'Detecting dependencies in smooth regression models', Biometrika **75**, 639–650.

Nadaraya, E.A. (1964) 'On estimating regression', Theory Prob. Applic. **9**, 141–142.

Nussbaum, B.W. (1985) 'Spline smoothing in regression models and asymptotic efficiency in L_2', Ann. Statist. **13**, 984–997.

Priestley, M.B. and Chao, M.T. (1972) 'Nonparametric function fitting', J. Roy. Statist. Soc. Ser. B **34**, 385–392.

Silverman, B.W. (1984) 'Spline smoothing: the equivalent variable kernel method', Ann. Statist. **12**, 898–916.

Speckman, P. (1985) 'Spline smoothing and optimal rates of convergence in nonparametric regression', Ann. Statist. **13**, 970–984.

Vieu, P. and Hart, J. D. (1988) 'Nonparametric regression under dependence: A class of asymptotically optimal data-driven bandwidths', Report.

Wand, M.; Herrmann, E.; Engel, J. and Gasser, Th. (1990) 'A bandwidth selector for bivariate kernel regression', Report, Universität Heidelberg.

Watson, G.S. (1964) 'Smooth regression analysis', Sankhyā Ser. A **26**, 359–372.

ON THE NONPARAMETRIC ESTIMATION OF THE ENTROPY FUNCTIONAL

László GYÖRFI
Hungarian Academy of Sciences
Technical University of Budapest
Stoczek u.2
H-1111 Budapest, Hungary

Edward C. VAN DER MEULEN
Department of Mathematics
Katholieke Universiteit Leuven
Celestijnenlaan 200B
B-3030 Leuven, Belgium

ABSTRACT. Let $f(x)$ be an unknown probability density function, $x \in R^d$. The differential entropy functional $H(f)$ is defined by

$$H(f) = -\int_{R^d} f(x)\log f(x)dx .$$

We consider the problem of estimating $H(f)$ nonparametrically, based on a random sample $X_1,\ldots,X_n$ from the underlying density. Several methods to estimate $H(f)$ have been put forward in the literature. Here, a new class of entropy estimators is considered. The common feature of these estimators is that they are based on an L_1-consistent density estimator $f_n(x)$. We show that a certain slow a.s. L_1 rate of convergence of $f_n(x)$ implies the a.s. convergence of the corresponding entropy estimators $H_n(f_n)$. One of the advantages of these new entropy estimators is that the only condition put on the density function $f(x)$ is that $H(f)$ be finite. Furthermore, we analyze the conditions on f_n and H_n which guarantee the a.s. convergence of these new entropy estimators.

1. Introduction

Let $X_1,\ldots,X_n$ be independent, identically distributed random vectors, taking values in R^d, $d \geq 1$, with common unknown probability density function $f(x)$. We wish to estimate the Shannon (or differential) entropy of $f(x)$, defined by

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 81–95.
© 1991 *Kluwer Academic Publishers.*

$$H(f) = -\int_{R^d} f(x)\log f(x)dx\,. \tag{1.1}$$

In the literature, various estimators for H(f), based on a random sample $X_1,\ldots,X_n$ from the underlying distribution, have been proposed and their properties studied, mostly for the case d = 1. For an exhaustive list of references in this regard see [8].

Based on a nonparametric density estimate $f_n(x)$, there are essentially two different ways to estimate H(f), either by

$$H_n = H_n(f_n) = -\frac{1}{n}\sum_{i=1}^{n} \log f_n(X_i)I_{\{X_i \in A_n\}} \tag{1.2}$$

or by

$$\tilde{H}_n = \tilde{H}_n(f_n) = -\int_{A_n} f_n(x)\log f_n(x)dx\,, \tag{1.3}$$

with $A_n \to R^d$ as $n \to \infty$.

In Györfi and van der Meulen [7], two new entropy estimators were recently introduced for the case $d \geq 1$, which are histogram-based in the sense that they involve a histogram-based density estimator $\hat{f}_n$. In [7] the a.s. consistency of estimators of type (1.2) and (1.3) based on $\hat{f}_n$ is proved, with as only condition on f(x) that H(f) is finite. There it turned out that the underlying histogram-based density estimator was L_1-consistent as well, due to the conditions assumed to guarantee the a.s. consistency of the entropy estimator.

In [7] we raised the following question : given any density estimator f_n which is L_1-consistent, what additional properties should it satisfy so that the empiric entropies of type (1.2) and (1.3) are a.s. consistent ? In this paper we address ourselves to this question concerning the entropy estimator of type (1.2) only, and provide such additional conditions. Our results are formulated in the Theorem below. (For empiric entropies of type (1.3) such conditions can be given as well, but we are omitting these results here, as in practice the integral in (1.3) will be extremely difficult to evaluate if f_n is not histogram-based.) We are thus able to extend the class of estimators of type (1.2) proposed in [7] considerably, viz. from those which are just histogram-based, to those which are L_1-consistent. The conditions under which our results hold are subsequently analyzed.

2. Almost Sure Consistency Result for Estimator H_n Based on L_1-Consistent f_n

We assume that the functional $H(f)$ to be estimated is well-defined and finite. Let be given a sample $X_1,\ldots,X_n$ of independent, identically distributed (i.i.d.) random vectors, taking values in R^d, $d \geq 1$, with as underlying probability density function the common unknown density $f(x)$. Let μ be the probability measure of X_1, so that

$$H(f) = -\int_{R^d} \log f(x)\mu(dx) . \tag{2.1}$$

Our estimator of $H(f)$ is constructed as follows.

We decompose the sample $\underset{\sim}{X} = \{X_1,\ldots,X_n\}$ into two subsamples $\underset{\sim}{Y} = \{Y_1,\ldots,Y_{[\frac{n}{2}]}\}$ and $\underset{\sim}{Z} = \{Z_1,\ldots,Z_{[\frac{n+1}{2}]}\}$, defined by

$$Y_i = X_{2i} \quad \text{for} \quad i = 1,\ldots,[\tfrac{n}{2}] , \tag{2.2}$$

and

$$Z_i = X_{2i-1} \quad \text{for} \quad i = 1,\ldots,[\tfrac{1}{2}(n+1)] , \tag{2.3}$$

where $[k]$ denotes the largest integer $\leq k$.

Now let $f_n = f_n(x) = f_n(x,\underset{\sim}{Z})$ be any density estimate based on $\underset{\sim}{Z}$ such that f_n is L_1-consistent, i.e., such that

$$\lim_{n\to\infty} \int_{R^d} |f_n(x) - f(x)|\,dx = 0 \quad \text{a.s.} \tag{2.4}$$

Our entropy estimator $H_n = H_n(f_n) = H_n(\underset{\sim}{Y},\underset{\sim}{Z})$ is of the form (1.2) and defined by

$$H_n(f_n) = -\frac{1}{[n/2]} \sum_{i=1}^{[n/2]} \log f_n(Y_i,\underset{\sim}{Z}) I_{\{f_n(Y_i,\underset{\sim}{Z}) \geq a_n\}}, \tag{2.5}$$

where $0 < a_n < 1$ and

$$\lim_{n\to\infty} a_n = 0 . \tag{2.6}$$

Furthermore, let

$$\delta_n = \max\{2, \text{ess sup } (\sup_x f_n(x))\} . \tag{2.7}$$

We then have the following theorem.

THEOREM. Assume that conditions (2.4) and (2.6) are satisfied, and that for each $c > 0$

$$\sum_{n=1}^{\infty} \exp\left(-\frac{cn}{\max^2\{\log \delta_n, -\log a_n\}}\right) < \infty . \tag{2.8}$$

Moreover, assume that

$$\lim_{n\to\infty} (-\log a_n) \int_{f_n(x) \geq a_n} |f_n(x) - f(x)|\, dx = 0 \quad \text{a.s.} \tag{2.9}$$

Then, if $H(f)$ is finite,

$$\lim_{n\to\infty} H_n(f_n) = H(f) \quad \text{a.s.} \tag{2.10}$$

Proof. We first introduce the notations

$$H_n^* = -\int_{f_n(x) \geq a_n} f(x) \log f_n(x) dx = E(H_n | \underset{\sim}{Z}) , \tag{2.11}$$

$$U_n = H_n - H_n^* , \tag{2.12}$$

$$V_n = \int_{f_n(x) \geq a_n} f(x) \log f(x) dx - \int_{f_n(x) \geq a_n} f(x) \log f_n(x) dx , \tag{2.13}$$

and

$$W_n = \int_{f_n(x) < a_n} f(x) \log f(x) dx . \tag{2.14}$$

Then clearly

$$H_n - H(f) = U_n + V_n + W_n . \tag{2.15}$$

The Theorem will be proven if we establish the a.s. convergence of each of the terms (U_n,V_n, and W_n) on the righthand side of (2.15). This is accomplished by the following three lemma's.

LEMMA 1. If (2.8) is satisfied, then

$$\lim_{n\to\infty} U_n = 0 \qquad \text{a.s.} \tag{2.16}$$

Proof. As in the proof of Theorem 1 of [7, p. 430], we apply Bennett's inequality to the sequence of random variables

$$\xi_i = -\log f_n(Y_i) I_{\{f_n(Y_i) \geq a_n\}} - \int_{f_n(x) \geq a_n} (-\log f_n(x)) f(x) dx . \tag{2.17}$$

Bennett's inequality (cf. [6, p. 160]) states that if $\xi_1,\ldots,\xi_n$ are i.i.d. zero mean variables with $|\xi_i| \leq t$, then, for each $\epsilon > 0$,

$$P(|\frac{1}{n}\sum_{i=1}^{n} \xi_i| > \epsilon) \leq 2 \exp\left(\frac{-n\epsilon^2}{4t^2}\right) . \tag{2.18}$$

Here we have

$$|\xi_i| \leq t = 2 \max\{-\log a_n, \log \delta_n\} , \tag{2.19}$$

and thus for any $\epsilon > 0$

$$P\left[|H_n - H_n^*| > \epsilon \middle| \underset{\sim}{Z}\right] \leq 2 \exp\left(-\frac{[\frac{n}{2}]\epsilon^2}{16 \max^2\{\log \delta_n, -\log a_n\}}\right). \tag{2.20}$$

Hence, by (2.8) we have that for any $\epsilon > 0$

$$\sum_{n=1}^{\infty} P(|H_n - H_n^*| > \epsilon) < \infty, \tag{2.21}$$

and now the Borel-Cantelli lemma implies that

$$\lim_{n\to\infty} U_n = 0 \qquad \text{a.s.} \tag{2.22}$$

LEMMA 2. If H(f) is finite, and (2.4) and (2.6) are satisfied, then

$$\lim_{n\to\infty} W_n = 0 \qquad \text{a.s.} \tag{2.23}$$

Proof. We put $\mathscr{F}_n = \{x : f_n(x) \geq a_n\}$ and let $\mathscr{A}_n = \mathscr{F}_n^c$, the complement of $\mathscr{F}_n$. Then

$$\begin{aligned}
\mu(\mathscr{A}_n) &= \int_{f_n(x) < a_n} f(x)dx \\
&\leq \int_{\frac{1}{2} f(x) < a_n} f(x)dx + \int_{f_n(x) < a_n \leq \frac{1}{2} f(x)} f(x)dx \\
&\leq \int_{\frac{1}{2} f(x) < a_n} f(x)dx + 2 \int_{R^d} |f_n(x) - f(x)|\, dx . \qquad (2.24)
\end{aligned}$$

Therefore, by (2.4) and (2.6) we have that

$$\lim_{n\to\infty} \mu(\mathscr{A}_n) = 0 \qquad \text{a.s.} \tag{2.25}$$

Since H(f) is finite, the indefinite integral

$$\nu(A) = \int_A |\log f(x)|\, \mu(dx) \tag{2.26}$$

is absolutely continuous with respect to μ. Now (2.25) and (2.26) imply that

$$\lim_{n\to\infty} |W_n| = 0 \qquad \text{a.s.} \tag{2.27}$$

so that Lemma 2 is proven.

LEMMA 3. If H(f) is finite, and (2.4), (2.6), and (2.9) are satisfied, then

$$\lim_{n\to\infty} V_n = 0 \qquad \text{a.s.} \tag{2.28}$$

Proof. (i) We first consider V_n^-, the negative part of V_n. By the so-called log-sum inequality (cf. [5, p. 48]) it follows that

$$-V_n = \int_{\mathscr{F}_n} f(x)\log\frac{f_n(x)}{f(x)}\,dx \le \int_{\mathscr{F}_n} f(x)d(x)\log\frac{\int_{\mathscr{F}_n} f_n(x)dx}{\int_{\mathscr{F}_n} f(x)dx}. \tag{2.29}$$

Now considering the obvious inequality

$$\left|\int_{\mathscr{F}_n} f_n(x)dx - \int_{\mathscr{F}_n} f(x)dx\right| \le \int_{R^d} |f_n(x) - f(x)|\,dx\,, \tag{2.30}$$

we conclude from (2.4) and (2.25) that

$$\lim_{n\to\infty}\int_{\mathscr{F}_n} f(x)dx = \lim_{n\to\infty}\int_{\mathscr{F}_n} f_n(x)dx = 1\,. \tag{2.31}$$

Hence, from (2.29) and (2.31) we conclude that

$$\lim_{n\to\infty} V_n^- = 0 \qquad \text{a.s.} \tag{2.32}$$

(ii) Next we consider V_n^+, the positive part of V_n. Let b, $0 < b < 1$, be arbitrary. Furthermore, let

$$\mathscr{B}_n = \left\{x : \frac{f(x)}{f_n(x)} \le \frac{1}{b}\right\}, \tag{2.33}$$

$$\mathscr{C}_n = \{x : f(x) \ge f_n(x)\}\,, \tag{2.34}$$

and

$$\mathscr{D} = \{x : f(x) \ge 1\}\,. \tag{2.35}$$

Then

$$V_n^+ \le \int_{\mathscr{C}_n \cap \mathscr{F}_n} f(x)\log\frac{f(x)}{f_n(x)}\,dx = \int_{\mathscr{F}_n \cap \mathscr{B}_n \cap \mathscr{C}_n} f(x)\log\frac{f(x)}{f_n(x)}\,dx$$

$$+ \int_{\mathscr{F}_n \cap \mathscr{B}_n^c} f(x)\log f(x)dx - \int_{\mathscr{F}_n \cap \mathscr{B}_n^c} f(x)\log f_n(x)dx \tag{2.36}$$

$$\leq -\log b + \int_{\mathscr{B}_n^c \cap \mathscr{D}} f(x)\log f(x)dx - (\log a_n) \int_{\mathscr{F}_n \cap \mathscr{B}_n^c} f(x)dx \,. \quad (2.37)$$

Moreover, observe that

$$\int_{\mathscr{F}_n \cap \mathscr{B}_n^c} f(x)dx = \frac{1}{1-b} \int_{\mathscr{F}_n \cap \mathscr{B}_n^c} (f(x) - bf(x))dx$$

$$\leq \frac{1}{1-b} \int_{\mathscr{F}_n \cap \mathscr{C}_n} (f(x) - f_n(x))dx$$

$$\leq \frac{1}{1-b} \int_{\mathscr{F}_n} |f_n(x) - f(x)|\, dx \,. \quad (2.38)$$

Condition (2.9) therefore implies that

$$\lim_{n\to\infty} (-\log a_n) \int_{\mathscr{F}_n \cap \mathscr{B}_n^c} f(x)dx = 0 \quad \text{a.s.} \quad (2.39)$$

Similarly,

$$\int_{\mathscr{B}_n^c \cap \mathscr{D}} f(x)dx \leq \frac{1}{1-b} \int_{\mathscr{C}_n \cap \mathscr{D}} (f(x) - f_n(x))dx$$

$$\leq \frac{1}{1-b} \int_{R^d} |f_n(x) - f(x)|\, dx \,. \quad (2.40)$$

Hence, condition (2.4) implies that

$$\lim_{n\to\infty} \mu(\mathscr{B}_n^c \cap \mathscr{D}) = 0 \,. \quad (2.41)$$

This, together with the assumption that $|H(f)| < \infty$, yields that

$$\lim_{n\to\infty} \int_{\mathscr{B}_n^c \cap \mathscr{D}} f(x)\log f(x)dx = 0 \quad \text{a.s.} \quad (2.42)$$

From (2.37), (2.39), and (2.42) it now follows that

$$\limsup_{n\to\infty} V_n^+ \leq -\log b \quad \text{a.s.} \quad (2.43)$$

Since $0 < b < 1$ was arbitrary, we have

$$\lim_{n\to\infty} V_n^+ = 0 \quad \text{a.s.} \tag{2.44}$$

This concludes the proof of Lemma 3.

3. Discussion of the Conditions

In this section we collect some comments and remarks regarding the meaning of and need for the various conditions assumed in the Theorem.

REMARK 1. It is somewhat surprising that the L_1-consistency of f_n can imply the a.s. consistency of the corresponding entropy estimator H_n, in view of the fact that by the Csiszár-Kemperman-Kullback inequality [5, p. 58] the I-divergence dominates the L_1-distance, whereas the converse is not true (cf. [3]).

Notice that (2.4) and (2.9) are satisfied if

$$\lim_{n\to\infty} (-\log a_n) \int_{R^d} |f_n(x) - f(x)|\,dx = 0 \quad \text{a.s.} \tag{3.1}$$

holds, together with (2.6).

Thus, a certain rate of a.s. convergence to zero of the L_1-error of the density estimate f_n implies the a.s. convergence of the corresponding entropy estimate H_n to $H(f)$, if one assumes finiteness of $H(f)$ and that conditions (2.6) and (2.8) are satisfied. If $a_n \to 0$ slowly, only a slow rate of a.s. convergence to zero of the L_1-error is required.

For example, if $a_n = (\log n)^{-1}$, then a rate of at least $o((\log\log n)^{-1})$ is needed for the a.s. convergence to zero of the L_1-error. Moreover, condition (2.8) is typically met if, for example, $a_n = (\log n)^{-1}$ or $a_n = n^{-\alpha}$, $\alpha > 0$, and $\delta_n < n^{\beta}$, $\beta > 0$.

For a typical density estimator one has $\delta_n \le n^{\beta}$ for some $\beta > 0$. Therefore, for the choice $a_n = (\log n)^{-1}$ the conditions of our Theorem are typically satisfied if

$$\lim_{n\to\infty} (\log\log n) \int_{R^d} |f_n(x) - f(x)|\,dx = 0 \quad \text{a.s.} \tag{3.2}$$

Hence, if (3.2) holds and $a_n = (\log n)^{-1}$, we have typically the a.s. consistency of H_n provided H(f) is finite.

One of the reasons, that the L_1-consistency of f_n can imply the a.s. consistency of the corresponding entropy estimator H_n is that small values of f_n are not taken into account in the definition of H_n, as can be seen from (2.5).

REMARK 2. One may ask whether condition (2.9) can possibly be omitted from the Theorem. However, we have constructed a counterexample which shows that this is not the case.

Example 1 : Take d = 1 and let

$$f(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{if } x \notin [0,1]\,. \end{cases} \tag{3.3}$$

Then clearly, H(f) = 0. Furthermore, let $a_n = (\log n)^{-1}$, so that (2.6) is satisfied, and define $c_n = (-\log a_n)^{-1/2}$. Next put

$$f_n(x) = \begin{cases} a_n & \text{if} \quad x \in [0, c_n] \\ \dfrac{1 - a_n c_n}{1 - c_n} & \text{if} \quad x \in (c_n, 1]\,. \end{cases} \tag{3.4}$$

Notice that this particular f_n does not depend on the sample $\underset{\sim}{Z}$. Also, in this case,

$$\lim_{n\to\infty} \delta_n = 2, \tag{3.5}$$

so that the sequence $\{\delta_n\}$ is bounded and condition (2.8) is satisfied. It is now easy to verify that, for this choice of f(x) and $f_n(x)$,

$$\int_R |f_n(x) - f(x)|\,dx = 2c_n(1 - a_n) \tag{3.6}$$

and

$$(-\log a_n) \int_{f_n(x) \geq a_n} |f_n(x) - f(x)|\,dx = 2c_n(1 - a_n)(-\log a_n) \tag{3.7}$$

so that

$$\lim_{n\to\infty} \int_R |f_n(x) - f(x)|\,dx = 0 \tag{3.8}$$

and

$$\lim_{n\to\infty} (-\log a_n) \int_{f_n(x) \geq a_n} |f_n(x) - f(x)|\,dx = \infty\,. \tag{3.9}$$

Hence, condition (2.4) is satisfied, but (2.9) is not. On the other hand, one has for this example that

$$H_n^* = -c_n \log a_n - (1 - c_n)\log\left(\frac{1 - a_n c_n}{1 - c_n}\right) \tag{3.10}$$

with H_n^* defined by (2.11). Hence

$$\lim_{n\to\infty} H_n^* = \infty\,. \tag{3.11}$$

Moreover, by Lemma 1,

$$\lim_{n\to\infty} U_n = 0 \qquad \text{a.s.} \tag{3.12}$$

and, since $H_n = U_n + H_n^*$ by (2.12), it follows that

$$\lim_{n\to\infty} H_n = \infty \qquad \text{a.s.} \tag{3.13}$$

Hence, for this example, the conclusion of the Theorem ((2.10)) is not arrived at. Thus, this example shows that (2.9) cannot be omitted as condition in our Theorem.

REMARK 3. Let us now consider a histogram-based density estimator $\hat{f}_n$ to be used for the role of f_n in (2.5). This estimator is constructed as follows. Let $\mathscr{P}_n = \{A_{n1}, A_{n2}, \ldots\}$ be a partition of R^d, where the A_{ni}'s are cubes of size h_n, $0 < h_n < 1$, $\lambda(A_{ni}) = h_n^d$ (i.e., the Lebesgue measure of A_{ni}), and $\mu_n(A_{ni}) = N_{in}(\underset{\sim}{Z})/[(n+1)/2]$, with $N_{in}(\underset{\sim}{Z})$ being the number of Z_i's falling in A_{ni} (i.e., μ_n is the empirical measure for $Z_1, \ldots, Z_{[(n+1)/2]}$ with Z_i as in (2.3)). Then the

histogram-based density estimator $\hat{f}_n$ is defined by

$$\hat{f}_n(x,\underset{\sim}{Z}) = \frac{\mu_n(A_{ni})}{\lambda(A_{ni})} \quad \text{if} \quad x \in A_{ni} \,. \tag{3.14}$$

The histogram-based entropy estimator considered in [7], and denoted here by $\hat{H}_n$, is obtained by substituting $\hat{f}_n$ for f_n in expression (2.5), i.e., $\hat{H}_n = H_n(\hat{f}_n)$.

The conditions imposed in [7] to guarantee the a.s. consistency of $H_n(\hat{f}_n)$ imply the L_1-consistency of $\hat{f}_n$, but do not involve condition (2.9) imposed here. However, the a.s. consistency result concerning $H_n(\hat{f}_n)$ derived in [7] cannot be regarded as a special case of the Theorem above. There, a.s. consistency is guaranteed without any smoothness conditions on f. On the other hand, condition (2.9) cannot be guaranteed without any smoothness conditions on f. In [7] though, we used direct properties of the histogram estimator to establish the a.s. consistency of the corresponding entropy estimator.

4. On the Existence of H(f)

In the following, let $\|x\|$ denote the Euclidean norm of $x = (x_1,\ldots,x_d) \in R^d$. In the statements of the Theorem the only condition imposed on f(x) is that $|H(f)| < \infty$. It is well-known (cf. Ash [1, p. 237], Berger [2, p. 108]) that the finiteness of H(f) is guaranteed if both $E\|X\|^2 < \infty$ (in which case $H(f) < \infty$) and f(x) is bounded (in which case $H(f) > -\infty$). Ash provides an example of a density on R for which $H(f) = +\infty$ and also of one for which $H(f) = -\infty$. In this section we establish rather mild sufficient conditions on f(x) in order that $H(f) < \infty$ (cf. Lemma's 4,5, and 6 below). In Lemma 7 we provide a sufficient condition on f(x) so that $H(f) > -\infty$. Due to lack of space, the proofs of these lemma's are omitted, but can be found in [9]. As a first result we have

LEMMA 4. If, for a random vector $X \in R^d$ with probability density function f(x),

$$E\{(\log\|X\|)^+\} < \infty \,, \tag{4.1}$$

then

$$H(f) < \infty \,. \tag{4.2}$$

REMARK 4. In Lemma 4, condition (4.1) is satisfied, for example, if $E\|X\|^\gamma < \infty$ for some $\gamma > 0$. In particular, this yields the well-known fact that if $E\|X\|^2 < \infty$, then $H(f) < \infty$. If $E\|X\|^2 < \infty$, a finite bound on H(f) can be obtained from

Theorem 4.5.1 of Berger [2, p. 108]. We also note that, in this absolutely continuous case, conditions (4.1) and (4.2) are not equivalent to each other.

This is demonstrated by the following example due to Csiszár [4], who derived it for a different purpose.

Example 2 (Csiszár [4]) : Let d = 1 and

$$f(x) = \begin{cases} \dfrac{c}{(\log\ k)^2} & \text{if} \quad k \le x \le k + \frac{1}{k} \quad (k = 2,3,\ldots) \\ \\ 0 & \text{otherwise}, \end{cases} \tag{4.3}$$

where c is a norming constant. Then it is easily verified that $H(f) < \infty$, but $E\{(\log\|X\|)^+\} = \infty$. In fact, for this example also $H(f) > -\infty$, as will follow from Lemma 7 below.

REMARK 5. Whereas (4.1) can be regarded as a tail condition on the probability density function f(x), we can give yet another condition in order for (4.2) to hold, just in terms of an explicit bound on the tail of f(x).

LEMMA 5. Given a probability density function f(x) on R^d, assume that there are positive constants C, r_0, and ϵ, such that

$$f(x) \le \frac{C}{\|x\|^d(\log\|x\|)^{2+\epsilon}} \tag{4.4}$$

for $\|x\| > r_0$. Then $H(f) < \infty$.

REMARK 6. Another condition, which is simpler but more restrictive than condition (4.1) and still implies (4.2), is provided by the following lemma.

LEMMA 6. If, for a random vector $X \in R^d$ with probability density function f(x),

$$\int_{R^d} (f(x))^{1-\epsilon}dx < \infty \tag{4.5}$$

for some ϵ, $0 < \epsilon < 1$, then $H(f) < \infty$.

REMARK 7. For d = 1, condition (4.5) is met with $\epsilon = 1/2$, if, for example, for some $\eta > 0$, $E|X|^{1+\eta} < \infty$. This follows from Lemma 1, part (ii), of Devroye and Györfi [6, p. 81], which in this particular case yields that

$$\int_R \sqrt{f(x)}dx \le K_\eta (E|X|^{1+\eta})^{\frac{1}{2(1+\eta)}} , \tag{4.6}$$

where K_η is a known positive constant depending only upon η.

Finally, we give a sufficient condition on f(x) in order that $H(f) > -\infty$. This result can be regarded as the dual of Lemma 6.

LEMMA 7. If, for a random vector $X \in R^d$ with probability density function f(x),

$$\int_{R^d} (f(x))^{1+\epsilon} dx < \infty \tag{4.7}$$

for some $\epsilon > 0$, then $H(f) > -\infty$.

5. Conclusion

In this paper we have considered the estimation of entropy, using an estimator of the form (2.5) when the estimator is based on an L_1-consistent density estimate. Sufficient conditions have been established on such an estimator so as to guarantee the a.s. consistency of the corresponding entropy estimator. L_1-consistent density estimates have many desirable properties [6], whereas a.s. consistency of the entropy estimator is a rather strong result to obtain, especially in view of the fact that all we assume regarding the underlying density is that H(f) is finite. This generality makes the class of entropy estimators proposed here (and in [7]) attractive, as compared to other entropy estimators previously proposed in the literature. As no results on the a.s. rate of convergence of any entropy estimator are available yet, a definitive comparison with other entropy estimators cannot be carried out at the moment. In fact, it remains an open problem to find results on the a.s. rate of convergence of any entropy estimator available in the literature. An application of our Theorem to the case of a kernel density estimate is given in [8].

References

[1] R.B. Ash, *Information Theory.* New York : Interscience, 1965.

[2] T. Berger, *Rate Distortion Theory : A Mathematical Basis for Data Compression.* Englewoods Cliffs, NJ : Prentice Hall, 1971.

[3] I. Csiszár, "Informationtheoretische Konvergenz-begriffe im Raum der Wahrscheinlichkeitsverteilungen," *Magyar Tudo. Akad. Mat. Kutató Intézek. Közl.*, vol. 7, pp. 137-158, 1962.

[4] I. Csiszár, "On generalized entropy," *Studia Sci. Math. Hungar.*, vol. 4, pp. 401-419, 1969.

[5] I. Csiszár and J. Körner, *Information Theory : Coding Theorems for Discrete Memoryless Systems.* Budapest : Akadémiai-Kiadó, and New York : Academic Press, 1981.

[6] L. Devroye and L. Györfi, *Nonparametric Density Estimation : The L_1 View.* New York : Wiley, 1985.

[7] L. Györfi and E.C. van der Meulen, "Density-free convergence properties of various estimators of entropy," *Comput. Statist. Data Anal.*, vol. 5, pp. 425-436, 1987.

[8] L. Györfi and E.C. van der Meulen, "An entropy estimate based on a kernel density estimation," in I. Berkes, E. Csáki, P. Révész (Eds.), *Limit Theorems in Probability and Statistics.* North-Holland Publishing Co., 1990 (to appear).

[9] L. Györfi and E.C. van der Meulen, "Entropy estimation based on L_1-consistent density estimates," Preprint Series, Department of Mathematics, K.U. Leuven, Vol. 2, No. 1, January 1990, pp. 75-105.

II. Curve and Functional Estimation (Continued)

ANALYSIS OF SAMPLES OF CURVES

THEO GASSER
Zentralinstitut für seelische Gesundheit
I5
6800 Mannheim 1
Germany

ALOIS KNEIP
Universität Bonn
Wirtschaftstheorie II
Adenauerallee 24-26
5300 Bonn 1
Germany

ABSTRACT. Most approaches to regression problems deal with single curves. However, often data are available representing a sample of curves, and one ought to wish to exploit the sample information. As a first goal, we want to estimate a valid average curve which reflects the individual-dynamic and intensity. To this end, we try to align individual curves such that similar events or structures take place at identical times: This can be achieved via individual extrema or inflection points ("structural pionts") which can be determined nonparametrically. The structural points then allow it to define shift functions to align the individual curves prior to any other statistical analysis. The new method is illustrated by application to growth curve data, and its asymptotic performance is evaluated. Briefly, we describe an alternative method which tries to determine the underlying model from the data ("Self-modeling, nonlinear regression"). It is more difficult to apply, needing in particular more apriori knowledge.

1. Introduction

Many classical applications of parametric and nonparametric regression in chemistry and physics assume that data for one experimental unit is given ("the one-curve problem"). In particular in biomedicine, data of a sample subjects is usually available ("the n-curve-problem"). Let us thus assume that data Y_{ij}

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 99–110.
© 1991 *Kluwer Academic Publishers.*

are collected to study some biological process across time, concentrations etc., t_{ij} ($j = 1, ..., n_i$; $i = 1, ..., m$). Thus the following nonparametric regression model is postulated:

$$Y_{ij} = f_i(t_{ij}) + \epsilon_{ij} \tag{1}$$

where the ϵ_{ij} denote the residuals and where the f_i are smooth individual regression functions. The design points t_{ij} satisfy $t_{ij} \in J = [a_0, a_1] \subset \mathbb{R}$.

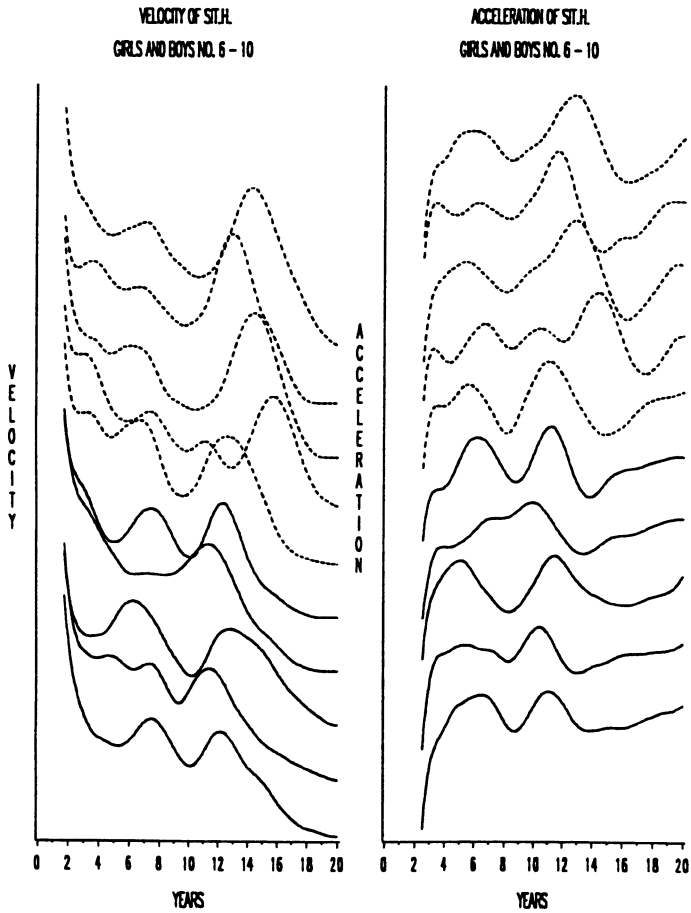


Figure 1: Smoothed velocity (left) and acceleration curves of sitting height of 5 boys (dashed) and 5 girls (solid) from 0 to 20 years.

Finding an appropriate parametric nonlinear model is often difficult for the one-curve-problem, which has led to the development of nonparametric regression estimators as documented in this volume. For the n-curve-problem, these difficulties become multiplied, since a model, fitting some subjects well, may fail with others. In such situations a nonparametric approach would be a good starting point. A meaningful estimator of an average curve could both summarize the data well and lead to hypotheses about the mechanisms generating $f_1, ..., f_m$. The cross-sectional average is usually inappropriate qualitatively and quantitatively due to shifts in individual curves. This is illustrated in figure 1, showing smoothed velocity and acceleration curves for sitting height (automatic function fitting using kernel estimators and a bandwidth according to Gasser, Kneip and Köhler (1990)).

Obviously, the growth spurts - defined as ages of maximal velocity - have substantial variation between individuals and also between boys and girls. Amplitude of these curves, indicating intensity of the respective growth processes, vary substantially. The procedure proposed to handle such situations involves the following steps:

(i) Define features which characterize the structure of individual curves via "structural functionals" (extrema, inflation points etc.).

(ii) Assign to each curve a number of "structural points" by applying the structural functionals to smoothed estimated regression functions.

(iii) The structural points are used to determine smooth monotone "shift functions" which allow it to synchronize the individual curves to an average dynamic.

(iv) The "structural average" is obtained as the pointwise average of synchronized individual curves.

The introduction of our approach is done both at a conceptual level and at an estimation level. Here, an outline is given and the reader is referred to a more complete and more technical treatment in Kneip and Gasser (1990). A brief discussion of relations with semiparametric techniques for curve samples ("self-modeling nonlinear regression") is given at the end of the paper (Kneip and Gasser, 1988).

2. How to Quantify Individual Dynamic

A basic assumption is that the regression functions $f_1, ..., f_m \in C^{(\nu)}(J)$ for some $\nu \geq 2$. The norm used is

$$\|f_i\|^{(\nu)} := \sum_{s=0}^{\nu} \sup_t |f_i^{(s)}(t)|.$$

A basic problem - in formal and computational terms - is that some structural point cannot be identified in some individual data set. Then we may declare it as missing and assign some value $\alpha \notin J$.

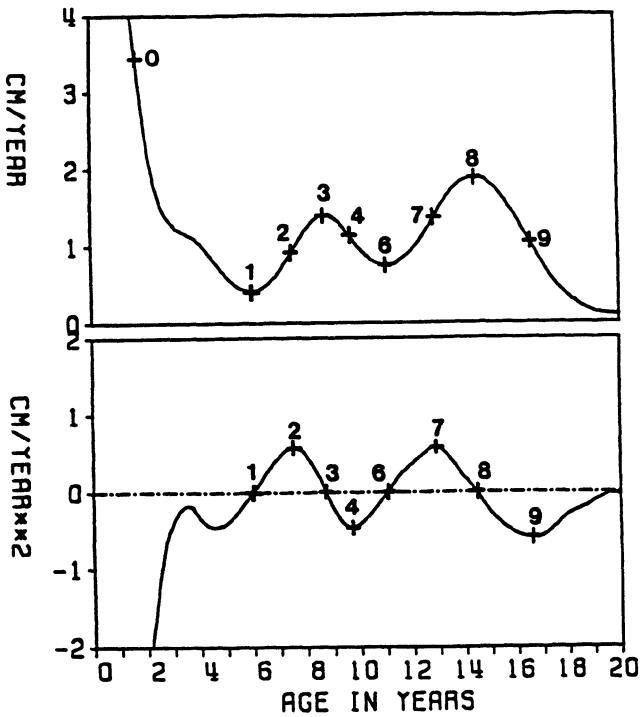


Figure 2: Structural functionals that might be used for analyzing growth data in terms of velocity (above) and acceleration (below).

Definition . *A functional $T : C^2(J) \longrightarrow [a_0, a_1] \cup \{\alpha\}$ is called a "first order structural e-functional" if the following holds:*

(i) *T is continous on $D_T := \{v \in C^2(J) | T(v) \neq \alpha\}$.*

(ii) *For all $v, w \in D_T, v(T(v))$ and $w(T(w))$ are either both local maxima or both local minima, and $sgn(v"(T(v)) = sgn(w"T(w)) \neq 0$.*

Similarly one might quantify the rate of increase or decrease in monotone segents of curves. Points where some percentage of increase/decrease is reached are then called "p-points" (p for percentage) and the respective functionals are called "first order structural p-functionals". When applying these concepts to the $\mu-$th derivative $\nu^{(\mu)}$ instead of the function itself, we obtain $(\mu+1)st$ order structural functions. The value $T(v)$ obtained by applying a structural functional T is called a "structural point".

In practice, we have first to define a number of structural functionals $T_1, ..., T_l$. This may be done by a visual analysis or by using apriori knowledge. An example is given in fig.2 for human height growth. There T_1, T_3, T_6 and T_8 are 2nd order e-functionals, T_2, T_4, T_7 and T_9 are third order e-functionals, and T_0 is a 2nd order p-functional demarcating the fast decline of velocity in infancy. The e-functionals reflect the dynamics of the so-called mid-growth spurt and of the pubertal spurt which represent our a priori knowledge. We have developed an automatic procedure for extracting structural functionals from the data (Gasser and Kneip, 1990) which for the growth data would lead to the same functionals as apriori knowledge.

Given structural functionals $T_1, ..., T_l$ we may estimate structural points $\hat{\tau}_{r_i} := T_r(\tilde{f}_i)$ by kernel estimatiors $\tilde{f}_1, ..., \tilde{f}_m$ of $f_1, ..., f_m$ (for details see Gasser and Herrmann, in this volume, and the literature cited there). This may often be best achieved by estimating derivatives (compare also Müller 1985,1989). Theory (section 4) and intuition indicates that a mildly undersmoothing bandwidth is appropriate at the step in order to optimize the estimate of a "structural average curve" later on. When applying these concepts to longitudinal growth data of various dimensions of the body, a subtantial variability in passing through somatic development in fact came up. Typically, the difference between the .05th and the .95th quantile of the structural point T_8 defined in fig. 2 ("peak velocity of the pubertal spurt") was about 3 years. This is larger compared to the sex difference and to ethnic differences.

3. How to Define Shift Functions and an Average Curve

Given structural functionals $T_1, ...T_l$, $(l \in \mathbb{N})$ we require that $T_r(v) < T_s(v)$ holds for $r < s$ when $T_r(v) \neq \alpha$ and $T_s(v) \neq \alpha$ (i.e. not missing). Set $\mathbf{T}(v) := (T_1(v), ..., T_l(v))$ and denote by H_l the set of l-tuples $x = (x_1, ..., x_l)' \in (J \cup \{\alpha\})^l$ with $x_r < x_s$ for $r < s$ and $x_r \neq \alpha, x_s \neq \alpha$.

The synchronization of the true curves $f_1, ..., f_m$ is based on the structural points τ_{ri} $(r = 1, ..., l;\ i = 1, ..., m)$ obtained in the previous step to quantify the individual dynamic. Mathematically, this is achieved by an operator G which assigns to the set of individual structural points a shift function g_i. The shift functions are defined in such a way that individual structural points τ_{ri} are aligned to fixed points x_r ('the common time scale"), most naturally the average locations $\bar{\tau}_r = \frac{1}{m}\sum_{i=1}^{m} \tau_{ri} := x_r$. The definitions are thus as follows:

Definition . *An operator $G : \mathcal{H}_l^2 \to C^1(\mathbb{R})$ is called a shift operator if*

(i) $G_{(\tau,x)}(\cdot), (\tau, x) \in \mathcal{H}_l^2$, *is strictly monotonically increasing and* $G_{(\cdot)}(t)$ *is continous in* $t \in J$.

(ii) *For any* $\tau = (\tau_1, ..., \tau_l)' \in H_l$ *and any* $x = (x_1, ..., x_l)' \in \mathcal{H}_l$ *with* $\tau_r \neq \alpha, x_r \neq \alpha$ *we have*

$$G_{(\tau,x)}(x_r) = \tau_r.$$

A "shift function" g_i for f_i is then obtained as

$$g_i(\cdot) := G_{(\mathbf{T}(f_i),\bar{\tau})}(\cdot)$$

where $\bar{\tau} := (\bar{\tau}_1, ..., \bar{\tau}_l)'$.

It would be desirable and in line with the whole set-up to have $g_i \in C^2(\mathbb{R})$. This poses, however, problems on the algorithmic side (see below). Applying the shift functions g_i to the regression functions f_i leads to a synchronized sample $\{f_i^*\} = \{f_i \circ g_i\}$ and straightforwardly to an average curve:

Definition . *A "structural average (curve)" f for the sample of curves $\{f_i\}$ is obtained as follows:*

$$f(t) := \frac{1}{m}\sum_{i=1}^{m} f_i(g_i(t)),\ t \in J_i^* = [\sup_i g_i^{-1}(a_0),\ \inf_i g_i^{-1}(a_1)]$$

When starting and end points a_S, a_E of the generating process are known, we may set $g(a_s) = a_S$ and $g(a_E) = a_E$ and then $J^* = J$. Intuitively, f may be considered as the basic process which is modulated by nature in individuals in various ways, leading to curves $\{f_i\}$. The statistician tries to drive this process backwards and to eliminate via shift functions part of the variation introduced by nature. When estimating shift functions g_i, it has to be kept in mind that they are fixed by defintion only at structural points. In between, the are restricted by the requirements of monotonicity and smoothness. Interpolation splines, for example, do not satisfy monotonicity. The following procedure garantees that $g_i \in C^1(\mathbb{R})$ but not more:

(i) for $l = 0$ set $g_i(t) := t$

(ii) for $l = 1$ set $g_i(t) := t + \tau_{1i} - \bar{\tau}_1$

(iii) for $l = 2$ let g_i be the straight line with $g_i(\bar{\tau}_1) = \tau_{1i}$; and $g_i(\bar{\tau}_2) = \tau_{2i}$

(iv) for $l > 2$ set $g_i(\bar{\tau}_1) = \tau_1, ..., g_i(\bar{\tau}_l) = \tau_{li}$. Use then monotone piecewise cubic interpolants as investigated in Fritsch and Butland (1984). When structural points are missing, apply this procedure to the ramaining ones.

Estimation is then straightforward, based on estimated structural points $\{\hat{\tau}_{r_i}\}$, leading to the "estimated structural average" $\hat{f}$:

$$\hat{f}(t) := \frac{1}{m} \sum_{i=1}^{m} \hat{f}_i(\hat{g}_i(t)),$$

where $\hat{f}_i$ is a mildly undersmoothed kernel estimate. Undersmoothing is advisable, since averaging curves in itself decreases the variance of $\hat{f}$. Applying the shift operator G splits the sample $\{f_i\}$ into a synchronized sample $\{f_i \circ g_i\}$ - which should be easier to analyze - and a sample of shift functions $\{g_i\}$.

Figure 3 shows the application of this concept to the growth data of sitting height, as exemplified in fig. 1.

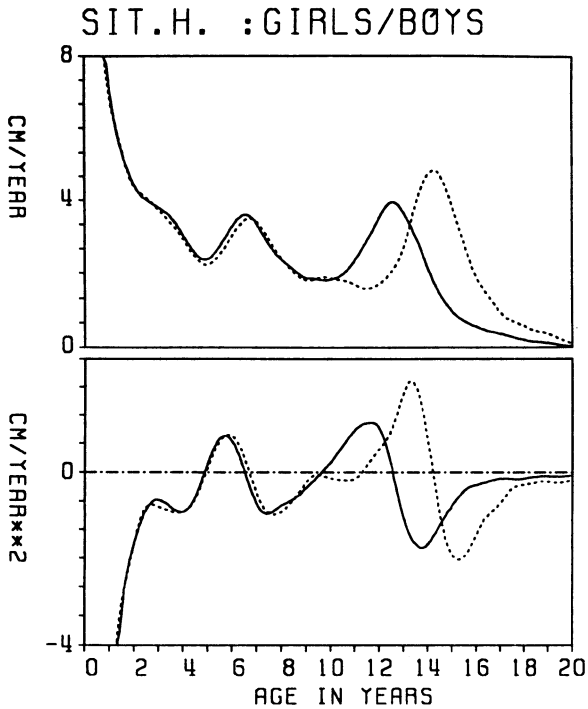


Figure 3: Estimated structural average velocity curve (above) and acceleration curve (below) of sitting height for boys (dashed) and girls (solid).

Depicted are the estimated structural average velocity and acceleration curves based on samples of $n = 112$ girls and $n = 120$ boys (Gasser et al., 1990). The estimates were obtained using the structural points T_0, T_2, T_4, T_7, T_8 and T_9 of fig.2. A smooth average curve with a clear and well- interpretable pattern arises. In addition to the strong pubertal spurt -peaking at about 12 years for girls and at about 14 years for boys - there is an accentuated mid-growth spurt at about age 7. From the endocrinological side it is interesting that this spurt appears about at the same age and with the same intensity for both sexes. This clear phenomenon is not part of any of the parametric functional models for growth and can thus presently not be quantified parametrically. Its omission in the parametric models also leads to a serious sex-dependent bias when quantifying the pubertal spurt.

4. Asymptotic Results

For the estimators introduced, rates of convergence have been derived in Kneip and Gasser (1990). The reader is referred to this paper for details and proofs. A number of assumptions have been made for technical convenience, such as equally paced data and the same number of data point per curve, i.e. $n_i = n \ \forall \ i$. The more basic assumptions are as follows:

A_1 The random variables ϵ_{ij} are independent with expectation 0 and variance $\sigma_i^2 < \infty$. Moreover, $E \ \epsilon_{ij}^{\rho} < C_\rho$ for some $C_\rho < \infty$ for all i and all $\rho \in \mathbb{N}$.

A_2 The f_i are assumed to be function valued random variables with a compact domain $\mathcal{R} \subset C^{\nu}(J)$, where $\nu \geq \mu + 1$ (μ = highest order of functionals $T_1, ..., T_l$).

A_3 There exists an open subspace $D_r \subset C^{\mu+1}$ such that $\mathcal{R} \subset D_r \subseteq D_{T_r}$ where $r \in \{1, ..., l\}$.

Assumption A_3 implies that no missing structural points are allowed for the true regression functions f_i. We now give a Theorem on rates of convergence for the estimates of e-structural points $\hat{\tau}_{r_i}$. It is assumed that a k-th order kernel K and a bandwidth h are used for the estimators $\tilde{f}_1, ..., \tilde{f}_m$ which serve for extracting the structural points. The subspace $\mathcal{R}$ is called "symmetric at T_r" if there exists an $\epsilon > 0$ such that $v^{(\mu_r - 1)}(T_r(v) - x) = v^{(\mu_r - 1)}(T_r(v) + x)$ for all $x < \epsilon$. The following statements are conditional that $f_1, ..., f_m$ are given.

Theorem . *Bias and Variance for structural points of e-functionals.*

Assume that $h \to 0$ as $n \to \infty$ in such a way that $n^{1/(2\mu r+3)} h / \log n \to \infty$. Then:

1. $E(\hat{\tau}_{r_i} - \tau_{r_i}) = \beta_{r_i} + O(1/(nh^{2\mu r+1}))$,
   where $\beta_{r_i} = O(h^{k-1}/\sqrt{nh^{2\mu r+1}})$ if $\mathcal{R}$ is symmetric at τ_r, or else

$$\beta_{r_i} = h^k \int_{-1}^{1} K^{(\mu_r)}(x) x^{\mu_r + k} dx \cdot \frac{(-1)^{\mu_r+1} f_i^{(\mu_r+k)}(\tau_{r_i})}{(k+\mu_r)! f_i^{(\mu_r+1)}(\tau_{r_i})} + o(h^k).$$

2. $Var(\hat{\tau}_{r_i}) = \nu_{r_i} + o(\beta_{r_i}^2) + o(1/(nh^{2\mu_r+1}))$
   where $\nu_{r_i} = \frac{\sigma_i^2}{nh^{2\mu_r+1}} \int_{-1}^{1} K^{(\mu_r)}(x)^2 dx f_i^{(\mu_r+1)}(\tau_{r_i})^2$

Convergence holds uniformly in i and m.

Similar to this theorem results for p-functionals and for shift functions can be derived. The asymptotic results have a plausible interpretation: $f_i^{(\mu_r+1)}(\tau_{r_i})$ gives some measure of the steepness of increase/decrease close to τ_{ri}. This influences both bias and variance of the structural point. As to bias the degree of assymmetry given by $f_i^{(\mu_r+k)}(\tau_{r_i})$ may also become influential.

When choosing the optimal rates for the bandwidth h, the following result holds for the estimated structural average $\hat{f}$:

$$E\int(\hat{f}(t)-f(t))^2dt = O(n^{-\frac{2k}{k+2\mu+1}} + (mn)^{-\frac{2k}{2k+2\mu+1}}).$$

When taking $k=4$ and $T_1,...,T_l$ as first order structural functionals, and when in addition $n^{4/7}=O(m)$, we obtain:

$$E\int(\hat{f}(t)-f(t))^2dt = O(n^{-8/7}).$$

This can be further improved to $O(n^{-8/5})$ when structural functionals are "symmetric". This result shows that the rates for the average curve are substantially better than for single curves estimated nonparametrically, so that the n-curve-problem is not just more difficult, but also more rewarding.

5. Semiparametric Model Building

We now give a sketch of some relation to semiparametric model building. The n-curve problem has been approached in a different way via a semiparametric regression problem in Kneip and Gasser (1988):

$$Y_{ij} = F(t_{ij};\underline{\theta_i}) + \epsilon_{ij}, \quad i=1,...,m;\ j=1,...,n_i.$$

Here, F is an unknown parametric model function and $\underline{\theta_i}$ is an unknown individual parameter vector of dimension p. Both F and $\underline{\theta}_i, i=1,...,m$, have to be estimated from Y_{ij}. Therefore, in addition to the dimension p, qualitative information on F is needed. In Kneip and Gasser (1988) a rather complex algorithm, related to back-fitting algorithms, was proposed to solve the estimation problem. The above concept contains an important class of nonlinear models arising from linear location-shift-transformation. Such "shape-invariant models" have first been considered in Lawton et al. (1972) :

$$F(t_{ij}, \underline{\theta_i}) = f_i(t) = \theta_{3i} \cdot \phi(\frac{t - \theta_{2i}}{\theta_{ii}}) + \theta_{4i}$$

with real parameters satisfying $\theta_{3i}, \theta_{1i} > 0$ and some real function $\phi \in C^{\nu}(\mathbb{R})(\nu \geq 2)$. A somewhat more comlicated model was successfully applied to growth data in Stützle et al. (1980). Härdle and Marron (1990) studied the problem of estimating parameters transforming f_i into f_j $(i \neq j)$.

The concept that the structural average curve should reflect the process generating individual realizations makes it plausible that such an average curve may be helpful in model building. For the shape-invariant model, the model or shape function μ can in fact be estimated nonparametrically via the structural average. The argument proceeds roughly as follows: when x is the location of an extremum of $\phi(\mu)$ $(\mu \leq \nu - 2)$, then each $f_i^{(\mu)}$ has a "corresponding" extremum at $T(f_i) = \theta_{1i}x + \theta_{2i}$. Via the determination of extrema, information about θ_{1i} and θ_{2i} can be gained even when ϕ is unknown. Let us assume that two structural e-functionals T_1, T_2 with $T_1(v) < T_2(v)$ can be defined. Then there exist x_1, x_2 such that

$$\begin{aligned} T_1(f_i) &= \theta_{1i}x_1 + \theta_{2i} \\ T_2(f_i) &= \theta_{1i}x_2 + \theta_{2i} \end{aligned}.$$

To make things unique, the average θ_{3i}, θ_{1i} may be nomalized to one, and the average θ_{4i}, θ_{2i} to zero. This leads to $x_1 = \bar{\tau}_1 = \frac{1}{m}\sum_{i=1}^{m} \tau_{1i}, x_2 = \bar{\tau}_2 = \frac{1}{m}\sum_{i=1}^{m} \tau_{2i}$, and $f(t) = \frac{1}{m}\sum_{i=1}^{m} f_i(\theta_{1i}t + \theta_{2i})$. The above relations now can be used to determine θ_{1i}, θ_{2i} from τ_{1i}, τ_{2i}, and to compute shift-functions $g_i(t) = \theta_{1i}t + \theta_{2i}$. Consequently

$$\phi(t) = \frac{1}{m}\sum_{i=1}^{m} f_i(g_i(t)).$$

Thus, the n-curve-problem does not only lead to better statistical properties in terms of an average curves as compared with individual curves but may also be very helpful in model building. The procedure just described contrasts well with usual trial-and-error search among some functional models taken either from the literature or postulated by the researcher after inspecting the data.

References

Fritsch, F.N. and Butland, I. (1984) : A method for constructing local monotone piecewiese cubic interpolants; *SIAM J. Sci. Stat. Computing, 5, 300-304.*

Gasser, Th., Kneip, A. and Köhler,W. (1990) : A flexible and fast method for automatic smoothing; *Manuscript, revised.*

Gasser, Th. and Kneip, A. (1990) : Searching for structure in samples of curves; *Manuscript (in preparation).*

Gasser, Th., Kneip, A., Binding,A., Prader, A. Largo, R. and Molinari, R. (1990) : The dynamics of linear growth in distance, velocity and acceleration; *Annals of Human Biology (to appear).*

Härdle, W. and Marron,, I.S. (1990) : Semiparametric comparison of regression curves; *Annals of Satistics , 18, 63-89.*

Kneip, A. and Gasser, Th. (1988) : Convergence and consistency properties for self-modeling nonlinear regression; *Annals of Satistics , 16, 82-113.*

Kneip, A. and Gasser, Th. (1990) : Statistical tools to analyze data representing a sample of curves; *Manuscript, revised.*

Lawton, W.H., Sylvestre, E.A. and Maggio, M.S. (1972) : Self modeling nonlinear regression; *Technometrics, 14, 513-532.*

Müller, H.G. (1985) : Kernel estimators of zeros and location and size of extrema of regression functions; *Scandinavian Journal of Statistics, 12, 221/232.*

Müller, H.G. (1989) : Adaptive nonparametric peak estimation; *Annals of Statistics, 17, 1053-1070.*

Stützle, W., Gasser, Th., Molinari, L, Largo, R.H., Prader, A. and Huber, P.I. (1980) : Shape-invariant modeling of Human growth.; *Annals of Human Biology, 7, 507-528 .*

BOOTSTRAP METHODS IN NONPARAMETRIC REGRESSION

W. HÄRDLE
CORE
34 voie du Roman Pays
Université Catholique de Louvain
B-1348 Louvain-la-Neuve
Belgium

E. MAMMEN
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
D-6900 Heidelberg
Germany

ABSTRACT. Bootstrap techniques naturally arise in the setting of nonparametric regression when we consider questions of smoothing parameter selection or error bar construction. The bootstrap provides a simple-to-implement alternative to procedures based on asymptotic arguments. In this paper we give an overview over the various bootstrap techniques that have been used and proposed in nonparametric regression. The bootstrap has to be adapted to the models and questions one has in mind. An interesting variant that we consider more closely is called the *Wild Bootstrap*. This technique has been used for construction of confidence bands and for comparison with competing parametric models.

1. Introduction.

In this paper we will study bootstrapping for estimating a nonparametric regression function m. The nonparametric regression model can be written as:

$$Y_i = m(X_i) + \varepsilon_i \quad (i = 1,\ldots,n)$$

where X_i are the design variables (for simplicity one dimensional) and ε_i are the error terms. Our aim is to consider statistics related to the estimation of the unknown regression function m. We pursue this aim in different models concerning the stochastic structure of the variables. For simplicity we consider three models.

MODEL 1.

The ε_i's *are independent, identically distributed random variables with* $E\,\varepsilon_i = 0$. *The* X_i *are deterministic.*

MODEL 2.

The pairs (X_i, Y_i) *are independent, identically distributed random variables with* $E(\varepsilon_i \mid X_i)$

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 111–123.
© 1991 *Kluwer Academic Publishers.*

$= 0$. *Then* $m(x) = E(Y_i \mid X_i = x)$.

MODEL 3.

The ε_i*'s are independent random variables with* $E\,\varepsilon_i = 0$. *The* X_i *are deterministic. The distribution of the errors may depend on the design variable.*

The model that has been most dominantly investigated is MODEL 1, the socalled 'fixed design' model, see Eubank (1988) and Härdle (1990). Note that MODEL 3 covers the class of models of type 1. Also if we condition on the design variables then MODEL 2 is contained in MODEL 3. Of course there is a wide range of possibilities between these above model classes. For example, with the variance function $\sigma^2(x) = \mathrm{var}(\varepsilon \mid X = x)$, one can assume that in MODEL 3 the additional assumption holds that $\varepsilon_i / \sigma(X_i)$ are i.i.d. . One such approach could be to parametrize the variance function as in Carroll (1982) and estimate the parameters before entering a bootstrap step.

Each of these models suggests a different resampling procedure. Furthermore one may use a resampling procedure motivated by a larger model (for instance MODEL 3) in a smaller model (for instance MODEL 1). This makes sense if one wants to safeguard oneself in MODEL 1 against deviations from MODEL 1. But clearly if one is more interested in efficiency than model robustness one should prefer resampling methods motivated by the assumed model. In the context of resampling procedures for linear models this point has also been made in Liu and Singh (1989). To our knowledge the first bootstrap study for making inference about m can be found in a film by McDonald(1982). He assumed MODEL 2 and resampled from the pairs of observations (X_i, Y_i). A recent bootstrap overview has been given in Mammen (1990c).

Throughout our paper we use the kernel estimator $\hat{m}_h$ with bandwidth $h = h_n$ and kernel K (Nadaraya, 1964; Watson, 1964)

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)\, Y_i}{\sum_{i=1}^{n} K_h(x - X_i)}, \tag{1.1}$$

$$K_h(\cdot) = h^{-1} K(\cdot / h).$$

For simplification of notation the dependence of h on n will be suppressed. Generalizations of the results presented here to higher dimensional design variables are straightforward. In the fixed design model with equidistant $X_i = i/n$ the denominator in (1.1) is often conveniently replaced by n, see Müller (1988).

We are mainly interested in the distribution of functionals of $\hat{m}_h(\cdot) - m(\cdot)$, for instance the L_2 norm of this function or the evaluations of this function at a set of points. But we have also other functionals such as shape parameters in mind (see Mammen, 1990a). We discuss the bootstrap procedures which have been proposed in the literature for MODEL 1 - 3 in the next section. We do not address in this paper the problem of computational feasibility of bootstrap in this context. To avoid the computer intensive direct computation of the

smoothing in (1.1) we propose the method of Weighted Averaging of Rounded Points (WARPing) of Härdle and Scott (1990) as a fast method for performing the resampling steps. Throughout the paper we call the computable simulated stochastic structure the 'bootstrap world'.

2. The bootstrap procedures.

2.1. I.I.D. ERRORS (MODEL 1), RESIDUAL RESAMPLING .

To mimic the stochastic nature of this model in the bootstrap world one proceeds as follows.

STEP 1. *Calculate residuals*

$$\hat{\varepsilon}_i = Y_i - m(X_i) \qquad (i = 1,\ldots,n).$$

STEP 2. *Centering.*

$$\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \hat{\varepsilon}_\bullet \text{, where } \hat{\varepsilon}_\bullet = n^{-1} \textstyle\sum_{i=1}^n \hat{\varepsilon}_i .$$

STEP 3. *Resampling.*

Draw randomly $\varepsilon_1^*, \ldots, \varepsilon_n^*$ *out of the set* $\{\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n\}$.

STEP 4. *Create bootstrap observations*

$$Y_i^* = \hat{m}_g(X_i) + \varepsilon_i^* .$$

Here a bandwidth g *is chosen which may be different from the initial bandwidth* h.

STEP 5. *Calculate a nonparametric kernel estimate in the bootstrap world.*

$$\hat{m}_h^*(x) = \frac{\sum_{i=1}^n K_h(x - X_i)\, Y_i^*}{\sum_{i=1}^n K_h(x - X_i)} . \tag{2.1}$$

STEP 6. *Bootstrap approximation.*

To approximate the distribution of the desired functional of $\hat{m}_h(\bullet) - m(\bullet)$ *use the computable conditional distribution of the functional of* $\hat{m}_h^*(\bullet) - \hat{m}_g(\bullet)$.

Let us discuss this bootstrap procedure for the case of the evaluation functional $\hat{m}_h(x) - m(x)$. It is common language to say that the 'bootstrap works', if in STEP 6 for suitable choice of h and g the bootstrap distribution tends to the same limit as $\hat{m}_h(x) - m(x)$ in probability. This says nothing about the finite sample behaviour of the bootstrap procedure. We will report

below (theorem 2.2) theoretical reasons for ameliorated finite sample approximations by bootstrapping.

Before entering into a discussion of existing results let us shortly remark that the centering STEP 2 is appropriate. By contrast to linear least squares regression (with intercept) the residuals do not add up to zero. Therefore a bias in the resampling stage would occur if we do not guarantee the bootstrap errors to have mean zero in the bootstrap world (although this does not affect a first order asymptotic analysis). STEP 2,3 is in practice done from residuals from an interior interval of the design space in order to avoid boundary effects. This has been done in the paper by Härdle and Bowman (1988) who showed that the bootstrap works.

THEOREM 2.1. *The conditional distribution of* $\sqrt{nh}\,(\hat{m}^*_h(x) - \hat{m}_g(x))$. *tends in probability to the same Normal limit as* $\sqrt{nh}\,(\hat{m}_h(x) - m(x))$, *provided the errors have finite variance* σ^2, m *is twice differentiable, standard assumptions on* K *hold and* $h \sim n^{-1/5}$, $g \to 0$, $g/h \to \infty$.

In fact the above theorem was originally proved without the technique of oversmoothing. Instead of a different bandwidth g in the resampling step Härdle and Bowman used the same $g = h$ which then lead to the necessity to estimate the bias of the bootstrap estimator explicitly. The oversmooth bandwidth g has been introduced exactly for that reason to deal with the bias implicitly. A similar observation has been made by Scott and Terrell (1987) when they tried to estimate the MISE expansion of density estimators for bandwidth selection. For this purpose they estimated the second derivative of the density. Also here the variance of the estimator for the second derivative is proportional to $n^{-1}\,g^{-5}$ which makes clear why the "optimal rate" of $g \sim n^{-1/5}$ does not work here. The above bootstrap procedure has also be used for bandwidth selection by the above authors. See also Hall (1990a) who investigated this bootstrap for general L_p distances with a so called uniform kernel.

The bootstrap can be used for the construction of confidence intervals. The accuracy of the bootstrap confidence intervals has been considered by Hall (1990b). Let

$$\gamma_x^2 = \mathrm{var}(\hat{m}_h(x)) = \sigma^2 \tau_x^2 [\textstyle\sum_{i=1}^n K_h(x - X_i)]^{-2} = \sigma^2 \beta_x^2$$

say where $\tau_x^2 = \sum_{i=1}^n K_h^2(x - X_i)$. The 'ordinary' and 'studentized' versions of $\hat{m}_h(x) - E\,\hat{m}_h(x)$ are

$$S = \tau_x^{-1}\,\sigma^{-1} \textstyle\sum_{i=1}^n K_h(x - X_i)\,\varepsilon_i \;,$$

$$T = \tau_x^{-1}\,\hat{\sigma}^{-1} \textstyle\sum_{i=1}^n K_h(x - X_i)\,\varepsilon_i \,.$$

Note that by undersmoothing one may center about $E\,\hat{m}_h(x)$. Then for $\mu_j = E[\,(\varepsilon_i/\sigma)^j\,]$ and polynomials p_j

$$P(S \le u) = \Phi(u) + (nh)^{-1/2}\,\mu_3\,p_1(u)\,\phi(u) + (nh)^{-1}\,\{\,(\mu_4 - 3)\,p_2(u) + \mu_3^2\,p_3(u)\,\}\,\phi(u)$$
$$+ \text{ lower order terms},$$

$$P(T\le u) = \Phi(u) + (nh)^{-1/2}\,\mu_3\, p_1(u)\,\phi(u) + (nh)^{-1}\,\{(\mu_4-3)\,p_2(u) + \mu_3^2\, p_3(u)\}\,\phi(u)$$
$$+ (h/n)^{1/2}\,\mu_3\, p_4(u)\,\phi(u) + \text{lower order terms.}$$

In the bootstrap world we have the complete analogue

$$P^*(S^*\le u) = \Phi(u) + (nh)^{-1/2}\,\hat{\mu}_3\, p_1(u)\,\phi(u) + (nh)^{-1}\,\{(\hat{\mu}_4-3)\,p_2(u) + \hat{\mu}_3^2\, p_3(u)\}\,\phi(u)$$
$$+ \text{lower order terms,}$$

$$P^*(T^*\le u) = \Phi(u) + (nh)^{-1/2}\,\hat{\mu}_3\, p_1(u)\,\phi(u) + (nh)^{-1}\,\{(\hat{\mu}_4-3)\,p_2(u) + \hat{\mu}_3^2\, p_3(u)\}\,\phi(u)$$
$$+ (h/n)^{1/2}\,\hat{\mu}_3\, p_4(u)\,\phi(u) + \text{lower order terms.}$$

Since the constants $\hat{\mu}_j$ converge to μ_j at the rate $n^{-1/2}$ we have the following Theorem.

THEOREM 2.2. *Under some regularity conditions the following expansions hold*

$$P^*(S^*\le u) - P(S\le u) = O_P(n^{-1}\,h^{-1/2}),$$

$$P^*(T^*\le u) - P(T\le u) = O_P(n^{-1}\,h^{-1/2}).$$

Note that this rate does not hold for the nonstudentized statistic

$$U = \beta_x^{-1}\,(\hat{m}_h - E\,\hat{m}_h)(x).$$

Then we observe a significantly weaker approximation

$$P^*(U^*\le u) - P(U\le u) = \Phi(u/\hat{\sigma}) - \Phi(u/\sigma) + \text{lower order terms,}$$

since this difference is only of the order $\hat{\sigma} - \sigma = O_P(n^{-1/2})$, see eg. Gasser et al.(1986). Also in other contexts (see for instants Hall (1988) and Mammen (1990b)) it is well known that studentizing gives a considerable improvement of coverage error.

2.2. AN APPLICATION WITH ALMOST I.I.D. ERRORS.

A direct application of the above bootstrap to kernel sprectral density estimates is not straightforward since the periodogram values become only "asymptotically independent" and also the error structure for this regression problem is of multiplicative nature. Consider a strictly stationary real valued process and let Y_i be the periodogram values at discrete equidistant frequencies in $[-\pi, \pi]$. The kernel spectral density estimate is of the form (1.1) (with denominator equal to $2\pi n$ since the design is equidistant on $[-\pi, \pi]$). The regression equation is mulitplicative, i.e. $Y_i = m(X_i)\,\varepsilon_i$. This makes a slight modification of the above

resampling steps necessary where we in fact replace residual contruction by $\hat{\varepsilon}_i = Y_i / \hat{m}_h(X_i)$ and the centering is done by $\tilde{\varepsilon}_i = \hat{\varepsilon}_i / \hat{\varepsilon}_.$. Details for this resampling procedure are found in Franke and Härdle(1990) who show an analogous result to theorem 2.1. and treat also bootstrap bandwidth selection . For bootstrapping a nonparametric, nonlinear autoregressive time series see Franke(1990).

2.3. THE RANDOM DESIGN MODEL 2, PAIRWISE RESAMPLING.

When the data are generated according to MODEL 2 it seems natural to resample from the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$. This has been considered in Dikta(1988) and McDonald(1982). However, we would like to argue that this sort of resampling does not reflect the stochastic structure of MODEL 2. Note that the bootstrap does not represent the conditional distribution of the observations given the design points. Indeed, in the bootstrap world the conditional expectation $E^*(Y_i^* \mid X_i^*)$ is equal to Y_i^* with probability one (if the design variables are pairwise different with probability one). Here $E^*(U \mid V)$ denotes $E(U \mid V, (X_1, Y_1), \ldots, (X_n, Y_n))$. Consider for instance the case that one bootstraps the distribution of the kernel estimate at a fixed point x . Denote the bootstrap sample by $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ and define now

$$\hat{m}_h^*(x) = \frac{\sum_{i=1}^n K_h(x - X_i^*)\, Y_i^*}{\sum_{i=1}^n K_h(x - X_i^*)} . \tag{2.2}$$

Then under the conditions of theorem 2.1 the bias of ($\hat{m}_h(x) - m(x)$) is of order $O(n^{-2/5})$ whereas the bootstrap bias estimate $E^*(\hat{m}_h^*(x) - \hat{m}_h(x))$ is of the lower order $O_P(n^{-4/5})$. Therefore here bootstrap works only after a separate bias estimation or undersmoothing as in Dikta(1988). To see why this bootstrap does not estimate the bias correctly recall that $\hat{m}_h(x) - m(x)$ is an asymptotically linear statistic and therefore the bootstrap bias estimate must be asymptotically zero. To appreciate why consider the followiung calculations. With

$$\hat{r}_h^*(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i^*)\, Y_i^* ,$$

$$\hat{f}_h^*(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i^*) , \text{ and}$$

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i)$$

one gets

$$E^* \hat{m}_h^*(x) = E^* \frac{\hat{r}_h^*(x)}{\hat{f}_h^*(x)}$$

$$= E^* \frac{\hat{r}_h^*(x)}{\hat{f}_h(x)} \left[1 + \frac{\hat{f}_h^*(x) - \hat{f}_h(x)}{\hat{f}_h(x)} \right] + \text{smaller order terms}$$

$$\approx \hat{m}_h(x) + \left(\hat{f}_h(x)\right)^{-2} E^* \hat{r}_h^*(x) \left(\hat{f}_h^*(x) - \hat{f}_h(x)\right).$$

Now apply $E^* K_h\left(x - X_i^*\right) = \hat{f}_h(x)$ to obtain

$$E^* \hat{r}_h^*(x) \left(\hat{f}_h^*(x) - \hat{f}_h(x)\right)$$

$$= E^* n^{-2} \sum_{i=1}^n K_h\left(x - X_i^*\right) Y_i^* \left(K_h\left(x - X_i^*\right) - \hat{f}_h(x)\right)$$

$$= n^{-2} \sum_{i=1}^n \left(K_h(x - X_i)\right)^2 Y_i \; - \; n^{-1} \left(\hat{f}_h(x)\right)^2 \hat{m}_h(x)$$

$$= O_P((nh)^{-1}).$$

Another example where the paired bootstrap fails is given in Härdle and Mammen (1990). There the L_2 - distance between the nonparametric kernel regression estimator $\hat{m}_h$ and a parametric regression estimator is proposed as goodness of fit test statistic of a parametric regression model. It is shown that here bootstrap does not estimate the distribution of the test statistic consistently on the hypotheses. The test statistic turns out to be asymptotically equivalent to a U - statistic $c_n + \sum_{i \neq j} H_n((X_i,Y_i), (X_j,Y_j))$, which is clean , i.e. $E[H_n((X_i,Y_i), (X_j,Y_j)) \mid X_j,Y_j] = 0$ for $i \neq j$. The following lemma, which we learned from van Zwet (1989), shows that bootstrap does not work for clean U-statistics $\hat{T}_n$ as an estimate of the distribution of $V_n = \hat{T}_n - E\hat{T}_n$. But we may remark here that bootstrap may also work for clean U-statistics after another more appropriate choice of the bootstrapped statistic V_n .

LEMMA 2.3. *For a sample* $U_1, \ldots, U_n$ *of i.i.d. random variables and a symmetric function* H $(H(x,y) = H(y,x))$ *assume*

$$E(H(U_1,U_2) \mid U_2) = 0 ,$$

$$E \; H^2(U_1,U_1) < \infty,$$

$$E \, H^2(U_1,U_2) < \infty .$$

We consider

$$S = \Sigma_{i \neq j} \; H(U_i, U_j),$$

$$S^* = \Sigma_{i \neq j} \; H(U_i^*, U_j^*)$$

where $(U_1^*, \dots, U_n^*)$ *is a resample drawn from* $\{U_1, \dots, U_n\}$. *Then*

$$n^{-2}\,[\mathrm{var}^*(S^*) - 3\,\mathrm{var}(S)\,] \to 0 \quad \textit{in probability.}$$

Note that under the assumption of the lemma n^{-2} is the correct norming factor because of

$$\mathrm{var}(S) = n(n-1)\ E\ H^2(U_1,U_2)\,.$$

2.4. SMOOTHED BOOTSTRAP IN THE RANDOM DESIGN MODEL 2

A pairwise resampling procedure may be contructed by bootstrapping from the two dimensional distribution function

$$\hat{F}_n(x,y) = n^{-1} \textstyle\sum_{i=1}^n 1_{\{Y_i \le y\}} \displaystyle\int_{-\infty}^{x} K_g(t-X_i)\,dt$$

as in Cao-Abad and Gonzalez-Manteiga (1989) where g has to be appropriately chosen. The smooth bootstrap observations can be generated from a pairwise resampling as in section 2.3 by adding independent variables with density $g^{-1}\,K_g$ to the design variables. Denote now the smooth bootstrap sample by $(X_1^*, Y_1^*), \dots, (X_n^*, Y_n^*)$ and construct the bootstrap analogue (2.2) to (1.1). The marginal density of X_i^* is $\hat{f}_g(x) = n^{-1}\sum_{i=1}^n K_g(x - X_i)$. The bias will be correctly reflected since

$$E^*(Y_i^* \mid X_i^* = x) = \hat{m}_g(x).$$

Cao-Abad and Gonzalez-Manteiga (1989) investigate the accuracy of the bootstrap approximation in this resampling scheme and obtain

THEOREM 2.4. *The bootstrap approximation (i.e. the conditional distribution of* $\sqrt{n\,h}\,(\hat{m}_h^*(x) - \hat{m}_g(x))$) *approximates in probability the law of* $\sqrt{n\,h}\,(\hat{m}_h(x) - m(x))$ *with order* $O_P(n^{-2/9})$, *i.e.*

$$\sup_{z\in\mathbf{R}} \mid P^*\{\sqrt{n\,h}\,(\hat{m}_h^*(x) - \hat{m}_g(x)) \le z\} - P\{\sqrt{n\,h}\,(\hat{m}_h(x) - m(x)) \le z\} \mid = O_P(n^{-2/9}),$$

provided some regularity conditions and $g \sim n^{-1/9}$.

An analogous result can be shown for resampling from the joint kernel density of the bivariate distribution of (X_i, Y_i). The regularity conditions assumed in this theorem entail that the conditional distribution of the errors given $X_i = x$ depend smoothly on x. This assumption is not necessary in the resampling discussed in the next section.

2.5. NOT IDENTICALLY DISTRIBUTED ERRORS (MODEL 3), WILD BOOTSTRAPPING.

For MODEL 3 the *wild bootstrap* procedure has been introduced in Härdle and Mammen (1990). The wild bootstrap is related to proposals in linear regression models of Wu (1986) (see Beran (1986), Liu (1988), Liu and Singh(1988), Mammen (1989)). Since in this approach one is going to use one *single residual* $\hat{\varepsilon}_i$ to estimate the 'conditional' distribution $\mathcal{L}(Y_i - m(X_i) \mid X_i)$ by an estimate $\hat{F}_i$ we are calling it the *wild bootstrap*. More precisely $\hat{F}_i$ is defined as an arbitrary distribution which fulfills

$$E_{\hat{F}_i} Z = 0,$$

$$E_{\hat{F}_i} Z^2 = (\hat{\varepsilon}_i)^2,$$

$$E_{\hat{F}_i} Z^3 = (\hat{\varepsilon}_i)^3.$$

For instance one may use a two point distribution which is uniquely determined by these requirements. Then $\hat{F}_i = \mathcal{L}(Z_i)$ where $Z_i = -(\sqrt{5}-1)\hat{\varepsilon}_i / 2$ with probability $(\sqrt{5}+1)/(2\sqrt{5})$ and $Z_i = (\sqrt{5}+1)\hat{\varepsilon}_i / 2$ with probability $1 - (\sqrt{5}+1)/(2\sqrt{5})$. Or in another construction one may put $Z_i = \hat{\varepsilon}_i U_i$ where $\mathcal{L}(U_i)$ does not depend on i and where $EU_i = 0$, $E U_i^2 = E U_i^3 = 1$, e.g. $U_i = V_i/\sqrt{2} + (V_i^2 - 1)/2$ where the V_i's are independent N(0,1) - distributed variables.

For the construction of the bootstrap observations one generates independent $\varepsilon_i^* \sim \hat{F}_i$. Note that STEP 2 of section 2.1 is not necessary since the ε_i^* have automatically mean zero by construction. STEP 4,5 though are identical, one uses $(X_i, Y_i^* = \hat{m}_g(X_i) + \varepsilon_i^*)$ as bootstrap observations (for an appropriate choice of the bandwidth g , see section 2.1). Then one creates $\hat{m}_h^*(\bullet)$ according to (2.1).

To avoid technical regularity conditions on the 'conditional' error distributions let us consider the asymptotic performance of wild bootstrap in the random design MODEL 2. For the distribution of $\hat{m}_h(x) - m(x)$ at a finite number of points x wild bootstrap has been studied in Härdle and Marron(1990) see section 2.6 below.

Let us consider wild bootstrap for the test statistic of Härdle and Mammen(1990) which we have mentioned in the last section. This test statistic tests the hypothesis of a parametric regression model and it is based on the distance between the parametric and the nonparametric kernel regression estimate $\hat{m}_h$.

For the regression function $m(\bullet) = E(Y_i \mid X_i = \bullet)$ a parametric model $\{m_\theta : \theta \in \Theta\}$ may be given. The parametric approach shall be compared with the nonparametric analysis which are only based on the assumption that $m(\bullet)$ is a 'smooth' function. To this account a parametric regression estimator $m_{\hat{\theta}}$ may be plotted against a kernel estimator $\hat{m}_h$. The question arises if visible differences between $m_{\hat{\theta}}$ and $\hat{m}_h$ can be explained by stochastic fluctuations or if they suggest to use nonparametric instead of parametric methods. One way

to proceed is to measure the difference between $m_{\hat{\theta}}$ and $\hat{m}_h$ by a distance and to use this distance as a test statistic for testing the parametric model. The L_2-distance between the nonparametric and parametric fits has been proposed in Härdle and Mammen (1990). More precisely let $\mathcal{K}_{h,n}$ denote the (random) smoothing operator

$$\mathcal{K}_{h,n}\, g(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)\, g(X_i)}{\sum_{i=1}^{n} K_h(x - X_i)}.$$

Because of $E(\hat{m}_h(x) \mid X_1, \ldots . X_n) = \mathcal{K}_{h,n}\, m(x)$ consider the following modification of the squared deviation between $\hat{m}_h$ and $m_{\hat{\theta}}$:

$$\hat{T}_n = n\sqrt{h} \int \left(\hat{m}_h(x) - \mathcal{K}_{h,n}\, m_{\hat{\theta}}(x)\right)^2 \pi(x)\, dx$$

where π is a weight function.

We propose to use $\hat{T}_n$ as a test statistic to test the parametric hypothesis:

$$m \in \{m_\theta : \theta \in \Theta\}.$$

Related tests for testing a parametric form of a density have been proposed by Neuhaus(1986, 1988), Cox, Koh, Wahba, and Yandell (1988), Cox and Koh(1989), Eubank and Spiegelman(1989). Related bootstrap tests have been considered in Azzalini, Bowman, and Härdle (1989) and Firth, Glosup, and Hinkley (1989). For an approximate calculation of critical values we have to determine the asymptotic distribution of $\hat{T}_n$ for a parametric $m = m_{\theta_0}$.

For simplicity we consider only the k-dimensional linear parametric model. Put

$$m_\theta(x) = \theta_1 g_1(x) + \ldots + \theta_k g_k(x) = \langle \theta, g(x) \rangle$$

where g is a $\mathbf{R}^k$-valued function (for some k). With a smooth weight function w the weighted least squares estimator $\hat{\theta}_n = \hat{\theta}$ is defined by

$$\hat{\theta} = \arg\min_\theta \sum_{i=1}^{n} w(X_i)(Y_i - m_\theta(X_i))^2.$$

$\hat{\theta}$ can easily be calculated

$$\hat{\theta} = \left(\sum_{i=1}^{n} w(X_i)\, g(X_i)\, g(X_i)^T\right)^{-1} \sum_{i=1}^{n} w(X_i)\, g(X_i)\, Y_i .$$

Now construct independent $\varepsilon_i^* \sim \hat{F}_i$ and use now $(X_i, Y_i^* = m_{\hat{\theta}}(X_i) + \varepsilon_i^*)$ as bootstrap

observations. Then create $\hat{T}_n^*$ like $\hat{T}_n$ by the squared deviance between the parametric fit and the nonparametric fit. From the Monte Carlo approximation of $\mathcal{L}^*(\hat{T}_n^*)$ construct the $(1-\alpha)$ quantile $\hat{t}_\alpha^W$ and reject the parametric hypothesis if $\hat{T}_n > \hat{t}_\alpha^W$. The following Theorem has been shown in Härdle and Mammen(1990).

THEOREM 2.5. *Assume, that m lies in the parametric hypotheses* $\{m_\theta: \theta \in \Theta\}$. *Then under some regularity conditions for a deterministic sequence* c_n *the conditional distribution of* $\hat{T}_n^* - c_n$ *converges weakly to the same limit as the distribution of* $\hat{T}_n - c_n$ *(in probability).*

2.6 SIMULTANEOUS ERROR BARS

Under the condition of theorem 2.1. the conditional distribution of $\sqrt{nh}\,(\hat{m}_h^*(x) - \hat{m}_g(x))$ tends in probability to the same Normal limit as $\sqrt{nh}\,(\hat{m}_h(x) - m(x))$. This convergence holds in fact uniformly over a grid of points, so it can be used for the construction of error bars with simultaneous coverage probability. The accuracy of these confidence intervals has been investigated by Cao - Abad (1990) and Härdle, Huet and Jolivet(1990).

The main advantage of bootstrapping in this context lies in the fact that the simulated distribution of $\sqrt{nh}\,(\hat{m}_h^*(\cdot) - \hat{m}_g(\cdot))$ at a finite number of points can be easily used to contruct confidence intervals with simultaneuous coverage probability. A conservative way of constructing confidence intervals at a finite number M of design points is the Bonferroni method where we use M pointwise confidence intervals each with coverage probability $1 - \alpha/M$. A more accurate method is to construct first pointwise confidence intervals with coverage probability $1 - \beta$, say, such that the uniform coverage is $1 - \alpha$. We suggest the following halving approach. First try individual (i.e. at each design point) coverage probabilty $1 - \beta = 1-\alpha/(2M)$ and calculate by simulation the resulting simultaneuous coverage $1 - \alpha_\beta$. If the result is more than α then try $\beta = \alpha/(4M)$ otherwise next try $\beta = 3\alpha/(4M)$. After stopping this halving approach find neighboring values β_* and β^* so that $\alpha_{\beta_*} < \alpha < \alpha_{\beta^*}$. Finally take the weighted average of the β_* and β^* intervals. For an application in an econometric context see Härdle and Marron (1990).

References

Azzalini, A., Bowman, A.W. and Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika.* **76** 1 - 11.

Beran,R.(1986). Discussion to Wu,C.F.J.: Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14** 1295-1298.

Cao - Abad, R. (1990). Rate of convergence for the wild bootstrap in nonparametric regression. *Ann. Statist.*, to appear.

Carroll, R.J. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist.* **10** 1224 - 1233.

Cox, D., Koh, E., Wahba, G. and Yandell, B.(1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Ann.of Statist.* **16** 113 - 119.

Cox, D. and Koh, E.(1989). A smoothing spline based test of model adequacy in polynomial regression. *Ann. Inst. Statist. Math.* **41** 383 - 400.

Dikta, G.(1988). Approximation of nearest neighbour regression function estimators *Technical report, University of Gießen.*

Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel - Dekker New York.

Eubank, R. and Spiegelman, C. (1989). Testing the goodness - of - fit of linear models via nonparametric regression techniques. *Unpublished manuscript.*

Firth, D., Glosup, J., and Hinkley, D.V. (1989). Nonparametric curves for checking model fit. *Unpublished manuscript.*

Franke, J.(1990). Bootstrapping of nonlinear autoregressive time series. Some preliminary remarks. *Proceedings of the Bootstrap Conference in Trier, Germany,* Springer Verlag, to appear.

Franke, J. and Härdle, W.(1990). On bootstrapping kernel spectral estimates. *Ann. Statist.* , to appear.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. Biometrika **73** 625 633.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Ann. Statist.* **16** 927 - 953.

Hall, P. (1990a). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multivar. Anal.* **32** 177 - 203.

Hall, P. (1990b). On bootstrap confidence intervals in nonparametric regression. *Unpublished manuscript.*

Härdle, W. (1990). *Applied Nonparametric Regression.* Econometric Society Monograph Series, Cambridge University Press.

Härdle, W. and Bowman, A. (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83** 102 - 110.

Härdle, W. and Mammen, E. (1990). Comparing nonparametric versus parametric regression fits. *Preprint SFB 123, Universität Heidelberg.*

Härdle, W. and Marron, J.S. (1990). Semiparametric comparison of regression curves. *Ann. Statist.* **18** 63 - 89.

Härdle, W. , Huet, S. and Jolivet, E.. (1990). Better bootstrap confidence intervals for nonparametric regression. *Manuscript.*

Härdle, W. and Scott, D.W. (1990) Smoothing in high and low dimensions by Weigthed Averagaing of Rounded Points. *J.Royal Stat.Soc., Series B, Discussion paper,* to appear.

Liu, R. (1988). Bootstrap procedures under some non i.i.d. models. *Ann. Statist.* **16** 1696 - 1708.

Liu, R. and Singh K. (1989). Efficiency and robustness in resampling. *Preprint.*

McDonald, J. A. (1982). Projection pursuit regression with the ORION I workstation. *A 20 minute, 16 mm color sound film, available for loan from J. Friedman, Stanford University.* .

Mammen, E. (1989). Bootstrap and Wild Bootstrap for high dimensional Linear Models. submitted to *Ann. Statist.*.

Mammen, E.(1990a). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* , to appear.

Mammen, E. (1990b). Higher order accuracy of bootstrap for smooth functionals. submitted to *Scand. J. Stat.*.

Mammen, E. (1990c). The bootstrap : consistency and applicability. *Habilitationsschrift*, in preparation.

Müller, H.-G. (1988). *Nonparametric regression analysis for longitudinal data.* Springer lecture notes in statistics 46.

Nadaraya, E.A. (1964). On estimating regression. *Theory Prob. Appl.* **10** 186 - 190.

Neuhaus, G. (1986). A class of quadratic goodness of fit tests. *Preprint.*

Neuhaus, G. (1988). Addendum to "Local asymptotics for linear rank statistics with estimated score functions" . *Ann. Statist.* **16** 1342 - 1343.

Scott, D.W. and Terrell, G.R. (1987). Biased and Unbiased crossvalidation in density estimation. *J. Amer. Statist. Assoc.* **82** 1131-1146.

van Zwet, W. (1989). Hoeffding's decomposition and the bootstrap. *Talk given at the conference on "Asymptotic methods for computer-intensive procedures in statistics" in Oberwolfach, West-Germany.*

Watson, G.S. (1964). Smooth regression analysis. *Sankhya, Series A* 359 - 372

Wu,C.F.J.(1986). Jackknife, bootstrap, and other resampling methods in regression analysis (with discussion). *Ann. Statist.* **14** 1261-1295.

ON THE INFLUENCE FUNCTION OF MAXIMUM PENALIZED LIKELIHOOD DENSITY ESTIMATORS.

V. K. KLONIAS
Department of Mathematics
University of Crete
Heraklion - Crete
P.O. Box 1470 - Greece

ABSTRACT. An approximation to the influence function of a class of maximum penalized likelihood estimators (MPLE) of a probability density function, is obtained. It can explain some of the known behaviour of these estimates, e.g., their "bump-hunting" abilities. A study of the influence function suggests a larger class of estimators, which contains as special cases, both the kernel estimates and known MPLE's. This is a two-parameter (p,h) class, where h is a smoothing and $p \in [0,1)$ a non-linearity parameter. The strong consistency of these estimates, in various metrics - including a Sobolev, supremum over $\mathbb{R}$ and Hellinger - can be established in a unified manner. The problem of the numerical evaluation of the estimates is also addressed.

1. Introduction

Let $X_1,\ldots,X_n$ be independent observations from a distribution function F with density f over $\mathbb{R}$, and let F_n denote the associated empirical distribution function. The nonparametric maximum penalized likelihood method of density estimation (MPLE), introduced by Good and Gaskins (1971) and (1980), produces as the estimator f_n of the density f the maximizer - over an appropriate space of functions H - of the log-likelihood minus a "roughness" penalty functional Φ:

$$\max\{n \int \log g dF_n - \Phi(u) : u \in H\}$$

$$\text{subject to: } \int g = 1, \quad g \geq 0 \tag{1}$$

where $u = g^{\frac{1}{2}}$. Examples of penalty functionals include: $\Phi(u) = \alpha\int(u')^2$ or $\beta\int(u')^2+\alpha\int(u'')^2$, with $\alpha > 0$ $\beta \geq 0$. The existence and uniqueness of such MPLE's were rigorously established, within the framework of Sobolev spaces $W^{2,\ell} \equiv \{u \in L_2(\mathbb{R}) : \|u^{(\ell)}\|_2 < +\infty\}$, by DeMontricher, Tapia and Thompson (1975), see also Tapia and Thompson (1978). For consistency properties of such estimators, and those of the class with

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 125–131.
© 1991 Kluwer Academic Publishers.

penalties $\Phi(u) = \lambda \|u\|^2$, where the norm is induced by the inner product:

$$\langle u,v\rangle = \int \tilde{u}(t)\overline{\tilde{v}(t)}\, m(ht)dt, \tag{2}$$

see Klonias (1982) and (1984). Here $\tilde{u}$ denotes the Fourier transform of u and $\lambda, h > 0$ and $m(\cdot)$ belongs to a wide class of nonnegative functions. For discretized MPLE's see Scott, Tapia and Thompson (1980). For MPLE's with penalties on the logarithm of the density see Silverman (1982).

The MPLE's with penalties on the root-density, i.e., solutions to Problem (1), are given by $f_n = u_n^2$, where u_n is defined implicitly by:

$$u_n(x) = \frac{n}{\lambda}\int \frac{1}{u_n(y)}\, \kappa_h(x-y)dF_n(y), \tag{3}$$

with

$$\kappa_h(x) = k(x/h)/h, \tag{4}$$

where $x,y \in \mathbb{R}$ and $\tilde{k}m = 1$.

Note that the non-penalized version of Problem (1) does not have a solution - the likelihood functional is unbounded, over the set of all densities. For example, the likelihood of the classical kernel estimate:

$$\hat{f}_n(x) = \int \kappa_h(x-y)dF_n(y), \tag{5}$$

with $k \geq 0$, $\int k = 1$, tends to infinity as the bandwidth h tends to zero.

In Section 2, we derive an approximation to the influence function of the MPLE's, defined by (3), and through it explain some of the "bump-hunting" abilities of these estimates. These considerations lead us to define a wider class of estimates, defined by (12), with improved performance at the "peaks" and "valleys" of the underlying probability density.

In Section 3, we deal with the consistency of these new estimators, defined by (12), and with their numerical evaluation.

2. On the influence function

Let us consider now, the influence of an observation at z, on the density estimate at x, measured by the influence function:

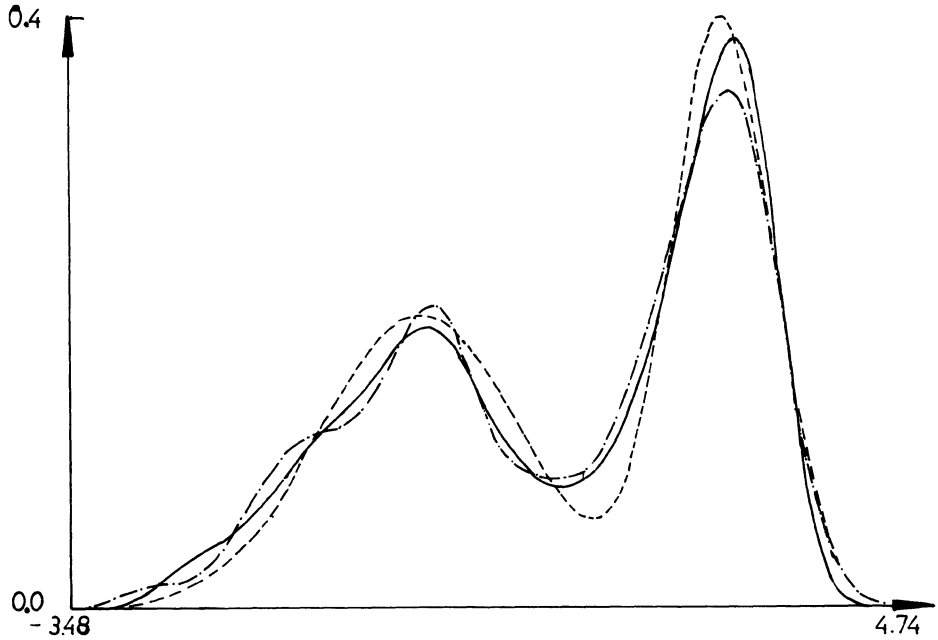


Figure 1: A: —— p = 0.9, h = 0.7; B: — • — p = 0.5, h = 0.4.

$$\psi(z|x,F) = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} f(x|F_\varepsilon), \tag{6}$$

where $F_\varepsilon = (1-\varepsilon)F + \varepsilon H_z$, $\varepsilon \in [0,1]$ and $H_z(\cdot)$ is the indicator of the interval $[z,+\infty)$. For the classical kernel estimate:

$$f(x|F_\varepsilon) = \int \kappa_h(x-y)dF_\varepsilon(y),$$

we have,

$$\psi_0(z|x,F) = \kappa_h(x-y) - f(x) \ . \tag{7}$$

In the case of an MPLE now, $f(x|F_\varepsilon) = u^2(x|F_\varepsilon)$, with

$$u(x|F_\varepsilon) = \int \frac{1}{u(y|F_\varepsilon)} \kappa_h(x-y)dF_\varepsilon(y), \tag{8}$$

we have that the influence function $\phi(z|x,F)$ of u, defined by (6) - with u in place of f - satisfies the integral equation:

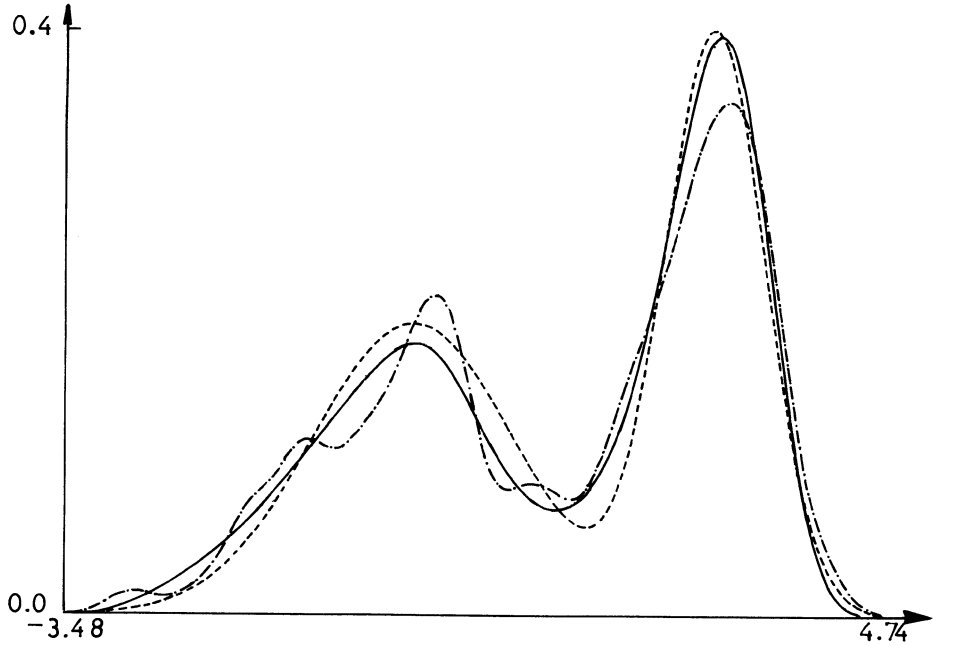


Figure 2: C: —— p = 0.75, h = 0.9; D: —·— p = 0, h = 0.25.

$$\phi(z|x,F)+\int \frac{1}{f(y|F)}\,\kappa_h(x-y)\phi(z|y,F)dF(y)$$
$$=\int \frac{1}{u(y|F)}\,\kappa_h(x-y)d(H_z-F)(y). \tag{9}$$

Note that, to a first order of approximation,

$$\phi(z|x,F)\approx\frac{1}{2}\int \frac{1}{u(y|F)}\,\kappa_h(x-y)d(H_z-F)(y), \tag{10}$$

and hence,

$$\phi(z|x,F)\approx\left[\frac{f(x|F)}{f(z|F)}\right]^{\frac{1}{2}}\kappa_h(x-z)-f(x|F). \tag{11}$$

Thus, an MPLE shifts probability mass from lower density points to higher density points, counteracting the effect of convolution - present in all methods of density estimation. In this way we can obtain a more "comfortable" fit - i.e., for larger h's - at the "peaks" and "valleys" of the curves simultaneously, which results in a better overall fit.

Based on these observations, we propose a generalization of the MPLE's considered above, defined through the following equation:

$$f_n^{1-p}(x) = \frac{n}{\lambda} \int \frac{1}{f_n^p(y)} \kappa_h(x-y) dF_n(y), \tag{12}$$

$p \in [0,1)$ and $\lambda > 0$ is computed so that $\int f_n = 1$. The influence function of the estimators in this class, is approximated - as in the case of the MPLE's considered above - by:

$$\psi_p(z|x,F) \approx \left[\frac{f(x|F)}{f(z|F)}\right]^p \kappa_h(x-z) - f(x|F). \tag{13}$$

Thus, the nonlinearity parameter p controls the shift of probability mass from low to high density points, further enhancing the diagnostic capabilities of these estimates. In this respect, compare the performance of estimates D - in Figure 2 - and B,A - of Figure 1 - corresponding to $p = 0, 0.5, 0.9$, i.e., to a kernel estimate, an MPLE and a member of the class defined through (12), respectively. The sample, of size $n = 100$, is from an equal mixture of a $N(0,1)$ and a $N(0,1/4)$ - its density being plotted with a dashed line in both figures - was generated using the standard random number generator of IMSL, with seed= 1141592653. The estimates A,B,D were based on kernel $k = \phi$ - the standard normal density - while estimate C was based on $k = \phi - \phi''$.

3. On consistency and computation

In this section, we consider first the question of the consistency of the estimators defined through (12). First, we define the auxiliary function:

$$\bar{f}_n^{1-p}(x) = \int \frac{1}{f^p(y)} \kappa_h(x-y) dF_n(y),$$

where f denoted the true underlying density.
Note that,

$$\| f_n^{1-p} - f^{1-p} \|^2 \leq \langle \bar{f}_n^{1-p} - f^{1-p}, f_n^{1-p} - f^{1-p} \rangle,$$

and hence,

$$\| f_n^{1-p} - f^{1-p} \|^2 \leq \| \bar{f}_n^{1-p} - f^{1-p} \|^2$$
$$= \frac{k(0)}{n^2 h_n} \Sigma_{i=1}^n \frac{1}{f^{2p}(X_i)} +$$

$$+\frac{1}{n^2}\sum_{i\neq j}\sum \kappa_{h_n}(X_i-X_j)/[f^p(X_i)f^p(X_j)]$$

$$-\frac{2}{n}\Sigma_{i=1}^n f^{1-2p}(X_i)+\|f^{1-p}\|^2 . \tag{14}$$

Thus, for $0\le p<1/2$, under a modest moment assumption on the underlying distribution, we have that

$$\|f_n^{1-p}-f^{1-p}\|^2=O_p((nh_n)^{-1})+O_p(h_n^{2(s+1)}),$$

where $s=0,1,2,\ldots$ denotes the number of even moments of the kernel k which are zero.

For $1/2\le p<1$, the situation is more demanding, primarily because the random variables $\{f(X_i)\}^{-2p}$ do not even have a finite first moment. The first term in (14), is then handled by Marcinkiewicz's version of the SLLN (see, e.g. Loève 1977), and for the second we can employ a truncation technique analogous to that in Klonias (1984).

We can thus conclude the strong consistency of these estimators, for each $p\in[0,1)$, with respect to the norm induced by (2). It is then straightforward to obtain the consistency of the estimators in the $L_2(\mathbb{R})$ and supremum over $\mathbb{R}$ norms. For $p\ge 1/2$ the rate of convergence depends on p and the rate at which $h=h_n$ vanishes to zero, depends on p as well. In most numerical examples, including the ones we present here, the best fits are obtained by estimators corresponding to large $p\ge 1/2$; while, at the moment, the rates we obtained are optimal for $p\le 1/2$ and seem very slow for large p.

For the numerical evaluation of the estimators, given implicitly by (12), we have to solve the following system of nonlinear equations in $q_i\equiv\{\lambda h f_n(X_i)\}^{-p}\ge 0$, $i=1,\ldots,n$:

$$q_i^{a-1}=\Sigma_{j=1}^n q_j\, k((X_i-X_j)/h), \qquad 1\le i\le n,$$

where $a\equiv 2-(1/p)<1$. Equivalently, we have to find the stationary point of the following optimization problem:

$$\min\{q^T\mathbb{K}q-(1/a)\Sigma_{i=1}^n q_i^a,\quad q\in\mathbb{R}_+^n\},$$

where $q^T\equiv(q_1,\ldots,q_n)$ and $\mathbb{K}$ denotes the $n\times n$, positive definite matrix, with $k((X_i-X_j)/h)$ as its $(i,j)^{th}$-element. This numerical optimization problem can be solved with the techniques in Klonias and Nash (1988).

Acknowledgment. This research was supported in part by a Grant from the Greek Government.

References

DeMontricher, G.F., Tapia, R.A. and Thompson, J.R. (1975). 'Nonparametric maximum likelihood estimation of probability densities by penalty function methods'. Ann. Statist. 3, 1329-1348.

Good, I.J. and Gaskins, R.A. (1971). 'Nonparametric roughness penalties for probability densities'. Biometrika 58, 255-277.

Good, I.J. and Gaskins, R.A. (1980). 'Density estimation and bumphunting by the penalized likelihood methods exemplified by scattering and meteorite data'. (Invited Paper). J. Amer. Statist. Assoc. 75, 42-73.

Klonias, V.K. (1982). 'Consistency of two nonparametric maximum penalized likelihood estimators of the probability density function'. Ann. Statist. 10, 811-824.

Klonias, V.K. (1984). 'On a class of nonparametric density and regreession estimators'. Ann. Statist. 12, 1263-1284.

Klonias, V.K. and Nash, S.G. (1988). 'On the numerical evaluation of a class of nonparametric estimators'. J. Statist. Computation Simulation.

Loève, M. (1977). Probability Theory. Springer-Verlag, Berlin. Graduate Texts in Mathematics 45.

Scott, D.W., Tapia, R.A. and Thompson, J.R. (1980). 'Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria'. Ann Statist. 8, 820-832.

Silverman, B.W. (1982). 'On the estimation of a probability density function by the penalized likelihood method'. Ann. Statist. 10, 795-810.

Tapia, R.A. and Thompson, J.R. (1978). Nonparametric Probability Density Estimation. The Johns Hopkins University Press, Baltimore and London.

NONPARAMETRIC CURVE ESTIMATION AND SIMPLE CURVE CHARACTERISTICS

E. MAMMEN
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
6900 Heidelberg
Germany

ABSTRACT. Many estimators of a nonparametric regression function m or a nonparametric density function f are based on some (implicit) smoothness assumptions on the curve m or f. In this paper we propose to use qualitative shape parameters of the curve instead of the usual quantitative measures of smoothness. Especially one may consider the number of extreme points, of inflection points, or of other characteristic points of the curve as shape parameter. These may be interpreted as a qualitative measure of smoothness. We consider the class of functions whose (p-1) st derivative consists of at most k monotone pieces. It is shown that in this class a regression function or density can be estimated with the same speed of convergence as in the classical approach of considering regression or density functions whose p-th derivatives lie in a ball in the L_∞ or L_2 norm. We show that in nonparametric regression this rate of convergence holds especially for the least squares estimator. For instance the least squares piecewise concave convex estimator converges with rate of convergence $O_p(n^{-2/5})$, i.e. with the same rate of convergence as classical nonparametric regression estimates under the usual conditions on the second derivative of m. The least squares estimator turns out to be a regression spline with estimated knot points.

1. Introduction.

In this paper we will study a new approach for estimating a nonparametric regression function m or a nonparametric density function f. Let us first consider the following nonparametric regression model :

$$Y_i = m(x_i) + \varepsilon_i \quad (i = 1,\ldots,n).$$

Given the observations (Y_i) and the deterministic design points (x_i), we want to estimate the unknown regression function m. The ε_i's are independent random variables with $E\,\varepsilon_i = 0$.

Many estimators in nonparametric regression are based on two steps:

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 133–139.
© 1991 *Kluwer Academic Publishers.*

♠ First estimate the average variance of the errors and the smoothness of m (for instance measured by the L_2-norm of a certain derivative of m).

♠ Then apply a nonparametric regression estimate which works well for regression functions having this amount of smoothness.

In this paper we want to make the following point: In this two step procedure it is possible to use qualitative shape parameters of the regression function instead of the usual quantitative measures of smoothness. For instance one may consider the number of extreme points, of inflection points, or of other characteristic points of the curve as shape parameter. For simplicity we suppose that the design points x_i lie in a compact subset K of **R**. We will consider

$T_p(m) = \inf \{ k \mid$ there exists a partition of K into k intervals $I_1,\dots,I_k$ such that $m^{(p-1)}$ is monotone on every I_j $(j = 1,\dots,k)\}$.

For $p \geq 2$ the quantity $T_p(m)$ is defined for all (p-2) times continuously differentiable functions m by defining $m^{(p-1)}$ as the (interval valued) generalized derivative of $m^{(p-2)}$. We consider classes $\mathcal{H}_{p,k,D}$ of similarly shaped regression functions :

$\mathcal{H}_{p,k,D} = \{m : K \to \mathbf{R} \mid m$ is (p - 2) times continuously differentiable, $T_p(m) \leq k$, $|m^{(p-1)}| \leq D\}$.

The dependence of $\mathcal{H}_{p,k,D}$ on D will be discussed further below . In particular for p=1 and p=2 $\mathcal{H}_{p,k,D}$ consists in piecewise monotone functions and piecewise concave/convex functions, respectively.

We propose the following procedure for selecting an estimator for m:

STEP 1 : *Choose* p . *This choice will be motivated by the goal of the data analysis and by a priori knowledge about* m *and will depend on the number* n *of observations.*

STEP 2 : *Estimate (the shape of* m*)* $T_p(m)$.

STEP 3 : *Estimate* m *using an estimator* $\hat{m}_n = \hat{m}_{n,p,k,D}$ *with values in* $\mathcal{H}_{p,k,D}$ *where* k *is the estimate of* $T_p(m)$.

The use of $T_p(m)$ has some advantages. First it can be easier interpreted than the usual abstract quantitative measures of smoothness. This is especially important if we are interested in the regression function m from a more data analytic point of view. Furthermore because $T_p(m)$ is discrete we can calculate $\hat{m}_{n,p,k,D}$ for the set of all interesting k. The following slight modification $\tilde{T}_p(m)$ of $T_p(m)$ may also be used. $\tilde{T}_p(m)$ is defined as $T_p(m)$ but $m^{(r)}$ is assumed to be monotone on every interval I_j $(j = 1,\dots,k)$ for *every* $r \leq p - 1$. This definition seems to be more appropriate for measuring changes of the structure of m. The estimator $\hat{m}_n = \hat{m}_{n,p,k,D}$ has been proposed by Holm and Frisén(1985) for p=2. In spirit related approaches are proposed in Kneip and Gasser (1988) and in Mächler(1989). Rates of convergence and

algorithms for a related estimate are discussed in Nemirovskii, Polyak, and Tsybakov(1984, 1985). For reference to further literature see Mammen(1990a).

2. Density estimation.

For a sample of i.i.d. observations $X_1,\ldots, X_n$ with values in a bounded interval of $\mathbf{R}$ we consider the following model assumption on the density f

$$f \in \mathcal{H}_{p,k,D,C} = \mathcal{H}_{p,k,D} \cap \{ g : \sup_x | g(x) | \leq C \}.$$

In the next theorem we show that f can be estimated in this model with the same rate of convergence as under the usual model assumptions on the p - th derivative of f.

THEOREM 2.1. *Fix* p,k,D, *and* $C > 0$. *There exists an estimator* $\hat{f}_n$ *of the density* f *such that to every* $\varepsilon > 0$ *there exists a* $b(\varepsilon)$ *with*

$$\sup_{f \in \mathcal{H}_{p,k,D,C}} P_f (d_1 (\hat{f}_n, f) > b(\varepsilon)\, n^{-p/2p+1})) < \varepsilon$$

where d_1 *denotes the* L_1 *- distance* $d_1(g_1, g_2) = \int | g_1(x) - g_2(x) | \, dx$.

The proof of the theorem will be given in the appendix. The Theorem is of only theoretical interest because it says nothing about the form of the estimate $\hat{f}_n$. For the special case of monotone densities (i.e. $p = k = 1$) it is well known that the rate of convergence of the Theorem holds for the Grenander estimate (see Groeneboom, 1985). The Grenander estimate has two interpretations (see Barlow et al. , 1972). First it is the maximum likelihood estimate in the class of monotone densities. Unfortunately we do not know if it is appropriate to apply the maximum likelihood principle for other choices of p and k although we conjecture that it is possible to prove the rates of convergence of Theorem 2.1 also for $p > 1$ by using the methods of van de Geer (1990b). Secondly the Grenander estimate is the weighted least squares estimate in the following 'regression' problem. The 'observations' are $(X_{(i+1)} - X_{(i)})^{-1}$ and the weights are $(X_{(i+1)} - X_{(i)})$ where $(X_{(1)}, \ldots, X_{(n)})$ is the order statistic of the sample $(X_1,\ldots, X_n)$. This estimate can be easily calculated, see the next section where least squares regression estimates are treated. Elsewhere the asymptotic results of the next section will be applied to this estimate.

3. Nonparametric regression.

Let us consider again the nonparametric regression model :

$$Y_{i,n} = m(x_{i,n}) + \varepsilon_{i,n} \quad (i = 1,\ldots,n).$$

Given the observations $(Y_{i,n})$ and the deterministic design points $(x_{i,n})$, we want to estimate the unknown regression function m . The $\varepsilon_{i,n}$'s are independent random variables with $E\, \varepsilon_{i,n} =$

0. We will consider the least squares estimate $\hat{m}_n$ of m , i.e.

$$\hat{m}_n = \hat{m}_{n,p,k,D} = \operatorname{argmin}_{\mu \in \mathcal{H}_{p,k,D}} \sum_{i=1}^{n} (\mu(x_{i,n}) - Y_{i,n})^2.$$

Let us cite the following asymptotic results for $\hat{m}_n$ from Mammen(1990a). Fix $D > 0$.

THEOREM 3.1 ($p = 1$, *piecewise isotone regression*). *Consider the* $\mathcal{H}_{1,k,D}$ *least squares estimator* $\hat{m}_n$. *Let the* $\varepsilon_{i,n}$*'s be independent and distributed according to a distribution with expectation* 0 *and continuous variance* $\sigma^2(x_{i,n})$. *Furthermore assume that* $x_{i,n} \approx F^{-1}(i/n)$ *for a distribution function* F. *Then under some regularity assumptions*

$$n^{1/3} \frac{(2\,F'(x_o))^{1/3}}{\sigma(x_o)^{2/3}\, m'(x_o)^{2/3}} \left(\hat{m}_n(x_o) - m(x_o) \right)$$

converges in distribution to the slope at zero of the greatest convex minorant of $W(t) + t^2$, *where* W *is the twosided Brownian motion.*

THEOREM 3.2 ($p = 2$, *piecewise convex/concave regression*). *Consider the* $\mathcal{H}_{2,k,D}$ *least squares estimator* $\hat{m}_n$. *Under some regularity conditions :*

$\hat{m}_n(x_o) - m(x_o) = O_P(n^{-2/5})$ *if* $m''(x_o)$ *exists and* $m''(x_o) \neq 0$.

and

$$\hat{m}_n(x_o) - m(x_o) = o_P(n^{-2/5}) \quad \textit{if } m''(x_o) = 0 .$$

THEOREM 3.3 (*arbitrary* p). *Consider the* $\mathcal{H}_{p,k,D}$ *least squares estimator* $\hat{m}_n$. *Then under some regularity conditions the following holds:*

$$\sup_{m \in \mathcal{H}_{p,k,D}} E_m \| \hat{m}_n - m \|_n^2 = O(n^{-2p/(2p+1)})$$

where $\| g \|_n^2 = n^{-1} \sum_{i=1}^{n} g(x_i)^2$.

The proofs of Theorems 3.1. - 3.3. are given in Mammen (1990a). Theorem 3.3. follows by application of the results of van de Geer (1990a). The proof is similar to the proof of Theorem 2.1. given in section 4.

$\hat{m}_n$ may be compared with other nonparametric regression estimators. For instance the smoothness classes $\mathcal{F}_{p,C}$ and $\mathcal{G}_{p,C}$ have been studied where

$$\mathcal{F}_{p,C} = \{ m : K \to \mathbf{R} \mid m^{(p)} \text{ exists and is absolutely bounded by } C \},$$

$$\mathcal{G}_{p,C} = \{ m : K \to \mathbf{R} \mid m^{(p)} \text{ exists and } S_p(m) \leq C \}$$

where

$$S_p(m) = \int_K m^{(p)}(x)^2\, dx.$$

There exist estimators $\tilde{m}_n$ (depending on m but not on C) such that under some regularity conditions on the design and on the distribution of the $\varepsilon_{i,n}$'s

$$\sup_m E_m \int_K (\tilde{m}_n(x) - m(x))^2 \, dx = 0_p(n^{-2p/(2p+1)}).$$

Here the supremum is taken over $\mathcal{F}_{p,C}$ or $\mathcal{G}_{p,C}$. It has been shown that the rate $n^{-2p/(2p+1)}$ is the best available (see Ibragimov and Hasminskii (1980), Stone (1982)). Therefore a regression function m can be estimated under the assumption that $T_p(m)$ is bounded as well as under the assumption that $S_p(m)$ is bounded as far as the order of convergence is concerned. It should be remarked that uniformly over $\mathcal{H}_{m,k,D}$ the rate of convergence in Theorem 3.3 cannot be achieved by a kernel estimator $\tilde{m}_n(x)$ with bandwidth which does *not* depend on x.

The estimator $\hat{m}_n$ has a nice interpretation. It turns out that $\hat{m}_n$ can be chosen as a a regression spline of order (p - 1) , i.e. $\hat{m}_n$ is an (p - 2) times continuously differentiable function and piecewise a polynomial of degree (p - 1). The knot points depend on the observations. Asymptotic analysis suggests that the number of knot points is locally (optimally) adapted to the variance of the observations at neighboring design points and to the local density of the design points. $\hat{m}_n$ is therefore a regression spline with data adaptive chosen knot points. Furthermore this form of $\hat{m}_n$ suggests algorithms of the following type for the calculation of $\hat{m}_n$. Choose a set of points as knot points and calculate the regression spline with these knot points. Then change the set of knot points according to the behaviour of the calculated regression spline and iterate. Details are discussed in Mammen(1990a).

We have compared the least squares estimator $\hat{m}_n$ with a kernel estimator $\tilde{m}_n$ in a simulation study (see Mammen, 1990a). We have considered regression functions in $\mathcal{H}_{2,2,D}$, i.e. regression functions which are convex/concave or concave/convex. The simulations can be summarized as follows.

♠ If the variance of the error variables is not too large the goodness-of-fit of the least squares estimator $\hat{m}_n$ is comparable to the kernel estimator.

♠ This is not true for very noisy data. Then the inflection point (i.e. the unknown boundary point of the intervals I_1 and I_2) is not estimated well by $\hat{m}_n$. This affects also the accuracy of $\hat{m}_n$ at points near the inflection point.

♠ The range of function values of $\hat{m}_n$, i.e. $\sup_x \hat{m}_n(x) - \inf_x \hat{m}_n(x)$, is larger for the least squares estimator than for the kernel estimator. This may be interpreted as follows. $\hat{m}_n$ tends to caricature the shape of the regression function m and to have a larger slope at the boundary of the intervals I_j whereas the kernel estimator may smooth features of the regression function away. One may see this as no principal disadvantage of $\hat{m}_n$ if one is more interested in the shape of m. Furthermore this point indicates that $\hat{m}_n$ is a good complement to the kernel estimate $\tilde{m}_n$ and that it makes sense to use in a data analysis $\hat{m}_n$ *and* $\tilde{m}_n$.

4. Summary.

We have shown that in nonparametric curve estimation problems the usual smoothness conditions can be replaced by more qualitative assumptions on simple characteristics of the

curve. In a data analysis this may be judged as a preferable approach

if one is more interested in the global form of the curve than in an accurate pointwise estimation of the curve or

if a priori knowledge on the shape of the curve is available.

We think that the here proposed estimates need some further analysis. Furthermore modifications of the estimates should be considered. For instance in the regression set up of section 3 we think of least squares smoothing splines or of least squares fits to a undersmoothed kernel estimate (see also Mammen(1990b) for the special case $p = 1$). Or one may consider modifications of the function classes $\mathcal{H}_{p,k,D}$. In a density estimation problem, for instance, it could make sense to assume that the density f lies in $\mathcal{H}_{p,k,D}$ after a known transformation of f.

5. Appendix. Proof of the Theorem in Section 2.

Denote the number of balls of diameter δ (measured by a distance d) which are necessary to cover a set $\mathcal{H}$ by $N(\delta, d, \mathcal{H})$. $\log(N(\delta, d, \mathcal{H}))$ is also called the Kolmogorov entropy of $\mathcal{H}$. We use the following result in Yatracos(1985).

THEOREM. *Suppose that a set* $\mathcal{H}$ *of densities is* L_1 *- totally bounded. Then there exists an estimator* $\hat{f}_n$ *of the density* f *such that for every* $\varepsilon > 0$ *there exists a* $b(\varepsilon)$ *with*

$$\sup_{f\in \mathcal{H}} P_f (d_1(\hat{f}_n, f) > b(\varepsilon)\, a_n) < \varepsilon$$

where a_n *satisfies* $n^{1/2} a_n = [\log N(a_n, d_1, \mathcal{H})]^{1/2}$.

Denote the sup norm by $d_\infty(g) = \sup_{0\le x\le 1} |g(x)|$ and the L_2 norm by $d_2^2(g) = \int g^2(x)\, dx$. Note that the sup norm d_∞ and the L_2 - norm d_2 bound the L_1 - norm. Our Theorem follows now from the following results in Babenko et al.(1979) (see also Birman and Solomjak, 1967).

Put $\mathcal{G}(p,M,L) = \{ g:[0,1] \to \mathbf{R} : \int_0^1 | g^{(p)}(x)| \le M , \sup_{0\le x\le 1} |g(x)| \le L \}$. Then

$$\log N(\delta, d_\infty, \mathcal{G}(p,M,L)) \le K\, (M/\delta)^{1/p} + p \log(L/\delta)$$

for a constant K if $p \ge 2$, and

$$\log N(\delta, d_2, \mathcal{G}(p,M,L)) \le K\, (M/\delta)^{1/p} + p \log(L/\delta)$$

for a constant K if $p = 1$.

References

Babenko, K.I.(editor) (1979). *Theoretical foundations and construction of numerical algorithms for the problems of mathematical physics.* Nauka, Moscow (in Russian).

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., and Brunk, H.D. (1972). *Statistical inference under order restrictions.* Wiley, New York.

Birman, M.S. and Solomjak, M.Z. (1967). Piecewise-polynomial approximations of functions of the classes W_p^α. *Math. USSR-Sbornik* **2** 295-317.

Birgé, L. (1987). Estimating a density under order restrictions: Nonasymptotic minimax risk. *Ann.Statist.* **15** 995 - 1012.

Groeneboom, P. (1985). Estimating a monotone density. *Proceedings of the Berkeley Converences in Honor of Neyman, J. and Kiefer, J. (ed. by LeCam, L.M. and Olshen, R.A.) II* 539 – 555.

Holm, S. and Frisén, M. (1985). Nonparametric regression with simple curve characteristics. *Research report* 1985: **4**, Univ. of Göteborg, Dept. of Statistics.

Ibragimov, I.A. and Hasminskii, R.Z. (1980). On nonparametric estimation of regression. *Soviet Math. Dokl.* **21** 810 – 814.

Kneip, A. and Gasser, T. (1988). Convergence and consistency results for self modeling nonlinear regression. *Ann. Statist.* **16** 82 - 112.

Mächler, M.B. (1989). 'Parametric' smoothing quality in nonparametric regression: shape control by penalizing inflection points. *Dissertation*, ETH Zürich.

Mammen, E. (1990a). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.*, to appear.

Mammen, E. (1990b). Estimating a smooth monotone regression function. *Ann. Statist.*, to appear.

Nemirovskii, A.S., Polyak, B.T., and Tsybakov, A.B. (1984). Signal processing via the nonparametric maximum likelihood method. *Problemy Peredachi Informatsii* **20** 29-46.

Nemirovskii, A.S., Polyak, B.T., and Tsybakov, A.B. (1985). Rate of convergence of nonparametric estimates of maximum-likelihood type. *Problemy Peredachi Informatsii* **21** 258-272.

Stone, C.J. (1982). Optimal rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040 – 1053.

Van de Geer, S. (1990a). Estimating a regression function. *Ann. Statist.* **18** 907 – 924.

Van de Geer, S. (1990b). Hellinger - consistency of certain nonparametric maximum likelihood estimators. *Preprint.*

Yatracos, Y.G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Ann. Statist.* **13** 768 – 774.

APPLICATIONS OF MULTIPARAMETER WEAK CONVERGENCE FOR ADAPTIVE NONPARAMETRIC CURVE ESTIMATION

HANS-GEORG MÜLLER and KATHY PREWITT
Division of Statistics
University of California
Davis CA 95616 USA

ABSTRACT. We give an overview on applications of weak convergence of stochastic processes to obtain adaptive nonparametric curve estimators through efficient data-based local bandwidth choices. We point out new developments based on multivariate time stochastic processes. Examples are multivariate curve estimates, where several local bandwidths are to be chosen for different coordinates, and estimates of local functionals of curves which can be expressed as maxima or zeros of local deviation processes and also depend on a bandwidth. As an illustration, we show that adaptive mode estimation for a probability density function is a consequence of weak convergence of a two-dimensional process in a bandwidth and a deviation coordinate. Various adaptive mode estimators are discussed.

1. Introduction

Curve estimation techniques depend on the choice of a smoothing parameter which often is crucial for practical applications. Two approaches are possible: global and local bandwidth choice. Since the latter contains the former as a special case, it can be expected to perform better, in particular for curves which exhibit large curvature in some areas and are relatively flat in others. Local bandwidth choice is also of particular interest for the estimation of local functionals, like coordinates of peaks, of peaks of a derivative, of zeros or of zeros of a derivative.

In this contribution, we illustrate how weak convergence techniques for multiparameter stochastic processes can be used to achieve efficient local bandwidth choices and adaptive estimates in many situations. The technique seems to have been applied for the first time in the univariate case by Abramson (1982) and Krieger and Pickands (1981) for kernel density estimation.

Given n i.i.d. data $X_1,\ldots,X_n$, having distribution function F and probability density function f, Abramson (1982) considers the kernel density estimator

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 141–166.
© 1991 *Kluwer Academic Publishers.*

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) \tag{1.1}$$

which was proposed by Rosenblatt (1956) and further discussed by Parzen (1962) and Rosenblatt (1971). Here, h_n is a sequence of bandwidths which asymptotically converges to zero, and K is a kernel function, satisfying $\int K(v)dv = 1$, $\int K(v)vdv = 0$.

Taking local bandwidth sequences of the form

$$h_n(s) = sn^{-1/5} \tag{1.2}$$

depending implicitly on the argument x where the density is to be estimated and a bandwidth parameter $s = [s_a, s_b]$ for some appropriately chosen $0<s_a<s_b<\infty$, Abramson showed that for fixed x, the sequence of stochastic processes in the univariate bandwidth parameter s,

$$Z_{1n}(s) = n^{2/5}\left(f_n(x,h_n(s)) - f(x)\right) \tag{1.3}$$

converges weakly on $\mathcal{C}([s_a,s_b])$ to a limiting Gaussian process $Z_1(s)$ with moment structure

$$EZ_1(s) = \frac{f^{(2)}(x)}{2} s^2 \int K(v)v^2 dv \tag{1.4}$$

$$\mathrm{cov}(Z_1(s_1),\ Z_1(s_2)) = \frac{f(x)}{s_1 s_2} \int K\left(\frac{v}{s_1}\right) K\left(\frac{v}{s_2}\right) dv. \tag{1.5}$$

A similar result was proven independently by Krieger and Pickands (1981).

For the fixed design nonparametric regression model

$$y_i = g(x_i) + \varepsilon_i, \quad 0 \le x_1 \le \ldots \le x_n \le 1, \quad i=1,\ldots,n, \tag{1.6}$$

where g is a smooth regression function with support [0,1], ε_i are i.i.d. errors with $E\varepsilon_i = 0$, $E\varepsilon_i^2 = \sigma^2 < \infty$, and y_i are noisy measurements of g obtained at asymptotically equidistant points x_i, a corresponding result was established in Müller and Stadtmüller (1987). The estimator considered there is a modification of an estimator due to Priestley and Chao (1972),

$$\tilde{g}_n(x,h_n) = \frac{1}{h_n} \sum_{i=1}^{n-1} (x_{i+1}-x_i) K\left(\frac{x-x_i}{h_n}\right) y_i \tag{1.7}$$

namely the convolution integral estimator

$$\tilde{g}_n(x,h_n) = \frac{1}{h_n} \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h_n}\right) du\ y_i, \tag{1.8}$$

which was proposed in Gasser and Müller (1979), see also Cheng and Lin (1981) for a closely related proposal. There again, K is a kernel function and h_n a sequence of bandwidths, and the s_i's are interpolating the x_i's, $s_o = 0$, $x_i \le s_i \le x_{i+1}$, $1 \le i \le n-1$, $s_n = 1$. A

common choice is $s_i = (1/2)(x_i + x_{i+1})$, $1 \le i \le n-1$. Again, if we assume certain properties for the kernel function and the regression function g, incorporating bandwidth choices (1.2) leads for each fixed $x \in (0,1)$ to consideration of the stochastic processes with univariate time

$$Z_{2n}(s) = n^{2/5}\left(g_n(x,h_n(s)) - g(x)\right), \tag{1.9}$$

where $s \in [s_a, s_b]$ for suitably chosen $0 < s_a < s_b < \infty$. Then it is shown that Z_{2n} converges weakly on $\mathcal{C}([s_a,s_b])$ to a Gaussian limit Z_2 with

$$EZ_2(s) = \frac{g^{(2)}(x)}{2} s^2 \int K(v) v^2 dv \tag{1.10}$$

$$\operatorname{cov}(Z_2(s_1), Z_2(s_2)) = \frac{\sigma^2}{s_1 s_2} \int K\left(\frac{v}{s_1}\right) K\left(\frac{v}{s_2}\right) dv. \tag{1.11}$$

The result was actually derived for higher order kernels, which allow further bias expansions, see below.

A third model considered frequently in curve estimation is the random design regression model, where it is assumed that n i.i.d. observations $(x_1,y_1),\ldots,(x_n,y_n)$ are available in order to infer the regression function

$$m(x) = E(y|X=x). \tag{1.12}$$

A common kernel estimation approach in this situation is the quotient-type Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964)

$$m_n(x,h_n) = \frac{\frac{1}{nh_n} \sum_{i=1}^{n} y_i K\left(\frac{x-X_i}{h_n}\right)}{\frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)} \tag{1.13}$$

which is motivated by estimating numerator and denominator of $m(x) = \int y f(x,y) dy / f(x)$ separately, provided that the joint density $f(x,y)$ of (x,y) exists.

For this situation, Mack and Müller (1987) investigated the univariate processes

$$Z_{3n}(s) = n^{2/5}\left(m_n(x,h_n(s)) - m(x)\right),$$

in the bandwidth parameter $s \in [s_a, s_b]$ for fixed x and suitably chosen values $0 < s_a < s_b < \infty$, where $h_n(s)$ is defined as in (1.2). It is shown that Z_{3n} converges weakly on $\mathcal{C}([s_a,s_b])$ to a Gaussian limit Z_3 with

$$EZ_3(s) = \frac{f(x)m^{(2)}(x) + 2f^{(1)}(x)m^{(1)}(x)}{2f(x)} s^2 \int K(v) v^2 dv,$$

$$\operatorname{cov}\left(Z_3(s_1), Z_3(s_2)\right) = \frac{\operatorname{var}(y|X=x)}{f(x)} \frac{1}{s_1 s_2} \int K\left(\frac{v}{s_1}\right) K\left(\frac{v}{s_2}\right) dv.$$

This result was also presented for the case of a multivariate independent variable X; the considered processes however remain univariate. Further weak convergence results for stochastic processes in smoothing parameters for nearest neighbor estimators in the random design nonparametric regression model were obtained in a series of articles by Bhattacharya and Mack (1985, 1987, 1989) and for x- and y-coordinates of peaks in a fixed design regression model by Müller (1989). The work of Bhattacharya and Mack interestingly leads to an asymptotically linear model, the parameters of which correspond to choices of smoothing parameters.

All these results essentially find their application in criteria for efficient local bandwidth choices which will be discussed in Section 3. It is clear that local bandwidth choices, i.e., choosing a different bandwidth at each point where an estimate is desired, potentially lead to smaller integrated mean squared errors (IMSE) than global bandwidth choice where the same bandwidth is employed for all points, if the respective mean squared error (MSE) optimal local bandwidths are employed. These issues are discussed in detail for the fixed design nonparametric regression model in Müller and Stadtmüller (1987), where it is shown that under certain conditions the IMSE of estimators with consistently estimated optimal local bandwidths also is smaller than the IMSE of estimators with consistently estimated optimal global bandwidths. This theoretical finding was corroborated in simulation studies. Therefore, local bandwidth choices are of great interest when curves as a whole are to be estimated, and even more so, when local functionals like function value at a point or coordinate of a zero or peak are aimed at.

In this contribution, we focus on applications of weak convergence of processes with multivariate time in curve estimation. In Section 2, a multivariate tightness criterion is reviewed which is a valuable tool for this enterprise. In Section 3 we discuss the notions of efficient bandwidth choice and adaptive curve estimation.

Then as an illustration of the general concept we develop a specific application to adaptive mode kernel estimation for probability density functions, complementing previous results by Romano (1988). This begins in Section 4 where mode estimation is defined and reviewed. In Section 5, relevant stochastic processes with two- dimensional time in a local deviation and a bandwidth coordinate are introduced and their weak convergence is demonstrated. Accordingly, a criterion for adaptive mode estimation is given in Section 6. Several bandwidth choices which lead to adaptive estimation procedures are discussed in Section 7, among them one which avoids estimation of derivatives by pilot estimators, a common method to obtain adaptive estimates.

Another class of stochastic processes with multivariate time can be defined in a multivariate bandwidth argument. This has applications in locally adaptive multivariate curve estimation whenever local bandwidths are to be chosen differently in various coordinate directions, which is reasonable in case of nonhomogeneous multivariate functions exhibiting for instance wavelike patterns. This application will be outlined in Section 8.

2. Tightness Criterion

We use tightness criterion (3) and Theorem 3 from Bickel and Wichura (1971). Consider a sequence of stochastic processes $Z_n: T \to \mathbb{R}$, where $T \subset \mathbb{R}^m$ is compact, $m \geq 1$. Let $B = \prod_{i=1}^{m} (t_{1i}, t_{2i}]$, $D = \prod_{i=1}^{m} (u_{1i}, u_{2i}]$ be *blocks* in $\mathbb{R}^m$ with $\lambda(B) > 0$, $\lambda(D) > 0$, where λ is the Lebesque measure. The *increment* of processes Z_n around a block B is defined as

$$Z_n(B) = \sum_{\mu_1=0,\ldots,\mu_m=0}^{1} (-1)^{m-\sum_i \mu_i} Z_n\big(t_{1i}+\mu_1(t_{2i}-t_{1i}),\ldots,t_{1m} + \mu_m(t_{2m}-t_{1m})\big).$$

The tightness criterion is

$$E\big(|Z_n(B)|^m |Z_n(D)|^m\big) \leq c\ \lambda(B)\lambda(D) \qquad (2.1)$$

for neighboring blocks B, D and for some constant c>0 independent of B, D. By the Cauchy-Schwarz inequality, it would suffice to have

$$E\big(|Z_n(B)|^{2m}\big) \leq c\ \lambda(B)^2 \ . \qquad (2.2)$$

In the following examples, we aim to establish (2.2). This is a direct extension of the moment tightness condition Theorem 12.3 of Billingsley (1968). In case that Z_n are in $D_q(T)$ (for the definition, see Bickel and Wichura) or in $\mathscr{C}(T)$ (i.e., continuous on T), weak convergence $Z_n \Rightarrow Z$ with a limit process Z requires that the Z_n are tight and that the finite-dimensional distributions of Z_n at any fixed number of points $t_1,\ldots,t_\ell \in T$ converge weakly to the corresponding distributions of Z.

3. Efficient local bandwidths and adaptive curve estimators

An *adaptive* curve estimator is fully data-dependent, but has the same asymptotic MSE at a fixed point as if the optimal local bandwidth at this point was used. Assume

$$n^{\alpha}\big(f_n(x,h_n)-f(x)\big) \xrightarrow{d} \mathcal{N}\big(\beta(x,r),\ V(x,r)\big) \ ,$$

where α is an appropriate constant (usually $0<\alpha<1/2$) and $r_n = rn^{-\gamma}$ for another constant $\gamma>0$ is the sequence of bandwidths. Then the locally optimal bandwidth,

$$h_n^* = r^* n^{-\gamma} \ , \ r^* = \operatorname*{argmin}_r \big(\beta(x,r)^2 + V(x,r)\big),$$

typically depends on unknowns like certain derivatives of the function to be estimated. An estimator is *adaptive*, if a data-dependent bandwidth sequence $\tilde{h}_n = \hat{r}_n n^{-\gamma}$ is used (where $\hat{r}_n$ depends on n and the

data) with the property

$$n^{\alpha}\left(f_n(x,\tilde{h}_n)-f(x)\right) \xrightarrow{d} \mathcal{N}\left(\beta(x,r^*),\ V(x,r^*)\right).$$

The bandwidth sequence $\tilde{h}_n$ is then *efficient*.

In order to find criteria for adaptive procedures, one can utilize tightness of the sequence of corresponding processes, $Z_n(h) \Rightarrow Z(h)$, in the bandwidth: If $\hat{r}_n \xrightarrow{p}$, r^* the tightness implies that $Z_n(r_n) \xrightarrow{d} Z(r^*)$. This means that any bandwidth choice satisfying $\hat{r}_n \xrightarrow{p} r^*$ is efficient, leading to a large class of possible adaptive procedures. In the following three sections, these concepts will be illustrated by means of adaptive mode estimation.

Another possibility to define adaptivity and efficiency is to base these notions on the actual MSE instead of the asymptotic MSE (the asymptotic MSE is the MSE as derived from the limiting distribution). In this sense,

$$\lim_{n\to\infty} n^{2\alpha}\mathrm{MSE}\left(f_n(x,h_n^*)\right) = \operatorname*{argmin}_h \lim_{n\to\infty} n^{2\alpha}\mathrm{MSE}\left(f_n(x,h)\right),$$

and for any efficient bandwidth sequence $\tilde{h}_n$,

$$\lim_{n\to\infty} n^{2\alpha}\ \mathrm{MSE}\left(f_n(x,\tilde{h}_n)\right) = \lim_{n\to\infty} n^{2\alpha}\ \mathrm{MSE}\left(f_n(x,h_n^*)\right).$$

We will not pursue this stronger notion of adaptivity here.

4. Mode estimators

Assume that i.i.d. observations X_i, $i=1,\dots,n$, are given; distributed according to a distribution function F which possesses a density f. The kernel estimator of f(x) (Rosenblatt, 1956; Parzen, 1962) is defined as

$$f_n(x,h_n) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right), \tag{4.1}$$

where h_n is a sequence of bandwidths and K a kernel function. We require that K is a *kernel of order k*, i.e., defining for $0\le\nu<k$, $k\ge 2$,

$$\mathcal{M}_{\nu,k} = \left\{g\in L^2([-1,1]):\ \text{support } g = [-1,1],\right.$$

$$\left.\int g(x)x^j dx = \begin{cases} 0 & 0\le j<k,\ j\ne\nu \\ (-1)^{\nu}\nu! & j=\nu \\ \ne 0 & j=k \end{cases}\right\},$$

that

$$K\in\mathcal{M}_{0,k},\ K\in\mathcal{C}^2(\mathbb{R}),\ K^{(2)} \text{ is Lipschitz continuous,} \tag{K}$$

denoting the derivatives of K by $K^{(1)}$, $K^{(2)}$.

This latter assumption could be relaxed to $K \in \mathcal{C}^1(\mathbb{R})$, $K^{(1)}$ is Lipschitz continuous, and possibly further, see Eddy (1982), but only at the expense of lengthy technical calculations in the proofs of the following section. Kernels satisfying (K) for k = 2,4,6 are, respectively,

$$K(x) = \frac{15}{16}(1-x^2)^2$$

$$K(x) = \frac{105}{64}(1-x^2)^2(1-3x^2)$$

$$K(x) = \frac{315}{2048}(1-x^2)^2(15-110x^2+143x^4), \text{ restricted to } [-1,1].$$

We further assume that f has a unique mode at θ, with the properties

$$f(\theta) > f(x) \quad \text{for all } x \neq \theta \tag{M1}$$

$$f^{(2)}(\theta) < 0, \tag{M2}$$

and

$$f^{(k+1)}(\theta) \neq 0, \; f \in \mathcal{C}^{k+1}. \tag{M3}$$

Various approaches for estimating the mode θ have been proposed by Parzen (1962), Chernoff (1964) and Grenander (1965), among others. An overview is given by Sager (1983). Given the kernel estimator (4.1), it is natural to take the mode of the estimated curve as estimator of the mode, as suggested by Parzen (1962):

$$\hat{\theta}(h_n) = \inf\{x: f_n(x,h_n) = \sup_y f_n(y,h_n)\}. \tag{4.2}$$

The measurability of this estimator was shown by Eddy (1980).

Obviously, $\hat{\theta}$ depends on the chosen bandwidth h_n. Adaptive mode estimators $\hat{\theta}(\hat{h}_n)$ were constructed by Romano (1988), who proved tightness through function indexed empirical processes. This approach works specifically for density estimation and might not be easily generalized to other curve estimation settings. In the following, we pursue an alternative method based on the weak convergence of an appropriate two-dimensional process in a bandwidth and a deviation coordinate and show how this leads to adaptive mode estimators. This serves as an illustration of the general principles outlined in the previous sections, and we do not aim to obtain the results under the weakest possible assumptions.

5. Weak convergence of deviation and bandwidth process

Define

$$V^{(1)} = \int K^{(1)}(v)^2 dv \tag{5.1}$$

and

$$\beta_k = \frac{(-1)^k}{k!} \int K(v) v^k dv. \tag{5.2}$$

We will see below that the optimal (in the sense of minimizing the asymptotic mean squared error) bandwidth sequence is

$$h_n^* = s^* n^{-1/(2k+3)}, \tag{5.3}$$

where

$$s^* = \left[\frac{3f(\theta)V^{(1)}}{2kf^{(k+1)}(\theta)^2\beta_k^2}\right]^{1/(2k+3)} . \tag{5.4}$$

Let $0<A<B<\infty$ be constants such that $A<s^*<B$. Assuming that $s\in[A, B]$, we consider bandwidths

$$h_n(s) = sn^{-1/(2k+3)}. \tag{5.5}$$

Let $\gamma_n = n^{k/(2k+3)}$ and define the kernel estimator

$$f_n(x,s) = f_n(x,h_n(s))$$

and the sequence of stochastic processes

$$Z_n(s,t) = \gamma_n^2\left(f_n(\theta+t\gamma_n^{-1}, s) - f_n(\theta,s)\right)$$

$$= n^{\frac{2k}{2k+3}} \frac{1}{nh_n} \sum_{i=1}^{n} \left[K\left(\frac{\theta+tn^{-k/(2k+3)}-X_i}{h_n(s)}\right) - K\left(\frac{\theta-X_i}{h_n(s)}\right)\right], \tag{5.6}$$

$(s,t) \in [A,B] \times [-C,C]$ for some $C > 0$.

Here, s is a "bandwidth" and t a "deviation" coordinate. The one-dimensional projection of this process on the t coordinate with s fixed was analysed by Eddy (1980, 1982). Observe that

$$Z_n(s,t) \in \mathcal{C}([A,B] \times [-C,C]). \tag{5.7}$$

<u>Theorem 5.1</u> *Under* (K), (M1)-(M3),

$$Z_n(s,t) \Rightarrow Z(s,t) \text{ on } \mathcal{C}([A,B] \times [-C,C]), \tag{5.8}$$

where Z is a Gaussian process characterized by the moment structure

$$EZ(s,t) = ts^k f^{(k+1)}(\theta)\beta_k + t^2 f^{(2)}(\theta)/2 \tag{5.9}$$

$$\operatorname{cov}(Z(s_1,t_1),\ Z(s_2,t_2)) = \frac{t_1 t_2}{(s_1 s_2)^2}\ f(\theta)\int K^{(1)}\left(\frac{v}{s_1}\right)K^{(1)}\left(\frac{v}{s_2}\right)dv, \quad (5.10)$$

and can be written as

$$Z(s,t) = ts^k f^{(k+1)}(\theta)\beta_k + t^2 f^{(2)}(\theta)/2$$

$$+ \frac{t}{s^2}\ (f(\theta))^{1/2}\int K^{(1)}\left(\frac{v}{s}\right)dW(v), \quad (5.11)$$

where W *stands for a standard Wiener process.*

The rest of this section is devoted to the proof which requires several auxiliary lemmas. As a first step, we analyze the moment structure of processes Z_n.

<u>Lemma 5.1</u>

$$EZ(s,t) = ts^k f^{(k+1)}(\theta)\beta_k + t^2 f^{(2)}(\theta)/2 + o(1), \quad (5.12)$$

where the o-term is uniform in s,t.

<u>Proof</u>. By a Taylor expansion,

$$E[Z_n(s,t)] = \frac{\gamma_n^2}{h_n(s)}\int\left[K\left(\frac{\theta+t\gamma_n^{-1}-x}{h_n(s)}\right) - K\left(\frac{\theta-x}{h_n(s)}\right)\right]f(x)dx$$

$$= \gamma_n^2\int\left[f\left(\theta+t\gamma_n^{-1}-yh_n(s)\right) - f\left(\theta-yh_n(s)\right)\right]K(y)dy$$

$$= \gamma_n^2\int f^{(1)}\left(\theta-yh_n(s)\right)\left(t\gamma_n^{-1}\right)K(y)dy$$

$$+ \gamma_n^2\int \frac{1}{2}\ f^{(2)}\left(\theta-yh_n(s)\right)\left(t\gamma_n^{-1}\right)^2K(y)dy$$

$$+ \gamma_n^2\ O\left(\gamma_n^{-3}\right)$$

$$= \mathrm{I}+\mathrm{II}+\mathrm{III}.$$

Now III = o(1), uniformly in s and t. Further,

$$\mathrm{II} = t^2 f^{(2)}(\theta)/2 + O(h_n(s))$$

$$= t^2 f^{(2)}(\theta)/2 + o(1), \text{ uniformly in } s.$$

By a second Taylor expansion, applying the moment conditions for the kernel function K, and observing $f^{(1)}(\theta) = 0$,

$$I = t\gamma_n \int \left[\sum_{i=1}^{k} \frac{f^{(i)}(\theta)(-yh_n(s))^{i-1}}{(i-1)!} + \frac{f^{(k+1)}(\xi)(-yh_n(s))^k}{k!} \right] K(y)dy$$

$$= t\gamma_n \frac{(-1)^k (h_n(s))^k}{k!} \int f^{(k+1)}(\xi) y^k K(y)dy$$

$$= t \frac{(-1)^k s^k}{k!} f^{(k+1)}(\theta) \int K(y) y^k dy + O(h_n(s))$$

$$= ts^k f^{(k+1)}(\theta)\beta_k + o(1), \text{ uniformly in } s,$$

where ξ is a suitable mean value. Combining these results yields (5.12).

<u>Lemma 5.2</u>

$$\operatorname{cov}\left(Z_n(s_1,t_1),\ Z_n(s_2,t_2)\right)$$

$$= \frac{t_1 t_2}{s_1^2 s_2^2} f(\theta) \int K^{(1)}\left(\frac{v}{s_1}\right) K^{(1)}\left(\frac{v}{s_2}\right) dv \left(1+o(1)\right). \tag{5.13}$$

<u>Proof</u>. Let

$$\phi_n(X_i,s,t) = \frac{1}{nh_n(s)} \left\{ K\left(\frac{\theta + t\gamma_n^{-1} - X_i}{h_n(s)}\right) - K\left(\frac{\theta - X_i}{h_n(s)}\right) \right.$$

$$\left. - E\left[K\left(\frac{\theta + t\gamma_n^{-1} - X_i}{h_n(s)}\right) - K\left(\frac{\theta - X_i}{h_n(s)}\right) \right] \right\}. \tag{5.14}$$

Then obviously $E\phi_n(X_i,s,t) = 0$ and

$$Z_n(s,t) - EZ_n(s,t) = \gamma_n^2 \sum_{i=1}^{n} \phi_n(X_i,s,t). \tag{5.15}$$

Observe

$$\operatorname{cov}(Z(s_1,t_1),\ Z(s_2,t_2)) = \gamma_n^4 \sum_{i=1}^{n} E\left[\phi_n(X_i,s_1,t_1)\phi_n(X_i,s_2,t_2)\right]$$

$$= \gamma_n^4 \frac{1}{n^2 h_n(s_1) h_n(s_2)} \sum_{i=1}^{n} \left\{ E\left[K\left(\frac{\theta + t_1\gamma_n^{-1} - X_i}{h_n(s_1)}\right) - K\left(\frac{\theta - X_i}{h_n(s_1)}\right) \right] \right.$$

$$\left[K\left(\frac{\theta+t_2\gamma_n^{-1}-X_i}{h_n(s_2)}\right) - K\left(\frac{\theta-X_i}{h_n(s_2)}\right)\right]$$

$$- E\left[K\left(\frac{\theta+t_1\gamma_n^{-1}-X_i}{h_n(s_1)}\right) - K\left(\frac{\theta-X_i}{h_n(s_1)}\right)\right]E\left[K\left(\frac{\theta+t_2\gamma_n^{-1}-X_i}{h_n(s_2)}\right) - K\left(\frac{\theta-X_i}{h_n(s_2)}\right)\right]\Bigg\}.$$

The second summand is seen to be

$$n\left(n^{-2}\ EZ_n(s_1,t_1)EZ_n(s_2,t_2)\right) = o(1), \tag{5.16}$$

uniformly according to Lemma 5.1.

For the first summand we obtain

$$\frac{n^{(2k-1)/(2k+3)}}{s_1s_2}\ E\left(\left[K\left(\frac{\theta+t_1\gamma_n^{-1}-X_1}{h_n(s_1)}\right)-K\left(\frac{\theta-X_1}{h_n(s_1)}\right)\right]\left[K\left(\frac{\theta+t_2\gamma_n^{-1}-X_1}{h_n(s_2)}\right)-K\left(\frac{\theta-X_1}{h_n(s_2)}\right)\right]\right)$$

$$= \frac{n^{(2k-2)/(2k+3)}}{s_1s_2}\int\left[K\left(\frac{t_1\gamma_n^{-1}}{h_n(s_1)}+\frac{v}{s_1}\right) - K\left(\frac{v}{s_1}\right)\right]\left[K\left(\frac{t_2\gamma_n^{-1}}{h_n(s_2)}+\frac{v}{s_2}\right) - K\left(\frac{v}{s_2}\right)\right]$$

$$f\left(\theta-vn^{-1/(2k+3)}\right)dv$$

$$= \frac{n^{2(k-1)/(2k+3)}}{s_1s_2}\int\left[K^{(1)}\left(\frac{v}{s_1}\right)\left(\frac{t_1\gamma_n^{-1}}{h_n(s_1)}\right) + O\left(\left(\frac{t_1\gamma_n^{-1}}{h_n(s_1)}\right)^2\right)\right]$$

$$\left[K^{(1)}\left(\frac{v}{s_2}\right)\left(\frac{t_2\gamma_n^{-1}}{h_n(s_2)}\right) + O\left(\left(\frac{t_2\gamma_n^{-1}}{h_n(s_2)}\right)^2\right)\right]f\left(\theta-vn^{-1/(2k+3)}\right)dv$$

$$= \frac{t_1t_2}{s_1^2s_2^2}\ f(\theta)\int K^{(1)}\left(\frac{v}{s_1}\right)K^{(1)}\left(\frac{v}{s_2}\right)dv\left(1+o(1)\right)$$

$$+ \frac{n^{2(k-1)/(2k+3)}}{s_1s_2}\ O\left(\left[\left(\frac{t_1\gamma_n^{-1}}{h_n(s_1)}\right)^2 + \left(\frac{t_2\gamma_n^{-1}}{h_n(s_2)}\right)^2\right]\left[\frac{t_1\gamma_n^{-1}}{h_n(s_1)} + \frac{t_2\gamma_n^{-1}}{h_n(s_2)}\right]\right)$$

$$= \frac{t_1t_2}{s_1^2s_2^2}\ f(\theta)\int K^{(1)}\left(\frac{v}{s_1}\right)K^{(1)}\left(\frac{v}{s_2}\right)dv + o(1),$$

where we have used the compactness of the support of K and observe that the o-term is uniform in s and t.

Next we consider the joint asymptotic normality of centered processes

$$\bar{Z}_n(s,t) = Z_n(s,t) - EZ_n(s,t) \tag{5.17}$$

at a fixed number $(s_1,t_1),\ldots,(s_\ell,t_\ell)$ of points in $[A,B] \times [-C,C]$.

<u>Lemma 5.3</u> *Let* $(t_1,s_1),\ldots,(t_\ell,s_\ell) \in [A,B] \times [-C,C]$. *Then*

$$\left(\bar{Z}_n(t_1,s_1),\ldots,\bar{Z}_n(t_\ell,s_\ell)\right)^T \xrightarrow{d} N(0,\Sigma), \tag{5.18}$$

where $\Sigma = (\sigma_{ij})_{1\le i,j\le \ell}$,

$$\sigma_{ij} = \frac{t_i t_j}{(s_i s_j)^2} f(\theta)\int K^{(1)}\left(\frac{v}{s_i}\right)K^{(1)}\left(\frac{v}{s_j}\right)dv.$$

<u>Proof</u>. Consider first the case $\ell=1$. We need to show asymptotic normality of

$$\bar{Z}_n(s,t) = \gamma_n^2 \sum_{i=1}^{n} \phi_n(X_i,s,t),$$

where ϕ_n is defined in (5.14). According to (5.13),

$$\operatorname{var}\left(\bar{Z}_n(s,t)\right) = \frac{t^2}{s^3} f(\theta)\int K^{(1)}(v)^2 dv + o(1),$$

$$= \sigma^2(s,t) + o(1), \text{ say.}$$

Therefore the Lindeberg condition for $\bar{Z}_n(s,t)$ is equivalent to

$$\sum_{i=1}^{n} E\left[\gamma_n^4 \phi_n^2\left(X_i,s,t\right) 1_{\{\gamma_n^2\phi_n(X_i,s,t) > \delta\sigma(s,t)\}}\right] \to 0 \tag{5.19}$$

for all $\delta>0$.

Observe that owing to the Lipschitz continuity of K,

$$|\gamma_n^2 \phi_n^2\left(X_i,s,t\right)| = O\left(n^{-(k+1)/(2k+3)}\right).$$

This implies that for sufficiently large n,

$$1_{\{\gamma_n^2\phi_n(X_i,s,t) > \delta\sigma(s,t)\}} = 0$$

and therefore, that (5.19) holds.

For $\ell\ge 2$, according to the Cramér-Wold device, one shows that any linear combination of $\bar{Z}_n(s_i,t_i)$, $1\le i\le \ell$, is asymptotically normal.

The proof remains essentially unchanged. The covariance matrix is obtained as in Lemma 5.2.

Since processes Z_n are in $\mathscr{C}([A,B] \times [-C,C])$, according to the criterion for weak convergence in Section 2 it remains to show tightness via (2.2), where m=2. We prove this in two steps, first showing tightness for centered processes $\bar{Z}_n$ in the deviation coordinate.

Lemma 5.4

$$E|\bar{Z}_n(s,t_2) - \bar{Z}_n(s,t_1)|^4 \le c|t_2-t_1|^4 \tag{5.20}$$

for some constant c>0.

Proof. Let

$$I_1(X,s,t) = 1_{\{|\theta-X| \le h_n(s)\}\cup\{|\theta+t\gamma_n^{-1}-X| \le h_n(s)\}} \tag{5.21}$$

and

$$I_2(X,s) = 1_{\{|\theta-X| \le 2h_n(s)\}} \cdot \tag{5.22}$$

Observe that for sufficiently large n, $I_1(X,s,t) \le I_2(X,s)$, uniformly in t. Observe

$$\begin{aligned} &E|\bar{Z}_n(s,t_2) - \bar{Z}_n(s,t_1)|^4 \\ &\le \sum_{i=1}^{n} E|\phi_n(X_i,s,t_2) - \phi_n(X_i,s,t_1)|^4 \\ &+ \sum_{i\ne j}^{n} E|\phi_n(X_i,s,t_2) - \phi_n(X_i,s,t_1)|^2 E|\phi_n(X_j,s,t_2) - \phi_n(X_j,s,t_1)|^2 \\ &= I + II. \end{aligned} \tag{5.23}$$

Now

$$\begin{aligned} &E|\phi_n(X_1,s,t_2) - \phi_n(X_1,s,t_1)|^4 \\ &= O\left(n^{-8/(2k+3)}\right) E\left|\left[K\left(\frac{\theta+t_2\gamma_n^{-1}-X_1}{h_n(s)}\right) - K\left(\frac{\theta-X_1}{h_n(s)}\right)\right] I_1(X_1,s,t_2)\right. \\ &\left. - \left[K\left(\frac{\theta+t_1\gamma_n^{-1}-X_1}{h_n(s)}\right) - K\left(\frac{\theta-X_1}{h_n(s)}\right)\right] I_1(X_1,s,t_1)\right|^4 \\ &= O\left(n^{-8/(2k+3)}\right) E\left\{\left|K\left(\frac{\theta+t_2\gamma_n^{-1}-X_1}{h_n(s)}\right) - K\left(\frac{\theta+t_1\gamma_n^{-1}-X_1}{h_n(s)}\right)\right|^4 I_2(X_1,s)\right\} \end{aligned}$$

$$= O\left(n^{-8/(2k+3)} n^{-4(k-1)/(2k+3)}\right) |t_2 - t_1|^4 E I_2(X_1, s),$$

$$= O\left(n^{-(4k+5)/(2k+3)}\right) |t_2 - t_1|^4$$

since

$$E\, I_2(X_1, s) = \int_{|\theta - x| \le 2h_n(s)} f(x)dx = O\left(n^{-1/(2k+3)}\right). \tag{5.24}$$

Therefore,

$$I = O\left(n^{-(2k+2)/(2k+3)}\right) |t_2 - t_1|^4.$$

Further,

$$E|\phi_n(X_1, s, t_2) - \phi_n(X_1, s, t_1)|^2$$

$$= O\left(n^{-4/(2k+3)}\right) E\left| \left[K\left(\frac{\theta + t_2 \gamma_n^{-1} - X_1}{h_n(s)}\right) - K\left(\frac{\theta - X_1}{h_n(s)}\right)\right] I_1(X_1, s, t_2)\right.$$

$$\left. - \left[K\left(\frac{\theta + t_1 \gamma_n^{-1} - X_1}{h_n(s)}\right) - K\left(\frac{\theta - X_1}{h_n(s)}\right)\right] I_1(X_1, s, t_1)\right|^2$$

$$= O(n^{-1}) |t_2 - t_1|^2 .$$

Therefore,

$$II = \sum_{i \ne j} O(n^{-2}) |t_2 - t_1|^4 = O(1) |t_2 - t_1|^4 .$$

The result follows.

Tightness for processes $\bar{Z}_n$ in the bandwidth coordinate is implied by

<u>Lemma 5.5</u>

$$E|\bar{Z}_n(s_2, t) - \bar{Z}_n(s_1, t)|^4 \le c|s_2 - s_1|^4 \tag{5.25}$$

for some constant $c > 0$.

<u>Proof</u>. As in (5.23), we have

$$E|\bar{Z}_n(s_2, t) - \bar{Z}_n(s_1, t)|^4$$

$$\leq \sum_{i=1}^{n} E|\phi_n(X_i,s_2,t) - \phi_n(X_i,s_1,t)|^4$$

$$+ \sum_{i\neq j}^{n} E|\phi_n(X_i,s_2,t) - \phi_n(X_i,s_1,t)|^2 E|\phi_n(X_j,s_2,t) - \phi_n(X_j,s_1,t)|^2$$

$$= I + II.$$

Defining

$$H_n(X,s,t) = \frac{1}{s}\left[K\left(\frac{\theta+t\gamma_n^{-1}-X}{h_n(s)}\right) - K\left(\frac{\theta-X}{h_n(s)}\right)\right],$$

we observe that

$$|H_n(X,s,t)| = O\left(n^{-(2k+2)/(2k+3)}\right) \tag{5.26}$$

uniformly in s,t, due to the Lipschitz continuity of K. Further, for n large enough,

$$\left|\frac{d}{ds} H_n(X,s,t)\right| = \left| - \frac{1}{s^2} H_n(X,s,t) + \left\{\frac{d}{ds}\left(\frac{\theta+t\gamma_n^{-1}-X}{h_n(s)}\right)\right\}\right.$$

$$\left. I_1(X,s,t)K^{(1)}\left(\frac{\theta+t\gamma_n^{-1}-X}{h_n(s)}\right) - \left\{\frac{d}{ds}\left(\frac{\theta-X}{h_n(s)}\right)\right\} I_1(X,s,0)K^{(1)}\left(\frac{\theta-X}{h_n(s)}\right)\right|$$

$$\leq O\left(n^{-(k-1)/(2k+3)}\right)\left|K^{(1)}\left(\frac{\theta+t\gamma_n^{-1}-X}{h_n(s)}\right)\right|\left|\frac{d}{ds}\left(\frac{\theta+t\gamma_n^{-1}-X}{h_n(s)}\right) - \frac{d}{ds}\left(\frac{\theta-X}{h_n(s)}\right)\right| I_2(X,s)$$

$$+ \left|\frac{d}{ds}\left(\frac{\theta-X}{h_n(s)}\right)\right|\left|K^{(1)}\left(\frac{\theta+t\gamma_n^{-1}-X}{h_n(s)}\right) - K^{(1)}\left(\frac{\theta-X}{h_n(s)}\right)\right| I_2(X,s)$$

$$= O\left(n^{-(k-1)/(2k+3)}\right) + O(1)\ O\left(\gamma_n^{-1} n^{1/(2k+3)}\right)$$

$$= O\left(n^{-(k-1)/(2k+3)}\right), \text{ uniformly in } s,t, \tag{5.27}$$

owing to the Lipschitz continuity of $K^{(1)}$ and to

$$\left\{\frac{d}{ds}\left(\frac{\theta-X}{h_n(s)}\right)\right\} I_2(X,s) = O(1).$$

As a consequence,

$$I = O\left(n^{-8/(2k+3)}\right)E\left(H_n(X_1,s_2,t) - H_n(X_1,s_1,t)\right|^4 \sup_s I_2(X_1,s)$$

$$= O\left(n^{-8/(2k+3)}(s_2-s_1)^4\right)\left|\left(\frac{d}{ds} H_n(X_1,s,t)\right|_{s=\xi_1}\right)\right| E\sup_s I_2(X_1,s)$$

$$= O\left(n^{-2(k+2)/(2k+3)}\right)(s_2-s_1)^4 \tag{5.28}$$

according to (5.27), observing that (5.24) remains valid for $E\sup_s I_2(X_1,s)$. Similarly,

$$E|\phi_n(X_i,s_2,t) - \phi_n(X_i,s_1,t)|^2$$

$$= O\left(n^{-4/(2k+3)}\right)(s_2-s_1)^2\left|\left(\frac{d}{ds} H_n(X_1,s,t)\right|_{s=\xi_2}\right)\right|^2 E\sup_s I_2(X_1,s)$$

$$= O(n^{-1})(s_2-s_1)^2 ,$$

and therefore

$$II = O(1)(s_2-s_1)^4, \tag{5.29}$$

where the O-term is uniform in s,t. The result follows from (5.28), (5.29).

<u>Proof of Theorem 5.1</u>. According to Section 2, we need to establish weak convergence of the finite-dimensional distributions and tightness of processes Z_n. The former follows from Lemmas 5.1-5.3, which also imply that the moment structure of any limit process would be given by (5.9), (5.10). For the latter, we need to show (2.2) for m=2. Consider the block $B = [s_1,s_2] \times [t_1,t_2]$. Then

$$Z_n(B) = Z_n(s_2,t_2) - Z_n(s_1,t_2) - Z_n(s_2,t_1) + Z_n(s_1,t_1).$$

We have to show

$$E\left(Z_n(B)\right)^4 \le c|s_2-s_1|^2|t_2-t_1|^2 \tag{5.30}$$

which implies (2.2). It follows from Lemmas 5.4 and 5.5 that

$$E\left(Z_n(B)\right)^4$$

$$\le c_1 \min\left\{\left(E\left(Z_n(s_2,t_2) - Z_n(s_1,t_2)\right)^4 + E\left(Z_n(s_1,t_1) - Z_n(s_2,t_1)\right)^4\right),\right.$$

$$\left.\left(E\left(Z_n(s_2,t_2) - Z_n(s_2,t_1)\right)^4 + E\left(Z_n(s_1,t_1) - Z_n(s_1,t_2)\right)^4\right)\right\}$$

$$\le c_2 \min\left(|t_2-t_1|^4,\ |s_2-s_1|^4\right)$$

$$\leq c_2|s_2-s_1|^2|t_2-t_1|^2,$$

which implies (5.30).

6. Adaptive mode estimation

Consider the mapping μ defined on $\mathscr{C}\big([A,B] \times [-C,C]\big)$ by

$$\mu(\psi)(s) = \inf\{t(s): \psi(s,t) = \sup_{r\in[-C,C]} \psi(s,r)\} \tag{6.1}$$

for $\psi\in\mathscr{C}([A,B]\times[-C,C])$. The mapping μ is measurable and its discontinuities have measure zero under the measure induced by Z, where Z is the process (5.11); we find

$$\mu(Z)(s) = \frac{-s^k f^{(k+1)}(\theta)\beta_k}{f^{(2)}(\theta)} - \frac{(f(\theta))^{1/2}}{s^2 f^{(2)}(\theta)} \int K^{(1)}\left(\frac{v}{s}\right) dW(v). \tag{6.2}$$

By the functional mapping theorem,

$$\mu(Z_n) \Rightarrow \mu(Z). \tag{6.3}$$

In terms of the bandwidth sequences $h_n(s) = sn^{-1/(2k+3)}$ considered here we rewrite the mode estimator $\hat{\theta}(h_n)$ (4.2) as follows:

$$\hat{\theta}(s) = \inf\{x: f_n(x, sn^{-1/(2k+3)}) = \sup_y f_n(y, sn^{-1/(2k+3)})\}. \tag{6.4}$$

Obviously, Z_n is maximized whenever

$$\theta + \hat{t}\gamma_n^{-1} = \hat{\theta}(s),$$

so that

$$\mu(Z_n)(s) = n^{k/(2k+3)}\left(\hat{\theta}(s)-\theta\right) \tag{6.5}$$

according to (6.1). This together with (6.3) and Theorem 5.1 implies (see also Romano, 1988):

<u>Corollary 6.1</u>

$$n^{k/(2k+3)}\left(\hat{\theta}(s)-\theta\right) \xrightarrow{d} \mathcal{N}\left(\frac{-s^k f^{(k+1)}(\theta)\beta_k}{f^{(2)}(\theta)}, \frac{f(\theta)}{s^3 f^{(2)}(\theta)^2} \int K^{(1)}(v)^2 dv\right). \tag{6.6}$$

Employing consistent estimators for f, $f^{(2)}$, $f^{(k+1)}$, asymptotic confidence intervals for θ can be constructed based on (6.6). An alternative are bootstrap confidence intervals (Romano, 1988). From (6.6), the asymptotic MSE of $\hat{\theta}(s)$ can be derived and minimized w.r. to s. The result is s* (5.4).

<u>Corollary 6.2</u> *Any global bandwidth choice* $h_n(\hat{s}) = \hat{s}n^{-1/(2k+3)}$ *satisfying*

$$\hat{s} \xrightarrow{p} s^* \tag{6.7}$$

is efficient, and $\hat{\theta}(\hat{s})$ (6.4) *is then adaptive.*

<u>Proof</u>. We need to show

$$Z_n(\hat{s},t) \Rightarrow Z(s^*,t) \tag{6.8}$$

which implies

$$\mu\big(Z_n(\hat{s},t)\big) \xrightarrow{d} \mu\big(Z(s^*,t)\big)$$

and

$$\lim_{n\to\infty} \mathcal{L}\big(n^{k/(2k+3)}\big(\hat{\theta}(\hat{s})-\theta\big)\big) = \lim_{n\to\infty} \mathcal{L}\big(n^{k/(2k+3)}\big(\hat{\theta}(s^*)-\theta\big)\big),$$

$\mathcal{L}(X)$ denoting the distribution of X.

Now, for any $\delta,\varepsilon > 0$,

$$P\big(\sup_t |Z_n(\hat{s},t)-Z_n(s^*,t)|>\varepsilon\big)$$

$$\le P\Big(\sup_{|s_1-s_2|\le\delta,t} |Z_n(s_1,t)-Z_n(s_2,t)|>\varepsilon\Big) + P(|\hat{s}-s^*|>\delta),$$

and therefore by tightness and the assumption,

$$\sup_t |Z_n(\hat{s},t)-Z_n(s^*,t)| = o_p(1),$$

which implies (6.8).

Estimating the unknowns in s* by preliminary estimators $\hat{f},\hat{f}^{(k+1)},\hat{\theta}$ for $f,f^{(k+1)}$ and θ yields (6.7) as long as $\hat{f},\hat{f}^{(k+1)}$ are uniformly consistent; these assumptions can be achieved, where uniform consistency can be obtained from Silverman (1978); one needs to establish in particular

$$E\hat{f}^{(k+1)}(x) = f^{(k+1)}(x) + o(1)$$

under $f\in\mathcal{C}^{k+1}$, where the remainder is uniform in x.

A promising alternative which will be investigated in more detail in the following section is the choice of local bandwidths

$$h_n(x) = s(x)n^{-1/(2k+3)}$$

for mode estimation. We then define the mode estimator based on estimates $\hat{s}(x)$ by

$$\hat{\theta} = \inf\left\{x:\ f_n\left(x,\hat{s}(x)n^{-1/(2k+3)}\right) = \sup_y f_n\left(y,\hat{s}(y)n^{-1/(2k+3)}\right)\right\}. \quad (6.9)$$

In terms of processes Z_n, such a local choice corresponds to

$$\hat{s}(t) = \hat{s}\left(\theta+tn^{-k/(2k+3)}\right),\ t\in[-C,C].$$

<u>Corollary 6.3</u> *If*

$$\sup_t|\hat{s}(t) - s^*| = o_p(1), \quad (6.10)$$

then $\hat{\theta}$ (6.9) *is adaptive.*

<u>Proof</u>. Again we need to show (6.8), replacing $\hat{s}$ by $\hat{s}(t)$. Following the proof of Corollary 6.2, we find that

$$P\left(\sup_t|Z_n(\hat{s}(t),t)-Z_n(s^*,t)|>\varepsilon\right)$$

$$\le P\left(\sup_{|s_1-s_2|\le\delta,t}|Z_n(s_1,t)-Z_n(s_2,t)|>\varepsilon\right)$$

$$+ P(\sup_t|\hat{s}(t)-s^*|>\delta),$$

implies the result.

Next we explore practical methods of efficient local bandwidth choice based on Corollary 6.3.

7. Efficient bandwidth choices for mode estimation

According to Corollary 6.1, any bandwidth sequence with

$$\hat{s}(t) \xrightarrow{p} s^* \text{ uniformly in } t \quad (7.1)$$

is asymptotically efficient. Considering kernel estimates

$$f_n^{(\nu)}(x,h_{n\nu}) = \frac{1}{nh_{n\nu}^{\nu+1}} \sum_{i=1}^{n} K_\nu\left(\frac{x-X_i}{h_{n\nu}}\right) \quad (7.2)$$

for the ν-th derivative $f^{(\nu)}(x)$, where $K_\nu\in\mathcal{M}_{\nu,k+1}$, we find for $\nu=1$ the MSE-optimal local bandwidth to be

$$h_n^*(x) = \left[\frac{3f(x)V^{(1)}}{2kf^{(k+1)}(x)^2\beta_k^2}\right]^{1/(2k+3)} n^{-1/(2k+3)} \tag{7.3}$$

$$= s^*(x)n^{-1/(2k+3)},$$

so that

$$h_n^*(\theta) = s^*(\theta)n^{-1/(2k+3)} = s^*n^{-1/(2k+3)}.$$

Therefore, given estimates $\hat{s}(x)$, for $s^*(x)$, it is a natural approach to define

$$\hat{\theta} = \inf\{x\colon f_n^{(1)}(x,\hat{s}(x)n^{-1/(2k+3)}) = 0\}. \tag{7.4}$$

Here, $\hat{s}(x)$ corresponds to local bandwidth choice for the estimate of the first derivative, and

$$\hat{s}(x) \xrightarrow{p} s^*(x) \tag{7.5}$$

to choosing consistent estimates for the optimal local bandwidths for the first derivative. According to Corollary 6.3, mode estimators (7.4), which equivalently could be defined as

$$\hat{\theta} = \inf\left\{x\colon f_n\left(x,\hat{s}(x)n^{-1/(2k+3)}\right) = \sup_y f_n\left(y,\hat{s}(y)n^{-1/(2k+3)}\right)\right\},$$

are adaptive, whenever (7.1) is satisfied.

We consider now two approaches to achieve this. A first and obvious approach consists in direct estimation of the unknowns appearing in (7.3),

$$\hat{s}(x) = \left[\frac{3\hat{f}(x)V^{(1)}}{2k\hat{f}^{(k+1)}(x)^2\beta_k^2}\right]^{1/(2k+3)}. \tag{7.6}$$

The following result summarizes this discussion.

<u>Corollary 7.1</u> *Any estimators* $\hat{f}$, $\hat{f}^{(k+1)}$ *for* f, $f^{(k+1)}$, *that satisfy*

$$\hat{f}(\theta+t\gamma_n^{-1}) \xrightarrow{p} f(\theta)$$

$$\hat{f}^{(k+1)}(\theta+t\gamma_n^{-1}) \xrightarrow{p} f^{(k+1)}(\theta)$$

uniformly in t *render the mode estimator* $\hat{\theta}$ (7.4) *employing data-driven bandwidths* (7.6) *adaptive.*

Specific examples can be easily found. Take estimators (7.2) with h_{no}, h_{nk+1} such that

$$h_{no} n^{k/(2k+3)} \to \infty, \quad h_{nk+1}^{k+2} \; n^{k/(2k+3)} \to \infty$$

and kernels K_0, K_{k+1} are Lipschitz continuous. Then

$$\hat{f}^{(k+1)}\left(\theta+t\gamma_n^{-1}\right) - \hat{f}^{(k+1)}(\theta) \xrightarrow{p} 0$$

and

$$\hat{f}\left(\theta+t\gamma_n^{-1}\right) - \hat{f}(\theta) \xrightarrow{p} 0,$$

uniformly in t, so that the assumptions of Corollary 7.1 are satisfied.

A disadvantage of the approach described above is the necessity to estimate $f^{(k+1)}$, which for the common value k=2 amounts to estimating the third derivative of f. This might introduce additional variability in small sample applications. An alternative which avoids the estimation of higher derivatives can be based on a scheme which was discussed in Müller and Wang (1990) for local bandwidth choice in density and hazard rate estimation. Efficient local bandwidths for the first derivative would be obtained as follows.

Let $\hat{f}(\cdot,h_{no})$, $\hat{f}^{(1)}(\cdot,h_{n1})$ be estimates of type (7.2) for f, $f^{(1)}$ with

$$\sup_x |\hat{f}(x)-f(x)| \xrightarrow{p} 0 \quad \text{and} \quad \sup_x |\hat{f}^{(1)}(x)-f^{(1)}(x)| \xrightarrow{p} 0;$$

sufficient conditions on kernels and bandwidths can be found in Silverman (1978). We then define the "resmoothing" or, as it has been termed in the context of nonparametric regression in Müller (1985), see also Staniswalis (1989), "direct pilot estimator" estimate (in contrast to the "indirect pilot estimator" estimate described above) of the MSE of $f_n^{(1)}$ to be

$$\hat{MSE}\left(f_n^{(1)}(x,h)\right) = \left(\frac{1}{h}\int K_1(y)\hat{f}(x-yb,h_{no})dy-\hat{f}^{(1)}(x,h_{n1})\right)^2$$

$$+ \frac{1}{nh^3}\int K_1(y)^2\hat{f}(x-hy,h_{no})dy, \qquad (7.7)$$

where the integration can be carried out numerically (e.g., by the fast Fourier transform algorithm). The first term corresponds to the bias squared, the second term to the variance. The corresponding bandwidth estimator is

$$\hat{s}(x) = n^{1/(2k+3)} \operatorname*{argmin}_{h} \hat{MSE}\left(f_n^{(1)}(x,h)\right). \tag{7.8}$$

Analogous to the proofs in Müller and Wang (1990) one shows $\hat{s}(x) \xrightarrow{p} s^*(x)$. For (7.1) we need to show

$$\sup_{h_1 \le h \le h_2} \left| \frac{\hat{MSE}\left(\hat{f}_n^{(1)}(\theta+t\gamma_n^{-1},h)\right) - \hat{MSE}\left(\hat{f}_n^{(1)}(\theta,h)\right)}{\hat{MSE}\left(\hat{f}_n^{(1)}(\theta,h)\right)} \right| = o_p(1), \tag{7.9}$$

where h_1, h_2 are suitable bandwidth sequences such that $h_1 < h_n^* < h_2$. According to the proofs in Müller and Wang, we can replace $\hat{MSE}$ in the denominator by MSE; and a careful analysis shows that (7.9) indeed holds.

Thus we have found the following second procedure for adaptive mode estimation based on consistently estimated optimal local bandwidths for the first derivative:

(1) Find pilot estimators $\hat{f}$, $\hat{f}^{(1)}$ (with slightly oversmoothing bandwidths).

(2) Calculate local bandwidths (7.8).

(3) Construct $f_n^{(1)}(x,\hat{s}(x))$ and locate the zero, which gives $\hat{\theta}$.

Local bandwidth choice was found quite promising for estimating the function itself in the context of fixed design nonparametric regression, according to simulation results in Müller and Stadtmüller (1987), and it can be expected to lead to improvements when estimating local functionals like modes as compared to global bandwidth choice.

One open remaining question is the sensitivity of the two presented approaches to the choice of the pilot bandwidths for $\hat{f}$, $\hat{f}^{(1)}$ respectively, $\hat{f}$, $\hat{f}^{(k+1)}$. In a different context very good performance was obtained with an iterative method, where one starts with slightly oversmoothing global bandwidths for the pilot estimators, obtains the corresponding local bandwidths and then uses the corresponding estimators to obtain new local bandwidths. Such an iterative procedure obviously is efficient; preliminary results suggest that 3-5 iterations produce good final curve estimators.

8. Adaptive multivariate smoothing

As a second illustration we sketch briefly the efficient choice of local bandwidths in fixed design multivariate nonparametric regression, allowing for different bandwidths in the various coordinate directions. This problem was investigated in Müller and Prewitt (1990) and we review the results.

Consider the fixed design regression model (1.6),

$$y_i = g(\bar{x}_i)+\varepsilon_i \quad i=1,\dots,n,$$

where the regression function is $g: A \to \mathbb{R}$, $A \subset \mathbb{R}^m$ for some $m \geq 1$. Assume i.i.d. errors ε_i with $E\varepsilon_i = 0$, $E\varepsilon_i^2 = \sigma^2 < \infty$, $E\varepsilon_i^{2m} = c < \infty$. Using multiindex notation $\alpha = (\alpha_1,\dots,\alpha_m)$, $x = (x_1,\dots,x_m)$ and assuming $A = \bigcup_{i=1}^{n} A_i$ such that $\bar{x}_i \in A_i$, consider the kernel estimator

$$\hat{g}(x) = \sum_{i=1}^{n} h^{-1} y_i \int_{A_i} K\left(\frac{x-y}{h}\right) dy, \tag{8.1}$$

where for the bandwidth vector $h = (h_1,\dots,h_m)$, $h^{-1} = \prod_{j=1}^{m} h_j^{-1}$, $(x-y)/h$ is to be interpreted coordinatewise, and $K:[-1,1]^m \to \mathbb{R}$ is a kernel function with

$$K(v) = \prod_{j=1}^{m} K_j(v_j), \text{ where } K_j \in \mathcal{M}_{0,k}.$$

Assuming further asymptotic equidistance, i.e.,

$$\max_{1\leq i\leq n} |\lambda(A_i) - \lambda(A)n^{-1}| = o(n^{-1})$$

and

$$\max_{1\leq i\leq n} \sup_{x,y\in A_i} \|x-y\| = O\left(n^{-1/m}\right),$$

where $\lambda(\cdot)$ denotes Lebesgue measure and $\|\cdot\|$ Euclidean norm in $\mathbb{R}^m$, and let $x \in \mathring{A}$ and $mk < 2k+m$.

Then we set

$$h_i = s_i n^{-1/(2k+m)}, \; 0 < a < s_i < b < \infty, \; 1 \leq i \leq n, \tag{8.2}$$

where constants a,b are chosen such that

$$a < s_j^* < b, \quad j=1,\dots,m$$

whenever

$$s_\ell^* = \left|\frac{1}{\eta_\ell(x)}\right|^{1/k}\left[\frac{\omega \prod_{j=1}^{m} |\eta_j(x)|^{1/k}}{2km}\right]^{1/(2k+m)} \tag{8.3}$$

exists, where

$$\eta_j(x) = \frac{1}{k!}\int K_j(v)v^k dv \frac{\partial^k}{\partial x_j^k} g(x) \quad \text{and} \quad \omega = \sigma^2\lambda(A) \prod_{j=1}^{m} K_j(v)^2 dv.$$

The estimator (8.1) with this special bandwidth choice is denoted by $\hat{g}(x,s)$.

The weak convergence result proved in Müller and Prewitt (1990) is for the multivariate parameter processes

$$Z_n(s) = n^{k/(2k+m)}\left(\hat{g}(x,s)-g(x)\right), \quad s\in[a,b]^m . \tag{8.4}$$

It then holds:

$$Z_n(s) \Rightarrow Z(s) \text{ on } \mathcal{C}([a,b]^m),$$

where Z is a multiparameter Gaussian process with

$$EZ(s) = \sum_{j=1}^{m} \eta_j(x)$$

$$\operatorname{cov}\left(Z(s),Z(t)\right) = \frac{\sigma^2\lambda(A)}{\prod_{j=1}^{m} s_j t_j} \prod_{j=1}^{m} \int K_j\left(\frac{v}{t_j}\right)K_j\left(\frac{v}{s_j}\right)dv. \tag{8.5}$$

As usual, the typical application of (8.4) is for multiple efficient local bandwidth choice:
If s* (8.3) exists (which then is the minimizer of the asymptotic mean squared error), any data-driven bandwidth choice satisfying $\hat{s} \xrightarrow{P} s^*$ is efficient.

Necessary and sufficient criteria for the existence of minimizers s* (8.3) are given in Müller and Prewitt. Such bandwidth choices can again be obtained through two approaches, one where we use consistent estimates of the unknowns $(\partial^k/\partial x_j^k)g(x)$ appearing in (8.3), and another employing the resmoothing method in analogy to the discussion in Section 7. The first approach was completely worked out in Müller and Prewitt, the second approach is numerically more complicated, since a multivariate minimizer $\hat{s}$ has to be located. Still, for practical purposes this second approach might be promising.

Obviously, gains achieved by choosing multiple different bandwidths will be the highest for wavelike patterns where the phase of the wave is parallel to one of the chosen coordinate axes. It is clear that the angle of the chosen coordinate system with respect to a reference coordinate system also plays a role.

We conjecture that the best angle would be one which yields the largest differences between the multiple bandwidths, i.e., with respect to which the surface g is maximally inhomogeneous in terms of $(\partial^k/\partial x_j^k)g(x)$.

References

Abramson, I. (1982) 'Arbitrariness of the pilot estimator in adaptive kernel methods', J. Mult. Anal. 12, 562-567.

Bhattacharya, P.K. and Mack, Y.P. (1985) 'A two-stage procedure for nonparametric estimation', Statistics and Decisions, Supplement 2, 143-153.

Bhattacharya, P.K. and Mack, Y.P. (1987) 'Linear functions of nearest neighbor estimators of a univariate regression function', Ann. Statist. 15, 976-994.

Bhattacharya, P.K. and Mack, Y.P. (1989) 'Effective data-driven multivariate k-NN function estimation', Technical Report.

Bickel, P.J. and Wichura, M.J. (1971) 'Convergence criteria for multi-parameter stochastic processes and some applications', Ann. Math. Statist. 42, 1656-1670.

Billingsley, P. (1968) Convergence of Probability Measures, Wiley, New York.

Cheng, K.F. and Lin, P.E. (1981) 'Nonparametric estimation of a regression function', Z. Wahr. verw. Geb. 57, 223-233.

Chernoff, H. (1964) 'Estimation of the mode', Ann. Inst. Statist. Math. 16, 31-41.

Eddy, W.F. (1980) 'Optimum kernel estimators of the mode', Ann. Statist. 8, 870-882.

Eddy, W.F. (1982) 'The asymptotic distributions of kernel estimators of the mode', Z. Wahr. verw. Geb. 59, 279-290.

Gasser, Th. and Müller, H.G. (1979) 'Kernel estimation of regression functions', Lecture Notes in Mathematics 757, 23-68, Springer-Verlag, Berlin.

Grenander, U. (1965) 'Some direct estimates of the mode', Ann. Math. Statist. 36, 131-138.

Krieger, A.M. and Pickands, J. III (1981) 'Weak convergence and efficient density estimation at a point', Ann. Statist. 9, 1066-1078.

Mack, Y.P. and Müller, H.G. (1987) 'Adaptive nonparametric estimation of a multivariate regression function', J. Mult. Anal. 17, 163-181.

Müller, H.G. (1985) 'Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators', Statistics and Decisions Supplement 2, 193-206.

Müller, H.G. (1989) 'Adaptive nonparametric peak estimation', Ann. Statist. 17, 1053-1069.

Müller, H.G. and Prewitt, K. (1990) 'Weak convergence of multi-parameter stochastic processes with application to multivariate

smoothing and differentiation', Technical Report.
Müller, H.G. and Stadtmüller, U. (1987) 'Variable bandwidth kernel estimates of regression functions', Ann. Statist. 15, 182-201.
Müller, H.G. and Wang, J.L. (1990) 'Locally adaptive hazard smoothing', Technical Report. Probability Th. Rel. Fields, in press.
Nadaraya, E.A. (1964) 'On estimating regression', Theor. Prob. Appl. 9, 141-142.
Parzen, E. (1962) 'On estimation of a probability density and mode', Ann. Math. Statist. 33, 1065-1076.
Priestley, M.B. and Chao, M.T. (1972) 'Nonparametric function fitting', J. Roy. Statist. Soc. B 34, 384-392.
Romano, J.P. (1988) 'On weak convergence and optimality of kernel density estimates of the mode', Ann. Statist. 16, 629-647.
Rosenblatt, M. (1956) 'Remarks on some nonparametric estimates of a density function', Ann. Math. Stat. 27, 642-649.
Rosenblatt, M. (1971) 'Curve estimates', Ann. Math. Statist. 42, 1815-1842.
Sager, T.W. (1983) 'Estimating modes and isopleths', Communications in Statistics A12, 529-557.
Silverman, B.W. (1978) 'Weak and strong uniform consistency of the kernel estimate of a density and its derivatives', Ann. Statist. 6, 177-184.
Staniswalis, J.G. (1989) 'Local bandwidth selection for kernel estimates', J. Am. Statist. Assoc. 84, 276-283.
Watson, G.S. (1964) 'Smooth regression analysis', Sankhya A26, 359-372.

ON ASYMPTOTIC EFFICIENCY OF AVERAGE DERIVATIVE ESTIMATES

Alexander SAMAROV
University of Lowell and MIT
CCREMS, E40-121, MIT
1 Amherst st.
Cambridge, MA 02139
USA

ABSTRACT. It is shown, using results of Koshevnik and Levit (1976), that the average derivative estimator of Haerdle and Stoker (1989) is the asymptotically efficient nonparametric estimator of the average derivative of regression function.

In many nonparametric density or regression estimation problems it is often the case that one is primarily interested not in the entire curve but just in certain characteristics of it, i.e. in the value of a functional of the unknown curve. For example, one may wish to estimate the value of the density or its derivative at a point, the mode of a unimodal distribution, or integral functionals like the Shannon entropy or Fisher information of the distribution, etc.

Recently Haerdle and Stoker (1989) (referred to as H-S in the rest of this paper) have considered a problem of estimating the average derivative functional (ADF)

$$\delta = E\,(m'(\mathbf{X})) \tag{1}$$

of the regression function $m(x)=E(Y|X=x)$, where (X,Y), $X \in R^d$, $Y \in R^1$ has a distribution with the density $p(x,y) = q(y|x)f(x)$, and $m'(x)$ is a d-vector of derivatives.

The primary motivation for estimating the ADF $\delta=\delta(p)$ comes from a semiparametric model where the regression function $m(x)$ has the form

$$m(x) = g(x^T\beta), \tag{2}$$

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 167–172.
© 1991 *Kluwer Academic Publishers.*

with some unknown (smooth) function g: $R^1 \to R^1$. For this model the ADF is proportional to β since

$$\delta = E[m'(\mathbf{X})] = E[dg/d(\mathbf{x}^T\beta)]\beta , \qquad (3)$$

and a nonparametric estimator of δ will estimate β up to a scale, even though g is unknown. The semiparametric model (2) covers such important models as probit and logit models, censored regression, Tobit models, etc. Stoker (1986) gives an extensive discussion and motivation for the models of the form (2).

We will assume $|\beta| = 1$, where $|\beta|$ is the Euclidean norm for $\beta \in R^d$.

Observe that integration by parts in (1) gives

$$\delta = \delta(p) = E(m'(\mathbf{X})) = E[l(\mathbf{X})Y] \qquad (4)$$

where $l(x) = - f'(x)/f(x)$, the negative score function vector of the marginal density f. The relation (4) is used in H-S to construct a natural estimator of δ, which is a "trimmed" sample version of (4):

$$\hat{\delta} = n^{-1}\sum \hat{l}(\mathbf{X}_i)Y_i\hat{I}_i \qquad (5)$$

where $\hat{l}(\mathbf{x}) = - \hat{f}'_h(\mathbf{x})/\hat{f}_h(\mathbf{x})$, $\hat{f}_h(\mathbf{x})$ and $\hat{f}'_h(\mathbf{x})$ are kernel estimators (the subscript h indicates a window width) of f and f' respectively, and $\hat{I}_i = I\{\hat{f}_h(\mathbf{X}_i) > b_n\}, b_n > 0$, serves as a technical device to prevent erratic behavior of $\hat{\delta}$ for very small values of f_h (here $I\{\cdot\}$ is the indicator function).

Under appropriate assumptions on the distribution p(x,y) and on the kernel and window width of the kernel estimator, it is proved in H-S that $n^{1/2}(\hat{\delta} - \delta(p))$ is asymptotically normal with the covariance matrix

$$\Sigma = \Sigma(p) = \mathrm{Var}_p[r(\mathbf{X}, Y)] \qquad (6)$$

where

$$r(\mathbf{x},y) = l(\mathbf{x})\, y + m'(\mathbf{x}) - l(\mathbf{x})\, m(\mathbf{x}) . \qquad (7)$$

Stoker (1987) considered also a "direct" estimator of (1), based

on the kernel estimator of m'(x), and showed that it has the same asymptotic covariance matrix (6).

The purpose of this note is to show that the average derivative estimator (ADE) (5) is, in some sense, asymptotically efficient. The main result is contained in the following asymptotic lower bound for the risk of an arbitrary estimator of the ADF $\delta(p)$. Cf. Levit (1978), Ibragimov and Has'minsky (1979), and recent work of Newey and Stoker (1988).

Denote **P** a class of distributions p(x,y) satisfying assumptions 1-5 and 6-9 of Haerdle and Stoker (1989), and let $V_\varepsilon(p) = \{s \in \mathbf{P}: \| s - p \|_H < \varepsilon \}$, where $\|\cdot\|_H$ is the Hellinger norm. Other topologies, e.g., those which are weaker than the one induced by the Hellinger norm, can also be used in the next theorem.

<u>Theorem.</u> Assume that $\Sigma=\Sigma(p)$, as defined in (6), is a positive definite matrix. Let $l(u)$ be an arbitrary non-negative, non-decreasing for $u > 0$ loss function, and $\|\cdot\|$ be any seminorm in R^d. Then the following inequality holds:

$$\lim_{\varepsilon\to\infty} \underline{\lim}_{n\to\infty} \inf_{T_n} \sup_{s\in V_\varepsilon(p)} E_s\, l(n^{1/2} \|T_n - \delta(s)\|) \geq E_p\, l(\|\Sigma^{-1/2}\eta\|), \quad (8)$$

where inf is taken over all possible estimates T_n of δ.

<u>Remark 1.</u> Note that taking $l(u)=u^2$ and $\|x\|=|(x^Ta)|$ for arbitrary $a \in R^d$, we obtain from (8) that the H-S estimator is asymptotically efficient in the sense that the matrix Σ in (6) is indeed the "smallest" (in the sense of partial order of non-negative definiteness) asymptotic covariance matrix of any ADF estimator

$$\lim_{\varepsilon\to\infty} \underline{\lim}_{n\to\infty} \inf_{T_n} \sup_{s\in V_\varepsilon(p)} n\, E_s\left[(T_n - \delta(s))(T_n - \delta(s))^T\right] \geq \Sigma. \quad (9)$$

<u>Proof</u> is based on the results of Koshevnik and Levit (1976) and is similar to (but does not follow from) Levit (1978).

Let $\xi(x,y)$ be an arbitrary, smooth function, $\xi: R^d \times R^1 \to R^d$, with compact support and such that $E_p\xi(X,Y) = 0$. Then for sufficiently small $|\theta|$, the expression

$$p_\theta(\mathbf{x},y) = p(\mathbf{x},y)\,(1 + \theta^T\xi(\mathbf{x},y)) \quad (10)$$

describes a family of distributions which belongs to $V_\varepsilon(p)$ and is continuous at $\theta = 0$. Integrating first with respect to y and

retaining only the first order terms for $|\theta| \to 0$, we have

$$\delta(p_\theta) - \delta(p) = \int\int \left(-\frac{f'(x)+\theta^T\left(\int \xi p dy\right)'_x}{f(x)+\theta^T\int \xi p dy}(1+\theta^T\xi(x,y)) + \frac{f'(x)}{f(x)} \right) y p(x,y) dx dy \tag{11}$$

$$= \left(\int \frac{f'(x)\int \xi^T p dy m(x) - f'(x)\int \xi^T y p dy - \left(\int \xi p dy\right)'_x \int y p dy}{f(x)} dx \right) \theta + o(|\theta|).$$

Interchanging now the order of integration and differentiation and integrating by parts in the third term in (11), we obtain

$$\delta(p_\theta)-\delta(p) = \int (-l(\mathbf{x})m(\mathbf{x})+l(\mathbf{x})\, y+m'(\mathbf{x}))\xi^T(\mathbf{x},y)p(\mathbf{x},y)d\mathbf{x}dy\theta+o(|\theta|), \tag{12}$$

which means, using the terminology of Koshevnik and Levit (1976), that the functional δ is weakly differentiable along the families of the type (10) with the gradient given by $r(x,y) - E_p r(X,Y) = r(x,y) - \delta(p)$ (recall that $E_p\xi(X,Y) = 0$). Now, since functions $\xi(x,y)$ are dense in the subspace $\{\eta \in L_2(p): E_p\, \eta = 0\}$ of $L_2(p)$, there exists a sequence $\xi_k(x,y)$ such that

$$\lim_{k\to\infty} E_p \left(\xi_k - r - \delta(p)\right)^T \left(\xi_k - r - \delta(p)\right) = 0 \tag{13}$$

and the distributions $p_\theta(x,y)$ in (10) belong to **P** for sufficiently small $|\theta|$ and are continuous at $\theta = 0$. Theorem now follows from Theorems 1, 2, and Remark 2 of Koshevnik and Levit (1976).

To claim the asymptotic efficiency of the estimator (5) in the locally asymptotically minimax (LAM) sense of Hájek (1972) and Koshevnik and Levit (1976), one needs to establish the asymptotic normality and convergence of moments of (5), uniformly over some neighborhoods of p. Examination of the proof of Theorem 3.1 of H-S, shows that the asymptotic normality and convergence of the second moments hold uniformly over $V_\varepsilon(p)$ if one assumes, for example, that a) the moments in the H-S assumption 4 and the derivatives in their assumption 5 are bounded on $V_\varepsilon(p)$; b) Lipschitz functions in their assumptions 7 and 9 are the same for all $s \in V_\varepsilon(p)$; and c) assumption 8 holds uniformly over $V_\varepsilon(p)$. We

have then

Corollary. Under the above assumptions the ADE (5) is LAM with respect to neighborhoods $V_\varepsilon(p)$ and loss functions l growing not faster than quadratic.

Remark 2. Similar result holds for estimation of the weighted ADF (see, e.g., Powell, Stock, Stoker (1989))

$$\delta_W = E\, w(\mathbf{X})\, m'(\mathbf{X}) \qquad (14)$$

with the corresponding weighted ADE $\hat{\delta}_w$. The asymptotic covariance matrix $\hat{\delta}_w$ is

$$\Sigma_w = \mathrm{Var}\,[m'(\mathbf{X})w(\mathbf{X})-(Y-m(\mathbf{X}))\frac{(w(\mathbf{X})f(\mathbf{X}))'}{f(\mathbf{X})}]\ . \qquad (15)$$

Remark 3. Similar lower bound and LAM results also hold for estimators of the second average derivative functional $\delta_2(p)=E(m''(X))$ which is very useful in selecting appropriate restrictions on the form of the general regression function $m(x)$ for the purposes of dimensionality reduction, see Samarov (1990).

Remark 4. There exist other functionals of the distribution $p(x,y)$ which could be used to estimate β in the semiparametric model (2), see, e.g., Ichimura (1988). The above result does not imply that the ADE (5) is also asymptotically efficient estimator of the parameter β in the model (2): conceivably there might exist a functional, other than the ADF (1) or (14), which leads to the estimator with the covariance matrix "smaller" than that of the ADE (5), cf. remarks in the introduction of Begun, et al (1983).

Acknowledgement. This research was supported in part by a grant from the National Science Foundation.

References

1. Begun, J., W. Hall, Huang W.-M., and Wellner J. (1983) "Information and Asymptotic Efficiency in Parametric - Nonparametric Models", The Annals of Statistics, 11,432-452.
2. Hájek, J. (1972) "Local Asymptotic Minimax and Admissibility in Estimation", Proceedings of the Sixth Berkeley Symposium, I, 175-194.

3. Haerdle, W. and Stoker T. (1989) "Investigating Smooth Multiple Regression by the Method of Average Derivatives", Journal of the American Statistical Association, vol. 84, 986-995.
4. Ibragimov, I. and Has'minsky R. (1981), Statistical Estimation: Asymptotic Theory, Springer Verlag, New York.
5. Ichimura, H. (1988), "Estimation of Single Index Models", Ph.D. dissertation, Dept. of Economics, MIT.
6. Koshevnik, Ya. and Levit B. (1976) "On a Nonparametric Analogue of Information Matrix", Theory of Probability and its Applications, 21, 738-753.
7. Levit, B. (1978) "Asymptotically Efficient Estimation of Nonlinear Functionals", Problems of Information Transmission, 14, 204-209.
8. Newey, W. and Stoker T. (1988) "Efficiency Properties of Average Derivative Estimators", paper presented at the summer meeting of the American Econometric Society, Minneapolis, June 1988.
9. Powell, J., Stock J., and Stoker T. (1989) "Semiparametric Estimation of Index Coefficients", Econometrica, 57, 637-648.
10. Samarov, A. (1990) " Exploring Regression Structure Using Nonparametric Functional Estimation", unpublished manuscript.
11. Stoker, T. (1986) "Consistent Estimation of Scaled Coefficients", Econometrica, 54, 1461-1481.
12. Stoker, T. (1987) "Equivalence of Direct and Indirect Estimators of Average Derivatives", WR# 1961-87, Sloan School of Management, MIT, November 1987.

NONPARAMETRIC ESTIMATION OF ELLIPTICALLY CONTOURED DENSITIES

W. STUTE and U. WERNER
Mathematical Institute University of Giessen
Arndtstraße 2
D-W 6300 Giessen, Germany

ABSTRACT. A density f on the d-dimensional Euclidean space is said to be elliptically contoured iff it admits a representation

$$f(x) = (\det \Sigma)^{-1/2} k((x-\mu)^t \Sigma^{-1} (x-\mu)) \quad .$$

We analyze the large sample behavior of a kernel-type estimator of f, when both the parametric component (μ, Σ) as well as the nonparametric transfer function k are unknown. It turns out that the rate of convergence is independent of d.

1. INTRODUCTION AND MAIN RESULTS

Assume that we have observed a sequence $X_1, X_2, \ldots, X_n, \ldots$ of independent identically distributed (i.i.d.) random vectors in the Euclidean space $\mathbb{R}^d$ with Lebesgue density f. It is well known that without further assumptions on f (up to regularity conditions) the accuracy of approximation of f by any nonparametric estimate f_n deteriorates as d increases.

Over the last years much attention has been given to a family of multivariate densities which aims at generalizing the multivariate normal family on the one hand and, on the other hand, is large enough to also include long-tailed distributions. See Anderson and Fang (1990), Cambanis et al. (1981), Kelker (1970) and Muirhead (1982), among others.

Here is a formal definition. A density f is said to be elliptically contoured iff it admits a representation

$$f(x) = (\det \Sigma)^{-1/2} k((x-\mu)^t \Sigma^{-1} (x-\mu)) \quad , \quad x \in \mathbb{R}^d \quad , \qquad (1.1)$$

where k is a so-called transfer function defined on the positive real line, $\mu \in \mathbb{R}^d$ is a location parameter and Σ is a symmetric positive definite d×d-matrix. The family of all elliptically contoured densities

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 173–190.
© 1991 Kluwer Academic Publishers.

constitutes a semiparametric family, with (μ,Σ) and k denoting the parametric and the nonparametric component, respectively. Since k is univariate and (μ,Σ) typically may be estimated at a rate not depending on d, it is likely that the same holds for f.

In this paper we propose a kernel-type estimate f_n for f. Compared with the classical Parzen-Rosenblatt estimator one major technical difficulty in the analysis of f_n comes from the fact that the summands are no longer independent.

A precise definition of f_n together with some kind of motivation will be given below. Also the main result of the paper will be stated in this section. Section 2 contains the proofs, and in section 3 some simulation results will be reported. To make the paper relatively self-contained, some technical tools from matrix-algebra and some facts about the polar-coordinate transformation are presented in the appendix.

Now, let K be a so-called smoothing kernel satisfying

$$\int K(u)du = 1 \quad .$$

Also, let $h_n > 0$ be a given bandwidth. To motivate f_n below, assume first that $\mu = 0$ and $\Sigma = I$, the d×d-identity matrix. Put $g(x) = ||x||^2$ for $x \in \mathbb{R}^d$, $||\cdot||$ always denoting the Euclidean norm of x. Set

$$f_n^1(x) = (nh_n)^{-1} \sum_{i=1}^{n} K\left[\frac{g(x)-g(X_i)}{h_n}\right] ,$$

i.e. f_n^1 is the Parzen-Rosenblatt estimate at g(x) of the density of $g(X_1)$. Let U be the d-dimensional Euclidean space $(d \geq 2)$ with $\{x \in \mathbb{R}^d : x_1 \leq 0 \text{ and } x_2 = 0\}$ removed. Let $T = (T_1,\ldots,T_d): U \to \mathbb{R}^d$ be given by

$$T_1(x) := \sum_{i=1}^{d} x_i^2 , \qquad T_2(x) := \text{atan } 2(x_1,x_2)$$

and

$$T_k(x) := \arctan x_k \left[\sum_{i=1}^{k-1} x_i^2\right]^{-1/2} , \quad 3 \leq k \leq d \quad .$$

Here, atan 2(x,y) is defined by

$$\operatorname{atan} 2(x,y) = \begin{cases} \arctan(y/x) & x > 0,\ y \in \mathbb{R} \\ \arctan(-x/y) + \frac{\pi}{2} & x < 0,\ y > 0 \\ \arctan(-x/y) - \frac{\pi}{2} & x < 0,\ y < 0 \\ \pi/2 & x = 0,\ y > 0 \\ -\pi/2 & x = 0,\ y < 0 \end{cases}$$

T constitutes a bijection between U and

$$V = \left\{ v \in \mathbb{R}^d : v_1 > 0,\ -\pi < v_2 < \pi,\ -\pi/2 < v_i < \pi/2 \text{ for } 3 \le i \le d \right\}.$$

Its inverse is given by $S(v) = (S_1(v),\ldots,S_d(v))$, with

$$S_1(v) = v_1^{1/2} \prod_{j=2}^{d} \cos v_j$$

and

$$S_i(v) = v_1^{1/2} \sin v_i \prod_{j=i+1}^{d} \cos v_j \quad , \quad 2 \le i \le d \quad .$$

Let $J(x) = \det(T'(x))$, where T' is the Jacobian of T. Change of variables yields

$$\mathbb{E}\, f_n^1(x) = h_n^{-1} \int_U K\left[\frac{g(x)-g(y)}{h_n} \right] k(g(y))\,dy$$

$$= h_n^{-1} \int_0^\infty \int_{-\pi}^{\pi} \int_{-\pi/2}^{\pi/2} \cdots \int_{-\pi/2}^{\pi/2} K\left[\frac{g(x)-g(S(v))}{h_n} \right] k(g(S(v)))\, |J(S(v))|^{-1} dv \ .$$

By Lemma 4.3 in the appendix, the last integral equals

$$c\, h_n^{-1} \int_0^\infty K\left[\frac{g(x)-v_1}{h_n} \right] k(v_1)\, v_1^{(d-2)/2} dv_1 \quad ,$$

with $c = \pi^{d/2}/\Gamma(d/2)$. If k is continuous at $g(x)$, then under appropriate conditions on K the last integral tends to

$$c\, f(x)\,[g(x)]^{(d-2)/2} \quad , \quad \text{as } h_n \to 0 \quad .$$

So, an estimate of $f(x)$ which is asymptotically unbiased, is obtained by setting

$$f_n(x) = \frac{1}{c\, nh_n ||x||^{d-2}} \sum_{i=1}^{n} K\left[\frac{g(x)-g(X_i)}{h_n} \right] \quad .$$

Similarly, we get

$$\operatorname{Var} f_n(x) = [c\, h_n ||x||^{d-2}]^{-2} n^{-1} \left[c \int_0^\infty K^2\left[\frac{g(x)-v_1}{h_n} \right] k(v_1) v_1^{(d-2)/2} dv_1 \right.$$

$$\left. - \left[c \int_0^\infty K\left[\frac{g(x)-v_1}{h_n} \right] k(v_1) v_1^{(d-2)/2} dv_1 \right]^2 \right].$$

Hence

$$\lim_{n\to\infty} nh_n \operatorname{Var} f_n(x) = \frac{\Gamma(d/2)}{\pi^{d/2}} f(x) ||x||^{-(d-2)} \int K^2(u)du \equiv \sigma^2 \quad .$$

If k is twice continuously differentiable in a neighborhood of $g(x) = ||x||^2$, $x \neq 0$, then (for a symmetric K with compact support)

$$\text{Bias} = \mathbb{E} f_n(x) - f(x) = h_n^2 \frac{z''(g(x))}{2||x||^{d-2}} \int u^2 K(u)du + o(h_n^2) \quad ,$$

where $z(v) = k(v)v^{(d-2)/2}$. This shows that the optimal bandwidth minimizing $D = \text{Var} + \text{Bias}^2$ is of the form $a_n = \text{const} \times n^{-1/5}$, with D being of the order $n^{-4/5}$.

The constant depends on k and is therefore unknown. It may be consistently estimated by plugging in an estimate of the unknown $z''(g(x))$. Note also that the rate at which D tends to zero is independent of the dimension d. Verification of Ljapunov's criterion gives

$$(nh_n)^{1/2}[f_n(x) - \mathbb{E} f_n(x)] \to \mathcal{N}(0,\sigma^2) \quad \text{in distribution} \quad .$$

We now consider estimation of a general elliptically contoured f when μ and Σ are known:

$$f(x) = (\det \Sigma)^{-1/2} k((x-\mu)^t \Sigma^{-1}(x-\mu)) \quad .$$

Of course, the representation of f is not unique. Putting $a = [\det \Sigma]^{-1/d}$ and setting $k^*(x) = a^{d/2}k(ax)$, $\Sigma^* = a\Sigma$, we get $\det \Sigma^* = 1$ and

$$f(x) = (\det \Sigma)^{-1/2} k((x-\mu)^t \Sigma^{-1}(x-\mu)) = k^*((x-\mu)^t \Sigma^{*-1}(x-\mu)) \quad .$$

In other words, we may and do assume that without loss of generality $\det \Sigma = 1$.

Write $\Sigma = U\Lambda U^t$, where Λ is the diagonal matrix of (positive) eigenvalues of Σ and U is an appropriate orthonormal matrix. Set $A = (U\Lambda^{-1/2})^t$ and put $Z_i = A(X_i-\mu)$. From Lemma 4.4, each Z has the density

$$f_0(z) = k(z^t z) \quad .$$

Furthermore,

$$f(x) = f_0(A(x-\mu)) \quad .$$

We are thus led to estimate f(x) by

$$f_n(x) := f_{0n}(A(x-\mu)) \quad ,$$

where f_{0n} is the aforementioned kernel estimate, but this time computed for $Z_1,\ldots,Z_n$. This gives

$$f_n(x) = \frac{1}{c\ nh_n[(x-\mu)^t\Sigma^{-1}(x-\mu)]^{(d-2)/2}} \times$$

$$\times \sum_{i=1}^{n} K\left[\frac{(x-\mu)^t\Sigma^{-1}(x-\mu) - (X_i-\mu)^t\Sigma^{-1}(X_i-\mu)}{h_n}\right] \quad .$$

In this general case,

$$\lim_{n\to\infty} nh_n \operatorname{Var} f_n(x) = \frac{\Gamma(d/2)}{\pi^{d/2}} f(x)[(x-\mu)^t\Sigma^{-1}(x-\mu)]^{-(d-2)/2}\int K^2(u)du \quad .$$

The expansion of the bias is the same as before, with x replaced by $A(x-\mu)$:

$$\text{Bias} = h_n^2 \frac{z''((x-\mu)^t\Sigma^{-1}(x-\mu))}{2[(x-\mu)^t\Sigma^{-1}(x-\mu)]^{(d-2)/2}} \int u^2K(u)du + o(h_n^2) \quad .$$

When Σ and μ are unknown, f_n needs to be modified so as to include estimates of Σ and μ, say, Σ_m and μ_m. For this we use a preliminary sample $X_{-1},\ldots,X_{-m}$ being independent of $X_1,X_2,\ldots$ We shall only investigate the case when Σ is unknown but μ is known, say, $\mu = 0$. The extension to the general case will be obvious. So, f_n results in

$$f_n(x) = \frac{1}{c\ nh_n[x^t\Sigma_m^{-1}x]^{(d-2)/2}} \sum_{i=1}^{n} K\left[\frac{x^t\Sigma_m^{-1}x - X_i^t\Sigma_m^{-1}X_i}{h_n}\right] . \quad (1.2)$$

Throughout the paper the following regularity conditions are assumed to hold:

K is symmetric at zero: $K(u) = K(-u)$ (1.3)

K is twice continuously differentiable (1.4)

K has compact support (say, $[-1,1]$) . (1.5)

As to k, we require that

$$k \text{ is twice continuously differentiable in a neighborhood of } x^t\Sigma^{-1}x \text{ (recall } \mu = 0) \quad . \tag{1.6}$$

Our main result will be stated for bandwidths satisfying

$$nh_n^5 \to 0 \quad \text{and} \quad nh_n \to \infty \tag{1.7}$$

$$\frac{nh_n}{m} \to 0 \quad \text{and} \quad \frac{n}{m^2h_n^3} \to 0 \quad . \tag{1.8}$$

Finally, as to Σ_m, we need

$$m^{1/2}||\Sigma_m - \Sigma||_E = O_{\mathbb{P}}(1) \quad . \tag{1.9}$$

Condition (1.7) guarantees, that the bias is negligible compared with the variance part. The case $nh_n^5 \to b > 0$ may be treated likewise. (1.8) is important for dealing with the error arising from estimation of Σ. (1.9) is a convenient condition. For example, under finite second moments we have (see (4.2) and (4.3))

$$\text{Cov } X = c\Sigma \quad \text{with} \quad \det[\text{Cov}(X)] = c^d \quad .$$

So we are led to put

$$\Sigma_m = c^{-1} \text{ Cov}(X_{-1},\ldots,X_{-m})$$

with

$$c = \left[\det[\text{Cov}(X_{-1},\ldots,X_{-m})]\right]^{-1/d} \quad .$$

Clearly, this Σ_m satisfies (1.9).

<u>Theorem.</u> Under (1.3) - (1.9),

$$(nh_n)^{1/2}[f_n(x) - f(x)] \to \mathcal{N}(0,\sigma^2) \quad \text{in distribution} \quad ,$$

with

$$\sigma^2 = \frac{\Gamma(d/2)}{\pi^{d/2}} f(x) \; [x^t\Sigma^{-1}x]^{-(d-2)/2} \int K^2(u)du \quad .$$

2. PROOFS

The conditional expectation of $f_n(x)$ given $X_{-1},X_{-2},\ldots,X_{-m}$ equals

$$\mathbb{E}[f_n(x)\,|X_{-1},\ldots,X_{-m}] =$$

$$= \frac{1}{c\, h_n [x^t \Sigma_m^{-1} x]^{(d-2)/2}} \int K\left[\frac{x^t \Sigma_m^{-1} x - y^t \Sigma_m^{-1} y}{h_n}\right] k(y^t \Sigma^{-1} y)\, dy \quad .$$

When $\Sigma = U \Lambda U^t$, the transformation $z = (U\Lambda^{-1/2})^t y$ leads to

$$\mathbb{E}[f_n(x) \,|X_{-1}, \ldots, X_{-m}]$$

$$= \frac{1}{c\, h_n [x^t \Sigma_m^{-1} x]^{(d-2)/2}} \int K\left[\frac{x^t \Sigma_m^{-1} x - z^t E_m^{-1} z}{h_n}\right] k(z^t z)\, dz \quad ,$$

where $E_m^{-1} = \Lambda^{1/2} U^{-1} \Sigma_m^{-1} U \Lambda^{1/2}$. We have

$$R_m \equiv E_m^{-1} - I = \Lambda^{1/2} U^{-1} [\Sigma_m^{-1} - \Sigma^{-1}] U \Lambda^{1/2}$$

and therefore (see the appendix)

$$||R_m||_E = ||E_m^{-1} - I||_E \le ||\Lambda^{1/2} U^{-1}||_E \, ||\Sigma_m^{-1} - \Sigma^{-1}||_E \, ||U \Lambda^{1/2}||_E \; .$$

Since Σ and Σ_m are symmetric and positive definite, $||\Sigma_m^{-1} - \Sigma^{-1}||_E$ will be (at least) of the order of $||\Sigma_m - \Sigma||_E$. Hence

$$||\Sigma_m^{-1} - \Sigma^{-1}||_E = O_{\mathbb{P}}(m^{-1/2}) \text{ whenever } ||\Sigma_m - \Sigma||_E = O_{\mathbb{P}}(m^{-1/2}) \; .$$

Now, write $u = u_m = x^t \Sigma_m^{-1} x$ for short. Under (1.3) - (1.5) we obtain from Taylor-expansion

$$\mathbb{E}[f_n(x) \,|X_{-1}, \ldots, X_{-m}] = \frac{1}{c\, u^{(d-2)/2}} [I_1 - I_2 + I_3] \quad ,$$

where

$$I_1 = h_n^{-1} \int K\left[\frac{u - z^t z}{h_n}\right] k(z^t z)\, dz \quad ,$$

$$I_2 = h_n^{-1} \int \frac{z^t R_m z}{h_n} K'\left[\frac{u - z^t z}{h_n}\right] k(z^t z)\, dz$$

and

$$I_3 = h_n^{-1} \int \left[\frac{z^t R_m z}{h_n}\right]^2 \frac{K''(\Delta)}{2} k(z^t z)\, dz$$

for some Δ between $h_n^{-1}(u - z^t z)$ and $h_n^{-1}(u - z^t E_m^{-1} z)$. As before,

$$I_1 = \frac{\pi^{d/2}}{h_n \Gamma(d/2)} \int_0^\infty K\left[\frac{u-v_1}{h_n}\right] k(v_1) v_1^{\frac{d-2}{2}} dv_1 \quad .$$

Similarly, writing $R_m = (r_{ij})$ and introducing polar coordinates, we get

$$I_2 = h_n^{-2} \sum_{i,j} r_{ij} \int z_i z_j \, K'\left[\frac{u-z^t z}{h_n}\right] k(z^t z) dz$$

$$= h_n^{-2} \sum_{i,j} r_{ij} \int_0^\infty \int_{-\pi}^{\pi} \int_{-\pi/2}^{\pi/2} \cdots \int_{-\pi/2}^{\pi/2} K'\left[\frac{u-v_1}{h_n}\right] k(v_1) v_1^{\frac{d}{2}} 2^{-1} S_{ij}(v_2,\ldots,v_d) dv_d \ldots dv_1 .$$

Here S_{ij} is an appropriate product of sin- and cos-functions, the form of which is completely determined by the (polar coordinate) transformation S and its Jacobian. Note that $|S_{ij}| \le 1$. It follows

$$I_2 = h_n^{-2} \sum_{i,j} r_{ij} F_{ij} \int_0^\infty K'\left[\frac{u-v}{h_n}\right] k(v) v^{d/2} dv$$

for some absolute constants F_{ij}. Thus, integration by parts gives

$$I_2 = \sum_{i,j} r_{ij} F_{ij} h_n^{-1} \int_0^\infty K\left[\frac{u-v}{h_n}\right] [k(v) v^{d/2}]' dv \quad .$$

For I_3, $K''(\Delta) = 0$ unless

$$|u_m - z^t z| \le h_n \quad \text{or} \quad |u_m - z^t \Sigma_m^{-1} z| \le h_n \quad . \tag{2.1}$$

But $h_n \to 0$ and $E_m^{-1} \to I$ with respect to $||\cdot||_E$. So all z's fulfilling (2.1), say, $z \in P_{n,m}$, are contained in a compact set. Moreover, K'' is bounded. Putting these facts together, we obtain (with λ = Lebesgue measure on $\mathbb{R}^d$).

$$|I_3| \le \text{const } h_n^{-3} \; ||R_m||_2^2 \; \lambda(P_{n,m}) \quad .$$

But $\lambda(P_{n,m}) = O(h_n)$. Here the constants may depend on the sequences involved. Under (1.3) - (1.5) we thus get

$$|I_2| = O_{\mathbb{P}}(m^{-1/2})$$

and

$$|I_3| = O_{\mathbb{P}}(h_n^{-2} m^{-1}) \quad .$$

Furthermore, $u_m \to x^t\Sigma^{-1}x$ in probability. Finally (extend k by setting $k(v) = 0$ for $v < 0$)

$$I_1 = c\left[z(u_m) + \frac{1}{2} z''(u_m)h_n^2\int u^2K(u)du + o(h_n^2)\right]$$

eventually. Collecting error terms, we thus get the following

<u>2.1 Lemma.</u> Under (1.3) - (1.6),

$$\mathbb{E}[f_n(x)\,|X_{-1},\ldots,X_{-m}] = f(x) + I_0 + O_{\mathbb{P}}(m^{-1/2}+h_n^{-2}m^{-1}) + o(h_n^2)$$

with

$$I_0 = \frac{h_n^2 z''(x^t\Sigma^{-1}x)}{2[x^t\Sigma^{-1}x]^{(d-2)/2}}\int u^2K(u)du \quad .$$

<u>2.2 Corollary.</u> Under $nh_n^5 \to 0$,

$$(nh_n)^{1/2}\left[\mathbb{E}[f_n(x)\,|X_{-1},\ldots,X_{-m}]-f(x)\right] = O_{\mathbb{P}}\left[\left(\frac{nh_n}{m}\right)^{1/2} + \frac{(nh_n)^{1/2}}{mh_n^2}\right] \quad .$$

Under suitable assumptions on m, the right-hand side is $o_{\mathbb{P}}(1)$. On the other hand, given $X_{-1},X_{-2},\ldots,X_{-m},\ldots$, f_n is a sum of i.i.d. random variables. With the same arguments as employed in the last lemma, one can prove

$$nh_n\ \mathrm{Var}[f_n(x)\,|X_{-1},X_{-2},\ldots,] \to \sigma^2 \quad \text{in probability} \quad .$$

Also Ljapunov's criterion, in a conditional setup, is easily checked. The following lemma is thus immediate.

<u>2.3 Lemma.</u> Under (1.3) - (1.9),

$$\lim_{n\to\infty} \mathbb{P}\left[(nh_n)^{1/2}\left[f_n(x) - \mathbb{E}[f_n(x)\,|X_{-1},X_{-2},\ldots]\right] \le b\,|X_{-1},\ldots\right]$$
$$= \mathbb{P}(\mathcal{N}(0,\sigma^2) \le b) \quad \text{in probability} \quad .$$

We are now in the position to give the

<u>Proof of the Theorem.</u> Integration of the conditional probability on the left-hand side in Lemma 2.3 and an application of the dominated convergence theorem yields

$$(nh_n)^{1/2}\left[f_n(x) - \mathbb{E}[f_n(x)\,|X_{-1},X_{-2},\ldots]\right] \to \mathcal{N}(0,\sigma^2)$$

in distribution. Under (1.7), the conditional expectation can be replaced by $f(x)$. This completes the proof of the Theorem.

□

3. SOME SIMULATION RESULTS

In this section we report on several simulation results, which show that f_n outperforms the d-dimensional kernel estimate (cf. Cacoullos (1964)) quite well when the underlying density has elliptical contours. For f we took a centered normal density on the plane with $\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1/4 \end{bmatrix}$. Sample size was n = 25, with

$$K(u) = \frac{15}{16}(1-u^2)^2 1_{[-1,1]}(u) \quad .$$

Though in our main result we only investigated f_n pointwise, it may be shown that $f_n \to f$ uniformly on $\{x: x^t\Sigma^{-1}x \geq \varepsilon\}$ in probability as $n \to \infty$ (when $\mu = 0$), each $\varepsilon > 0$. The point $x = \mu$ presents some difficulties for two reasons. Firstly, for $d \geq 3$, the variance of $f_n(x)$ increases to infinity as $x \to \mu$. Secondly, since each $X_i^t\Sigma_m^{-1}X_i$ is nonnegative, even asymptotically there is a bias when estimating k at zero. For d = 2 the first phenomenon does not occur. To eliminate the second, we used Schuster's (1985) modification (in a neighborhood of μ) to get a uniformly convergent estimator on the whole of the plane.
In the displays below we present plots of the

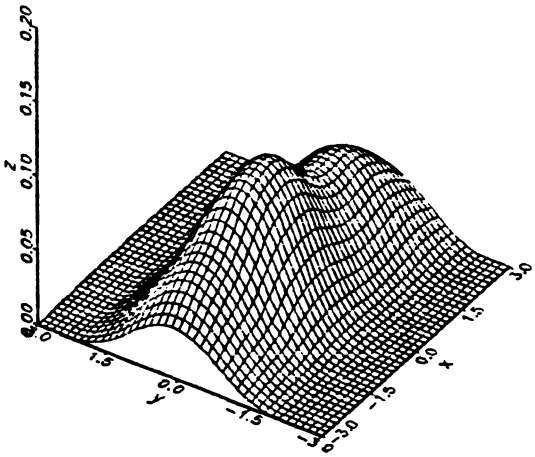
(a) two-dimensional kernel estimate

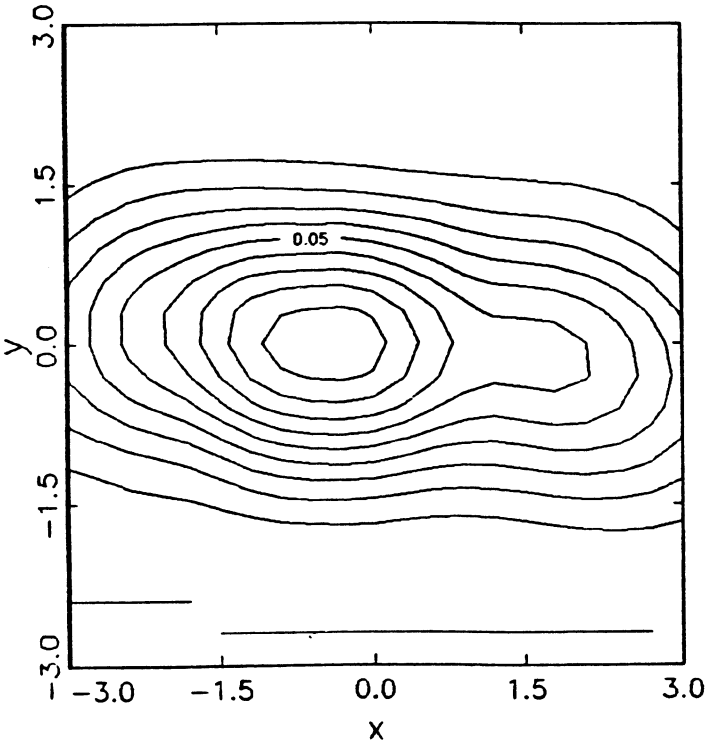
(b) f_n as defined in this paper

(c) Schuster's modification (adapted to f_n).

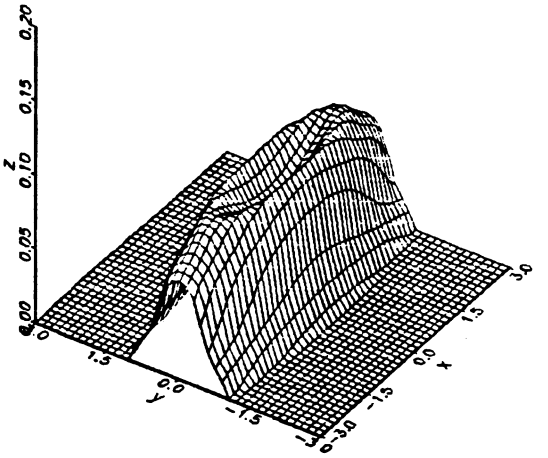
Because of lack of space we have to restrict ourselves to one bandwidth, namely h = 1.75. Compared with other choices of h it seemed to us that this choice of h gave the most reasonable result for the two-dimensional kernel-estimate. For each h considered by us the two-dimensional kernel-estimate was unable to exhibit the symmetry and unimodality of f (display 1). In contrast to this, f_n by definition is elliptically contoured. The relative large bias in a neighborhood of μ becomes apparent in terms of a volcano-effect (display 2). The Schuster modification of f_n estimates f quite well (display 3). Display 4 presents the estimate of k together with its Schuster modification. Display 5 gives the true density together with its contour lines.

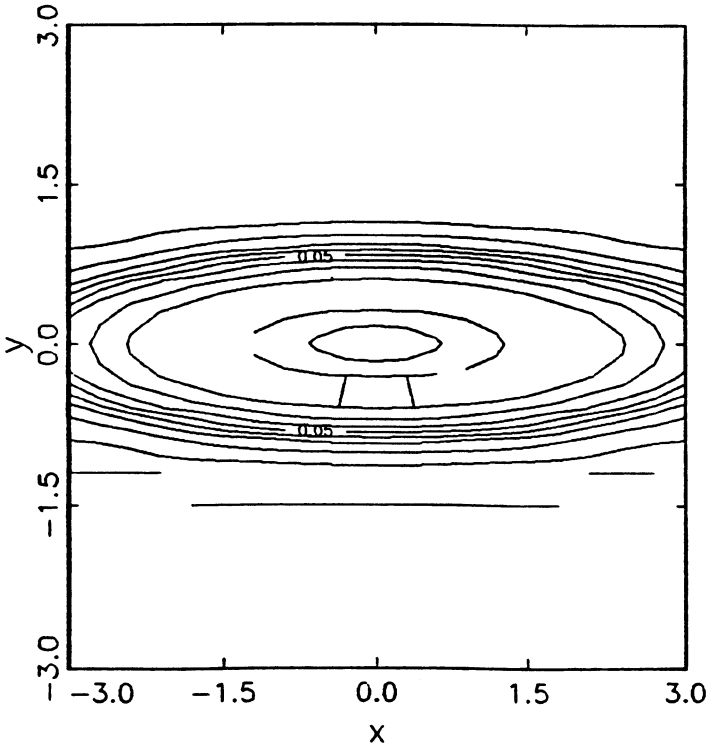
Display 1


0.20
0.15
0.10
0.05
0.00
z
0
1.5
0.0
y
-1.5
-3
-3.0
-1.5
0.0
x
1.5
3.0

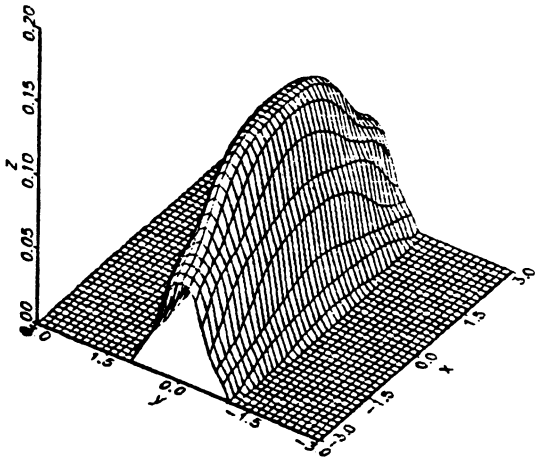

3.0
1.5
0.0
-1.5
-3.0
y
0.05
-3.0
-1.5
0.0
1.5
3.0
x

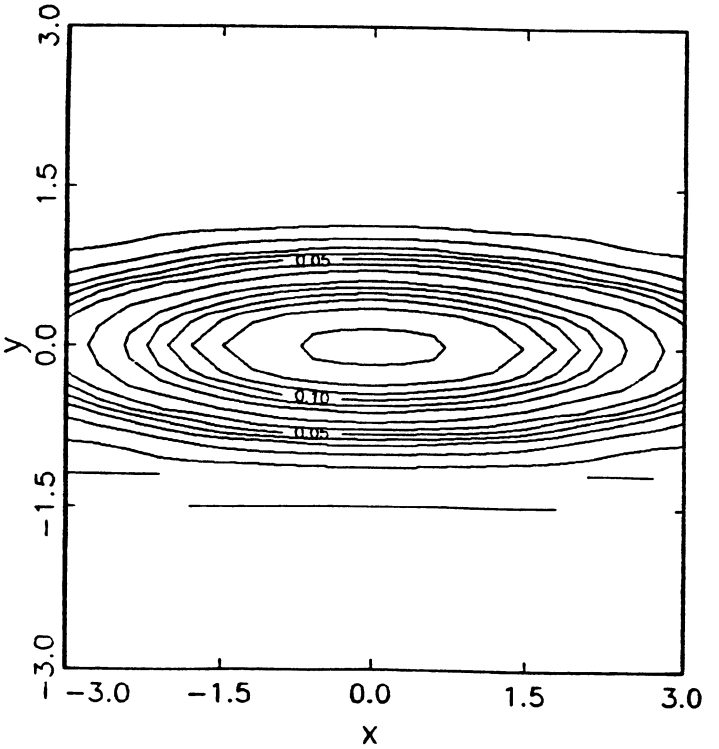
Display 2


0.00
0.05
0.10
0.15
0.20
z
0
1.5
0.0
-1.5
-3.0
y
-3.0
-1.5
0.0
1.5
3.0
x

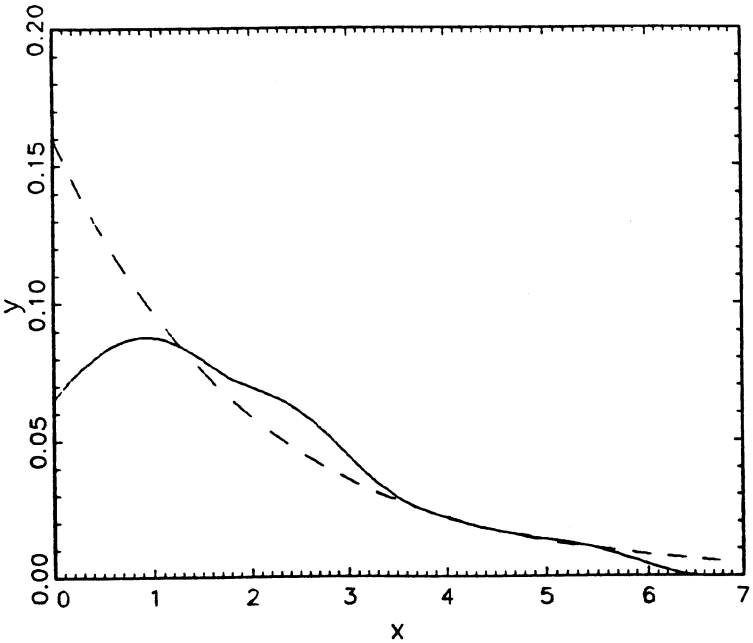

0.05
0.05
y
3.0
1.5
0.0
-1.5
-3.0
-3.0
-1.5
0.0
1.5
3.0
x

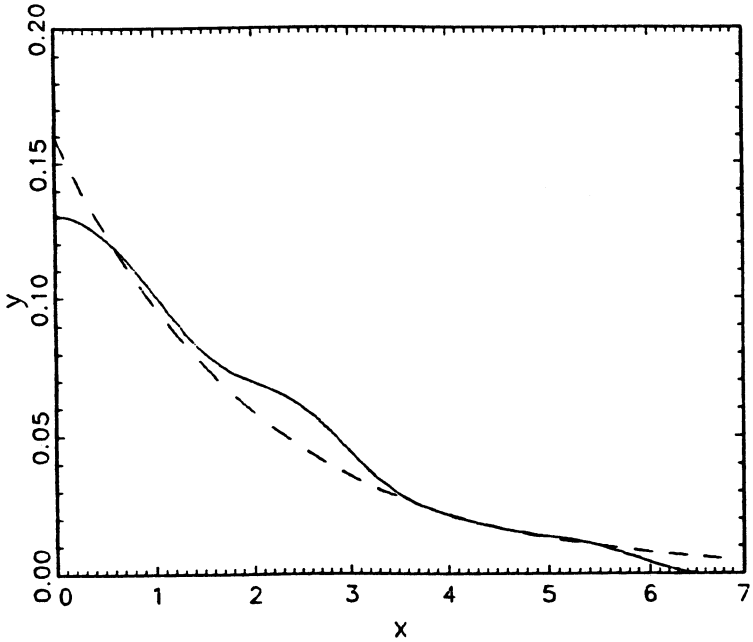
Display 3


0.00
0.05
0.10
0.15
0.20
z
0
1.5
y
0.0
-1.5
-3.0
-3.0
-1.5
0.0
x
1.5
3.0

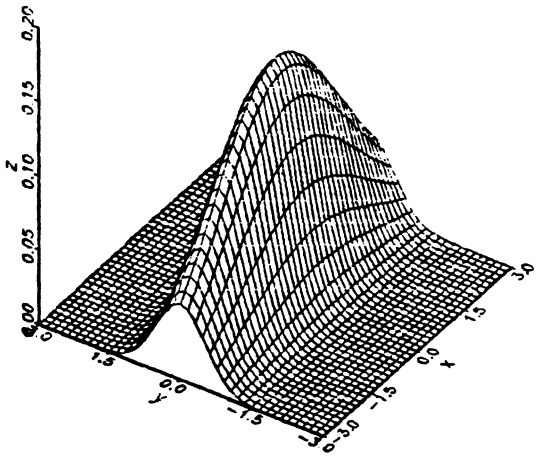

3.0
1.5
0.0
-1.5
-3.0
y
0.05
0.10
0.05
-3.0
-1.5
0.0
1.5
3.0
x

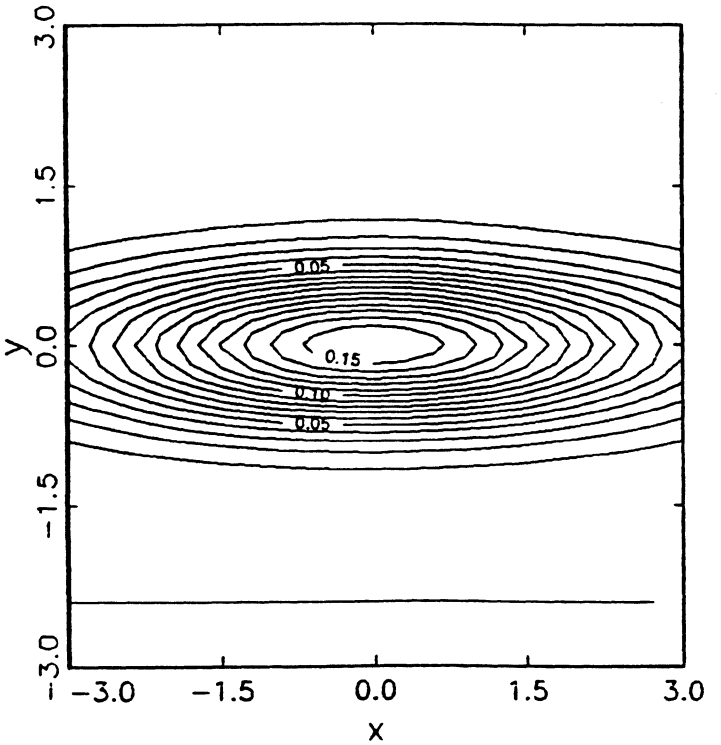
Display 4


0.20
0.15
0.10
0.05
0.00
y
0
1
2
3
4
5
6
7
x


0.20
0.15
0.10
0.05
0.00
y
0
1
2
3
4
5
6
7
x

Display 5


0.20
0.15
0.10
0.05
0.00
z
1.5
0.0
-1.5
y
-3.0
-1.5
0.0
1.5
3.0
x


3.0
1.5
0.0
-1.5
-3.0
y
0.05
0.15
0.10
0.05
-3.0
-1.5
0.0
1.5
3.0
x

4. APPENDIX

Recall $J(x) = \det(T'(x))$. Lemmas 4.1 - 4.3 below are easy to check.

<u>4.1 Lemma.</u> For $x \in U$,

$$J(x) = 2 \prod_{i=2}^{d-1} \left[\sum_{j=1}^{i} x_j^2 \right]^{-1/2} .$$

<u>4.2 Lemma.</u> We have

$$G := \pi \prod_{i=1}^{d-2} \int_{-\pi/2}^{\pi/2} \cos^i(x)\,dx = \pi^{d/2}/\Gamma(d/2) .$$

Finally, as to S from section 1, we need

<u>4.3 Lemma.</u> For $v \in V$,

(i) $\quad ||S(v)||^2 = v_1 \qquad (||\cdot|| = \text{Euclidean norm})$

(ii) $\quad [J(S(v))]^{-1} = 2^{-1} v_1^{\frac{d-2}{2}} \prod_{j=3}^{d} \cos^{j-2} v_j .$

Applied to spherically contoured densities $f = k \circ g$, $g(x) = ||x||^2$, we get by change of variables

$$\begin{aligned} 1 &= \int f(x)\,dx = \int k(g(x))\,dx \\ &= \int_0^\infty \int_{-\pi}^{\pi} \int_{-\pi/2}^{\pi/2} \cdots \int_{-\pi/2}^{\pi/2} k(g(S(v)))\,|J(S(v))|^{-1} dv_d \ldots dv_1 \\ &= G \int_0^\infty k(v_1) v_1^{\frac{d-2}{2}} dv_1 . \end{aligned}$$

In other words, the transfer function necessarily satisfies

$$\int_0^\infty k(v) v^{\frac{d-2}{2}} dv = \pi^{-d/2} \Gamma(d/2) .$$

The following lemma enables one to reduce the analysis of elliptically contoured densities to the spherical case. For this, let $\Sigma = U \Lambda U^t$, where Λ is the diagonal matrix of (positive) eigenvalues of Σ and U is

an appropriate orthonormal matrix. Write $X \in E_d(\mu,\Sigma)$ for short if f satisfies (1.1).

<u>4.4 Lemma.</u> Set $A = (U\Lambda^{-1/2})^t$ and $B = U\Lambda^{1/2}$. Then

(i) $X \in E_d(\mu,\Sigma)$ implies $Z := A(X-\mu) \in E_d(0,I)$

(ii) $Z \in E_d(0,I)$ implies $X := BZ + \mu \in E_d(\mu,\Sigma)$,

the transfer function being unchanged in each case.

<u>Proof.</u> Use $\Sigma = U\Lambda U^t$ and change of variables. □

Change of variables also shows that for any orthonormal matrix C

$$Z \in E_d(0,I) \text{ implies } X = CZ \in E_d(0,I) \quad ,$$

with the same k, i.e. Z = X in distribution. In particular,

$$(Z_{\sigma_1},\ldots,Z_{\sigma_d}) = (Z_1,\ldots,Z_d) \quad \text{in distribution}$$

for any permutation $(\sigma_1,\ldots,\sigma_d)$. As a consequence, $Z_1,\ldots,Z_d$ are identically distributed. Furthermore,

$$(Z_1,\ldots,Z_i,\ldots,Z_d) = (Z_1,\ldots,-Z_i,\ldots,Z_d) \quad \text{in distribution .}$$

So, under finite second moments,

$$\mathbb{E}(Z_iZ_j) = 0 \text{ for } i \neq j \text{ and } \mathbb{E}(Z_i) = 0 \quad .$$

Conclude

$$\mathrm{Cov}(Z) = cI \quad , \quad c = \mathrm{Var}(Z_1) = \mathbb{E}(Z_1^2) \quad .$$

Lemma 4.4 (ii) implies that for $X \in E_d(\mu,\Sigma)$

$$\mathbb{E}(X) = \mu$$

$$\mathrm{Cov}(X) = B\ \mathrm{Cov}(Z)B' = cBB' = c\Sigma$$

and

$$\det[\mathrm{Cov}(X)] = c^d \det \Sigma \quad .$$

We end this section by recalling a fact from matrix-algebra. For any d×d matrix $A = (a_{ij})$ let

$$||A||_E = \left[\sum_{i=1}^{d} \sum_{j=1}^{d} a_{ij}^2 \right]^{1/2} ,$$

respectively,

$$\|A\|_2 := \sup_{x \neq 0} \|Ax\| \,/\, \|x\|$$

denote the Euclidean, respectively, spectral norm of A, with $\|x\|$ denoting the Euclidean norm of x. Graybill (1983) discusses some important relationships between $\|\cdot\|_E$ and $\|\cdot\|_2$. From these and the Wielandt-Hoffman Theorem (cf. Wilkinson (1972)) we get the following result, which will be needed in the proof of the Theorem.

4.5 Lemma. Assume A_n and A are symmetric and nonsingular. Then

$$\|A_n^{-1} - A^{-1}\|_E \leq K\|A_n - A\|_E,$$

where $K < \infty$ is a constant (depending on A_n and A).

REFERENCES

Anderson, T.W. and Fang, K.T. (1990), eds. Statistical inference in ellipticaly contoured and related distributions. Allerton Press, New York.

Cacoullos, T. (1964). Estimation of a multivariate density. Ann. Inst. Statist. Math. 18, 179-189.

Cambanis, S., Huang, S. and Simons, G. (1981). On the theory of elliptically contoured distributions. J. Multivariate Anal. 11, 368-385.

Graybill, F.A. (1983). Matrices with applications in statistics. Wadsworth, Belmont.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. Sankhyā A 32, 419-430.

Muirhead, R.J. (1982). Aspects of multivariate statistical theory. Wiley, New York.

Schuster, E.F. (1985). Incorporating support constraints into nonparametric estimators of densities. Comm. Statist. Theor. Math. 14, 1123-1136.

Wilkinson, J.A. (1972). The algebraic eigenvalue problem. Clarendon Press, Oxford.

UNIFORM DECONVOLUTION: NONPARAMETRIC MAXIMUM LIKELIHOOD AND INVERSE ESTIMATION

A. J. van ES
Faculty of Mathematics and Computer Science
University of Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
The Netherlands

ABSTRACT. We consider the Uniform Deconvolution problem. Having a sample from the convolution of the uniform [0,1) distribution and an unkown distribution function F, we want to estimate F. In this problem a maximum likelihood estimator of F can be derived, provided an extra support condition on F is satisfied. The problem can also be viewed as an inverse estimation problem. Since the transformation which maps the unknown distribution function F on the distribution of the data can be inverted, we can compute the inverse value of the empirical distribution function of the data and thus obtain estimates of F. The estimators we consider are based on derivatives of convex minorants of certain step functions computed from the data. We discuss the pointwise asymptotic distribution of the estimators. It turns out that the asymptotic variance of the maximum likelihood estimator is smaller than the asymptotic variance of the inversion type estimators considered.

1. Introduction

The general deconvolution problem can be described as follows. Suppose that we are interested in the distribution of iid random variables $Y_1, ..., Y_n$, and that these random variables can not be observed directly. Instead we can observe $X_1, ..., X_n$ where the X's are equal to the Y's plus independent errors e with a known distribution, i.e. $X_i = Y_i + e_i$. In other words: we have a sample from a distribution $G = K * F$, where the distribution function K is known, and we want to estimate the unknown distribution function F. Here $*$ denotes convolution. We assume that F is uniquely determined by G and K. Since we don't want to assume that F belongs to some specific parametric family of distributions we treat the deconvolution problem as a nonparametric estimation problem.

The most popular approach to this problem applies characteristic functions (Fourier transforms) and kernel smoothing to obtain estimates of F or its density. The characteristic function (chf) of the X's is the product of the chf of K, which is known, and the chf of F. Since we can estimate the chf of the X's from the data we can also, by dividing this estimate by the chf of K, obtain an estimate of the chf of F. By kernel smoothing and Fourier inversion we can then obtain an estimate of F or its density. For refinements of this method see Stefanski and Carroll (1987, 1989), Fan (1988), Carroll and Hall (1988), Devroye (1989), Liu and Taylor (1989, 1990) and Stefanski (1990).

The deconvolution problem as stated above is an example of an *inverse problem.* The distribution of the observations is a known transformation of the distribution we want to estimate. This means that if we can invert the transformation and if we can estimate the distribution of the data, then we can estimate the unknown distribution by computing the inverse of the estimate of

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 191–198.
© 1991 *Kluwer Academic Publishers.*

the distribution of the data. The estimators based on Fourier transforms and kernel smoothing can be viewed as special examples of these inversion estimators. For discussions of inverse problems in this more general setting see O'Sullivan (1986) and Carroll et al. (1990).

Now let us consider nonparametric maximum likelihood estimation of the distribution function (df) F. Assuming the df K has a density k the df G of the data has a density $g = k * F$. This means that we can try to find a df F which maximizes the loglikelihood

$$L_n(F) = \sum_{i=1}^{n} log(g(X_i)) = \sum_{i=1}^{n} log(\int k(X_i - u)dF(u)). \tag{1.1}$$

In general there exist no explicit expressions for these nonparametric maximum likelihood estimators (NPMLE's) of F, so in most cases we have to resort to iterative methods such as the EM algorithm or smooth versions of the EM algorithm, see Silverman et al. (1990). For two choices of K explicit expressions for NPMLE's of F are available. If K is the standard exponential df the NPMLE is derived by Vardi (1989) and Grübel (1990). The other case where an explicit expression is available is *Uniform Deconvolution.* If K is equal to the uniform $[0,1)$ df, under an additional support restriction on F, an NPMLE is given in Van Es (1990) and Van Es and Van Zuijlen (1990). A nice feature of the Uniform Deconvolution problem is that it is a problem where in certain cases the NPMLE of the df F is explicitly known and that it is an inverse problem where the inversion of the transformation which maps the df F on the df G of the data is extremely simple. Since it is easy to invert this transformation, inversion type estimators can also be computed in this problem. It is our purpose to compare the NPMLE with two inversion estimators of F.

Since we are estimating a distribution function we only consider estimates which are non-decreasing and right-continuous. All the estimates in this paper are essentially equal to some "monotone derivative of a step function" H_n, which estimates the primitive of F. We obtain these "monotone derivatives of step functions" by computing the derivatives of their convex minorants on some interval $[0,T]$. By the convex minorant of a function H_n on $[0,T]$ we mean

$$\hat{H}_n(x) = max\{f(x) : f : [0,T] \to \mathrm{R} \textit{ is convex and } \quad f \leq H_n\}, \quad x \in [0,T].$$

Note that the convex minorant of a step function is a piecewise linear convex function, so its right-continuous derivative is again a right-continuous non-decreasing step function. These estimators based on convex minorant derivatives are related to the Grenander maximum likelihood estimator of a monotone density (see Grenander (1956), Prakasa Rao (1969), Groeneboom (1985) and Birgé (1989)), and to maximum likelihood estimators in certain interval censoring problems (see Groeneboom (1987), (1990)).

In Section 2 we discuss Uniform Deconvolution. Subsection 2.1 is devoted to nonparametric maximum likelihood estimation while Subsection 2.2 introduces two inversion type estimators. Both subsections contain the pointwise asymptotic distribution of the estimators. The estimators are compared in Section 3.

2. Uniform Deconvolution

Let K be equal to the uniform $[0,1)$ distribution function, i.e. $K(x) = \int_{-\infty}^{x} k(u)du$, where $k(u) = I_{[0,1)}(u)$ denotes the uniform $[0,1)$ density. In this deconvolution problem the observations have a density g, given by

$$g(x) = \int_{-\infty}^{\infty} k(x-u)dF(u) = F(x) - F(x-1). \tag{2.1}$$

Note that this relation defines a map of the unknown df F onto the density g of the observations.

The following device transforms the convolution sample to a sample from the uniform distribution. Note the absence of conditions on F.

> If $\delta_i = [X_i]$ (the integer part) and $Z_i = X_i - [X_i] = X_i \; mod \, 1$, then $Z_1, ..., Z_n$ is a sample from the uniform $[0, 1)$ distribution.

In a certain sense this procedure can also be reversed. Given a uniform $[0, 1)$ random variable Z with some extra randomization we can construct a random variable X which has the convolution distribution function G.

> If Z is uniformly [0,1) distributed and if, given $Z = z$, we decide with probability $g(z + k)$ whether $\delta = k$, then $X = Z + \delta$ has df G.

For the proofs of these two statements, which are quite elementary, we refer to Van Es and Van Zuijlen (1990).

We need some further notation. If $Z_1, ..., Z_n$ is the uniform sample obtained from the observations $X_1, ..., X_n$ by the procedure above, then $0 \leq Z_{1:n} \leq ... \leq Z_{n:n} \leq 1$ denote its order statistics and U_n its empirical distribution function.

2.1. NONPARAMETRIC MAXIMUM LIKELIHOOD

Let us consider maximum likelihood estimation of F. For K equal to the uniform [0,1) distribution function the loglikelihood (1.1) equals

$$L_n(F) = \sum_{i=1}^{n} log(g(X_i)) = \sum_{i=1}^{n} log(F(X_i) - F(X_i - 1))$$

$$= \sum_{i=1}^{n} \{(1 - \delta_i) log F(Z_i) + \delta_i log(1 - F(Z_i))\}. \tag{2.2}$$

The last equality is valid provided we assume $F(0) = 0$ and $F(1) = 1$. A loglikelihood of this form also appears in certain interval censoring problems. Groeneboom (1987), (1990) derives an NPMLE for these problems. The next theorem gives an NPMLE in our Uniform Deconvolution problem, assuming that the support restriction above is satisfied. The proof is an adapted version of a proof in a preliminary version of Groeneboom (1990). Note that because the loglikelihood depends on the values of F in the points Z_i only, this maximizer is not unique.

Theorem 1. *Let $H_n : \{0, 1, 2, ..., n\} \to [0, \infty)$ be defined by $H_n(0) = 0$, and*

$$H_n(i) = \sum_{j=1}^{n} I_{\{0 \leq Z_j \leq Z_{i:n}, \delta_j = 0\}}, \tag{2.3}$$

for $i = 1, 2, ..., n$. Let $\bar{H}_n : [0, n + 1] \to \infty$ denote the right-continuous step function which has value $H_n(i)$ on the intervals $[i, i + 1)$, for $i = 0, 1, ..., n$, and value $H_n(n)$ at the point $n + 1$. Further let h_n denote the right-continuous derivative of the convex minorant of $\bar{H}_n$ on $[0, n + 1]$. Then

$$F_n^{ML}(u) = h_n(nU_n(u)) \tag{2.4}$$

maximizes the loglikelihood (2.2).

Proof. Let y be an arbitrary distribution function on $[0, n+1]$ which is constant on the intervals $[i, i+1)$, $i = 0, 1, ..., n-1$ and $[n, n+1]$. First we derive a maximizer of the expression

$$\int log(y(u))d\bar{H}_n(u) + \int log(1 - y(u))d(u - \bar{H}_n(u)). \tag{2.5}$$

Because y is non-decreasing and satisfies $\hat{H}_n(x) \leq \bar{H}_n(x)$, on $[0, n+1]$, as well as $\hat{H}_n(n+1) = \bar{H}_n(n+1)$, we have

$$\int log(y(u))d\bar{H}_n(u) + \int log(1 - y(u))d(u - \bar{H}_n(u))$$

$$\leq \int log(y(u))d\hat{H}_n(u) + \int log(1 - y(u))d(u - \hat{H}_n(u))$$

$$= \int log(y(u))h_n(u)du + \int log(1 - y(u))(1 - h_n(u))du,$$

where $\hat{H}_n$ denotes the convex minorant of $\bar{H}_n$ on $[0, n+1]$. Note that we have $0 \leq h_n \leq 1$. This expression is maximal for y equal to h_n. This follows from the "inequality of the means" ($log(0) = -\infty$ by convention)

$$h_n(u)logx + (1 - h_n(u))logy \leq log\{xh_n(u) + y(1 - h_n(u))\}, \text{for } x \geq 0, y \geq 0$$

applied with x equal to $y(u)/h_n(u)$ and y equal to $(1 - y(u))/(1 - h_n(u))$. Summarizing we have

$$\int log(y(u))d\bar{H}_n(u) + \int log(1 - y(u))d(u - \bar{H}_n(u))$$

$$\leq \int log(h_n(u))d\hat{H}_n(u) + \int log(1 - h_n(u))d(u - \hat{H}_n(u)).$$

Since $\bar{H}_n$ and $\hat{H}_n$, as well as $u - \bar{H}_n$ and $u - \hat{H}_n$, assign equal mass to the intervals of maximal length where h_n is constant, equality is attained for y equal to h_n. This shows that h_n maximizes (2.5).

Now note that the expression (2.5) equals

$$\sum_{i=1}^{n}\{(1 - \Delta_i)log(y(i)) + \Delta_i log(1 - y(i))\},$$

where Δ_i denotes the δ_i belonging to $Z_{i:n}$, and that the loglikelihood (2.2) equals

$$\sum_{i=1}^{n}\{(1 - \Delta_i)log(F(Z_{i:n})) + \Delta_i log(1 - F(Z_{i:n}))\}.$$

Therefore the loglikelihood is maximized by taking $F(Z_{i:n})$ equal to $h_n(i)$. The function F_n^{ML} defined in (2.4) satisfies this requirement. □

The pointwise asymptotic distribution of the NPMLE is given by the next theorem. For a proof see Van Es and Van Zuijlen (1990). The asymptotics are similar to the asymptotics of the Grenander estimator of a monotone density.

Theorem 2. *Let u_0 be such that $0 < F(u_0) < 1$ and let F be differentiable at u_0, with strictly positive derivative $f(u_0)$. Then we have, as $n \to \infty$,*

$$n^{1/3}\frac{F_n^{ML}(u_0) - F(u_0)}{\{4f(u_0)F(u_0)(1 - F(u_0))\}^{1/3}} \xrightarrow{\mathcal{D}} Z, \tag{2.6}$$

where Z is the last time that two-sided Brownian motion minus the parabola $y(t) = t^2$ reaches its maximum.

The assumptions $F(0) = 0$ and $F(1) = 1$ are crucial to obtain the simpler likelihood (2.2), for which a maximizer is explicitly known. In cases where this assumption is not satisfied there are no explicit expressions for NPMLE's. However, the methodology of the NPMLE given by Theorem 1 can be extended to distribution functions F with arbitrary bounded support. Van Es and Van Zuijlen (1990) construct estimators in those more general cases which have the same asymptotic distribution as the NPMLE above, i.e. for these estimators Theorem 2 also holds. It should be noted however that the estimators do not maximize the likelihood if the support restriction on F is not satisfied.

2.2. TWO ESTIMATORS BASED ON INVERSION

Since the distribution of the observations is a transformation of the distribution function F which we want to estimate, Uniform Deconvolution can be seen as an inverse or ill-posed problem. The transformation which maps the df F onto the density g of the observations is given by (2.1). Inverting this transformation we get

$$F(x) = g(x) + g(x-1) + g(x-2) + ..., \tag{2.7}$$

and therefore

$$\int_0^x F(u)du = G(x) + G(x-1) + G(x-2) + ... \tag{2.8}$$

This shows that we can easily obtain root n consistent estimators of the primitive of F. Given such an estimate we can compute a monotone estimate of F itself, using the convex minorant of the estimate of its primitive. Assuming that we know that $F(0) = 0$ and $F(T) = 1$ for some $T > 0$, we can define an inversion estimator $F_n^{I,1}$ by replacing G by G_n, the edf of the sample, and by taking $F_n^{I,1}$ equal to the right-continuous derivative of the convex minorant on $[0, T]$ of the step function

$$H_n^{I,1}(x) = G_n(x) + G_n(x-1) + G_n(x-2) + ... \tag{2.9}$$

A similar estimator, $F_n^{I,2}$ say, can be constructed from the sample $-X_1, ..., -X_n$. It turns out that these estimators have an asymptotic distribution similar to the asymptotic distribution of the NPMLE in the previous section.

Theorem 3. *Let F be stricly increasing on $[0, T]$, with $F(0) = 0$ and $F(T) = 1$, for some $T > 0$. Further, let u_0 be such that $0 < F(u_0) < 1$ and that F is differentiable at u_0, with strictly positive derivative $f(u_0)$. Then we have , as $n \to \infty$,*

$$n^{1/3}\frac{F_n^{I,1}(u_0) - F(u_0)}{\{4f(u_0)F(u_0)\}^{1/3}} \xrightarrow{\mathcal{D}} Z, \tag{2.10}$$

and

$$n^{1/3}\frac{F_n^{I,2}(u_0)-F(u_0)}{\{4f(u_0)(1-F(u_0))\}^{1/3}}\xrightarrow{\mathcal{D}} Z, \tag{2.11}$$

where Z is the last time that two-sided Brownian motion minus the parabola $y(t)=t^2$ reaches its maximum.

Proof. We give the proof of (2.10) for the case where there exists an integer k such that $0<k<u_0<k+1<T$. The other cases can be treated similarly.

Note that by the assumptions of the theorem we have $F(u)<1$ on $[k,k+1)$, which implies $\int_k^{k+1}F(u)du<1$. Let $t_0>k+1$ be such that $\int_k^{t_0}F(u)du=1$. We define random variables $T_1,...,T_n$ by

$$T_i=k+Z_i \quad \text{if} \quad 0\le X_i<k \quad \text{and} \quad T_i=V_i \quad \text{otherwise},$$

where $V_1,...,V_n$ are independent of $X_1,...,X_n$ with

$$P(V_i\le v)=\frac{\int_{k+1}^{v}F(u)du}{1-\int_k^{k+1}F(u)du} \quad \text{, for} \quad k+1\le v\le t_0.$$

Now, for $k\le t<k+1$, we have

$$\begin{aligned}P(T_i\le t)&=P(0\le X_i\le t-k)+P(1\le X_i\le 1+t-k)+...+P(k\le X_i\le t)\\&=\int_0^{t-k}\{F(u)-F(u-1)\}du+\int_1^{1+t-k}\{F(u)-F(u-1)\}du\\&\quad ...+\int_k^{t}\{F(u)-F(u-1)\}du=\int_k^t F(u)du,\end{aligned}$$

and for $k+1\le t<t_0$

$$\begin{aligned}P(T_i\le t)&=P(T_i<k+1)+P(k+1\le T_i\le t)\\&=\int_k^{k+1}F(u)du+P(T_i\ge k+1)P(k+1\le T_i\le t|T_i\ge k+1)\\&=\int_k^{k+1}F(u)du+P(X_i\ge k+1)P(k+1\le V_i\le t)=\int_k^t F(u)du.\end{aligned}$$

This shows that T_i has distribution function $P(T_i\le t)=\int_k^t F(u)du$ and density $F(t)$ on $[k,t_0)$. Let H_n^T denote the edf of the sample $T_1,...,T_n$.

We complete the proof of (2.10) by arguing below that the asymptotic distribution of $F_n^{I,1}(u_0)$ is the same as the asymptotic distribution of $h_n^T(u_0)$, where h_n^T denotes the Grenander estimator of the increasing density F of $T_1,...,T_n$. Recall that this maximum likelihood estimator of an increasing density on an interval is equal to the right-continuous derivative of the convex minorant of the edf of the sample. Since the results of Prakasa Rao (1969) and Groeneboom (1985) concerning the pointwise asymptotic distribution of the Grenander estimator in our case imply that (2.10) holds with $F_n^{I,1}(u_0)$ replaced by $h_n^T(u_0)$, the proof is then complete.

The asymptotic distributions of $F_n^{I,1}(u_0)$ and $h_n^T(u_0)$ are the same because on the interval $[k,k+1)$ both step functions H_n^I and H_n^T jump only at the points T_i which belong to the interval.

Moreover these jumps are of equal size. This means that if we could disregard the values of the step functions outside the interval, then the derivatives of the convex minorants of H_n^I and H_n^T would be the same on the interval $[k, k+1)$. Indeed, it is known that the asymptotic behavior of convex minorant derivatives of processes like the ones we consider here, is determined by the local behavior of these processes. More specifically, it is determined by the processes on intervals of length of order $n^{-1/3}$ around the point u_0. For examples see Groeneboom (1985), (1987) and Van Es and Van Zuijlen (1990). This shows that the asymptotic distributions of $F_n^{I,1}(u_0)$ and $h_n^T(u_0)$ are indeed the same. □

3. Discussion

Considering the asymptotic distributions, given by Theorems 2 and 3, of the estimators presented in the previous sections, a striking feature is the cube root n rate of convergence, which differs from the usual root n rate of convergence. However, this is not an unusual rate of convergence. For instance, the Grenander estimator of a monotone density also shares this rate of convergence. It is shown in Birgé (1989) that the cube root n rate is the optimal rate for estimating a monotone density. In the Uniform Deconvolution problem, from a minimax point of view, this rate of convergence is the optimal one for estimating the distribution function F, see Fan (1988) and Chapter 4 in Van Es (1990).

Although the estimators all have a cube root n rate of convergence the asymptotic variance of the maximum likelihood estimator turns out to be smaller than the asymptotic variances of the two inversion type estimators considered. This shows that, even if we compute inverse values of good estimators of the distribution of the data, such as the empirical distribution function of the sample, we can get estimators which behave suboptimally. In the Uniform Deconvolution problem we gain in precision by using the maximum likelihood estimator instead.

Acknowledgement.
Part of this work was prepared with the support of the Netherlands Organization for Scientific Research.

References

[1] Birgé, L. (1989) 'The Grenander estimator: A nonasymptotic approach', Ann. Statist. 17, 1532-1549.

[2] Carroll, R.J. and P. Hall (1988) 'Optimal rates of convergence for deconvolving a density', J. Amer. Statist. Assoc. 83, 1184-1186.

[3] Carroll, R.J., Van Rooij, A.C.M. and F.H. Ruymgaart (1990) 'Theoretical aspects of ill-posed problems in statistics', Report 9007, Catholic University Nijmegen, The Netherlands.

[4] Devroye, L. (1989) 'Consistent deconvolution in density estimation', Canadian J. Statist. 17, 235-239.

[5] Van Es, A.J. (1990) 'Aspects of Nonparametric Density Estimation', CWI tract 77, CWI Amsterdam.

[6] Van Es, A.J. and M.C.A. van Zuijlen (1990) 'Convex minorant estimators of distributions in nonparametric deconvolution problems', Report 9020, Catholic University Nijmegen, The Netherlands.

[7] Fan, J (1988) 'On the optimal rates of convergence for nonparametric deconvolution problem',

Stat. Techn. Rept. 157, University of California at Berkeley.

[8] Grenander, U. (1956) 'On the theory of mortality measurement , Part II', Skand. Akt. 39, 125-153.

[9] Groeneboom, P. (1985) 'Estimating a monotone density', in L.M. LeCam and R.A. Olshen (eds.), Proceedings of the Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer, Wadsworth, Hayward, 539-555.

[10] Groeneboom, P. (1987) 'Asymptotics for incomplete censored observations', Report 87-18, University of Amsterdam, The Netherlands.

[11] Groeneboom, P. (1990) 'Asymptotics for interval censored observations', Report Delft University of Technology, in preparation.

[12] Grübel, R. (1990) 'Two properties of uniform distributions, with applications to some nonparametric estimation problems', Preprint Delft University of Technology, Delft, The Netherlands.

[13] Liu, M.C. and R.L. Taylor (1989) 'A consistent nonparametric density estimator for the deconvolution problem', Canadian J. Statist. 17, 427-438.

[14] Liu, M.C. and R.L. Taylor (1990) 'Simulations and computations of nonparametric density estimates for the deconvolution problem', J. Stat. Comp. & Simul. 35, 145-167.

[15] O'Sullivan, F. (1986) 'A statistical perspective on ill-posed inverse problems', Statistical Science 1, 502-527.

[16] Prakasa Rao, B.L.S. (1969) 'Estimation of a unimodal density ', Sankya Ser. A 31, 23-36.

[17] Silverman, B.W., Jones, M.C., Wilson, J.D. and D.W. Nychka (1990), 'A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography (with Discussion)', J. R. Statist. Soc. B 52, 271-324.

[18] Stefanski, L.A. (1990) 'Rates of convergence of some estimators in a class of deconvolution problems', Statist. and Probab. Lett. 9, 229-235.

[19] Stefanski, L.A. and R.J. Carroll (1987) 'Deconvoluting kernel density estimators', Statistics, to appear.

[20] Stefanski, L.A. and R.J. Carroll (1989) 'Deconvolution and score tests in generalized linear measurement error models', Ann. Statist., to appear.

[21] Vardi, Y. (1989) 'Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation', Biometrika 76, 751-761.

III. Parameter Selection, Smoothing

SMOOTHING PARAMETER SELECTION IN IMAGE RESTORATION

K. P.- S. CHAN and J.W. KAY
Department of Statistics,
University of Glasgow,
Glasgow, G12 8QQ
Scotland, U.K.

ABSTRACT. We consider the problem of the automatic selection of the smoothing parameter in image restoration using the method of regularisation. We consider two new smoothing parameter selectors based on the estimation cross-validation function and compare their performance with some others proposed in the literature and also with some optimal methods. These selectors are assessed with regard to their ability to restore two-dimensional images which have been subjected to different degrees of blur and noise. For most blur-noise combinations, the new methods produced satisfactory restorations except when the blur and noise were both high. The existing methods produced seriously undersmoothed restorations in the high/medium blur and low noise combinations. Although the simulated distributions of the estimated smoothing parameter often did not overlap with the optimal distribution, nevertheless, they were often within the window of acceptable values. The restorations obtained by minimising the estimation and prediction errors, respectively, were of a similar quality which suggests that the choice of error criterion is not a practical issue for the restoration of large degraded images.

1. Introduction

The problem of image restoration is to estimate an unknown function x(s) given data in the form of a function y(t). The data are assumed to have arisen from the model

$$y(t) = \int G(t,s)x(s)ds + \epsilon(t), \tag{1}$$

where G is a blurring function and ϵ is a signal-independent random noise process. We assume that G is known. Two-dimensional images usually consist of a large square array of pixels, typically 2^{12}–2^{20} in number, and an observation is available at each pixel. In the sequel we develop the material using a discrete formulation of model (1)

$$y = Hx + \epsilon$$

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 201–211.
© 1991 Kluwer Academic Publishers.

where y, x and ϵ are n-dimensional vectors, H is a known n×n matrix and the components of ϵ are independent $N(0, \sigma^2)$ random variables. Note that we are using one-dimensional notation which is made possible by stacking the images in lexicographic order; see Gonzales and Wintz (1987)and Kay (1988) for further details. Image restoration is an example of an ill-posed problem where the maximum-likelihood solution is highly oscillatory in nature. The solution process requires to be stabilised and this is achieved using the method of regularisation which reduces variability at the expense of introducing bias into the estimate of x; see Poggio et al (1985) for further details.

Employing this method with a quadratic roughness penalty, we take our estimate of x to be the solution of the optimisation problem

$$\inf_{x} \left\{\frac{1}{n} \|y - Hx\|^2 + \lambda\, x^T C x\right\} \tag{2}$$

where C is a non-negative definite smoothing matrix. The parameter λ is termed a smoothing parameter because it controls the compromise between fidelity to the data and the smoothness of the estimate of x; the extremes $\lambda = 0$ and $\lambda = \infty$ yield, respectively, an ultra-noisy, unstable solution and a constant, over-smoothed estimate of x. The purpose of this article is to discuss some methods for the automatic selection of λ. In section 2 we express the problem in a canonical form. We introduce two new methods for the selection of λ in section 3 based on the idea of estimation cross-validation. In section 4 we describe the other methods used in our numerical experiments. The practical performance of the different selectors is discussed in section 5 and in section 6 we present our conclusions.

2. Canonical Structure

In image restoration it is often realistic to assume that the matrices H and C in (2) have block-circulant structure. Hence they are simultaneously diagonalised by the block Fourier matrix (Gonzales and Wintz (1987)) and so we may employ two-dimensional FFTs to perform the computation and to analyse the smoothing parameter selectors. We denote by $\{h_i\}$, $\{c_i\}$ and $\{u_i\}$ the eigenvalues of H and C respectively and the common eigenvectors. The problem (2) has solution ('*' denotes complex conjugate transpose)

$$\hat{x}(\lambda) = (H^T H + n\lambda C)^{-1} H^T y$$

$$= \sum_{i=1}^{n} \frac{h_i^* (u_i^* y)}{(|h_i|^2 + n\lambda c_i)} u_i$$

Some other useful quantities are

$$RSS(\lambda) = \frac{1}{n} \|y - H\hat{x}(\lambda)\|^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{n^2\lambda^2 c_i^2 |u_i^* y|^2}{(|h_i|^2 + n\lambda c_i)^2}$$

$$t(\lambda) = \frac{1}{n} \operatorname{tr}[H(H^TH + n\lambda C)^{-1}H^T] = \frac{1}{n} \sum_{i=1}^{n} \frac{|h_i|^2}{(|h_i|^2 + n\lambda c_i)}$$

In order to assess the ability of the various smoothing parameter selectors to produce a satisfactory restoration, we consider some optimal measures of performance which are clearly not available in practice. The mean squared errors for estimation and prediction are defined by

$$\begin{aligned} EMSE(\lambda) &= n^{-1}E\,(||\hat{x}(\lambda) - x||^2) \\ &= n^{-1} \sum_{i=1}^{n} \frac{(n\lambda c_i)^2 |u_i^* x|^2}{(|h_i|^2+n\lambda c_i)^2} + n^{-1}\sigma^2 \sum_{i=1}^{n} \frac{|h_i|^2}{(|h_i|^2+n\lambda c_i)^2} \end{aligned}$$

and
$$\begin{aligned} PMSE(\lambda) &= n^{-1}\, E(||H\hat{x}(\lambda) - Hx||^2) \\ &= n^{-1} \sum_{i=1}^{n} \frac{|h_i|^2 (n\lambda c_i)^2 |u_i^* x|^2}{(|h_i|^2+n\lambda c_i)^2} + n^{-1}\sigma^2 \sum_{i=1}^{n} \frac{|h_i|^4}{(|h_i|^2+n\lambda c_i)^2} \end{aligned}$$

We also consider the actual average squared errors for estimation and prediction defined by

$$AESE(\lambda) = n^{-1}\, ||\hat{x}(\lambda) - x||^2$$

$$APSE(\lambda) = n^{-1}\, ||H\hat{x}(\lambda) - Hx||^2.$$

3. Estimation Cross-Validation

Methods for smoothing parameter selection such as GCV and many others (Golub et al (1979); Härdle et al (1988)) aim to minimise the prediction error, whereas, as pointed out by Rice (1986), it is more appropriate to minimise the estimation error in problems involving deconvolution. He proposed a method which requires to use an estimate of σ^2. GCV was developed as a rotationally-invariant form of the cross-validation function for prediction

$$n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_{-i}(\lambda))^2$$

where $\hat{y}_{-i}(\lambda)$ is the estimate of $E(y_i)$ obtained when the ith observation is omitted. We therefore consider the cross-validation function for estimation defined by

$$ECV(\lambda) = n^{-1} \sum_{i=1}^{n} (\hat{x}_i(\lambda;\, D) - \hat{x}_i(\lambda;\, D_{-i}))^2$$

where $\hat{x}_i(\lambda;\, D)$ and $\hat{x}_i(\lambda;\, D_{-i})$ denote, respectively, the estimate of the ith component of x, x_i, based on all of the data and based on all of the data except the ith observation. Now using some standard matrix results (Kay (1990b)) it follows that

$$ECV(\lambda) = n^{-1} \sum_{i=1}^{n} w_i(\lambda)\ (y_i - H\hat{x}_i(\lambda))^2$$

where $w_i(\lambda) = \dfrac{a_{2ii}(\lambda)}{(1 - a_{1ii}(\lambda))^2}$ and $a_{rii}(\lambda)$ denotes the (i,i)th entry of $K_r(\lambda) = H(H^TH + n\lambda C)^{-r}H^T$. In the case where H and C are circulant,

$$ECV(\lambda) = n^{-1}\ tr[K_2(\lambda)]\ GCV(\lambda),$$

which does not have suitable theoretical properties unlike the following criterion.

$$ECV1(\lambda) = n^{-1} \sum_{i=1}^{n} \frac{|h_i|^2\ n^2\lambda^2 c_i^2 |u_i^* y|^2}{(|h_i|^2 + n\lambda c_i)^4} /(1-t(\lambda))^2$$

Here $a_{2ii}(\lambda)$ has been replaced by the ith eigenvalue of $K_2(\lambda)$ and $a_{1ii}(\lambda)$ has been replaced by $n^{-1}Tr[K_1(\lambda)]$. The only reason for considering $ECV1(\lambda)$ is that it leads to a choice for λ which has an asymptotic optimality property. Now when $H = I$, $ECV1(\lambda) \neq GCV(\lambda)$ and so it is interesting to consider also

$$ECV2(\lambda) = n^{-1} \sum_{i=1}^{n} \frac{|h_i|^2 n^2\lambda^2 c_i^2 (1+n\lambda c_i)^2 |u_i^* y|^2}{(|h_i|^2 + n\lambda c_i)^4} /(1-t(\lambda))^2 ,$$

which does reduce to $GCV(\lambda)$ in the 'no-blur' case (H=I). We now define the other choices of λ used in our experiments.

4. Smoothing Parameter Selectors

In our notation, and ignoring terms which don't involve λ, the criteria proposed by Rice (1986) are defined by

$$RICE1(\lambda) = 2n^{-1}\sigma^2 \sum_{i=1}^{n} \frac{|h_i|^2}{(|h_i|^2 + n\lambda c_i)}$$

$$- n^{-1} \sum_{i=1}^{n} \frac{(|h_i|^4 + 2n\lambda |h_i|^2 c_i) |u_i^* y|^2}{(|h_i|^2 + n\lambda c_i)^2}$$

$$RICE2(\lambda) = 2n^{-1}\sigma^2 \sum_{i=1}^{n} \frac{1}{(|h_i|^2 + n\lambda c_i)}$$

$$- n^{-1} \sum_{i=1}^{n} \frac{(|h_i|^2 + 2n\lambda c_i) |u_i^* y|^2}{(|h_i|^2 + n\lambda c_i)^2} .$$

The criterion $RICE1(\lambda)$ is actually equivalent to Mallow's C_L (Mallows (1973)). Both criteria produce asymptotically optimal values of λ in

the sense that the minimisers of their expectations are the minimisers of PMSE(λ) and EMSE(λ), respectively. In order to apply them we must estimate σ^2. We also consider

$$\mathrm{GCV}(\lambda) = \frac{\mathrm{RSS}(\lambda)}{(1 - t(\lambda))^2}$$

which produces an asymptotically optimal λ in terms of the prediction error. It is shown by Kay (1990a) that the minimisers of PMSE(λ) and EMSE(λ) are different, asymptotically, in that they tend to zero at different rates and with different constants. So we might expect the choice of error criterion to be important. The criterion ECV1(λ) produces an asymptotically optimal λ in terms of the estimation error (Kay (1990b)) and so is theoretically comparable with RICE2(λ) with the advantage that an estimator for σ^2 is not required. We now proceed to describe our experiments and to discuss the comparative performance of the selectors. Where an estimate of σ^2 was required, we used the estimator proposed by Kay (1988) which is based on local neighbourhood contrasts of the observed image. This estimator is consistent, provided that Hx is bounded, and performs well when n is large and/or the signal-to-noise ratio is small.

5. A Simulation Study

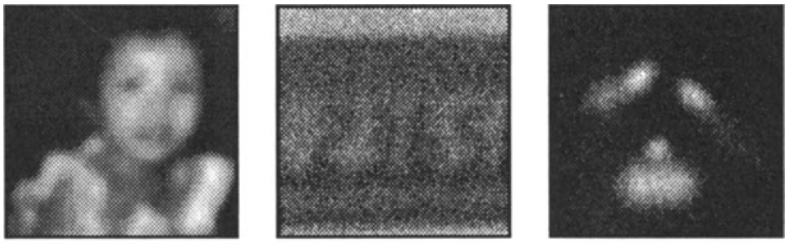
5.1 *Description of the study*

The objective of this small simulation study is to examine the behaviour of the new methods proposed in Section 3 and to compare them with some existing methods described in Section 4, in terms of quality of restoration of some degraded images.

In this study we have chosen three test images, two of which are of size 128×128 and were obtained by frame-grabbing a photograph of a girl and an object which is a calculator, respectively. The third image is of size 64×64 and was made up from a mixture of seven bivariate Gaussian distributions. These images are known as 'Kayin', 'Calc' and 'Astro', respectively, hereafter. Each image was blurred using three different blur matrices. The entries of the blur matrices were acquired from the function values of a bivariate Gaussian distribution with covariance matrix $\tau^2 I$. The size of the blur was fixed at 5×5, 9×9 and 15×15, which we refer to as small, medium and large blur, respectively, so that the values of τ are approximately 0.9, 1.8 and 3.15. Then, each blurred image was degraded further by adding independent Gaussian noise. The noise values were chosen to be 1 and 10 which represent low and high noise levels, respectively. Thus the degradation process applied to the images can be summarised in the following scheme:

		Noise	
		low	high
	small	B1S1	B1S2
Blur	medium	B2S1	B2S2
	large	B3S1	B3S2

Examples of some of the degraded images are shown in Fig. 1. Using an 'eight nearest neighbours' system with wrap-around, the entries of Q were taken as $q_{ii} = 8$ and $q_{ij} = -1$, if pixel j is a neighbour of pixel i, where $C = Q^T Q$. In the study, the true noise was used; in addition a noise estimate, namely the neigbourhood noise estimate, was used with the RICE1 and RICE2 methods.



(I) (II) (III)

Figure 1. Some degraded images of (I) 'Kayin', (II) 'Calc' and (III) 'Astro', which were distorted by B3S1, B3S2 and B1S2, respectively.

A criterion is required to measure discrepancies between the true image x and its estimate $\hat{x}$ obtained from various methods described earlier. The Mean Absolute Difference (MAD) was used and is defined as

$$\frac{1}{n} \sum_{i=1}^{n} |x_i - \hat{x}_i| .$$

There is a certain window of λ values within which the restoration of a degraded image is acceptable. In the case of small blur, this window of λ values is wide and hence the choice of λ is less crucial. However in the medium/large blur situation, especially in the presence of high noise the window of λ values is narrow so the λ value is important in order to achieve a reasonable restoration; see Fig. 2.

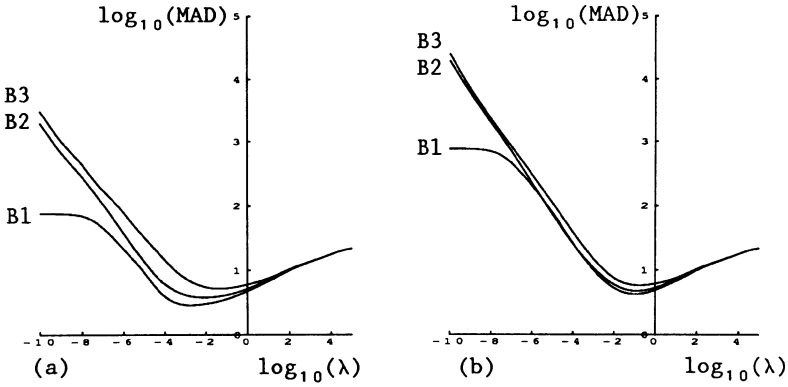


Figure 2. Plot of $\log_{10}(\text{MAD})$ against $\log_{10}(\lambda)$ under (a) low noise and (b) high noise with different blurs for a realisation of 'Kayin'.

5.2 *Results of the simulation study*

The results presented here are based on 100 simulations. An example of the distributions of $\hat{\lambda}$ which were obtained from the simulation are summarised in Fig 3. The following conclusions can be drawn from the simulation. First of all, the medians of the $\hat{\lambda}$s are generally larger in the high noise case than in the low noise case which implies that more smoothing is required in the presence of high noise. For all test images, APSE, AESE, PMSE and EMSE produce comparable estimates for λ in all blur and noise combinations. In general, all the $\hat{\lambda}$s obtained from the methods are comparable in the small blur situation and lie within the window of acceptable values of λ and hence the restorations are very similar. Examples of some of the restorations, given the small blur and high noise combination, are shown in Fig. 4. The estimates of λ obtained from the ECV1 and ECV2 methods are similar to those estimates obtained by the mean squared error criteria, except in the presence of large blur where the ECV1 and ECV2 estimates of λ are relatively large. Hence restorations using the λ produced by the ECV1 and ECV2 methods are, in general, slightly oversmoothed (see Fig. 5 (II) and (III)).

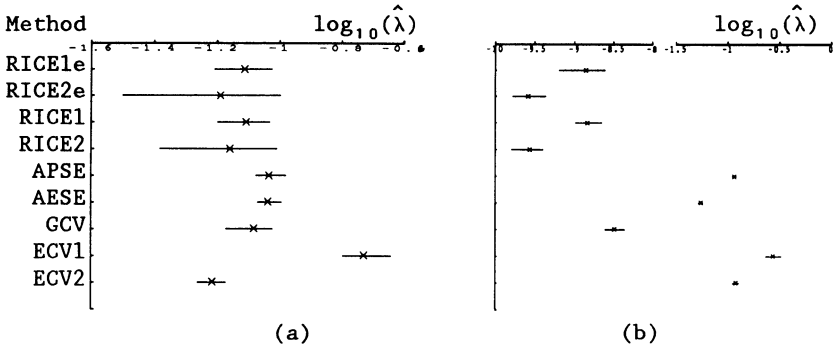
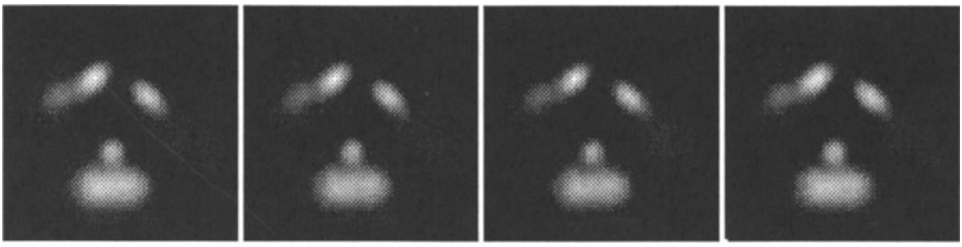


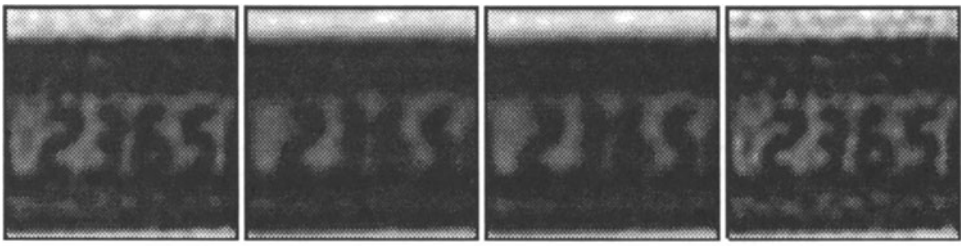
Figure 3. Distributions of $\hat{\lambda}$ obtained for the 'Kayin' image under (a) B1S2 and (b) B3S1. The median is marked by a '×' and the straight line gives the range of values of $\log_{10}(\hat{\lambda})$. An acceptable window of $\log_{10}(\hat{\lambda})$ is approximately (-1.75,0.0) in case (a) and (-1.5,-1.0) in case (b).

For the 'Kayin' and 'Calc' images, in the presence of high noise, the differences in order of magnitude of the medians of $\hat{\lambda}$ between the methods are not great. In contrast, the RICE1, RICE2, GCV methods produce much smaller values of $\hat{\lambda}$ than the other methods in the medium/large blur and low noise combinations. These small $\hat{\lambda}$ values given by these fore-mentioned methods reflect the poor quality of the restorations (see Fig. 6 for example). The median $\hat{\lambda}$s of the GCV, RICE1 and RICE2 methods decrease as the blur increases. For the 'ASTRO' artifical image, the RICE1 and GCV methods now produce comparable values of λ in the medium/large blur and low noise situations but the restorations using these methods still tend to be undersmoothed. In the high noise situations the medians of $\hat{\lambda}$ are still similar amongst the methods. The range of $\hat{\lambda}$ obtained by the RICE2 method in the medium and



(I) λ=0.089 (II) λ=0.049 (III) λ=0.039 (IV) λ=0.056

Figure 4. Restoration of B1S2 degraded 'Astro' image using (I) AESE, (II) ECV1, (III) ECV2 asnd (IV) GCV methods.



(I) λ=0.20 (II) λ=1.51 (III) λ=1.07 (IV) λ=0.085

Figure 5. Restoration of B3S2 degraded 'Calc' image using (I) AESE, (II) ECV1, (III) ECV2 and (IV) GCV methods.

(I) λ=0.053 (II) λ=0.12 (III) λ=3.5e-9 (IV) λ=2.8e-10

Figure 6. Restoration of B3S1 degraded 'Kayin' image using (I) AESE, (II) ECV2, (III) GCV and (IV) RICE2e methods.

large blur situations, with either the true or estimated noise, are extremely wide. This suggests that RICE2 can produce inconsistent estimates of λ. In addition the RICE2 method is extremely sensitive to the choice of noise estimate. For example, with the 'ASTRO' image in the medium blur and low noise situation, a scatter plot of the $\hat{\lambda}$s with the true noise against the $\hat{\lambda}$s with neigbourhood noise estimates are shown in Fig. 7(a). In one of the realisations, circled in the figure, RICE2 with the true noise estimate produced a λ value of 0.136e-7, whereas with the neigbourhood noise estimate, the value of λ was 0.395e-1. Fig. 7(b) shows a plot of the RICE2 criterion as a function λ using the true and the neigbourhood noise estimates. It shows clearly that the two minima of the RICE2 method using the true noise and estimated noise are quite different. On other occasions, using the true noise can produce a sensible value of λ and vice versa. This phenomenon is also found in the 'Kayin' and 'Calc' images.

Some cases of multiple minima were found in 2% of the simulations, with 'Kayin' and 'Calc' images for some blur-noise combinations. The methods concerned are GCV and ECV1. An example of these is illustrated here by plotting their functions. Fig. 8 shows the function of GCV and ECV1, together with the function of the mean squared error methods, RICE1 and ECV2, plotted against a range of λ in the small blur and low noise situation. Both the function values and the $\hat{\lambda}$s are in logarithms to base 10. The two minima can be easily seen for the ECV1 function. The function of GCV is flat near the region of the minimum; the inset picture shows where the two minima are. The function of RICE1 is also generally flat near the minimum in the presence of low noise. In addition, the RICE1 function is insensitive to the choice of noise estimate. Note also the similarities between the GCV, PMSE and APSE functions. The ECV2 function is similar to the EMSE and AESE functions. The ECV1 function behaves very similarly to ECV2 to start with, but as λ increases the function becomes very unstable around the local minimum and eventually it decreases monotonically.

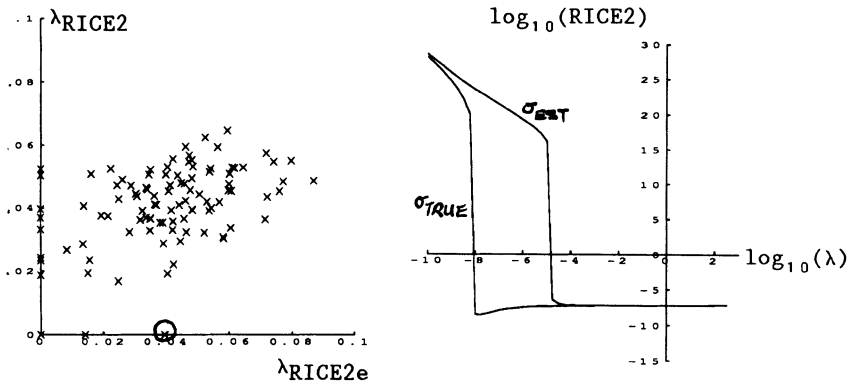


Figure 7. (a) Scatter plot of the $\hat{\lambda}$s obtained by RICE2 method with true noise against the $\hat{\lambda}$s obtained using the estimated noise. (b) Plot of RICE2 criterion as a function of λ, given the true or estimated noise.

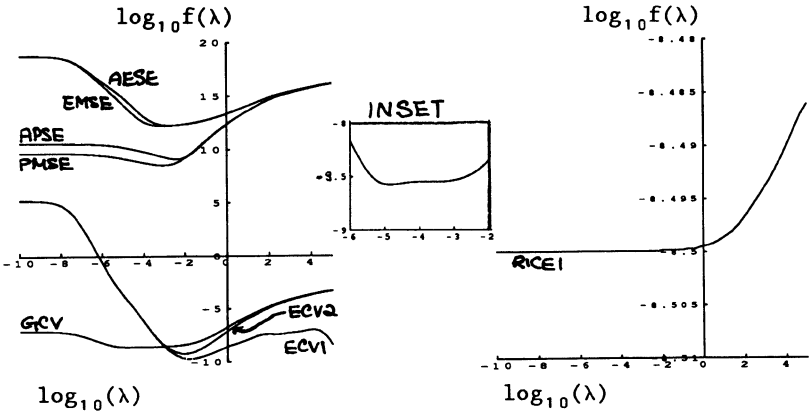


Figure 8. Plot of the smoothing parameter selection criteria as a function of λ.

6. Conclusions and discussion

In the simulation study, we showed that PMSE, EMSE, APSE and AESE perform well in all blur and noise combinations. This suggests that the choice as to whether to base the estimation of λ on a prediction or estimation criterion does not matter in practice since the value of λ and the quality of the restorations are similar for both criteria. However there is a huge difference between the GCV and the ECV methods of choosing the smoothing parameter λ in the presence of medium/large blur and low noise situations. ECV1 and ECV2 methods provide reasonable values of λ for all three test images (Fig. 9) under most of the blur and noise combinations. In one or two exceptions they can, especially ECV1, produce a large value of λ and hence a restoration which is slightly oversmoothed, particularly, in the large blur situation. The ECV2 function behaves rather like the EMSE and AESE functions, whereas the ECV1 function can occasionally be rather unstable around local minimum and can produce multiple minima as shown in the simulation study. ECV2 produce reliable estimates for λ in the simulation study for all test images.

RICE1, RICE2 and GCV can produce $\hat{\lambda}$s which are far too small in the medium /large blur and low noise situations and consequently the restorations using these methods are totally undersmoothed. This was also found for the GCV method by Thompson et al (1990). The breakdown of these methods in the medium/large blur and low noise combination is worth investigating further. Perhaps our choice of blur is too severe for these methods to cope. Or are these methods image dependent, i.e. do they only perform well with less complicated images? Sensitivity of RICE2 to the noise estimate is also worth further investigation.

(I) Kayin (II) Calc (III) Astro

Figure 9. True scene of the three test images.

7. Acknowledgement

We are grateful for financial support under Grant GR/F/32622 of the SERC's Initiative in Complex Stochastic Systems.

8. References

Golub, G.H, Heath, M. and Wahba, G. (1979) 'Generalised cross-validation as a method for choosing a good ridge parameter', Technometrics, 21, pp. 215-223.

Gonzales, R.C. and Wintz, P. (1987) Digital Image Processing, Second Edition, Addison-Wesley.

Hardle, W., Hall, P. and Marron, S.J. (1988) 'How far are automatically chosen regression smoothing parameters from their optimum?', J. Amer. Statist. Assoc., 83, pp. 86-101.

Kay, J.W. (1988) 'On the choice of regularisation parameter in image restoration, Springer Lecture Notes in Computer Science, 301, pp. 587-596.

Kay, J.W. (1990a) 'Asymptotic comparison factors for smoothing parameter choices in regression problems' Revised version submitted.

Kay, J.W. (1990b) 'A new choice of smoothing parameter for deconvolution problems based on estimation cross-validation' In preparation.

Mallows, C.L. (1973) 'Some comments on C_p.' Technometrics, 15, pp. 661-675.

Poggio, T., Torre, V. and Koch, C. (1985) 'Computational vision and regularisation theory, Nature, 317, pp. 314-319.

Rice, J. (1986) 'Choice of smoothing parameter in deconvolution problems', Contemporary Mathematics, 59, pp. 137-151.

Thompson, A.M., Brown, J.C., Kay, J.W. and Titterington, D.M. (1990) 'A study of methods of choosing the smoothing parameter in image restoration by regularisation' Revised version submitted.

ESTIMATING THE QUANTILE-DENSITY FUNCTION

MIKLÓS CSÖRGŐ
Department of Mathematics and Statistics
Carleton University
Ottawa, Ontario, Canada K1S 5B6

PAUL DEHEUVELS
L.S.T.A., Université Paris VI
4 Place Jussieu
75252 Paris Cedex, France

LAJOS HORVÁTH
Department of Mathematics
The University of Utah
Salt Lake City, Utah 84112, U.S.A.

ABSTRACT. Let $X_1, \ldots, X_n$ be i.i.d. r.v.'s with an absolutely continuous distribution function F, quantile function $Q = F^{-1}$, the inverse of F, and strictly positive density function $f = F'$. Define the empirical quantile function Q_n by $Q_n(t) = X_{k,n}$ if $(k-1)/n < t \leq k/n$ $(k = 1, 2, \ldots, n)$, where $X_{1,1} \leq \ldots \leq X_{n,n}$ are the order statistics of the above random sample. We study the problem of estimating the quantile-density function $p(x) = Q'(x) = 1/f(Q(x))$, $0 < x < 1$, by a sequence of kernel-quantile-density estimators p_n, defined as

$$p_n(x) = h_n^{-1} \int_{1/(n+1)}^{n/(n+1)} K((x-t)/h_n) dQ_n(t) = h_n^{-1} \sum_{i=1}^{n-1} K((x - i/n)/h_n)(X_{i+1,n} - X_{i,n}),$$

where $h_n \to 0$, $nh_n \to \infty$, for smooth kernel and quantile functions K and Q.

1. Introduction

Let $X_1, X_2, \ldots$ be i.i.d. r.v.'s with a continuous distribution function F, *quantile function* $Q = F^{-1}$, defined by

$$Q(t) = \inf\{x : F(x) \geq t\}, \quad 0 < t < 1, \tag{1.1}$$

and put $x_F = \sup\{x : F(x) = 0\}$, $x^F = \inf\{x : F(x) = 1\}$, $-\infty \leq x_F < x^F \leq \infty$. Thus $F(Q(t)) = t \in (0,1)$, and if we put $U_1 = F(X_1)$, $U_2 = F(X_2), \ldots$, then these r.v.'s are independent uniform-(0,1) r.v.'s. We define the *empirical quantile function* Q_n by

$$Q_n(t) = X_{k,n}, \quad (k-1)/n < t \leq k/n, \quad (k = 1, \ldots, n), \tag{1.2}$$

where $X_{1,n} \leq \cdots \leq X_{n,n}$ are the order statistics of the first n i.i.d. r.v.'s $X_1, \ldots, X_n$ and, accordingly, the *uniform empirical quantile function* U_n by

$$U_n(t) = U_{k,n} = F(X_{k,n}), \quad (k-1)/n < t \leq k/n, \ (k = 1, \ldots, n). \tag{1.3}$$

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 213–223.
© 1991 *Kluwer Academic Publishers.*

Let $f = F'$ be the density function of the distribution function F, and define the *quantile process* ρ_n by

$$\rho_n(t) = n^{1/2} f(Q(t))(Q_n(t) - Q(t)), \quad 0 < t < 1, \tag{1.4}$$

and the corresponding *uniform quantile process* u_n by

$$u_n(t) = n^{1/2}(U_n(t) - t), \quad 0 < t < 1. \tag{1.5}$$

M. Csörgő and Révész (1978) proved

THEOREM A. *Let $X_1, X_2, \ldots$ be i.i.d. r.v.'s with a continuous distribution function F and assume*

(a) *F is twice differentiable on (x_F, x^F),*

(b) *$F'(x) = f(x) > 0$ on (x_F, x^F),*

(c) *for some $0 < \gamma < \infty$ we have*

$$\sup_{0<t<1} t(1-t)|f'(Q(t))|/f^2(Q(t)) \le \gamma < \infty.$$

Then, as $n \to \infty$, with $\delta_n = 25n^{-1}\log\log n$ we have

$$\sup_{\delta_n \le t \le 1-\delta_n} |\rho_n(t) - u_n| \overset{a.s.}{=} \mathcal{O}(n^{-1/2}\log\log n). \tag{1.6}$$

If, in addition to (a), (b) *and* (c), *we assume also*

(d) *$A = \lim_{x \downarrow x_F} f(x) < \infty, \quad B = \lim_{x \uparrow x^F} f(x) < \infty$,*

(e) *one of*

(e,α) *$A \wedge B > 0$,*

(e,β) *if $A = 0$ (respectively $B = 0$) then f is nondecreasing (respectively nonincreasing) on an interval to the right of x_F (respectively to the left of x^F),*

then, if (e, α) obtains,

$$\sup_{0 \le t \le 1} |\rho_n(t) - u_n(t)| \overset{a.s.}{=} \mathcal{O}(n^{-1/2}\log\log n), \tag{1.7}$$

and if (e, β) obtains,

$$\sup_{0 \le t \le 1} |\rho(t) - u_n(t)| \overset{a.s.}{=} \begin{cases} \mathcal{O}(n^{-1/2}\log\log n), & \text{if } \gamma < 1 \\ \mathcal{O}\{n^{-1/2}(\log\log n)^2\}, & \text{if } \gamma = 1 \\ \mathcal{O}\{n^{-1/2}(\log\log n)^{\gamma}(\log n)^{(1+\varepsilon)(\gamma-1)}\}, & \text{if } \gamma > 1, \end{cases} \tag{1.8}$$

where $\varepsilon > 0$ is arbitrary, and γ is as in (c).

In fact, using only the assumptions (a), (b) and (c), the statement of (1.6) can be extended over a wider range than $[\delta_n, 1 - \delta_n]$ as follows (cf. Theorem 3.1 of M. Csörgő, S. Csörgő, Horváth and Révész(1984)):

$$\sup_{1/(n+1) \le t \le n/(n+1)} |\rho_n(t) - u_n(t)| \overset{a.s.}{=} \begin{cases} \mathcal{O}\{n^{-1/2}(\log\log n)^{1+\gamma}\}, & \text{if } \gamma \le 1 \\ \mathcal{O}\{n^{-1/2}(\log\log n)^{\gamma}(\log n)^{(1+\varepsilon)(\gamma-1)}\}, & \text{if } \gamma > 1, \end{cases} \tag{1.9}$$

where again $\varepsilon > 0$ is arbitrary, and γ is as in (c).

In the light of (1.9) it is seen that the tail monotonicity assumptions of (e) play a role only in extending the interval $[1/(n+1), n/(n+1)]$ to $[0,1]$ when comparing ρ_n and u_n uniformly in $t \in [0,1]$. Hence the conditions of (e) have nothing to do with the data, the empirical parts of ρ_n and u_n (cf. (1.2) and (1.3)).

M. Csörgő and Révész(1975, 1978) also proved

THEOREM B. *For a sequence of i.i.d. r.v.'s $X_1, X_2, \ldots$ having a continuous distribution function F, there exists a probability space with a sequence of Brownian bridges $\{B_n(t),\ 0 \le t \le 1\}$ on it such that*

$$\sup_{0 \le t \le 1} |u_n(t) - B_n(t)| \overset{a.s.}{=} \mathcal{O}(n^{-1/2} \log n). \tag{1.10}$$

Consequently, under the conditions of Theorem A, with the same sequence of Brownian bridges we have as well

$$\sup_{0<t<1} |\rho_n(t) - B_n(t)| = \begin{cases} \mathcal{O}(n^{-1/2} \log n), & \text{if } \gamma < 2, \\ \mathcal{O}\{n^{-1/2} (\log \log n)^{\gamma} (\log n)^{(1+\varepsilon)(\gamma-1)}\}, & \text{if } \gamma \ge 2, \end{cases} \tag{1.11}$$

where $\varepsilon > 0$ is arbitrary, and γ is as in (c).

The statement of (1.11) is an analogue of Kolmogorov's classical theorem on the empirical distribution function. The latter gives a confidence band for the unknown distribution function F in terms of the empirical distribution function $F_n(x) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]} \cdot (X_i)$. However, (1.11) does not give an immediate confidence band for the quantile function Q. Indeed, its direct application yields only the "confidence band"

$$Q_n(t) - n^{-1/2} c / f(Q(t)) \le Q(t) \le Q_n(t) + n^{-1/2} c / f(Q(t)), \quad 0 < t < 1, \tag{1.12}$$

which depends on the unknown quantile-density function $Q'(t) = 1/f(Q(t))$. The constant c here depends on the size of the band, i.e. $c = c(\alpha)$ is determined by $P\{\sup_{0 \le t \le 1} |B(t)| \le c\} = 1 - \alpha$, $\alpha \in (0,1)$, where $B(\cdot)$ is a Brownian bridge.

Taking the route of (1.12) for constructing confidence bands for the quantile function Q requires the estimation of the quantile- density function $Q'(t) = 1/f(Q(t))$. M. Csörgő and Révész(1981, Theorem 5.5.2) proposed to estimate the latter function by $1/f_n(Q_n(t))$, where f_n is any appropriate estimator of f and Q_n is the empirical quantile function (cf. (1.2)). However, the use of two estimators (that of f and that of Q) may seem to be a somewhat artificial solution to the problem at hand. Rather, one may want to estimate $1/f(Q(\cdot))$ directly. The next two theorems were proved in M. Csörgő and Révész(1984).

THEOREM C. *Let $X_1, X_2, \ldots$ be i.i.d. r.v.'s with a continuous distribution function F satisfying the conditions of Theorem A. Assume also*

$$(f) \qquad \sup_{x_F < x < x^F} \frac{(F(x)(1-F(x)))^2}{f(x)} = \sup_{0<t<1} \frac{(t(1-t))^2}{f(Q(t))} \le C$$

with some $C > 0$. Then

$$\lim_{n\to\infty} \omega(n) \sup_{\varepsilon_n \le t \le 1-\varepsilon_n} \left| \frac{Q_n(t+h_n) - Q_n(t-h_n)}{2h_n} - \frac{1}{f(Q(t))} \right| \overset{a.s.}{=} 0, \tag{1.13}$$

where $h_n = n^{-\alpha}$, $\varepsilon_n = n^{-\beta}$, $\omega(n) = n^{\delta}$, and $3\beta + \delta < \alpha < 1/2$, $2\delta + 4\beta + \alpha < 1$.

The estimator $(2h_n)^{-1}(Q_n(t+h_n) - Q_n(t-h_n))$ of (1.13) for $p(t) = 1/f(Q(t))$ (cf. also Hájek and Šidak (1967)) is a natural analogue of the estimator $(2h_n)^{-1}(F_n(x+h_n) - F_n(x-h_n))$ for the density function f. A more sophisticated, Rosenblatt (1956)-Parzen (1962)-type kernel-quantile-density estimator p_n for p is defined as

$$\begin{aligned} p_n(t) &= h_n^{-1} \int_{1/(n+1)}^{n/(n+1)} K((t-y)/h_n) dQ_n(y) \\ &= h_n^{-1} \sum_{i=1}^{n-1} K((t-i/n)/h_n)(X_{i+1,n} - X_{i,n}). \end{aligned} \tag{1.14}$$

Concerning the latter, M. Csörgő and Révész(1984) proved

THEOREM D. *Let $X_1, X_2, \ldots$ be i.i.d. r.v.'s with a continuous distribution function F satisfying the conditions* (a) - (f), *and assume also that we have*

$$(g) \qquad \sup_{x_F < x < x^F} |f''(x)| = \sup_{0<t<1} |f''(Q(t))| \le C$$

with some $C > 0$. Let $K(\cdot)$ be an absolutely continuous density function, vanishing outside of the interval $(-1/2, 1/2)$ and such that

$$(h) \qquad \int_{-1/2}^{1/2} xK(x)dx = 0, \quad \int_{-1/2}^{1/2} x^2 K(x)dx < \infty, \quad \sup_x |K'(x)| < \infty.$$

Then

$$\lim_{n\to\infty} \omega(n) \sup_{\varepsilon_n \le t \le 1-\varepsilon_n} |p_n(t) - p(t)| \overset{a.s.}{=} 0, \tag{1.15}$$

where $p_n(t)$ is defined as in (1.14), $p(t) = 1/f(Q(t))$, $h_n = n^{-\alpha}$, $\varepsilon_n = n^{-\beta}$, $\omega(n) = n^{\delta}$ with $\delta + 10\beta < 2\alpha$ and $\alpha + 2\beta + \delta < 1/2$.

For (1.15) to hold true, we can for example have $\alpha = 1/5$ with $\delta < 11/40$ and $\beta < 1/80$.

As a consequence of (1.11), (1.15) and the law of the iterated logarithm for the quantile process ρ_n (cf. M. Csörgő and Révész(1978)) we conclude (cf. M. Csörgő and Révész(1984))

COROLLARY A. *Let $X_1, X_2, \ldots$ be i.i.d. r.v.'s with a continuous distribution function F, satisfying also the conditions (a) - (g) and K be a kernel function as in (h). Then with the sequence of Brownian bridges $\{B_n(t),\ 0 \le t \le 1\}$ of Theorem B we have*

$$\sup_{\varepsilon_n \le t \le 1-\varepsilon_n} \left| \frac{n^{1/2}}{p_n(t)} (Q_n(t) - Q(t)) - B_n(t) \right| \overset{a.s.}{=} o((\log\log n)^{1/2}/\omega(n)), \tag{1.16}$$

and hence also

$$\lim_{n\to\infty} P\Big\{Q_n(t) - cp_n(t)/n^{1/2} \le Q(t) \le Q_n(t) + cp_n(t)/n^{1/2};\ \varepsilon_n \le t \le 1-\varepsilon_n\Big\}$$
$$= P\Big\{\sup_{0\le t\le 1} |B(t)| \le c\Big\} = 1 - \sum_{k\neq 0}(-1)^{k+1}\exp(-2k^2c^2) = \tilde{K}(c),\ c \ge 0, \tag{1.17}$$

where $p_n(t)$ is as in (1.14), $\{B(t), 0 \le t \le 1\}$ is a Brownian bridge, $h_n = n^{-\alpha}$, $\varepsilon_n = n^{-\beta}$, $\omega(n) = n^{\delta}$ with $\delta + 10\beta < 2\alpha$ and $\alpha + 2\beta + \delta < 1/2$.

We note that (1.16) implies (1.17) because $\sup_{0\le t\le \varepsilon_n} |B(t)|$ and $\sup_{1-\varepsilon_n\le t\le 1} |B(t)|$ tend to zero almost surely as $n \to \infty$.

A similar corollary can be based also on Theorem C, and various one-sided confidence bands for Q can, of course, be constructed as well, using any one of Theorems D and C.

Estimation of the quantile-density function $p(t) = 1/Q(t)$ is, naturally, not the only way to produce confidence bands for the quantile function Q. While it is true, of course, that a direct application of (1.11) produces only the "confidence band" of (1.12), assuming only the conditions (a), (b), (c) for F, M. Csörgő and Révész(1984) also deduced density-estimation-free confidence bands for Q like, for example, the one we state now.

COROLLARY B. *Let $X_1, X_2, \ldots$ be i.i.d. r.v.'s with a continuous distribution function F, satisfying also the conditions (a), (b), (c) of Theorem A. Then*

$$\lim_{n\to\infty} P\{Q_n(t - n^{-1/2}c) \le Q(t) \le Q_n(t + n^{-1/2}c); \varepsilon_n \le t \le 1-\varepsilon_n\}$$
$$= P\Big\{\sup_{0\le t\le 1} |B(t)| \le c\Big\} = \tilde{K}(c), \quad c \ge 0, \tag{1.18}$$

where $\varepsilon_n \downarrow 0$, $n^{1/2}\varepsilon_n \to \infty$ as $n \to \infty$, $B(\cdot)$ is a Brownian bridge and $\tilde{K}(c)$ the distribution function of the random variable $\sup_{0\le t\le 1} |B(t)|$, is as in (1.17).

There are five corollaries along these lines in M. Csörgő and Révész(1984). M. Csörgő and Horváth(1989) modified the proofs of these corollaries so that all of them remain true on assuming only that F is continuous. For example, their version of Corollary B reads as follows:

COROLLARY C. *Let $X_1, X_2, \ldots$ be i.i.d. r.v.'s with a continuous distribution function F, and let $\{\varepsilon_n\}_{n\ge 1}$ be a sequence of positive numbers such that as $n \to \infty$*

$$\varepsilon_n \downarrow 0 \quad \text{and} \quad n^{1/2}\varepsilon_n \to \infty. \tag{1.19}$$

Then

$$\lim_{n\to\infty} P\{Q_n(t - n^{-1/2}c) \le Q(t) < Q_n(t + n^{-1/2}c), \varepsilon_n \le t \le 1-\varepsilon_n\} = \tilde{K}(c), \tag{1.20}$$

where the distribution function $\tilde{K}(c)$ is as in (1.17).

We call attention to the fact that the only formal difference between the confidence bands of (1.18) and (1.20) is that in the latter, for the upper band of Q we have $<$ instead of $\leq$ for that of Q in (1.18). This slight modification is crucial for the sake of dropping the conditions (a), (b), (c) and still having (1.20) as stated.

REMARK 1. It is clear from the discussion on confidence bands for the quantile function in Section 4.2 of M. Csörgő (1983) that if ε_n is not a fixed constant in (0,1), then a condition like $n^{1/2}\varepsilon_n \to \infty$ as $n \to \infty$, is, in general, also necessary as well for the asymptotic validity of bands like, for example, that of (1.20), and none of this type of bands remains valid with $\varepsilon_n \equiv 0$. For details and proof of statements like Corollary C we refer to M. Csörgő and Horváth(1989).

Back to estimates of the quantile density function like, for example, those of Theorems C and D, they are of interest on their own. For example, depending on the approach one takes, they can also play a role in estimating the score function $(-1)J$ of the density function f, defined by

$$J(t) = f'(Q(t))/f(Q(t)) = \frac{d}{dt}f(Q(t)), \quad 0 < t < 1.$$

For thoughts and results along these lines we refer to Parzen (1979), M. Csörgő and Révész(1986), Burke and Horváth(1989), as well as to further references in these papers. Estimating the quantile-density function at a point is also of interest. For such results and their applications we refer to Babu and Rao (1990) and references therein.

In Section 2 we present some mean integrated L_1 and L_2 errors for estimating $p(t)$ by $p_n(t)$ as in (1.14) when the underlying quantile function Q is assumed to be very smooth.

In Section 3 we consider a kernel estimator $\hat{Q}_n$ for Q, and present a weak convergence result for the kernel-quantile process $n^{1/2}(\hat{Q}_n(\cdot) - Q(\cdot))$.

2. Mean Integrated L_1 and L_2 Errors

Throughout this section we assume that the *bandwidth* h_n of the quantile-density estimator $p_n(\cdot)$ as defined in (1.14) *is such that* $h_n \to 0$ *and* $nh_n \to \infty$ *as* $n \to \infty$.

We list now the *conditions* we assume here *on kernel* K:

(i) $\int_{-\infty}^{\infty} K(x)dx = 1$,

(ii) $K(\cdot)$ is a bounded function,

(iii) $K(x) = 0$ if $|x| \geq A > 0$,

(iv) The derivative $K'(\cdot)$ of $K(\cdot)$ is bounded,

(v) $\int_{-\infty}^{\infty} x^j K(x)dx = \begin{cases} 0 & \text{for } 1 \leq j \leq r-1, \\ r!k_r & \text{for } j = r. \end{cases}$

Conditions on Q:

(I) Q has $r+1$ derivatives on $(a-\varepsilon, b+\varepsilon)$, $0 < a < b < 1$, and they are continuous and bounded on $(a-\varepsilon, b+\varepsilon)$ for some $\varepsilon > 0$ and $r > 1$,

(II) $\sup_{0<t<1} t(1-t)|f'(Q(t))|/f^2(Q(t)) \leq \gamma < \infty$.

Condition (II) is identical to (c) of Theorem A. We inserted it in here for the sake of more convenient reference.

Next we define

$$p_{(n)}(t) = h_n^{-1} \int_{1/(n+1)}^{n/(n+1)} K((t-y)/h_n) dQ(y).$$

Similarly to Hall (1982), it can be shown that we have

$$\sup_{a \le t \le b} |p_{(n)}(t) - p(t) - (-1)^r h_n^r k_r Q^{(r+1)}(t)| = o(h_n^r), \tag{2.1}$$

where $p(t) = 1/f(Q(t))$, the quantile-density function.

Let

$$\psi(c,d) = E|dN + c|, \quad d > 0, \quad -\infty < c < \infty,$$

where N stands for a standardized normal random variable. Then

$$\psi(c,d) = 2d \int_{c/d}^{\infty} u d\Phi(u) + c(2\Phi(c/d) - 1),$$

where Φ is the standard normal distribution function.

Next we present two theorems without proof.

THEOREM 2.1. *If conditions (i) - (v) and (I), (II) hold true, then*

$$\begin{aligned} &E \int_a^b |p_n(t) - p(t)| dt \\ &= \int_a^b \psi\left\{(-1)^r h_n^r k_r Q^{(r+1)}(t), (nh_n)^{-1/2} p(t) \left(\int_{-A}^{A} K^2(u) du \right)^{1/2} \right\} dt \\ &\quad + o((nh_n)^{-1/2}) + o(h_n^r). \end{aligned} \tag{2.2}$$

THEOREM 2.2. *If conditions (i) - (v) and (I), (II) hold true, then*

$$\begin{aligned} E \int_a^b (p_n(t) - p(t))^2 dt &= h_n^{2r} k_r^2 \int_a^b (Q^{(r+1)}(t))^2 dt \\ &+ (nh_n)^{-1} \int_a^b p^2(t) dt \int_{-A}^{A} K^2(u) du + o((nh_n)^{-1}) + o(h_n^{2r}). \end{aligned} \tag{2.3}$$

Two lemmata, which may be of interest on their own, play a crucial role in proving these theorems.

LEMMA 2.1. *If conditions (I) and (II) hold true, then for ρ_n of (1.4) and u_n of (1.5) we have*

$$E \sup_{a_1 \le t \le b_1} |\rho_n(t) - u_n(t)| = \mathcal{O}(n^{-1/2})$$

and

$$E \sup_{a_1 \le t \le b_1} (\rho_n(t) - u_n(t))^2 = \mathcal{O}(n^{-1}),$$

for all $a - \varepsilon \le a_1 \le b_1 \le b + \varepsilon$.

The proof of this lemma is along the lines of proving Theorem A (M. Csörgő and Révész1978) using also inequalities of Wellner (1978), and Dvoretzky, Kiefer and Wolfowitz (1956).

LEMMA 2.2. *Let* $Z_1, Z_2, \ldots$ *be independent random variables with* $EZ_i = 0$ *and* $E|Z_i|^3 < \infty$, $i = 1, 2, \ldots$. *Then*

$$\sup_{-\infty<c<\infty} \left| E\left| \left(\sum_{i=1}^n EZ_i^2 \right)^{-1/2} \sum_{i=1}^n Z_i + c \right| - \psi(c,1) \right| = \mathcal{O}\left(\sum_{i=1}^n E|Z_i|^3 \Big/ \left(\sum_{i=1}^n EZ_i^2 \right)^{3/2} \right).$$

Using the method of the proof of Lemma 8 on page 90 of Devroye and Györfi (1985), the proof of this lemma is based on estimates of the remainder term of the central limit theorem in Bikelis (1966) (cf. also Petrov, 1975, page 132).

Multiplying through by $(nh_n)^{1/2}$ and nh_n in (2.2) and (2.3) respectively, we conclude that the optimal choice of h_n for our mean integrated L_1 and L_2 errors of smooth estimation of the quantile density function $p(t) = 1/f(Q(t))$ by $p_n(t)$ as in (1.14) is $h_n = n^{-1/(2r+1)}$.

3. A Smoothed Quantile Process

Let

$$\hat{Q}_n(t) = h_n^{-1} \int_{1/(n+1)}^{n/(n+1)} K((t-y)/h_n) Q_n(y) dy, \tag{3.1}$$

and consider the smoothed quantile process

$$\hat{q}_n(t) = n^{1/2}(\hat{Q}_n(t) - Q(t)), \quad 0 < t < 1, \tag{3.2}$$

where Q_n in (3.1) is the empirical quantile function of (1.2). We have

$$\begin{aligned} \hat{q}_n(t) &= n^{1/2} h_n^{-1} \int_{1/(n+1)}^{n/(n+1)} K((t-y)/h_n)(Q_n(y) - Q(y)) dy \\ &\quad + n^{1/2} \left\{ h_n^{-1} \int_{1/(n+1)}^{n/(n+1)} K((t-y)/h_n) Q(y) dy - Q(t) \right\} \\ &= \tilde{q}_n(t) + I_n(t), \quad 0 < t < 1. \end{aligned} \tag{3.3}$$

The kernel-quantile process $\tilde{q}_n$ of (3.3) is a smoothed version of

$$q_n(t) = n^{1/2}(Q_n(t) - Q(t)), \quad 0 < t < 1, \tag{3.4}$$

and, under suitable conditions, as $n \to \infty$, we should like to have

$$\sup_{0<t<1} |\tilde{q}_n(t) - q_n(t)| = o_P(1). \tag{3.5}$$

The numerical, non-stochastic process $I_n(t)$ of (3.3) represents the bias introduced by estimating the quantile function Q by the smoothed empirical quantile function $\hat{Q}_n$ in the $\hat{q}_n$ process of (3.2). In order to make the latter approach workable, under suitable conditions, as $n \to \infty$ we should want to have

$$\sup_{0<t<1} |I_n(t)| = o(1), \tag{3.6}$$

for, if (3.5) and (3.6) were true, then as $n \to \infty$ we would have also

$$\sup_{0<t<1} |\hat{q}_n(t) - q_n(t)| = o_P(1). \tag{3.7}$$

The latter would be desirable in that one would hope that the asymptotic behaviour of $\hat{q}_n$ should be the same as that of q_n. Unfortunately, due to the bias term I_n, the latter goal is not easily achievable.

Concerning the $\tilde{q}_n(\cdot)$ process term of (3.3) and $q_n(\cdot)$ of (3.4), we have

THEOREM 3.1. *On K assume conditions* (i) - (iii), *and for F assume only condition* (b) *of Theorem A with continuous $f(\cdot)$ on (x_F, x^F). Then with $0 < a < b < 1$ we have, as $n \to \infty$,*

$$\sup_{a \le t \le b} |\tilde{q}_n(t) - q_n(t)| = o_P(1). \tag{3.8}$$

PROOF. By (1.10) we have, as $n \to \infty$,

$$\sup_{a \le t \le b} |q_n(t) - B_n(t)/f(Q(t))| \overset{a.s.}{=} \mathcal{O}(n^{-1/2} \log n), \tag{3.9}$$

since $q_n(t) = u_n(t)/f(Q(\theta_{t,n}))$, where $U_n(t) \wedge t < \theta_{t,n} < U_n(t) \vee t$. Also,

$$\begin{aligned}
&\sup_{a \le t \le b} |\tilde{q}_n(t) - q_n(t)| \\
&\quad \le \sup_{a \le t \le b} \sup_{|y| \le A h_n} |q_n(t) - q_n(t+y)| \int_{(t-\frac{n}{n+1})/h_n}^{(t-\frac{1}{n+1})/h_n} |K(u)| du \\
&\quad + \sup_{a \le t \le b} |q_n(t)| \cdot \left| \int_{(t-\frac{n}{n+1})/h_n}^{(t-\frac{1}{n+1})/h_n} |K(u)| du - 1 \right| \\
&\quad = o_P(1)\mathcal{O}(1) + \mathcal{O}_P(1) o(1) = o_P(1),
\end{aligned} \tag{3.10}$$

where in the last line we have also utilized (3.9) together with the a.s. continuity of a Brownian bridge process.

Combining (3.8) and (3.9) we obtain

COROLLARY 3.1. *Under the conditions of Theorem 3.1 and with the Brownian bridges $B_n(\cdot)$ of (1.10) we have, as $n \to \infty$,*

$$\sup_{a \le t \le b} |\tilde{q}_n(t) - B_n(t)/f(Q(t))| = o_P(1). \tag{3.11}$$

COROLLARY 3.2. *Assume that as $n \to \infty$ we have*

$$\sup_{a \le t \le b} |I_n(t)| = o(1). \tag{3.12}$$

Then under the conditions of Theorem 3.1 and with the Brownian bridges $B_n(\cdot)$ of (1.10) we have, as $n \to \infty$,

$$\sup_{a \le t \le b} |\hat{q}_n(t) - B_n(t)/f(Q(t))| = o_P(1). \tag{3.13}$$

PROOF. By (3.3), (3.11) and (3.12) we get (3.13).

While the assumption of (3.12) appears to be innocent enough, we should note that if $h_n = n^{-1/(2r+1)}$, the optimal choice of h_n in Section 2, then (3.12) does not hold. Nor would it hold true when estimating $p(t) = 1/f(Q(t))$ by $p_n(t)$ as in Theorem D. Thus under conditions like in Sections 1 and 2, statements like (1.16) and (1.17) do not hold true for the process $\{\hat{q}_n(t)/p_n(t),\ 0 < t < 1\}$ even if we were to study this process in t only over $t \in [a, b]$, $0 < a < b < 1$. Hence, we can only conclude by saying that uniform estimation in t of the quantile function $Q(t)$ by the smoothed empirical quantile function $\hat{Q}_n(t)$ as in (3.1) appears to be difficult by our methods. Further work will have to be done along these lines before we can come to more reasonable conclusions. Estimation of the quantile-density function $p(t)$ by kernel-quantile-density estimators like $p_n(t)$ of (1.14), as summarized in Sections 1 and 2, is of independent interest, of course.

Acknowledgement. This research was supported by an NSERC Canada Grant of M. Csörgő at Carleton University.

References

[1] Babu, G.J. and Rao, C.R. (1990) 'Estimation of the reciprocal of the density quantile function at a point', J. Multivariate Anal. 33, 106-124.

[2] Bikelis, A. (1966) 'Estimates of the remainder term in the central limit theorem', Litovsk. Mat. Sb. 6, 323-346 (in Russian).

[3] Burke, M.D. and Horváth, L. (1989) 'Large sample properties of kernel-type score function estimators', J. Statist. Plann. Inference 22, 307-321.

[4] Csörgő, M. (1983) Quantile Processes with Statistical Applications, SIAM, Philadelphia.

[5] Csörgő, M. Csörgő, S., Horváth, L. and Révész, P. (1984) 'On weak and strong approximations of the quantile process', in M. Josifescu (ed.), Proceedings of the Seventh Conference on Probability Theory, Editura Academii Republicii Socialiste România, Bucuresti, pp. 81-95.

[6] Csörgő, M. and Horváth, L. (1989) 'On confidence bands for the quantile function', Tech. Rep. Ser. Lab. Res. Statist. Probab., No. 131, Carleton University - University of Ottawa, pp. 29-43.

[7] Csörgő, M. and Révész, P. (1975) 'Some notes on the empirical distribution function and the quantile process', in P. Révész (ed.), Limit Theorems of Probability Theory, Colloq. Math. Soc. J. Bolyai, Vol. 11, North Holland, Amsterdam, pp. 59-71.

[8] Csörgő, M. and Révész, P. (1978) 'Strong approximations of the quantile process', Ann. Statist. 6, 882-894.

[9] Csörgő, M. and Révész, P. (1981) Strong Approximations in Probability and Statistics, Academic Press, New York & Akadémiai Kiadó, Budapest.

[10] Csörgő, M. and Révész, P. (1984) 'Two approaches to constructing simultaneous confidence bounds for quantiles', Probab. Math. Statist. 4, 221-236.

[11] Csörgő, M. and Révész, P. (1986) 'A nearest neighbour-estimator for the score function', Probab. Th. Rel. Fields 71, 293-305.

[12] Devroye, L. and Györfi, L. (1985) Nonparametric Density Estimation: The L_1 View, Wiley, New York.

[13] Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956) 'Asymptotic minimax character of sample distribution function and of the multinomial estimator', Ann. Math. Statist.27, 642-669.

[14] Hall, P. (1982) 'Limit theorems for stochastic measures of the accuracy of density estimators', Stochastic Process. Appl. 13, 11- 25.

[15] Hájek, J. and Šidák, Z. (1967) Theory of Rank Tests, Academic Press, New York.

[16] Parzen, E. (1962) 'On estimation of a probability density function and mode', Ann. Math. Statist. 33, 1065-1076.

[17] Parzen, E. (1979) 'Nonparametric statistical data modeling', J. Am. Stat. Assoc. 74, 105-131.

[18] Petrov, V.V. (1975) Sums of Independent Random Variables, Springer, Berlin.

[19] Rosenblatt, M. (1956) 'Remarks on some nonparametric estimates of density function', Ann. Math. Statist. 27, 832-837.

[20] Wellner, J.A. (1978) 'Limit theorems for the ratio of the empirical distribution function to the true distribution function', Z. Wahrsch. Verw. Gebiete 45, 73-88.

DATA-DRIVEN SMOOTHING BASED ON CONVEXITY PROPERTIES

ANTONIO CUEVAS and WENCESLAO GONZALEZ-MANTEIGA
Dpto. de Matemáticas, Universidad Autónoma
28049-Madrid, Spain.
Dpto. de Estadística, Universidad de Santiago de Compostela
15771-Santiago de Compostela, Spain

ABSTRACT. A theorem on uniform convergence for densities with a fixed number of inflection points is proved. As an application, a method of automatic choice of bandwidth parameters in density estimation is proposed. It is based on the idea of making equal the numbers of inflection points of the estimates and that of the target density. The resulting estimates are almost surely uniformly consistent. This method is closely connected with previous ideas of Silverman (1981, 1983) and provides, in many practical situations, a natural procedure for the *bump-hunting* problem. A Monte Carlo study is included.

1. Introduction

The main purpose of this paper is to outline a new method of automatic choice of the bandwidth parameters in univariate kernel density estimates (see Section 3). The consistency properties of this method rely on a purely probabilistic result, of independent interest, concerning the convergence of density functions. This result is proved in Section 2. Section 4 is devoted to discussion and comments. Some simulation results are given in Section 5. In the sequel, all limits are taken as $n \to \infty$, unless otherwise specified.

Next, we introduce some preliminaries and notation and summarize the basic ideas.

Let f_n, f be probability density functions (on $\Re$) associated with absolutely continuous distribution functions F_n, F. Let P_n, P denote the corresponding probability measures. It is well-known (Scheffé's Theorem; see, e.g., Billingsley, 1968, p. 224) that

$$f_n \longrightarrow f \ a.e. \implies \int \mid f_n - f \mid d\mu_L = 2 \sup_{B \in \mathcal{B}} \mid P_n(B) - P(B) \mid \longrightarrow 0,$$

where *a.e.* denotes almost everywhere convergence with respect to the Lebesgue measure μ_L.

This result provides an interesting probabilistic interpretation for the a.e. convergence of densities in terms of uniform convergence (on the Borel σ-algebra $\mathcal{B}$) of

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 225–240.
© 1991 *Kluwer Academic Publishers.*

P_n to P. Such uniform convergence is sometimes called *strong* ; it has some interesting applications in statistics (Sethuraman, 1961a, b; Bickel, 1965; Hettmansperger, 1968). Of course, since strong convergence entails weak convergence, we have $f_n \to f \Longrightarrow F_n \to F$. The converse implication clearly does not hold in general, as the counterexample $f_n(x) = 1 + cos\, 2\pi nx, x \in [0,1]$ shows. The problem of obtaining sufficient conditions for $f_n \to f$ (in some sense), assuming the weak convergence $F_n \to F$, has been studied by Reiss (1973), Hettmansperger and Klimko (1974) (under hypotheses of unimodality) and Boos (1985) (using hypotheses of equicontinuity). Sweeting (1986) has proved that Boos' conditions are also necessary.

In Theorem 1 below we will prove a new sufficient condition for $f_n \longrightarrow f$, in terms of piecewise convexity of the densities f_n, f. Roughly speaking, this theorem establishes that, when f_n, f are *piecewise convex-concave*, all having the same (finite) number of inflection points, then the weak convergence $F_n \longrightarrow F$ entails $f_n \longrightarrow f$, uniformly on $\Re$. Moreover, the inflection points of f_n converge to the corresponding ones of f.

The application of this result to density estimation arises now in a straightforward way: let $X_1, ..., X_n$ be a random sample taken from an (unknown) absolutely continuous distribution with density f. Let

$$\hat{f}_n(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right) \tag{1}$$

be a kernel estimate of f, where $\{h_n\}$ is a sequence of positive constants called smoothing (or bandwidth) parameters. Denote by $I(\hat{f}_n), I(f)$ the number of inflection points of $\hat{f}_n$ and f, respectively. If the sequence $\{h_n\}$ could be chosen in such a way that $I(\hat{f}_n) = I(f)$, for all n, and $\hat{F}_n$ converged (weakly, almost surely) to F, then (from Theorem 1) we would have $\hat{f}_n \longrightarrow f$, uniformly almost surely. This idea will be formalized in Theorem 2. In intuitive terms, the bandwidth parameters h_n should be small enough as to ensure $\hat{F}_n \longrightarrow F$ and large enough as to avoid that $\hat{f}_n$ becomes *too noisy*. This conflict is typical in density estimation (see, e.g., Boneva, Kendall and Stepanov, 1971). In our case, if we assume that $I(f)$ is known in advance, the equation $I(\hat{f}_n) = I(f)$ provides a plausible criterion about *how much to smooth* the data, and the values of h_n with $I(\hat{f}_n) > I(f)$ would lead to estimates which could be considered too noisy. Moreover, the almost sure convergence (obtained in Theorem 1) of the inflection points of $\hat{f}_n$ to those of f has an interesting consequence: the *bumps* and the *dips* (i.e., the intervals of concavity and convexity) of f can be estimated consistently from those of $\hat{f}_n$. So, the method appears to be especially well-suited for the so called *bump-hunting problem*, which has received some attention in the literature (Good and Gaskins, 1980; Silverman, 1986), since the bumps of a density have interesting statistical and physical interpretations (see also Orear and Cassel, 1971, p. 281).

Of course, the effective resolution (with respect to h_n) of the equation $I(\hat{f}_n) = I(f)$ could be difficult in many cases. There is, however, an important particular situation in which this can be achieved by a simple algorithmic procedure, based upon ideas proposed by Silverman (1981, 1983) in a rather different context: if the kernel function K in (1) is the standard normal density, the number of inflection points in $\hat{f}_n = \hat{f}_n(.; h_n)$ is a decreasing right continuous function of the bandwidth parameter. This allows to assign h_n as the infimum of the values h where $f_n(.; h)$ has at most $I(f)$ inflection points.

This method (throughout called IP-method) is easily implemented in an algorithm. Furthermore, it automatically provides uniformly consistent estimates by incorporating in a natural way some prior information on f: its number of bumps (which in regular cases is $I(f) - 1$). On the other hand, in Theorem 3 we show that the IP-method can be adapted to the case that $I(f)$ is not exactly known: the idea is to use a sequence $\{\hat{g}_n\}$ of *pilot* density estimates with deterministic bandwidth parameters in order to obtain a consistent sequence $I_n = I(\hat{g}_n)$ of estimates of $I(f)$; then, the automatic IP-choice would be the infimum of $\{h : I_n = I(\hat{f}_n(.; h))\}$. However, this work is mainly concerned with the case where prior information on $I(f)$ is available (in particular, all the simulation results of Section 5 correspond to this case). This would be something in the spirit of the so-called *tailor-designed* methods (see Devroye, 1987 and Section 4 below). There are several questions which remain open, concerning both theoretical and practical aspects. They will be also commented in Section 4.

At the present stage, we make no claim of superiority for the IP-method. We feel, however, that this procedure deserves some attention since it corresponds to a natural idea and presents some interesting qualitative and quantitative advantages (see Sections 4 and 5).

2. A theorem on convergence of densities

The proof of the main result of this section (Theorem 1) is essentially based on a convexity argument which is established in the next lemma. It is valid for the general k-variate case although, in fact, we will only use it for $k = 1$.

LEMMA 1. *Let $\{f_n\}$ be a sequence of density functions on $\Re^k$. Denote by $\{F_n\}$ the corresponding sequence of distribution functions. Suppose that*

(i) $F_n \longrightarrow F_0$ on $\Re^k$, where F_0 is an absolutely continuous distribution function.

(ii)There exists an open, convex, non-empty set $A \subset \Re^k$ such that every f_n is concave (respectively, convex) on A.

Then, for every compact and convex set $B \subset A$,

$$f_n \longrightarrow f_0, \ uniformly \ on \ B,$$

where f_0 is the density function of F_0, continuous and concave (respectively,convex) on B.

PROOF: Let $\mathcal{C}$ and d_H denote, respectively, the class of compact convex non-empty sets in $\Re^{k+1}$ and the Hausdorff metric (see, e.g., Lay, 1982) defined on $\mathcal{C}$:

$$\begin{aligned} d_H(A,B) &= \inf\{\epsilon > 0 : A \subset B^\epsilon,\ B \subset A^\epsilon\} \\ &= \max\{\sup_{\mathbf{x}\in A}\inf_{\mathbf{y}\in B}\|\mathbf{x}-\mathbf{y}\|, \sup_{\mathbf{y}\in B}\inf_{\mathbf{x}\in A}\|\mathbf{x}-\mathbf{y}\|\}, \end{aligned}$$

for all $A, B \in \mathcal{C}$ (A^ϵ is the union of all open balls around points of A).

Let $B \subset A$ be compact and convex. We may assume that $\mu_L(B) > 0$. Since B is compact and A is open, there exists $\tau > 0$ such that $\overline{B^\tau} \subset A$. Then, $D = \overline{B^\tau}$ is a compact and convex set such that $B \subset int(D)$ ($\equiv$ interior of D) and the functions f_n are concave and continuous on D (since they are concave on A). Furthermore, the sequence $\{f_n\}$ is uniformly bounded on D: otherwise, the concavity of f_n would imply the existence of a subsequence in $\{f_n\}$ (which we also denote by $\{f_n\}$) such that $\int_A f_n(\mathbf{x})d\mathbf{x} \to \infty$, which would contradict the fact that the f_n are density functions. Hence, let us take $M > 0$ such that $f_n(\mathbf{x}) < M$, $\forall \mathbf{x} \in D$, $\forall n$. Define, for every n,

$$C_n = \{(\mathbf{x}, y) \in \Re^{k+1} : \mathbf{x} \in D,\ 0 \le y \le f_n(\mathbf{x})\}.$$

The sets C_n are compact and convex and, furthermore, the sequence $\{C_n\}$ is uniformly bounded since $C_n \subset D \times [0, M]$ for every n. Hence, by applying Blaschke's selection theorem (see, e.g., Lay, 1982, p. 98), there exists a subsequence $\{C_n\}$ such that

$$d_H(C_n, C) \longrightarrow 0, \tag{2}$$

for some compact and convex set C which must necessarily be of the form

$$C = \{(\mathbf{x}, y) : \mathbf{x} \in D,\ 0 \le y \le f_0(\mathbf{x})\},$$

where $f_0 : D \longrightarrow \Re$ is a concave function (observe that the sets C_n have a common basis $\{(\mathbf{x}, y) : \mathbf{x} \in D\}$).

The convergence (2) implies that

$$\forall \epsilon > 0\ \exists n_0 : C_n \subset C^\epsilon,\ C \subset C_n^\epsilon,\ \forall n \ge n_0. \tag{3}$$

If we take $\epsilon < \tau$, then $B^\epsilon \subset D$, and (3) implies

$$f_n \longrightarrow f_0,\ \textit{uniformly on } B. \tag{4}$$

This convergence, together with hypothesis (i), implies that f_0 coincides with a density function of F_0 on B: observe that (4) implies $\int_E f_n(\mathbf{x})d\mathbf{x} \to \int_E f_0(\mathbf{x})d\mathbf{x}$, for all Borelian sets $E \subset B$. Moreover, f_0 is continuous on B since it is the uniform limit of continuous functions.

If the densities f_n are assumed to be convex (instead of concave), the proof is similar: it is easy to prove that the f_n are uniformly bounded on D; then, the above reasoning is repeated by replacing the sets C_n by

$$C_n^- = \{(\mathbf{x}, y) \in \Re^{k+1} : \mathbf{x} \in D,\ f_n(\mathbf{x}) \leq y \leq M\},$$

where M is a upper bound of each f_n.

REMARK 1. An alternative proof for this lemma could be obtained by combining the differentiability properties of convex functions with Theorem 1 in Sweeting (1986). This would require to prove the equicontinuity of $\{f_n\}$, which should arise as a consequence of the existency and boundedness of subgradients (or lateral derivatives, in the univariate case) of f_n.

As indicated above, the hypothesis of *piecewise concavity-convexity* plays a central role in our approach. Next, we formally define this concept and state the main result of this section.

DEFINITION 1. *A function $f : \Re \longrightarrow \Re$ is said to be piecewise convex-concave (PCC) of order q if $\Re$ can be partitioned into $q+1$ intervals $I_0 = (-\infty, x_1]$, $I_1 = (x_1, x_2], \ldots, I_{q-1} = (x_{q-1}, x_q]$, $I_q = (x_q, \infty)$ such that f is strictly convex in $I_0, I_2, I_4, \ldots$, and strictly concave in $I_1, I_3, I_5, \ldots$*

THEOREM 1. *Let $\{f_n\}$, $n = 0, 1, \ldots$ be a sequence of univariate continuous density functions. Let $\{F_n\}$, $n = 0, 1, \ldots$ be the corresponding sequence of distribution functions. Let us assume that*

(i) $F_n \longrightarrow F_0$ on $\Re$.

(ii) f_n is PCC of order q, for every $n = 0, 1, \ldots$.

Then, $f_n \longrightarrow f_0$, uniformly on $\Re$.

PROOF: Let $x_1^n < \ldots < x_q^n$ be the inflection points of f_n, for $n = 0, 1, \ldots$. The proof shall be divided into five steps:

Proposition 1. $x_i^n \longrightarrow x_i^0$, as $n \to \infty$, for all $i = 1, \ldots, q$.

Proof: We may assume $i = 1$ (the proof would be analogous in the other cases). By contradiction, suppose that there exist $\epsilon > 0$ and a subsequence of $\{x_1^n\}$ (which we also denote by $\{x_1^n\}$) such that $x_1^n < x_1^0 - \epsilon$ for every n. Then, there are only two possibilities:

(a) For large enough n, f_n is strictly concave in the open interval $(x_1^0 - \epsilon, x_1)$.

(b) $x_1^n < x_2^n < x_1^0$, for infinitely many n.

The first assumpion leads to a contradiction with the hypothesis of weak convergence as an immediate consequence of Lemma 1.

Under the assumption (b), f_n should have, for infinitely many n, at most $q-2$ inflection points in the interval (x_1^0, ∞), whereas f_0 has exactly $q-1$. This implies the existence of an interval $J = (a, b) \subset [x_1^0, \infty)$ such that f_0 is strictly concave

(resp., convex) on J and f_n is strictly convex (resp., concave), infinitely often, on the same interval. This contradicts again (by Lemma 1) the hypothesis of weak convergence.

The symmetric assumption that $x_1^n > x_1^0 + \epsilon$ infinitely often leads to a similar contradiction.

Proposition 2. $f_n \longrightarrow f_0$, uniformly on each closed interval I containing no inflection point of f_0.

Proof: Define $J_0 = (-\infty, x_1^0)$, $J_i = (x_i^0, x_{i+1}^0)$, for $i = 1, \ldots, q-1$, and $J_q = (x_q^0, \infty)$. If the closed interval I does not contain inflection points of f_0, then necessarily $I \subset J_i$, for some $i \in \{0, \ldots, q\}$. Then, as a consequence of Proposition 1, I does not contain inflection points of f_n (for large n). Therefore, if $I \subset J_i$ is compact (which is necessarily the case when $1 \leq i \leq q-1$) f_n must be either strictly concave or strictly convex on I, for large n, according to the convexity character of f_0 on J_i. Then, from Lemma 1, $f_n \to f_0$ uniformly on I. If I is not compact, then it must be either $I = (-\infty, a] \subset (-\infty, x_1^0)$ or $I = [b, \infty) \subset (x_q^0, \infty)$, and f_n is either decreasing or increasing on I, for large n. Then, the uniform convergence follows from the simple monotonicity argument used by Hettmansperger and Klimko (1974).

Proposition 3. $f_n(x_i^n) \longrightarrow f_0(x_i^0)$, as $n \to \infty$, for all $i = 1, \ldots, q$.

Proof: We limit ourselves again to the case $i = 1$. Let us assume that the conclusion were not true. Then, there exists $\epsilon > 0$ such that $x_1^0 + \epsilon < x_2^0$, $f_0(x_1^0) - \epsilon > 0$, and either

$$f_n(x_1^n) - f_0(x_1^0) > \epsilon, \ \textit{for infinitely many } n \tag{5}$$

or

$$f_0(x_1^0) - f_n(x_1^n) > \epsilon, \ \textit{for infinitely many } n \tag{6}$$

On the other hand, from the continuity of f_0 and Propositions 1 and 2, the value of ϵ in (5) and (6) can be chosen small enough as to verify also the following conditions (7) to (10), for some n_0:

$$x_1^n \in (x_1^0 - \epsilon/2, x_1^0 + \epsilon/2), \ \forall n \geq n_0 \tag{7}$$

$$f_n \textit{ is strictly convex on } (x_1^0 - \epsilon, x_1^n)$$
$$\textit{and strictly concave on } (x_1^n, x_1^0 + \epsilon), \ \forall n \geq n_0. \tag{8}$$

$$|f_0(x_1^0) - f_0(x)| < \epsilon/4, \ \forall x \in (x_1^0 - \epsilon, x_1^0 + \epsilon). \tag{9}$$

$$|f_n(x_1^0 \pm \epsilon/2) - f_0(x_1^0 \pm \epsilon/2)| < \epsilon/4, \ \forall n \geq n_0. \tag{10}$$

Now, we will prove that assumption (5) leads to a contradiction: the concavity of f_n on the interval $(x_1^n, x_1^0 + \epsilon/2)$ implies that the graph of f_n on this interval is over the straight line connecting the points $(x_1^n, f_n(x_1^n))$ and $(x_1^0 + \epsilon/2, f_n(x_1^0 + \epsilon/2))$. In particular, this implies, from (7)-(10), that the graph of f_n on the interval

$(x_1^n, x_1^0+\epsilon/2)$ is over the straight line connecting the points $P_1 = (x_1^0-\epsilon/2, f_0(x_1^0)+\epsilon)$ and $P_2 = (x_1^0 + \epsilon/2, f_0(x_1^0) - \epsilon/2)$. Then, we have

$$\int_I f_n d\mu_L - \int_I f_0 d\mu_L \geq \int_I s d\mu_L - \int_I f_0 d\mu_L, \tag{11}$$

for infinitely many n, where s is the affine function whose graph is the straight line determined by P_1, P_2, and I is a closed interval $I = [x_1^0 + \delta, x_1^0 + \delta']$, where $0 < \delta < \delta' < \epsilon/2$ are chosen in such a way that $f_0(x) < s(x)$, $\forall x \in I$ (this choice is always possible since $s(x_1^0) = f_0(x_1^0) + \epsilon/4$ and f_0 is continuous). In short, the right-hand side of (11) is greater than zero which contradicts the hypothesis of weak convergence.

Assumption (6) leads to a similar contradiction.

Proposition 4. $f_n \longrightarrow f_0$, uniformly on each interval $[x_i^0, x_{i+1}^0]$, $i = 1, \ldots, q-1$. The convergence is also uniform on sets of the form $[a, x_i^0] \cup [x_q^0, b]$, for all $a < x_i^0$, $b > x_q^0$.

Proof: Let us fix $i \in \{1, \ldots, q-1\}$ and suppose that f_0 is concave on $[x_i^0, x_{i+1}^0]$ (the convex case is analogous). Define

$$C_i^n = \{(x, y) \in \Re^2 : \ x \in [x_i^n, x_{i+1}^n], \ 0 \leq y \leq f_n(x)\}.$$

From Propositions 1 and 3 we obtain (similarly to the proof of Lemma 1):

$$d_H(C_i^n, C_i^0) \longrightarrow 0, \ as \ n \to \infty.$$

Now, the uniform convergence on $[x_i^0, x_{i+1}^0]$ follows immediately as a consequence of the definition of Hausdorff distance and the structure of the sets C_i^n.

The second part of the statement can be proved in an analogous way.

Proposition 5. $f_n \longrightarrow f_0$, uniformly on $\Re$.

Proof: For each $a < x_1^0$, $b > x_q^0$ we have (from Proposition 2) $f_n \longrightarrow f_0$ uniformly on $(-\infty, a] \cup [b, \infty)$. So, the uniform convergence on $\Re$ is a direct consequence of this fact together with Proposition 4.

3. A bandwidth choice method based on convexity properties

3.1 A RESULT ON STRONG UNIFORM CONSISTENCY

Let $\hat{f}_n(t; h)$ be a univariate kernel density estimate of type (1) where $h = h_n$ denotes the bandwidth parameter. The choice of h is one of the most important practical problems in density estimation. So far, the literature on this topic has been mainly focussed on cross-validation techniques (see, e.g., Stone, 1984, Hall, 1987, Marron, 1987). More recently, other procedures based on the bootstrap methodology have also been studied (Hall, 1990).

As commented in the introduction, the method we propose here is based in a different idea: we choose the bandwidth parameter in such a way that the estimate $\hat{f}_n$ have the same number of inflection points than the target density f (assuming that this number is known).

Our main result is the following

THEOREM 2. *Let $X_1, \ldots, X_n$ be i.i.d. random variables with a PCC density f having q inflection points $x_1^0 < \ldots < x_q^0$.*

Let $\hat{f}_n = \hat{f}_n(t; h_n)$ a sequence of kernel density estimates as defined in (1), where K and $h_n = h_n(X_1, \ldots, X_n)$ have been chosen in order that, for each n, $\hat{f}_n(t; h_n)$ is a PCC density having exactly q inflection points $x_1^n < \ldots < x_q^n$.

If $h_n \to 0$ almost surely, then

$$\sup_{t \in \Re} |\hat{f}_n(t; h_n) - f(t)| \longrightarrow 0 \text{ almost surely}$$

and

$$x_i^n \longrightarrow x_i^0 \text{ almost surely}, \ \forall i = 1, \ldots, q. \tag{12}$$

PROOF: Let $F_n^*(t) = \frac{1}{n} \sum_{i=1}^n \delta(t - X_i)$ be the empirical distribution function associated with the sample $X_1, \ldots, X_n$ (δ degenerated distribution at 0), and let $\hat{F}_n$ be the distribution function corresponding to the density $\hat{f}_n(.; h_n)$. Observe that $\hat{F}_n$ may be expressed in the form $\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n V(t - X_i; h_n)$, where V(t;h) is the distribution function corresponding to the density $\frac{1}{h} K(\frac{t}{h})$.

In order to use Theorem 1 we only need to check the hypothesis (i), that is, the weak almost sure convergence

$$\hat{F}_n \longrightarrow F \text{ almost surely}, \tag{13}$$

where F is the distribution of the X_i. We will use the Bounded Lipschitz distance given by

$$d_{BL}(F, G) = \sup \left| \int \psi(t) dF(t) - \int \psi(t) dG(t) \right|,$$

where the supremum is taken over the class of functions $\psi : \Re \to \Re$ satisfying the condition $|\psi(t) - \psi(s)| \leq |t - s|/(1 + |t - s|)$, for all $t, s \in \Re$. It is known that d_{BL} metrizes the weak topology (see Huber, 1981, p. 33).

Since we clearly have

$$\sup_{a \in \Re} d_{BL}(V(. - a; h), \delta(. - a)) = d_{BL}(V(.; h), \delta) \longrightarrow 0, as \ h \to 0, \tag{14}$$

and, by hypothesis, $h_n \to 0$ almost surely, we may ensure that, given $\epsilon > 0$, there exists $n_0 \in \mathbf{N}$ such that

$$d_L(V(. - X_i; h_n), \delta(. - X_i)) < \epsilon, \ \forall i \in \mathbf{N}, \ \forall n \geq n_0.$$

Hence,

$$d_{BL}(\hat{F}_n, F_n^*) \leq \frac{1}{n} \sum_{i=1}^n d_{BL}(V(. - X_i; h_n), \delta(. - X_i)) < \epsilon, \text{ almost surely } \forall n \geq n_0. \tag{15}$$

Now, (13) immediately follows from (15) and the fact that, with probability one, F_n^* converges weakly (and, hence, also in d_{BL}) to F.

Finally, the conclusions of the theorem follow as direct consequences of Theorem 1.

3.2 Estimating the number of inflection points

If $I(f)$ is not exactly known, the IP-method can still be applied: it would suffice to obtain a sequence $\{I_n\}$ of estimates such that

$$I_n \to I(f), \ a.s. \tag{15}$$

Then, the smoothing parameter h_n would be chosen in order to equalize $I_n = I(f_n(.;h))$. The estimates I_n could be defined in a natural way as the numbers of inflection points corresponding to a sequence of *pilot estimates* of f whose smoothing parameters would be deterministically assigned to fulfill (15) (which, of course, implies $I_n = I(f)$ eventually a.s.). We make concrete these ideas in the following

THEOREM 3. *Let f be a PCC-density having q inflection points $x_1 < \ldots < x_q$. Let $T = [A, B]$ be a compact interval such that $x_i \in (A, B)$ for every $i = 1, \ldots, q$.*

Assume that

(i) f has three continuous derivatives.

(ii) The second derivative $f^{(2)}$ has no degenerate zero, that is, $f^{(3)}(x) \neq 0$ whenever $f^{(2)}(x) = 0$.

Let $\{\hat{g}_n\}$ be a sequence of density estimates such that

(iii) $\hat{g}_n^{(s)} \to f^{(s)}$, uniformly on $[A, B]$, for $s = 2, 3$.

Then, $I_T(\hat{g}_n) \to I(f)$, a.s., where $I_T(\hat{g}_n)$ denotes the number of inflection points of $\hat{g}_n$ in the interval T.

PROOF: From (iii), there exist $\epsilon > 0$ and $n_0 \in \mathbf{N}$ such that $(x_i - \epsilon, x_i + \epsilon) \in (A, B)$ for all $i = 1, \ldots, q$ and $\hat{g}_n^{(2)}$ has (a.s.) exactly one zero in $(x_i - \epsilon, x_i + \epsilon)$, for all $n \geq n_0$. Since $|f^{(2)}| > 0$ in the compact $T_\epsilon = [A, B] \setminus \bigcup_{i=1}^{q}(x_i - \epsilon, x_i + \epsilon)$, we have

$$|f^{(2)}| > a \ on \ T_\epsilon, \tag{16}$$

for some $a > 0$. Thus, the uniform convergence implies $|\hat{g}_n^{(2)}| > a$ (a.s.) for large enough n and, hence, $I_T(\hat{g}_n) = I_n$ eventually a.s.

REMARK 2. The estimation is restricted to a compact interval $[A, B]$ by technical reasons: inequality (16) need not hold on the whole real line outside of the intervals $(x_i - \epsilon, x_i + \epsilon)$ and so, $f_n^{(2)}$ could converge uniformly to $f^{(2)}$ showing many little bumps in the tails in such a way that $I_T(\hat{g}_n)$ need not converge to $I(f)$. This restriction to a compact interval represents no serious limitation from a practical point of view since, in most cases, we will be able to determine in advance an interval large enough as to contain all the inflection points of the target density.

REMARK 3. If $f^{(3)}$ is uniformly continuous and the sequence of pilot density estimates $\hat{g}_n$ is of kernel type, with $K \equiv N(0, 1)$ and deterministic bandwidths ϵ_n, a sufficient condition for the almost sure uniform convergences $\hat{g}_n^{(s)} \to f^{(s)}$, $(s = 2, 3)$ is $\epsilon_n \to 0$ and $n^{-1}\epsilon_n^{-7} log(\epsilon_n^{-1}) \to 0$; (see Silverman, 1978, Theorem C).

3.3 The effective computation of h_n

The smoothing method proposed in Subsection 3.1 suggests a problem of practical implementation: we need an algorithmic procedure for determining h_n in such a way that $I(\hat{f}_n(.;h_n)) = I(f)$. If the kernel function K is the standard normal density, this important problem can be solved by using the following result of Silverman (1981): let $\nu_m(h)$ denote the number of sign changes as t varies in $\partial^m \hat{f}_n(t;h)/\partial t^m$. Then, for each fixed sample and $m \geq 1$, $\nu_m(h)$ is a right continuous decreasing function of h.

We are interested in the case $m = 2$. As a consequence of the above result, h_n may be defined by

$$h_n = \inf\{h : \hat{f}_n(.;h) \textit{ has at most } q \textit{ inflection points}\}. \tag{17}$$

A similar choice (with *modes* instead of *inflection points*) has been proposed earlier by Silverman (1981, 1983) as a critical value for testing the null hypothesis that f has q modes against the alternative that f has more than q modes.

From the monotonicity of $\nu_2(h)$, it is straightforward to obtain an algorithm for computing the value h_n defined in (17) with any required precision.

4. Discussion and comments

4.1 Comparison with cross-validation techniques

In the first place, it should be pointed out that the basic idea behind the IP-method is quite different from that of cross-validation (CV). Whereas CV could be considered as a *universal* smoothing method, the IP-technique is rather oriented to those cases where some prior information is available on the number of inflection points of the target density. As a matter of fact, in many practical situations it is not unrealistic to assume, for example, that the density is bell-shaped or a mixture of two bell-shaped densities. The IP-method provides a natural procedure to incorporate this prior knowledge in the estimation for any given sample size. So, we think that, using Devroye's terminology, the IP-method might be placed within the *tailor-designed procedures.* In fact, as this author (1987, p. xviii) claims, *we feel that the major practical break-throughs in density estimation will not be on nonparametric estimates in their general forms, but on tailor designed estimates.*

On the other hand, it is known that CV techniques have some properties of asymptotic optimality. More precisely, the least squares CV and the pseudo- likelihood method asymptotically minimize the mean square integrated error (Stone, 1984) and the expected Kullback-Leibler loss (Hall, 1987), respectively. Since the IP choice forces some similarity between the derivatives $\hat{f}_n^{(s)}$ and f^s (for $s = 1, 2$), we conjecture that *it could also be a suitable choice for the estimation of the derivatives* and, in particular, it could asymptotically minimize some suitable distance such as $\sum_{s=0}^{2} \|\hat{f}_n^{(s)} - f^{(s)}\|_\infty$. So far, we have no formal proof of such optimality

result, but this situation is typical in the theory of automatic smoothing where the sequence of achievements usually is

(1) Intuitively based definition,
(2) consistency theorems,
(3) simulation results,
(4) convergence rates,
(5) optimality results.

This paper is only concerned with stages (1) to (3).

In short, we think that any comparison between IP and CV should take into account their different approaches. By now, our main (provisional) conclusion is that, when $I(f)$ is known in advance (a typical case would be $I(f) = 2$), the IP-method allows to use this prior information in order to obtain some qualitative advantages, concerning bump-hunting, estimation of the derivatives and noise reduction in the estimate. Moreover, the simulation results of Section 5 show also some quantitative improvements in the efficiency with respect to CV, in spite of a trend to oversmoothing under heavy-tailed distributions.

4.2 Modes and convexity

As mentioned in Section 3 above, Silverman's (1981, 1983) proposal, which we will call S-method, is mainly oriented to the problem of hypothesis testing on the number of modes. However, it could also be useful as a suitable bandwidth choice in density estimation when the number of modes is known in advance. We conjecture that, in this case, Silverman's choice would also lead to almost surely uniformly consistent estimates. In the corresponding proof, the convexity arguments used in Theorem 1 should obviously be replaced by others based on monotonicity. A similar reasoning would also allow to prove the a.s. uniform convergence of $\{f_n^{(1)}\}$ to $\{f^{(1)}\}$ when the bandwidth is chosen by the IP-method. This would represent a further example about how the ideas on modes and convexity appear closely related showing interesting applications in probability and statistics. The book of Dharmadhikari and Joag-dev (1988) provides a comprehensive view of these subjects.

4.3 The multivariate case

The direct extension of IP-method to the multivariate case does not seem straightforward. The main problem is that a hypothesis of piecewise convexity is considerably more restrictive when several variables are involved. For instance, the multivariate normal density is neither concave nor convex outside of a central ellipsoid. Nevertheless, this counterexample also suggests a possible approach: the normal, as well as other important multivariate densities (Wishart, Dirichlet,...), is logconcave (see Dharmadhikari and Joag-dev, 1988). We could perhaps adapt the convexity argument of Lemma 1 in order to obtain a version of IP-method for the logconcave case.

4.4 Further open problems and proposals for research

(a) The computation method given in (17) poses two mathematical problems. First, we need sufficient conditions (on f and K) to the almost sure convergence $h_n \to 0$. This problem has been solved by Silverman (1983) for the S-choice. As this author claims, his results (which also include convergence rates) could be adapted to the case in which h_n is defined in terms of sign changes in the second derivative. It should be noted, however, that Silverman's results require compact support for f. The unboundedness of the support could lead to inconsistency. For instance, if $X_1, \ldots, X_n$ is a sample from $f(x) = exp(-x)$ for $x \geq 0$, it is not difficult to prove that $nX_{(1)}$, $(n-1)(X_{(2)} - X_{(1)}), \ldots, X_{(n)} - X_{(n-1)}$ are iid random variables with density f. From this, it seems intuitively clear that the IP-bandwidth does not converge to zero for densities with heavier-than-exponential tails. The formalization of this result, as well as a proof of consistency for thinner tails, appear as interesting targets for further work; perhaps the techniques used by Broniatowski, Deheuvels and Devroye (1989) could be useful in this respect. This phenomenon of inconsistency under heavy tails would be an additional reason (see also Remark 2) for using the IP-method in a *modified* version, in order to make equal the number of inflection points in a fixed (compact) interval and disregard those outside of this interval.

Second, from a formal point of view, it would be interesting to ensure the accessibility of h_n, that is, to prove that the set $\{h\colon \hat{f}_n(t;h)$ has q inflection points$\}$ is almost surely non empty.

(b) Alternative versions of Theorem 2 could probably be obtained for density estimates based on orthogonal expansions. This would require, as a first step, a modified version of Theorem 1, since these estimates need not be density functions for each given sample. However, they present the advantage that the smoothing parameter varies discretely, which would facilitate its choice according to IP-method.

(c) Results on convergence of densities are also interesting in the Bayesian framework when studying convergence of posterior distributions (see Cuevas, 1988) or approximations of the likelihood function (see Boos, 1985; Sweeting, 1986). Theorem 1 could be potentially useful in this context.

5. A Monte Carlo study

5.1 Description of the study. Technical features

The performance of the IP-method has been compared with that of the ordinary pseudo likelihood cross-validation (henceforth denoted by CV) by using three different error measures: the supremum distance, the L_1-norm and the square of the L_2-norm. In all cases the density estimate is of kernel type with Gaussian kernel function. Three underlying densities have been used: the standard normal $N(0,1)$, the mixture $\frac{1}{2}N(0,1) + \frac{1}{2}N(6,2)$, and the standard Cauchy density. In each case we have considered sample sizes n=25, 50. The number of replications is 1000.

TABLE 1. Values of $E\|\hat{f}_n - f\|_\infty$ and $V\|\hat{f}_n - f\|_\infty$.

	Normal		Mixture		Cauchy	
	IP	CV	IP	CV	IP	CV
n=25	0.407298	0.425505	0.201247	0.205258	0.295291	0.296966
	6.4630E-4	4.0511E-3	5.3188E-5	2.9216E-4	3.1374E-3	3.0252E-3
n=50	0.402914	0.416573	0.200652	0.205923	0.270533	0.272566
	1.8745E-4	2.1773E-3	1.2525E-5	3.0436E-4	6.6429E-3	6.1403E-3

TABLE 2. Values of $E\|\hat{f}_n - f\|_1$ and $V\|\hat{f}_n - f\|_1$.

	Normal		Mixture		Cauchy	
	IP	CV	IP	CV	IP	CV
n=25	0.240567	0.267267	0.376117	0.365855	0.963849	0.676643
	1.0547E-2	1.3113E-2	9.0852E-3	9.99826E-3	9.0376E-2	6.2276E-2
n=50	0.195394	0.205668	0.320705	0.286585	1.14384	0.602842
	6.4399E-3	6.7240E-3	6.9889E-3	5.3691E-3	6.8641E-2	4.5860E-2

TABLE 3. Values of $E\|\hat{f}_n - f\|_2^2$ and $V\|\hat{f}_n - f\|_2^2$.

	Normal		Mixture		Cauchy	
	IP	CV	IP	CV	IP	CV
n=25	0.013443	0.018426	0.016015	0.015663	0.0825125	0.050172
	1.6419E-4	4.9284E-4	5.6910E-5	8.4679E-5	2.0401E-3	1.2002E-3
n=50	0.008906	0.010644	0.011884	0.009487	0.091687	0.036348
	6.1776E-5	1.2403E-4	2.9070E-5	3.0407E-5	2.4992E-3	1.0040E-3

Obviously, this study is not comprehensive; further sample sizes and underlying distributions should be considered and other smoothing procedures should be used for comparison. The choice of the pseudo likelihood croos-validation deserves some words of explanation: it is known (see Hall, 1987) that this method is not advisable when large sample sizes and heavy-tailed distributions are involved. However, Bowman (1984) has reported that, for small and moderate sample sizes (which is our case), the pseudo likelihood method provides slightly better results in both the normal and the mixture case. The Cauchy case was of a lesser importance for us because, as we point out above, the poor behaviour of the IP-method in this case could be conjectured in advance (a modified version is suggested in Subsection 4.4).

The study has been carried out by using a *MV20000 Data General* computer. The normal and Cauchy distributions and the proportion of the mixture have been generated by using, respectively, the *IMSL*-subroutines *GGNML*, *GGCAY*, and *GGUD*. The *seed numbers* have been 123456 for n=25 and 1234567 for n=50. Further details are available from us.

5.2 COMMENTS ON THE SIMULATION RESULTS

The numerical outputs are shown in Tables 1, 2, 3. The top and the bottom numbers in each cell are, respectively, the estimated mean value and the estimated variance of the corresponding error. In view of these results, we could make the following remarks:

(a) As we have expected (from intuitive considerations and from our consistency theorem), the IP-method shows an especially good behaviour with respect to the supremum norm. In fact, the IP-errors in Table 1 are consistently smaller for all three distributions and both sample sizes. The IP-variances are also smaller (except in the Cauchy case).

(b) In the normal case, the IP-method provides also better results (not only in mean error, but also in variance) with respect to all the three criteria. The improvements in variance are particularly interesting from a practical point of view, since the high variability is known to be one of the most important problems in the theory of automatic smoothing (see Scott and Terrell, 1987).

(c) In the mixture case, the IP-errors with respect to $\|.\|_1$ and $\|.\|_2$ are larger than those of CV-method. However, the IP-variances are smaller (except for $\|.\|_1$ with n=50).

(c) Not surprinsingly the results in the Cauchy case are worse for the IP choice (except, perhaps, with $\|.\|_\infty$). The point is that, under such a heavy-tailed distribution, the kernel estimates tend to show many little bumps in the tails (see 4.4 above). The IP-method, in its simplest form, is intended to remove all these unimportant bumps paying a price in efficiency, since the removal of the inflection points is achieved at the expense of oversmoothing.

(d) Of course the IP-method, as used here, assumes the prior knowledge of $I(f)$. Hence, the numerical results of tables 1-3 should be considered in the appropri-

ate perspective: they reflect the modifications in efficiency which can be obtained by incorporating this prior knowledge through the IP-choice. The study of the performance of the *pilot* version of IP (*i.e.* with previous estimation of $I(f)$; see Subsection 3.2) remains as an open problem.
(e) Finally, it should be recalled that the basic aims of IP-method are not mainly concerned with gains in efficiency, but with qualitative targets (bump -hunting, estimation of the derivatives,...). Hence, the IP-method should not be considered as a direct competitor of CV, because the different purposes and settings of both methods.

ACKNOWLEDGEMENTS. The research of the first author has been partially supported by DGICYT grant PB88-0178 (Spain). We are grateful to Luc Devroye for interesting suggestions concerning Subsection 4.4, and to Ignacio García Jurado and José Luis Domínguez Salgueiro for their help in the simulation study.

REFERENCES

BICKEL, P.J. (1965). On some robust estimates of location. *Ann. Math. Statist.* **36**, 847-858.
BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.
BONEVA, L.I., KENDALL, D.G. and STEFANOV, I. (1971). Spline transformations. *J. Roy. Statist. Soc.* **B 33**, 1-70.
BOOS, D.D. (1985). A converse to Scheffé's theorem. *Ann. Statist.* **13**, 423-427.
BOWMAN, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353-360.
BRONIATOWSKI, M., DEHEUVELS, P. and DEVROYE, L. (1989). On the relationship between stability of extreme order statistics and convergence of the maximum likelihood kernel density estimate. *Ann. Statist.* **17**, 1070-1086.
CUEVAS, A. (1988). On stability in a general decision problem. *Statistica* **XLVIII**, 9-14.
DHARMADHIKARI, S. and JOAG-DEV, K. (1988). *Unimodality, Convexity, and Applications.* Academic Press, New York.
DEVROYE, L. (1987). *A Course in Density Estimation.* Birkhäuser, Boston.
GOOD, I.J. and GASKINS, R.A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75**, 43-73.
HALL, P. (1987). On Kullback-Leibler loss and density estimation. *Ann. Statist.* **15**, 1491-1519.
HALL, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multiv. Anal.* **32**, 177-203.
HETTMANSPERGER, T.P. (1968). On the trimmed Mann-Whitney statistic. *Ann. Math. Statist.* **39** , 1610-1614.
HETTMANSPERGER, T.P. and KLIMKO, L.A. (1974). A note on the strong convergence of distributions. *Ann. Statist.* **2**, 597-598.

HUBER, P.J. (1981). *Robust Statistics.* Wiley, New York.
LAY, S.R. (1982). *Convex Sets and Their Applications.* Wiley, New York.
MARRON, J.S. (1987). A comparison of cross-validation techniques in density estimation. *Ann. Statist.* **15**, 152-162.
OREAR, J. and CASSEL, D. (1971). *Applications of statistical inference to physics.* In "Foundations of Statistical Inference", V.P. Godambe and D.A. Sprott, eds., pp. 280-288. Holt, Rinehart and Winston of Canada, Toronto.
REISS, R.D. (1973). On the measurability and consistency of maximum likelihood estimates for unimodal densities. *Ann. Statist.* **1**, 888-901.
SCOTT, D.W. and TERRELL, G.R. (1987). Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.* **82**, 1131-1146.
SETHURAMAN, J. (1961a). Some limit theorems for joint distributions. *Sankhyā* **A 23**, 379-386.
SETHURAMAN, J. (1961b). Some limit theorems connected with fractile graphical analysis. *Sankhyā* **A 23**, 79-90.
SILVERMAN, B.W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* **6**, 177-184.
SILVERMAN, B.W. (1981). Using kernel density estimates to investigate multimodality. *J. R. Statist. Soc.* **43**, 97-99.
SILVERMAN, B.W. (1983). *Some properties of a test for multimodality based on kernel density estimates.* In "Probability, Statistics and Analysis", J.F.C. Kingman and G.E.H. Reuter eds., pp. 248-259. Cambridge University Press.
SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.
STONE, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12**, 1285-1297.
SWEETING, T.J. (1986). On a converse to Scheffé's theorem. *Ann. Statist.* **14**, 1252-1256.

PROSPECTS FOR AUTOMATIC BANDWIDTH SELECTION IN EXTENSIONS TO BASIC KERNEL DENSITY ESTIMATION

M.C. JONES
Department of Statistics
The Open University
Walton Hall
Milton Keynes
MK7 6AA UK

ABSTRACT. Recently, much progress has been made on understanding the bandwidth selection problem in kernel density estimation. Here, analogous questions are considered for extensions to the basic problem, namely, for estimating derivatives, using "better" kernel estimators, and for the multivariate case. In basic kernel density estimation, recent advances have resulted in considerable improvements being made over "moderate" methods such as least squares cross-validation. Here, it is argued that, in the first two extension cases, the performance of moderate methods deteriorates even more, so that the necessity for "improved" methods — and indeed their potential in theory if not necessarily in practice — is greatly increased. Rather extraordinary things happen, however, when higher dimensions are considered.

1. Introduction

Considerable recent advances on the problem of data-based smoothing parameter selection are much in evidence at this meeting. In the particular context of bandwidth choice in kernel density estimation where performance is judged by the mean integrated squared error (MISE) criterion, it is now understood how to obtain methods that go (far) beyond the $n^{-1/10}$ relative error rate of convergence (RERoC) enjoyed by the well-known "least squares cross-validation" (CV) and a plethora of other methods to be found in the literature (Jones and Kappenman, 1990). (That is, $\hat{h}_{CV}/h_0 = 1 + O_p(n^{-1/10})$ as $n \to \infty$, where $\hat{h}_{CV}$ is the cross-validatory bandwidth choice, h_0 is the value of the bandwidth that minimises MISE, and n is the sample size). Indeed, Steve Marron discusses elsewhere in this volume how methods with the much better (and, in fact, best possible) RERoC of $n^{-1/2}$ can be attained. On the way to $n^{-1/2}$ RERoC, yet other methods have emerged that have intermediate theoretical properties (e.g. Park and Marron, 1990, Sheather and Jones, 1990) but which seem to work particularly well in practice (especially Sheather and Jones, 1990).

Virtually all of this work concerns the basic problem of estimating a univariate density, f, itself, using a nonnegative constant kernel estimate. In this paper, we investigate the potential performance (in RERoC terms) of the natural extensions of some of these methods to three important variations on the basic problem. The first of these variations is the question of estimating derivatives of f; these will be indexed by $\nu = 0, 1, \ldots$. The second variation concerns the use of higher order kernels

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 241–249.
© 1991 *Kluwer Academic Publishers.*

in the density estimates. A kernel K of order k for estimating $f^{(\nu)}$ has the properties

$$\int x^j K(x)dx = \begin{cases} (-1)^\nu \nu! & \text{if } j = \nu, \\ 0 & \text{if } 0 \le j \le \nu - 1 \text{ or } \nu + 1 \le j \le \nu + k - 1, \\ s \neq 0 & \text{if } j = \nu + k. \end{cases}$$

When $\nu = 0$, such kernels cannot be nonnegative unless $k = 2$ which is the basic case. Here, we also consider other even values $k = 4, 6, \ldots$ and tentatively relate our results to use of variable kernel density estimates. The third and final variation on the basic theme allows the multivariate case; dimensionality d might be $2, 3, \ldots$ rather than just 1. Note that here we follow the usual trick of employing the same single bandwidth in all directions, implicitly assuming some preliminary transformation of the data to make this at least approximately appropriate as, for example, in Silverman (1986), Section 4.2.1. Notice that the first and third variations are harder problems than the basic one, and that there is a consequent degradation in kernel estimation performance. Conversely, the modified estimates of the second variation are designed to enhance performance, at least in theory. (All three extensions necessitate use of larger bandwidths, however).

The $n^{-1/10}$ RERoC of CV and related methods in the basic case is usually described as being "extremely slow". A major point of this paper is that extensions of CV and other "$n^{-1/10}$ methods" (a representative of which will be called $\hat{h}_{n10}$) to the first two of the above situations always results in even worse performance in RERoC terms. Therefore, it seems that the need to develop alternatives to CV is even more crucial in these cases. Prospects for two such alternatives are also discussed in this paper. We should at this stage make it clear that criticism of CV is not principally aimed at the cross-validatory "leaving-one-out" aspect of it itself; rather, the main inadequacy of CV and related methods is their (implicit) use of too small a "pilot bandwidth" (for estimation of MISE). Remedying this by employing a more appropriate pilot bandwidth is the key to the improved performance of $\hat{h}_{PM}$, Park and Marron's (1990) bandwidth, which in fact has a RERoC of order $n^{-4/13}$ in the basic situation. Applying a little cunning while also dropping the cross-validatory aspect that remains in $\hat{h}_{PM}$ leads to Sheather and Jones's (1990) further improvement, $\hat{h}_{SJ}$, whose RERoC is of order $n^{-5/14}$ (and which also has a much reduced constant multiplier). It is the natural extensions of $\hat{h}_{PM}$ and $\hat{h}_{SJ}$ to the variations on the basic theme that are also considered here. We will see that although, in the first two situations, these too (usually, but not always) get worse in RERoC terms, they tend not to deteriorate so much and hence hold out hope for continued good bandwidth selectors in these other circumstances. Note that we do not, however, address practical issues here except to make some tempering comments in Section 6.

Our general, but simple, framework, and results for the first two extensions are presented in Section 2. To make more of the full results, each variation on the basic problem is considered individually in Section 3. A discussion of the multivariate case, which admits surprising results that are difficult to explain, is presented in Section 4. In Section 5, we ask if all this work in terms of RERoC's of $\hat{h}$'s to h_0 tells the whole story in regard of performance of associated $\hat{f}$'s, and argue that the answer is "yes". The current paper can be seen as an extension of Section 4 of Marron (1986) which, we claim, answers and clarifies some of the points made there.

The paper is couched in density estimation terms, but there is considerable relevance also to kernel regression and spectral density estimation.

2. General Framework and Results

Consider first the basic kernel density estimation problem in which $\nu = 0$, $k = 2$ and $d = 1$. The usual asymptotic expansion of MISE for large n and small (but not too small) bandwidth h looks

like

$$CR(f'')h^4 + D(nh)^{-1}$$

from which the well-known asymptotic approximation to h_0 is

$$h_* = C\{R(f'')n\}^{-1/5}. \tag{1}$$

Throughout the paper, the letters C and D are reserved for (various) constants, which depend on K but not on f, and which are of no particular interest here. Also, $R(\gamma)$, say, means $\int \gamma^2(x)dx$. An obvious strategy to obtain practicable $\hat{h}$'s is thus to replace $R(f'')$ on the right-hand side of (1) by an estimate of it. We will consider basing that estimate on a kernel density estimate $\hat{f}_g$, say, using the same kernel K (or at least a kernel with the same order k) as the final density estimate itself, but with a potentially different bandwidth g. Thus, we look at estimates rather like

$$\hat{h} = C\{R(\hat{f}_g'')n\}^{-1/5}. \tag{2}$$

To define $\hat{h}_{n10}$ and versions of $\hat{h}$ *asymptotically equivalent* (in RERoC terms) to $\hat{h}_{PM}$ and $\hat{h}_{SJ}$, the following choices are made in (2). For $\hat{h}_{n10}$ and $\hat{h}_{PM}$ it is not precisely $R(\hat{f}_g'')$ but rather $R(\hat{f}_g'')$ minus that term, independent of the data, corresponding to "$i = j$" terms in the double summation (over i and j) involved, which is used in the right-hand side of (2). Then, for $\hat{h}_{n10}$, set $g = h$. Now, h is intended to be chosen appropriately for estimating f itself and is, in fact, rather too small for estimating the functional $R(f'')$ well. Taking $g = g_0$, where g_0 is MSE-optimal for estimating $R(f'')$, defines, essentially, $\hat{h}_{PM}$. Finally, to obtain $\hat{h}_{SJ}$, reinstate the previously deleted $i = j$ terms and choose $g = g_1 \neq g_0$ optimally for the re-modified estimator. The trick here is to use g to cancel the positive $i = j$ terms with the leading smoothing bias term which is negative. It turns out that this bias reduction technique reduces MSE more than balancing squared bias and variance does when $i = j$ terms are not present.

Aside. While these specifications are sufficient for RERoC calculations, they do not provide fully practicable implementations, of course. Also, $\hat{h}_{CV}$ shares with $\hat{h}_{n10}$ the notion of using h in place of g in, effectively, loss function estimation; in fact, in our extensions, other $\hat{h}$'s in Jones and Kappenman (1990) display exactly the same REROC's as $\hat{h}_{n10}$, while $\hat{h}_{CV}$ itself has a very small advantage. Employing a pilot estimate of $g = h$ in (2) actually gives the earliest of plug-in ideas, while solving (2) for h, when $\hat{h} = h$ and $g = h$, makes for another popular method of the 1970's. In fact, these $\hat{h}$'s can have a better constant multiplier of RERoC than does $\hat{h}_{CV}$ (Jones and Kappenman, 1990). Questions such as how g's should depend on h and n, and how proportionality constants, which here depend on further functionals of f, should be estimated are extremely important in practice but will not be considered in detail here. Let us just say that $\hat{h}_{PM}$ and $\hat{h}_{SJ}$ do resemble their descriptions above rather closely, with g_0 and g_1 being thought of as multiples of appropriate powers of h and further kernel and normal-based estimation (involving scale estimation) being used at various stages to complete the methods.

From (2), performance of $\hat{h}$'s is directly linked to that of corresponding estimates of $R(f'')$. From Hall and Marron (1987), the MSE of $R(\hat{f}_g'')$ minus $i = j$ terms is of the form

$$E_f g^4 + Dn^{-2}g^{-9}, \tag{3}$$

where E_f is an unspecified (for our purposes) function of constants, K *and* f, but not of g or n. Setting $g = h \propto n^{-1/5}$ in (3) yields an MSE of order $\max(n^{-4/5}, n^{-1/5}) = n^{-1/5}$. Optimal choice

of $g = g_0$ is, however, $O(n^{-2/13})$ corresponding to an optimal MSE of order $n^{-8/13}$. Likewise, from Jones and Sheather (1990), the MSE of $R(\hat{f}_g'')$ itself looks like

$$(E_f g^2 - Cn^{-1}g^{-5})^2 + Dn^{-2}g^{-9} \tag{4}$$

where E_f and C are, usually, of the same sign, in which case $g = g_1$ can be chosen to cancel the first two terms, making $g_1 \propto n^{-1/7}$ and the MSE (from the variance term) $O(n^{-5/7})$. (Note that $g = h$ yields $O(1)$ MSE in (4), so taking $g = h$ and using cross-validatory $i = j$ deletion go hand in hand). From (2), these MSE rates translate directly to RERoC's by square rooting, giving orders $n^{-1/10}$, $n^{-4/13}$ and $n^{-5/14}$ for $\hat{h}_{n10}$, $\hat{h}_{PM}$ and $\hat{h}_{SJ}$, respectively, as above.

It is now a relatively simple matter to extend this analysis to the general case of arbitrary ν, k and d, although we do this initially only for general ν and k (leave $d = 1$). Unless otherwise stated, in this and the following section, $\hat{h}_{n10}$ and $\hat{h}_{CV}$ have identical performance, and we will use the notation $\hat{h}_{n10}$ to refer to either. Our starting point is the asymptotic expansion of MISE once more, which now has the form

$$CR(f^{(\nu+k)})h^{2k} + Dn^{-1}h^{-(2\nu+1)}$$

so that

$$h_* = C\{R(f^{(\nu+k)})n\}^{-1/(2(\nu+k)+1)}. \tag{5}$$

Kernel-based estimates of $R(f^{(\nu+k)})$ remain definable both with and without $i = j$ terms, and the general forms of both Hall and Marron's (1987) and Jones and Sheather's (1990) MSE results are straightforwardly obtainable. In the former case, and again using bandwidth g, we get

$$E_f g^{2k} + Dn^{-2}g^{-(4(\nu+k)+1)}. \tag{6}$$

Thus, using $g = h \propto n^{-1/(2(\nu+k)+1)}$ yields an MSE of order $n^{-1/(2(\nu+k)+1)}$ also, and thus a RERoC for (generalised) $\hat{h}_{n10}$ of order $n^{-1/\{2(2(\nu+k)+1)\}}$ $(= O(h_*^{1/2}))$. It is immediately clear that the RERoC performance of $\hat{h}_{n10}$ deteriorates from $n^{-1/10}$, which is best, for each of ν and k increasing. The general version of $\hat{h}_{PM}$ would take $g = g_0 \propto n^{-2/(4\nu+6k+1)}$ which results in a RERoC of order $n^{-2k/(4\nu+6k+1)}$. Reinserting $i = j$ terms gives an MSE of "$R(\hat{f}_g^{(\nu+k)})$" of the form

$$(E_f g^k - Cn^{-1}g^{-(2(\nu+k)+1)})^2 + Dn^{-2}g^{-(4(\nu+k)+1)}. \tag{7}$$

Thus $g = g_1 \propto n^{-1/(2\nu+3k+1)}$ and the RERoC of "$\hat{h}_{SJ}$" is $n^{-(2k+1)/\{2(2\nu+3k+1)\}}$. Thus, both $\hat{h}_{PM}$ and $\hat{h}_{SJ}$ also deteriorate with increasing ν (but from a better starting point and, in a sense, less quickly), while the performance of the former actually improves with increasing k. Rather than dwell further on these general formulae here, we will attempt to get a feel for the amount of deterioration or improvement in each case separately in the following section.

Aside. Throughout this paper, we assume "smooth enough" f. In the basic case, for the given RERoC's to hold, f must have smoothness of order $p \geq 2\frac{1}{4}$ for $\hat{h}_{n10}$ (but only 2 for $\hat{h}_{CV}$), $p \geq 3$ for $\hat{h}_{PM}$ and $p \geq 3\frac{1}{4}$ for $\hat{h}_{SJ}$. Here, p is defined by $|f^{(l)}(x) - f^{(l)}(y)| \leq M|x - y|^\epsilon$ for all x and y, for $M > 0$ a constant and $p = l + \epsilon$ where l is an integer and $0 < \epsilon \leq 1$. In general, we need $p \geq \nu + k + \frac{1}{4}$ for $\hat{h}_{n10}$ ($\nu + k$ for $\hat{h}_{CV}$), $p \geq \nu + \frac{3}{2}k$ for $\hat{h}_{PM}$ and $p \geq \nu + \frac{3}{2}k + \frac{1}{4}$ for $\hat{h}_{SJ}$. Thus, $\hat{h}_{CV}$, in particular, does have an advantage in terms of robustness to some lack of smoothness of f. Although the above p's are required for the RERoC's given here to obtain, $\hat{h}_{PM}$ and $\hat{h}_{SJ}$ actually maintain some RERoC advantage over $\hat{h}_{CV}$ until p reaches values (not given) intermediate to those quoted above.

The performance of $\hat{h}_{n10}$ deteriorates even more quickly with increasing k than it did for increasing ν. The RERoC of $n^{-1/18}$ when $k = 4$ appears to be applicable also to the important case of variable kernel density estimation (which shares fourth order kernel estimation rates) and to which Silverman (1986) and Hall (1989a) have suggested applying cross-validatory bandwidth choice. The behaviour of α_{PM} and α_{SJ} is quite different than before. PM's RERoC actually improves (slowly) with increasing k, while SJ's α decreases at much the same rate, both α_{PM} and $\alpha_{SJ} \to 1/3$ as $k \to \infty$.

To reiterate, the main reason for $\hat{h}_{n10}$'s poor performance is its implicit use of $h \sim n^{-1/(2(\nu+k)+1)}$ in place of g in estimating (5). This bandwidth is too small to afford a good estimate of $R(f^{(\nu+k)})$; recall that $g_0 \sim n^{-2/(4\nu+6k+1)}$ and $g_1 \sim n^{-1/(2\nu+3k+1)}$ are optimal for their respective estimators. The extent of the difference between h and g_0, g_1 is greater when k increases (rather than ν which we again set to 0). Table 3 below gives numerical evidence of this. Values in this table are of β, when h, g_0 and g_1 are written $n^{-\beta}$; in fact, $\beta = 1/(2k+1)$, $2/(6k+1)$ and $1/(3k+1)$ for these bandwidths in this situation. (Remember that smaller β's mean bigger bandwidths).

k	n10	PM	SJ
2	1/5 = .200	2/13 = .154	1/7 = .143
4	1/9 = .111	2/25 = .080	1/13 = .077
6	1/13 = .077	2/37 = .054	1/19 = .053

All β's decrease with increasing k. Row-wise, the decrease from β_{n10} to β_{PM} and β_{SJ} becomes more accentuated with increasing k, $\beta_{PM}/\beta_{n10} \to 2/3$, as does β_{SJ}/β_{n10}, as $k \to \infty$. We will not bore the reader with a similar table for increasing ν. If we had, there would also be differences apparent, although they get less substantial as ν increases (in fact, in those cases, $\beta_{PM}/\beta_{n10} \to 1$, and likewise for β_{SJ}/β_{n10}).

3. The Multivariate Case

For arbitrary d as well as ν and k, the asymptotic MISE expansion has the form

$$CR(f^{(\nu+k)})h^{2k} + Dn^{-1}h^{-(2\nu+d)}$$

and

$$h_* = C\{R(f^{(\nu+k)})n\}^{-1/(2(\nu+k)+d)}. \tag{8}$$

(There is, of course, an abuse of notation here when $d > 1$, $f^{(\nu+k)}$ referring then to $(\nu+k)$'th order partial derivatives of f and $R(.)$ meaning the (multiple) integral of the square of appropriate sums of partial derivatives). Kernel-based estimates of $R(f^{(\nu+k)})$ can again be defined both with and without $i = j$ terms, and multivariate generalisations of both Hall and Marron (1987) and Jones and Sheather (1990) can be obtained. In the first case, where g is the bandwidth again, the relevant MSE is

$$E_f g^{2k} + Dn^{-2}g^{-(4(\nu+k)+d)}. \tag{9}$$

Therefore, $g = h \propto n^{-1/(2(\nu+k)+d)}$ implies an MSE of order $n^{-\min(2k,d)/(2(\nu+k)+d)}$, so that the RERoC for (multivariate) $\hat{h}_{n10}$ is like $n^{-\min(2k,d)/\{2(2(\nu+k)+d)\}}$ $(= O(h_*^{\min(k,d/2)}))$. This is where $\hat{h}_{CV}$ differs

from $\hat{h}_{n10}$, however, since, at least if we up the smoothness requirement on f, for $\hat{h}_{CV}$ the g^{2k} bias term can be reduced, and the "bounding" k in the numerator of the fraction in its RERoC disappears. The generalised version of $\hat{h}_{PM}$ takes $g = g_0 \propto n^{-2/(4\nu+6k+d)}$ giving a RERoC of order $n^{-2k/(4\nu+6k+d)}$. Reinserting $i = j$ terms gives an MSE of the form

$$(E_f g^k - Cn^{-1}g^{-(2(\nu+k)+d)})^2 + Dn^{-2}g^{-(4(\nu+k)+d)}, \tag{10}$$

so $g = g_1 \propto n^{-1/(2\nu+3k+d)}$ and the RERoC of $\hat{h}_{SJ}$ is $n^{-(2k+\min(4,d))/\{2(2\nu+3k+d)\}}$.

A brief perusal of the above rates shows us that there is something quite different going on here from what was seen in Sections 2 and 3. To clarify, look at the $\nu = 0$, $k = 2$ case in more detail. Reverting to the notation $n^{-\alpha}$ for the RERoC's, we see that $\alpha_{n10} = \min(4, d)/\{2(4 + d)\}$, $\alpha_{CV} = d/\{2(4 + d)\}$ (as in Marron, 1986), $\alpha_{PM} = 4/(12 + d)$ and $\alpha_{SJ} = \{4 + \min(4, d)\}/\{2(6 + d)\}$. Numerical values of these α's are given for $d = 1, 2, 4, 6, 10$ in Table 4 below.

d	n10	CV	PM	SJ
1	1/10 = .100	1/10 = .100	4/13 = .308	5/14 = .357
2	1/6 = .167	1/6 = .167	2/7 = .286	3/8 = .375
4	1/4 = .250	1/4 = .250	1/4 = .250	2/5 = .400
6	1/5 = .200	3/10 = .300	2/9 = .222	1/3 = .333
10	1/7 = .143	5/14 = .357	2/11 = .182	1/4 = .250

All but $\hat{h}_{PM}$ exhibit improved performance at least initially. The watershed is then at $d = 4$. This is where the best RERoC's of both $\hat{h}_{n10}$ and $\hat{h}_{SJ}$ occur. It is interesting to note that at $d = 4$, n10, CV (which has monotonically improving RERoC) and PM (for which RERoC is monotonically getting worse) have exactly the same value of α, while SJ remains the best of the four. For $d > 4$, n10, PM and SJ steadily decline, while holding their accustomed relative merits. CV, on the other hand, without a bias term to stop it (for smooth enough f), continues to improve, and indeed α_{CV} actually tends to $\frac{1}{2}$ as $d \to \infty$ ($\hat{h}_{CV}$ "overtakes" $\hat{h}_{SJ}$ at $d = 7$). We might note, however, that the importance of what happens for a multivariate kernel estimate beyond about dimension 4 may not be of any great concern in practice because of the generally poor performance of the method in high dimensions.

Describing this extraordinary behaviour is one thing, explaining it is quite another. PM's performance is most easily understood, since it continues with the usual type of bias / variance tradeoff; its RERoC thus necessarily decreases with increasing d. MSE's of n10, CV and SJ are all dominated by variance early on, and use g's such that this variance decreases with increasing d. At $d = 4$, however, the squared bias term cuts back in for both n10 and SJ, and continues to dominate (with disadvantageous consequences) as d goes on. This effect is just not there for CV. Any real intuitive insights into these phenomena would be much appreciated! Note that Marron's (1986) speculation, based on the general k and d cases, that there is an inverse relationship between problem hardness and worse CV performance, is not supported by the general ν situation, where a harder problem and a poorer CV performance go together.

4. Does it all Matter? I

We have, thus far, framed this paper entirely in terms of how well $\hat{h}$'s estimate h_0. Is this an appropriate measure of performance, given that interest surely lies in how well $\hat{f}_{\hat{h}}$ estimates f, at least relative to how well $\hat{f}_{h_0}$ does?

Of course, it doesn't matter that $\hat{h} \neq h_0$ exactly; certainly, $\hat{f}_{\hat{h}}$ will, in practice, closely resemble $\hat{f}_{h_0}$ for $\hat{h}$ "near enough" to h_0. Ideally, one would like to quantify this notion and hence identify an interval around h_0 within which we would be happy for our $\hat{h}$ to lie. In the absence of such a measure, a particularly convenient proxy is provided by the MISE curve: $\hat{h}$ can be defined to be close to h_0 if MISE($\hat{h}$) is little increased over MISE(h_0) i.e. degree of importance of getting very close to h_0 is linked to flatness, or otherwise, of the MISE curve. (Note that we are taking care to avoid a direct interpretation of MISE($\hat{h}$) — rather meaninglessly defined by inserting the data-dependent quantity $\hat{h}$ in the place of a fixed h — as a relevant loss function for $\hat{f}_{\hat{h}}$).

Now,

$$\mathrm{MISE}(\hat{h}) = \mathrm{MISE}(h_0) + \frac{1}{2}(\hat{h} - h_0)^2 \mathrm{MISE}''(h_m)$$

for some h_m between $\hat{h}$ and h_0. Rearranging, and approximating h_m by h_0, yields

$$\frac{\mathrm{MISE}(\hat{h})}{\mathrm{MISE}(h_0)} - 1 \simeq \frac{1}{2}\frac{h_0^2\,\mathrm{MISE}''(h_0)}{\mathrm{MISE}(h_0)}\left(\frac{\hat{h}}{h_0} - 1\right)^2. \tag{11}$$

Clearly, $h_0^2\,\mathrm{MISE}''(h_0)/\mathrm{MISE}(h_0) = O(1)$ and, in fact, a little manipulation shows that the appropriate constant multiple of the squared RERoC term $(\hat{h}/h_0 - 1)^2$ in (11) is simply $k(2\nu + d)$. Thus (i) RERoC's of $\hat{h}$ to h_0 do indeed give full rate information about MISE($\hat{h}$)'s relative error comparison with MISE(h_0) and (ii) the constant multiplier in this relationship increases with each of ν, k and d.

The (theoretical) issue of whether we are addressing the right question in dealing exclusively with RERoC's of $\hat{h}$'s to h_0 is thus answered in the affirmative. This conflicts with the impression given on p.175 of Marron (1986) where $\mathrm{MISE}''(h_0) \sim n^{-(2k-2)/(2(\nu+k)+d)}$ itself, essentially, is cited, since it is a measure of curvature of the MISE curve at its minimum. MISE$''(h_0)$, of course, increases with increasing ν and d and decreases with k so, on this basis, one might expect the larger ν, d problem(s) to be less sensitive, and the increased k one more so, to poor choice of h. But $\mathrm{MISE}''(h_0) \sim h_0^2\,\mathrm{MISE}(h_0)$ is an *absolute* measure of sensitivity and (overwhelmingly) reflects the "problem hardness" rates of MISE we saw before. In contrast, *relative* values of MISE, as in (11), appear to us to be the more meaningful measure of sensitivity to choice of $\hat{h}$, and hence the results of this paper obtain.

Aside. Let h_2 be the bandwidth that minimises MISE$''(h)$. Then, in general,

$$h_2 \simeq \left\{\frac{(2\nu + d + 1)(2\nu + d + 2)}{(2k - 1)(2k - 2)}\right\}^{1/(2k+2\nu+d)} h_0, \tag{12}$$

and if $k = \nu + \frac{1}{2}(d + 3)$ (which is the popular case of $k = \nu + 2$ when $d = 1$), then $h_2 \simeq h_0$. This insight comes from Chiu (1990), in (essentially) the case $\nu = 0$. It means that (asymptotically), in these cases, $\mathrm{MISE}''(h_0) \leq \mathrm{MISE}''(h)$ and we are underestimating the constant in (11).

5. Does it all Matter? II

Colleagues have been quick to remind us that asymptotic convergence rates are not the whole story. What about practice? The theory tells us that the functionals involved in the extended problems are harder to estimate. It is unclear exactly how this translates to practice; it may even that "hard" becomes "impossible" for realistic sample sizes. Slightly less similar bandwidth selection problems make the"CV/not CV" dispute yet more unclear in practice. The case of variable kernel estimation is a good example. While CV is, here as elsewhere, relatively straightforward to extend (Silverman, 1986, Hall, 1989a), PM/SJ lookalikes require estimation of the functional

$$\int w(x) \left[\frac{d^4}{dx^4} \left\{ \frac{1}{f(x)} \right\} \right]^2 dx,$$

(e.g. Hall, 1989b) where one needs to choose a suitable weight function w (Hall, 1989a). It has to be said that estimation of this functional looks to be a practical nightmare. (We hope to investigate an alternative, however, based on directly estimating the bias itself, not its asymptotic representation). The answer to the question of consequences for practice, then, is that we do not, as yet, know.

For all that, we hope that the RERoC analysis of this paper is useful. The message is that one cannot afford to be complacent about using cross-validatory bandwidth choice because its performance may well prove to be quite poor. Consequently, the development of alternative methods in extensions to the basic problem seems to be particularly important, although success in such a venture may well not be guaranteed in practice.

Acknowledgements

I am very grateful to Peter Hall for his wise words on practicability, which underlie Section 6, and to Steve Marron for reminding me of his very relevant 1986 paper.

References

Chiu, S.-T. (1990) 'Why bandwidth selectors tend to choose smaller bandwidths, and a remedy', Biometrika 77, 222-226.

Hall, P. (1989a) 'On bandwidth selection for variable bandwidth density estimation', to appear.

Hall, P. (1989b) 'On the bias of variable bandwidth curve estimators', Technical Report CSTR-011-89, Statistics Research Section, Australian National University.

Hall, P. and Marron, J.S. (1987) 'Estimation of integrated squared density derivatives', Statist. Probab. Lett. 6, 109-115.

Hall, P., Sheather, S.J., Jones, M.C. and Marron, J.S. (1989) 'On optimal data-based bandwidth selection in kernel density estimation', to appear.

Jones, M.C. and Kappenman, R.F. (1989) 'On a class of kernel density estimate bandwidth selectors', to appear.

Jones, M.C. and Sheather, S.J. (1990) 'Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives', Statist. Probab. Lett., to appear.

Marron, J.S. (1986) 'Will the art of smoothing ever become a science?', Contemp. Math. 59, 169-178.

Marron, J.S. (this volume) 'Root-n bandwidth selection'.

Park, B.U. and Marron, J.S. (1990) 'Comparison of data-driven bandwidth selectors', J. Amer. Statist. Assoc. 85, 66-72.

Sheather, S.J. and Jones, M.C. (1990) 'A reliable data-based bandwidth selection method for kernel density estimation', J. Roy. Statist. Soc. Ser. B, to appear.

Silverman, B.W. (1986) Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.

ROOT N BANDWIDTH SELECTION

J. S. Marron
Department of Statistics
University of North Carolina
Chapel Hill, N. C. 27514
U. S. A.

ABSTRACT: For various data–based bandwidth selectors for a kernel density estimator, the relative rate of convergence of the selected bandwidth is considered. Several methods have recently been found which have the very fast rate of convergence of the square root of the sample size. Such a fast rate of convergence is quite surprising, in view of those usually encountered in nonparametric curve estimation problems. A survey and comparison of methods with this very fast rate of convergence is given.

1. Introduction and Notation

Nonparametric curve estimation is becoming increasingly recognized as a powerful data analytic tool. See Silverman (1986), Müller (1988), Eubank (1988) and Härdle (1990) for many interesting examples. All major types of nonparametric curve estimators, including histograms, kernels, splines, orthogonal series, etc., in many different settings, including regression, density, spectral density, and hazard estimation, depend on a smoothing parameter. In this paper smoothing parameter selection is studied in what may be considered the simplest of these many contexts: kernel density estimation.

A common mathematical means of formulating the density estimation problem, see Devroye and Györfi (1984) or Silverman (1986) for early references, is to assume that $X_1,\ldots,X_n$ are a set of independent, identically distributed observations from a density (with respect to Lebesgue measure) $f(x)$. Given a "bandwidth", or smoothing parameter $h > 0$, and a "kernel function" K (assumed here to be a probability density), the kernel estimator is defined by

$$f_n(x) = n^{-1} \sum_{i=1}^{n} K_h(x-X_i),$$

where $K_h(\cdot) = K(\cdot/h)/h$, is this rescaling of K.

The choice of bandwidth h is crucial to the performance of this estimator. Hence the focus of this paper is on data based bandwidth selection procedures. An

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 251–260.
© 1991 Kluwer Academic Publishers.

important way to study this problem, and assess various procedures, is by considering error criteria which indicate in some sense how well f_n is estimating f. One choice to be made in the selection of a criterion is pointwise vs. global. A pointwise criterion considers error only at a single point x, while a global one concerns error over a much larger set, e.g. the entire line is considered. In this paper, only the global criteria are considered, because that seems the most relevant to the way density estimators are used in practice: for the estimation of an entire curve. Given a global approach is to be made, the next issue to address is which norm should be used. There are a number of important intuitive reasons for using the L^1 norm, nicely made clear in Devroye and Györfi (1984). However in this paper, only the L^2 will be considered for reasons of tractability. Because of its simple inner product structure, this norm provides both more easily accessible intuitive insight, and also allows far deeper mathematical results with much less effort. In particular this is very clear in bandwidth selection, where the L^1 theory currently seems at least six years behind that developed for L^2. While the strong reasons given for use of the L^1 norm do indeed make the process of extending important ideas to that case worthwhile, it seems most appropriate to first understand these big ideas through consideration of the L^2 case.

Given that we will study the bandwidth selection problem from the L^2 viewpoint, there is still one other important choice to be made. One possibility is to use the Integrated Squared Error,

$$\mathrm{ISE}(h) = \int (f_n - f)^2,$$

and another is to use its expected value, the Mean Integrated Squared Error,

$$\mathrm{MISE}(h) = E(\mathrm{ISE}(h)).$$

The bandwidths which minimize these criteria (with ties arbitrarily broken by taking the smallest minimizer) will be denoted h_{MISE} and h_{ISE}.

The choice between ISE and MISE is currently a much more controversial issue than that of L^1 vs. L^2. See Hall and Marron (1989), Mammen (1988) and Jones (1989) for various viewpoints on the issues involved here. In this paper attention is restricted to MISE. This is because the main point here is study and comparison of various methods of obtaining an $n^{-1/2}$ relative rate of convergence to h_{MISE}, while it is known that no possible estimator (either one already proposed, or in fact any other function of the data) can possibly have a faster rate of convergence than the far slower $n^{-1/10}$ to h_{ISE}, in various minimax senses. See Hall and Marron (1987) for a special case, and section 2.2 of Hall and Marron (1989) for much more powerful and general versions of such lower bound results.

Sections 2, 3, 4 and 5 all describe various methods that have been discovered for

finding bandwidth selectors with an $n^{-1/2}$ rate of convergence. Also given are original references, together with discussion and comparison.

In all cases, important technical assumptions (other assumptions vary, depending on which case, see the cited papers for details) are that K is a probability density, and that the underlying density f has 4.25 derivatives, in the sense that derivatives of order up to 4 exist, and the fourth derivative satisfies the Lipschitz condition (i.e. Hölder continuity condition),

$$|f(x) - f(y)| \leq C|x-y|^{1/4}, \qquad \text{for all } x, y,$$

for some constant $C > 0$. The number 4.25 has been shown to be the best possible in Theorem 3.2 of Hall and Marron (1989).

A philosophical argument could be made against these assumptions on the grounds that if one feels "4.25 derivatives are really present", then one should use a higher order kernel (see for example section 3.6 of Silverman 1986), to obtain a faster rate of convergence of f_n to f. If this is done then stronger smoothness assumptions are required for $n^{-1/2}$ convergence of the data based bandwidth, which (if this is really useful) one might feel again should more properly be used by an even higher order kernel in f_n, and in fact this cycle could be continued. However the viewpoint taken in this paper is that this line of thought is not so relevant because higher order kernels have nearly never been used in practice, largely because most of the appealing and simple intuition of a kernel estimator (which is the usual reason given by those data analysts who find the kernel more compelling than say a spline or orthogonal series method) is lacking, but also for less important reasons such as the resulting estimator typically takes on negative values somewhere. Hence even when there are more than "two derivatives present" the nonnegative kernel estimator, is still the most important one to study, although it is inefficient in any asymptotic sense, because this is what people use to analyze data.

Another objection to these assumptions, which also needs discussion, is that even if 4.25 derivatives are present, there may not be enough information in the data to make full use of them. For example when $n = 50$, it seems unlikely one can do a good job of estimating the 4th derivative. This is an important issue from the viewpoint of applications. However careful study through simulations indicates that this effect is not as bad as one might expect. The reason is that while one can not recover the 4th derivative well, there is a "cancellation" that occurs in the sense that one only really needs "as much of the 4th derivative as can be obtained". It is planned to develop this admittedly nebulous idea in later papers.

To stay consistent with the notation f_n, the symbol h_n will be used to denote a data based bandwidth. All of the selectors in the following sections can be shown to have limiting normal distributions of the type:

$$n^{1/2}(h_n - h_{MISE}) / h_{MISE} \xrightarrow{d} N(\mu, \sigma^2),$$

for various asymptotic means and variances, which depend on the particular

selector. Explicit forms for μ and σ^2 can be found in the references. Hence this limiting distribution will not be repeated each time.

An important point is that such limiting distributions may not appear at first glance to say too much about the more important behavior of f_n when the bandwidth h_n is used. In fact this result is directly relevant, by a simple Taylor expansion argument, as discussed in Remark 3.7 of Park and Marron (1990).

Section 6 provides deeper comparison through describing experience with simulations, and also points out that the very fast asymptotic rate of $n^{-1/2}$ is mostly of theoretical interest, because in practice it usually does not "kick in" until the sample size becomes prohibitively large. A viewpoint on this is that to make effective use of the large "numbers of derivatives" assumed above, one needs a very large sample.

2. Direct Plug–In

The first method one might attempt to use in constructing an $n^{-1/2}$ bandwidth selection method is based on the asymptotic approximation

$$\text{MISE} = \text{AMISE} + o(\text{AMISE}),$$

where

$$\text{AMISE} = n^{-1}h^{-1} + h^4C_1 + C_2h^6,$$

for C_1 and C_2 appropriate constants, which depend on f through $\int(f'')^2$ and $\int(f''')^2$ respectively, and on K in a way that is not important here. Usually only the first two terms of AMISE are the only ones of importance and the third is not given because it is negligible. However in the present case, the two term approximation is insufficient, because the resulting bandwidth turns out to have only the rate of convergence $n^{-2/5}$.

Hall, Sheather, Jones and Marron (1990) pursue this approach. Note that explicit form of the minimizer of the three term AMISE is not generally available because it involves the root of a polynomial of degree 7. In that paper it is pointed out that one can either minimize numerically, or else use a simple, explicit $n^{-1/2}$ approximant of it. For practical implementation of either of these methods, a good (in fact $n^{-1/2}$) estimate of $\int(f'')^2$ is required. A sensible way to proceed with this is to use a kernel estimator of the form $\int(f_n'')^2$. Variations on this theme, and also the important question of how to choose the bandwidth for this pilot estimator are discussed in that paper. An important consideration relevant to the present discussion is that for $n^{-1/2}$ estimation here, it is required to use a higher order kernel in this pilot estimator. In fact the next highest order of 4 is

not sufficient either, with the even less appealing order 6 being required.

The two major intuitive weaknesses of this direct plug–in approach are the fact that two terms are needed in the AMISE expansion, and also the requirement of higher order kernels.

3. Smoothed Cross–Validation

Smoothed cross–validation, as introduced by Hall, Marron and Park (1990), is a method which overcomes the first objection given at the end of the last section, to the plug–in ideas. One way of motivating this idea follows from the representation

$$\text{MISE} = V(h) + B^2(h),$$

where

$$V(h) = n^{-1}\{h^{-1}\int K^2 + \int (f*K_h)^2\},$$

$$B^2(h) = \int (f*K_h - f)^2,$$

for $*$ denoting convolution. The smoothed cross–validation idea can be viewed as minimizing estimates of this which replace f by a kernel estimate f_n.

As discussed in that paper, this seems very natural because it can also be motivated in two other completely different ways. First one may think about "stabilizing" the large degree of sample variability inherent to the conventional least squares cross–validation, by thinking of "presmoothing" the pair wise differences of the observations which go into it, which is where the name smoothed cross–validation came from. Second it is in fact exactly the smoothed bootstrap estimate (and can be calculated without the usual simulation step) of the MISE. See Marron (1990) for more discussion on this connection to bootstrapping.

For $n^{-1/2}$ bandwidth selection, this exact MISE representation is much more palatable, because it does not involve the dubious two term expansion. However it also has the drawback that the pilot estimator must also use a kernel of order 6.

4. Bandwidth Factorization

A means of overcoming the objection concerning this need to use the unintuitive higher order kernels in the pilot (which is especially objectionable when one does not want to use a higher order kernel to construct f_n) has been discovered by Jones, Marron and Park (1990).

The essential idea is to use smoothed cross–validation, but to study carefully connections between the bandwidth of the pilot estimator, call this g, and the bandwidth of the main estimator, h. In particular, when g depends on h as

$$g = C\, n^p\, h^{-2}$$

for suitable constants C and p (see that paper for specifics), a type of cancellation occurs which enables the smoothed cross–validation bandwidth to yield an $n^{-1/2}$ rate of convergence, even when no higher order kernels are used in the pilot.

This bandwidth selection method overcomes both of the objections raised to the direct plug–in method, and also the higher order kernel problem of ordinary smoothed cross–validation.

5. A modification of BCV

In the pioneering paper Scott and Terrell (1987), it was pointed out that ordinary least squares cross–validation is too variable to be recommended for reliable practical usage. They proposed the idea of Biased Cross–Validation to overcome this problem. This idea can be viewed as one type of combination of the ideas in the preceding sections.

It is based on an AMISE representation as discussed in section 2, involving minimization of the criterion

$$\mathrm{BCV}(h) = n^{-1} h^{-1} \int K^2 + h^4 \int (f_n'')^2.$$

This is different from the direct plug–in because the variable h also appears in $\int (f_n'')^2$. This idea is not so far from the bandwidth factorization ideas of Section 4. The one important difference is that there g is a more complicated function of h. The original BCV choice $g = h$ turns out to be quite inefficient, giving only the slow rate of convergence $n^{-1/10}$, and this has also been born out in its poor simulation performance, see Park and Marron (1990). However by computations very similar to those in Jones, Marron and Park (1990) (reported in private correspondence by M. C. Jones), the same choice $g = C\, n^p\, h^{-2}$ gives the same type of cancellation, and again results in an $n^{-1/2}$ rate of convergence. As in the above section, this also needs only a nonnegative kernel in the pilot estimator.

Bandwidth factorization ideas do not seem to work, i.e. give $n^{-1/2}$ rate of convergence even with nonnegative pilot estimators, for the direct plug–in selectors or for the "solve the equation" type of selector (see Park and Marron (1990), because the "cancellation effect" only applies to those methods which involve minimization.

6. Simulations and Practial Considerations

The above results are all completely asymptotic in character, so it is sensible to consider such issues as: "when do the asymptotic effects describe the true

situation?" and "will there be any benefit to using an $n^{-1/2}$ selector for my set of data?" Such things can be addressed through simulations and other methods.

The first point specifically addressed here is whether or not there is much to be gained from use of the higher order final term in the AMISE expansion in Section 2. Figure 1 provides strong indication that there is probably very little to be gained in practice through the use of higher order terms in the AMISE expansion. This shows what happens when both f and K are N(0,1) and n = 1000. The picture is an overlay of the curve MISE(h), calculated exactly using the formulae in Marron and Wand (1990), compared with higher and higher order AMISE approximations. The curve corresponding to "2 terms" is that where the first two terms of the bias expansion are used, which is exactly the AMISE form given in section 2. Observe that there appears to be very little practical potential to using the second term that appears in Section 2, and indeed this is the case when much higher order terms are used as well. Since practical implementation requires the addition of noise through estimation of the coefficients, it seems highly unlikely that there will be net gain from inclusion of these highre order terms. This fact has been born out in the simulation study of Hall, Sheather, Jones and Marron (1990). Similar pictures for other sample sizes and other densities look much the same.

Figure 1

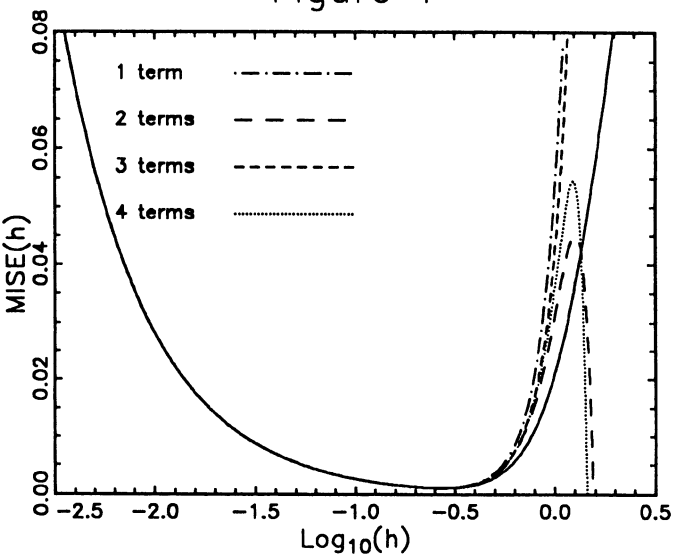


Figure 1. MISE(h) and various refined versions of AMISE(h), for f and K both N(0,1) and for n = 1000. Number of terms is for squared bias expansion.

A situation where the fast rate of convergence of the bandwidth factorized $n^{-1/2}$ method described in section 4 actually is clearly superior to other methods is shown in Figure 2. This is again for K and f both N(0,1), and n = 1000. This

time 500 pseudo data sets were simulated, and various data driven bandwidths were calculated. To understand the behavior of these selectors, kernel estimates of their distribution are shown. Because these distributions are all asymptotically normal, the normal reference distribution bandwidth given at (3.28) of Silverman (1986) was used to construct these. To give a feeling for how effective these are, the curve MISE(h) is also overlayed.

Figure 2

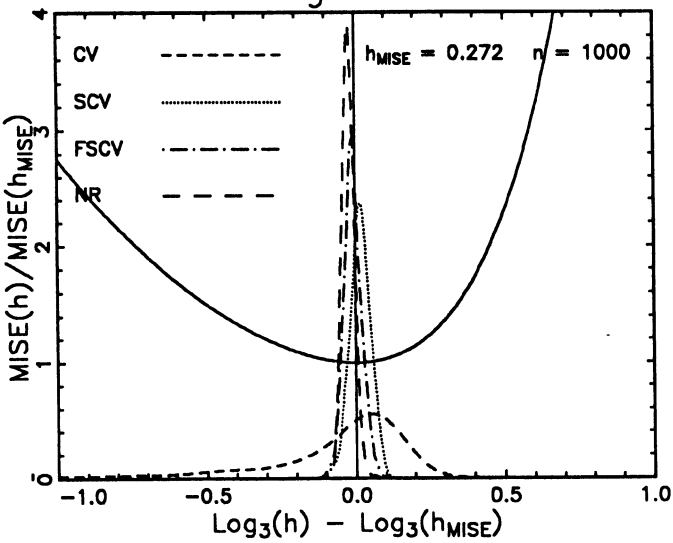


Figure 2. MISE(h), for f and K both N(0,1), sample size n = 1000, together with density estimates of the populations of 500 selected bandwidths, for Least Squares Cross–Validation, simple Smoothed Cross–Validation, bandwidth factorized Smoothed Cross–Validation and the Normal Reference Method.

Note that the population for CV is very spread relative to the others. This shows clearly why this bandwidth is not acceptable in practice: sometimes it is much too big, and sometimes much too small. The simple smoothed cross–validation bandwidth uses a pilot bandwidth g which does not depend on h. Higher order kernels have not been implemented in the pilot, because of intuitive qualms about these as expressed above, so this is not an $n^{-1/2}$ method. However note that the distribution clusters much more tightly about the optimum than for CV.

The factorized Smoothed Cross–Validation bandwidth is an actual $n^{-1/2}$ method, and as expected performs better, although not dramatically so. Note that the normal reference distribution bandwidth, defined at (3.28) of Silverman (1986), has an even tighter distribution, i.e. is less variable. However this also has a slight bias, which is larger than that for the bandwidth factorized Smoothed Cross–Validated bandwidth. For this reason the latter is superior in the sense of plugging the bandwidths into MISE(h) and averaging. This is not surprising

because when the normal reference bandwidth is analyzed by the methods of section 2, it is seen to have the slower rate $n^{-2/5}$. This is a situation which the asymptotics describe effectively.

A number of other estimation settings have been considered in the same way, and in fact the one shown in Figure 2 is the best of these in terms of effectiveness of $n^{-1/2}$ bandwidth selection. For most other settings there tends to be a bias towards oversmoothing. This bias is worse depending on the "difficulty", i.e. the number of features such as bumps, etc. in the underlying density. This bias in fact often severely limits the effectiveness of the Smoothed Cross–Validation bandwidths. However the bandwidth factorization version usually exhibits slightly less of this bias than the simple one, so there is at least some slight gain to be realized from using the $n^{-1/2}$ version. In such situations though it seems clear that the benefits of the $n^{-1/2}$ technology are in fact not strongly realized for moderate sample sizes.

7. References

Devroye, L. and Györfi, L. (1984) Nonparametric Density Estimation: the L1 View, Wiley, New York.

Eubank, R. L. (1988) Spline Smoothing and Nonparametric Regression, Dekker, New York.

Härdle, W. (1990) Applied Nonparametric Regression, Oxford University Press, Boston.

Hall, P. and Marron, J. S. (1987) 'On the amount of noise inherent in bandwidth selection for a kernel density estimator', Probability Theory and Related Fields 74, 567–581.

Hall, P. and Marron, J. S. (1989) 'Lower bounds for bandwidth selection in density estimation', tentatively accepted by Probability Theory and Related Fields.

Hall, P., Sheather, S. J., Jones, M. C. and Marron, J. S. (1990) 'On optimal data based bandwidth selection in kernel density estimation', unpublished manuscript.

Hall, P., Marron, J. S. and Park, B. U. (1990) 'Smoothed cross–validation', unpublished manuscript.

Jones, M. C. (1989) 'The roles of ISE and MISE in density estimation', unpublished manuscript.

Jones, M. C., Marron, J. S. and Park, B. U. (1990) 'A simple root n bandwidth selector', unpublished manuscript.

Mammen, E. (1988) 'A short note on optimal bandwidth selection for kernel estimators', Statistics and Probability Letters 9, 23–25.

Marron, J. S. (1990) 'Bootstrap Bandwidth Selection', Proceedings of the IMS Conference on Bootstrapping, East Lansing, Michigan.

Marron, J. S. and Wand, M. P. (1990) 'Exact mean integrated squared error calculations for density estimation', unpublished manuscript.

Müller, H.–G. (1988) Nonparametric Analysis of Longitudinal Data, Springer–Verlag, Berlin.

Park, B. U. and Marron, J. S. (1990) 'Comparison of data–driven bandwidth selectors', Journal of the American Statistical Association 85, 66–72.
Scott, D. W. and Terrell, G. R. (1987) 'Biased and unbiased cross–validation in density estimation', Journal of the American Statistical Association 82, 1131–1146.
Silverman, B. W. (1986) Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.

ESTIMATING SMOOTH DISTRIBUTION FUNCTIONS

P. SARDA
ENSA.M-INRA-Montpellier II
Department of Biometry
9, Place Pierre Viala
34060 Montpellier Cédex 1
France

ABSTRACT. The estimation of a smooth distribution function is investigated. The so-called kernel distribution estimate first introduced by Nadaraya (1964) provides a nonparametric alternative to the empirical distribution function (e.d.f.). This estimate is the same as the one obtained when locally fitting a polynomial on the e.d.f. using weighted L2-norm as a minimization criterion. Bias and variance of estimates are derived. As a by-product, it is shown that the loss in bias with respect to the e.d.f. is compensated by a gain in variance. Two different weights are compared on the basis of their bias. Finally, the problem of choosing the smoothing parameter is studied. A cross-validation procedure is introduced for which asymptotic optimality of the selected parameter is stated with respect to quadratic measures of error.

1. INTRODUCTION

Let $X, X_1, ..., X_n$ be independent real random variables with common distribution function F and density f. The most commonly used estimator for F is the empirical distribution function F_n defined by

$$F_n(x) = n^{-1}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i). \tag{1.1}$$

There are at least two advantages in connection with the empirical distribution function. First of all it is very easy to implement, requiring few computations when the sample has been sorted. Secondly, F_n has good statistical properties being for instance unbiased.

However, there has been pointed out in the literature that it could be an advantage to have a rather smooth estimator of the function F in practical situations (see Azzalini (1981) or Lejeune and Sarda (1989)).

To get a smooth nonparametric estimator for F, it is natural to take an integrand of an estimate f_n of the density function f. Nadaraya (1964) proposes to use for f_n the well-known kernel density estimator defined from a kernel K and a bandwith h_n by

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 261–270.
© 1991 *Kluwer Academic Publishers.*

$$f_n(x) = (nh_n)^{-1} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right). \tag{1.2}$$

The corresponding kernel distribution estimator $\hat{F}_n$ is given by

$$\hat{F}_n = \int_{-\infty}^{x} f_n(t)dt = n^{-1} \sum_{i=1}^{n} k\left(\frac{x-X_i}{h_n}\right), \qquad \text{where } k(t) = \int_{-\infty}^{t} K(u)du. \tag{1.3}$$

Under the assumption that f is continuous, Nadaraya (1964) proves that asymptotically $\hat{F}_n$ is unbiased and its variance is the same as F_n. Moreover he obtaines the almost sure uniform convergence of $\hat{F}_n$ to F (see also Winter (1973) and Yamato (1973)). Winter (1979) shows that the uniform rate of convergence is $(2n/\log\log n)^{-1/2}$. Reiss (1981) proves that F_n is asymptotically deficient with respect to $\hat{F}_n$ i.e., if i(n) denotes the sample size which is needed so that $F_n(x)$ has the same (or smaller) mean squared error as $\hat{F}_n(x)$, it is possible to choose h_n so that $i(n)-n \geq cn^{4/5}$. In the same paper asymptotic expressions for bias and variance are derived.

A second nonparametric alternative estimator of F can be obtained by smoothing the empirical distribution function. Such a method is introduced in Lejeune and Sarda (1989) and presented in Section 2. A distribution function estimator is obtained by fitting a polynomial to the e.d.f near the point of estimation. The minimization criterion is the weighted L2-norm. It is shown that this procedure is equivalent at the one of Nadaraya in the sense that it leads to a kernel estimator of the type $\hat{F}_n(x)$ with a certain kernel linked to the weights. In Section 3, bias and variance for estimators are derived. We discuss the problem of the choice of weights on the basis of bias minimization. In Section 4 we introduce a cross-validation procedure to choose the bandwidth h_n. It is shown that this cross-validated bandwidth is asymptotically optimal with respect to some quadratic measure of error.

2. SMOOTHING THE EMPIRICAL DISTRIBUTION FUNCTION

Let w be a nonnegative weighing function with support [-1,1]. The estimator $\hat{F}_{n,p}$ of F proposed in Lejeune and Sarda (1989) and given below, is obtained by smoothing F_n; i.e.,

$$\hat{F}_{n,p}(x) = P_{p,x}(x), \quad \text{where } P_{p,x} \text{ minimizes } \int_{-\infty}^{\infty} w\left(\frac{x-u}{h_n}\right)(F_n(u)-P_p(u))^2 du, \tag{2.1}$$

over the set of polynomials P_p with degree p, h_n being the bandwidth. Here, the role of the weights are of two orders. They induce a local or moving criterion since the values of the weights are changing with the point of estimation. On the other hand, the weights together with the bandwidth h_n and the dgree p of polynomials controls the smoothness of $\hat{F}_{n,p}$. For example, a weigt function w which is differentiable on the boundary points of

its support will give an estimator not as rough as than weights which are discontinuous. Note that the assumption of a compact support for w can be avoided in the general case. We can think, for example, of Gaussian weights. This assumption is made here since it is the most commonly used case in practice and it simplifies the writtings.

Note that contrary to F_n, $\hat{F}_{n,p}$ as $\hat{F}$ are not automatic estimators since one needs in practice to choose three quantities: the degree p of polynomials, the weight function w and the bandwidth h_n.

The procedure of estimation of F by $\hat{F}_{n,p}$ traces back to the equivalent technique in regression smoothing: Locally Weighted Regression (LWR) (Cleveland (1979) also called Moving Polynomial Regression (Lejeune (1984). In this last context the minimization criterion is replaced by weighted least squares. Asymptotic equivalence between LWR and kernel regression estimation has been proved by Lejeune (1984) (see also Müller (1987)). In the distribution function estimation context such a correspondence is not asymptotic but holds true for finite sample sizes.

Let us define $P_{p,x}$ by

$$P_{p,x}(u) = \sum_{i=0}^{p} \hat{a}_{x,i} u^i.$$

Then taking, whithout loss of generality x=0, the $\hat{a}_{x,i}$'s are solutions of the equations

$$\sum_{s=0}^{p} a_s w_{i+s} = b_i, \quad i=0,\dots,p, \tag{2.2}$$

$$\text{where } w_r = h_n^{r+1} \int_{-1}^{1} w(-t)t^r dt, \ r=0,\dots,2p \tag{2.3}$$

and

$$b_i = \int_R w(\frac{-u}{h_n}) u^i F_n(u) du = n^{-1} \sum_{j=1}^{n} \int_{X_j}^{\infty} u^i w(\frac{-u}{h_n}) du. \tag{2.4}$$

Finally, $\hat{F}_{n,p}(0)$ is the first solution $\hat{a}_{x,0}$ of the system (2.2). A simple computation (see Lejeune and Sarda (1989)) shows that $\hat{F}_{n,p}(x)$ can be written as

$$\hat{F}_{n,p}(x) = n^{-1} \sum_{i=0}^{n} k\left(\frac{x-X_i}{h_n}\right) = \hat{F}(x),$$

$$\text{where} \quad k(u) = \int_{-\infty}^{u} K(t)dt, \quad K(t) = (w(t)/D) \begin{vmatrix} 1 & v_1 & \dots & v_p \\ u & v_2 & \dots & v_{p+1} \\ \vdots & \vdots & & \vdots \\ u^p & v_{p+1} & \dots & v_{2p} \end{vmatrix}, \tag{2.5}$$

$$\text{and} \quad v_r = \int_{-1}^{1} w(-t)t^r dt,$$

and D is the determinant of a square matrix C of order p+1 with generic element $c_{j,s} = v_{j+s-2}$. It is easy to see from (2.5) that K has support [-1,1] and that

$$\int_{-1}^{1} K(u)du = 1, \qquad \int_{-1}^{1} u^r K(u)du=0, \ r=1,\dots,p.$$

Hence $\widehat{F}_{n,p}$ can be written as the integrand of a kernel density estimator whose kernel K is of order greater or equal to p+1. It can be shown (Gasser and Müller (1979) or Lejeune (1984)) that if $w = (1-u^2)I_{[-1,1]}$, the kernel K minimizes $\int K^2$ and if p=0 we obtain the Epanechnikov's kernel which has been shown to minimize the variance in density estimation for a given level of bias (Epanechnikov (1969)). Müller (1984) has shown that the kernels obtained with $w(u)=(1-u^2)^2 I_{[-1,1]}$ have the same optimality among kernels having a polynomial form and being differentiable at the boundary points of the support.

3. BIAS AND VARIANCE, CHOICE OF DEGREE AND WEIGHTS

In the following we will take a symmetric function w, since this is a natural choice in practice. This implies, by (2.3) that w_r vanishes for all odd values. This choice also leads to a symmetric kernel (see expression (2.5)). Then, the bias of $\widehat{F}_{n,p}(x)$ can be written as

$$E\widehat{F}_{n,p}(x) = \frac{1}{2}\left[F(x+h_n)+F(x-h_n)\right] + h_n\int_{-1}^{1} g(u)f(x-h_n u)du,$$

where $\qquad g(u) = k(u) - \frac{1}{2}, \ u\in [-1,1]$

Then assuming that f is (p+1)-times differentiable, and using Taylor expansions for F and f near the point x, it is easy to derive the following expression for the bias

$$\frac{h_n^{p+2}}{(p+2)!} f^{(p+1)}(x)\left[1-(p+2)\int_{-1}^{1} u^{p+1}k(u)du\right] + O(h_n^{p+4}). \tag{3.1}$$

See Reiss (1981) or Lejeune and Sarda (1989) for details. For an expression of the variance we get

$$\mathrm{Var}\widehat{F}_{n,p}(x) = n^{-1}F(x)(1-F(x)) + \frac{h_n}{n}f(x)\left[\int_{-1}^{1} k^2(u)du-1\right] + O\left(\frac{h_n^2}{n}\right). \tag{3.2}$$

From expressions (3.1) and (3.2) it follows that if a bias appears in the estimators $\widehat{F}_{n,p}$, the variance is improved when the empirical distribution function is employed. In Reiss (1981) it is shown that for some function F and a given kernel k, specific choice of the bandwidth leads to an improvment of the total mean squared error of the empirical distribution function. It is clear also that, increasing the degree p, will reduce the bias (for a function F sufficiently smooth), but it will cause a deterioration of the variance. In practice, a choice that seems to be accurate is p=2, estimating a rather smooth distribution function, since it corresponds to a minimization of the bias while the variance remains

acceptable (see Lejeune (1984) in regression estimation). In Lejeune and Sarda (1989) two different weight functions w(u) are studied. The first one is the indicator function w_1 on [-1,1] $w_1(u) = I_{[-1,1]}(u)$, and the second is Tukey's biweight w_2: $w_2(u) = (1-u^2)^2 I_{[-1,1]}(u)$.

These functions lead by (2.5) to the following "cumulative kernels" k_1 and k_2,respectively,

$$k_1(u) = \begin{cases} 0 & u > -1 \\ (4+9u-5u^3)/8 & |u| \leq 1, \\ 1 & u > 1 \end{cases}$$

and $$k_2(u) = \begin{cases} 0 & u > -1 \\ (32+105u-173u^3+147u^5-45u^7)/64 & |u| \leq 1. \\ 1 & u > 1 \end{cases}$$

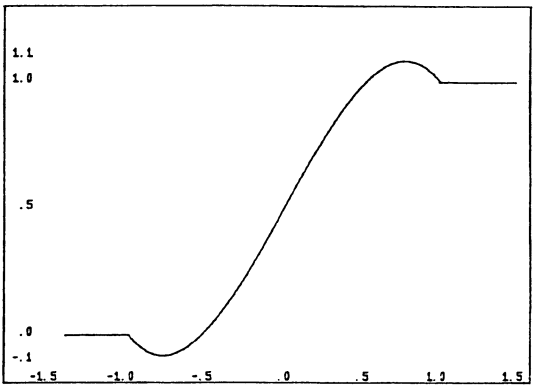


Figure 1. Cumulative kernel k_1.

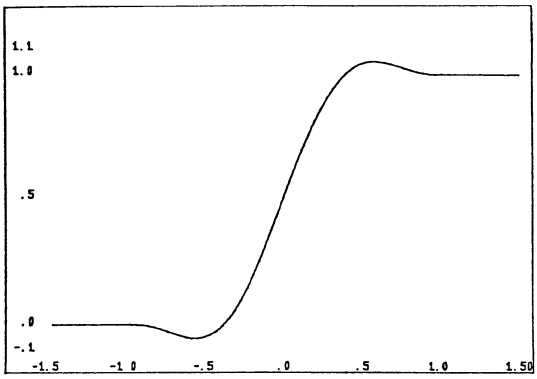


Figure 2. Cumulative kernel k_2.

The function k_2 will produce an estimator not only continuous but also differentiable. Note also that, by taking p=2, we obtain estimators $\hat{F}_{n,2}$ which are not necessarily increasing, but this will be only minor perturbation in practice. The estimator will be kept constant when such pertubations will occur.

Expressions for biases and variances of estimators can be easity derived from (3.1) and (3.2). The bias for estimators, using k_1 (respectively k_2) is

$$-(h^4/280)f^{(3)}(x) + O(h^6) \quad \text{(respectively } -(h^4/792)f^{(3)}(x) + O(h^6)\text{)},$$

and the variance is

$$n^{-1}F(x)(1-F(x)) - \frac{3\,h_n}{28\,n}f(x) + O(\frac{h_n^2}{n})$$

$$\text{(respectively } n^{-1}F(x)(1-F(x)) - \frac{35\,h_n}{429\,n}f(x) + O(\frac{h_n^2}{n})\text{)}.$$

If we now want to compare biases of estimators, this has to be done for the same amount of smoothing: i.e., for different bandwidths $h_{1,n}$ and $h_{2,n}$. One way to do this is to look at biases for given variance (or an approximation to it). For the two estimators $\hat{F}_{n,2}$, based on k_1 and k_2, the ratio of the second terms in the variances is equal to (35/429)/(3/28)=.762. So in order to have approximatively equal variances, one has to choose two bandwidths $h_{1,n}$ and $h_{2,n}$ with the same ratio. Note that to produce the same smoothing effect the bandwidth has to be greater when using weight w_2. For this choice of bandwidths, we get a ratio of biases equal to .952, which shows that the biases are almost equivalent for the two estimators. Finally, in practice we can choose the function w_1 as a weight function since the computations are simpler and the bandwidth needed to obtain a given amount of smoothing is less than that with the weights w_2. However, when differentiating $\hat{F}_{n,2}$ in order to get a kernel density estimator, these results are no longer true (see Lejeune and Sarda (1989)).

Figure 3 illustrates these results. The sample is drawn from a mixture of two normal distributions. Then the differences between the two estimators are not perceptible.

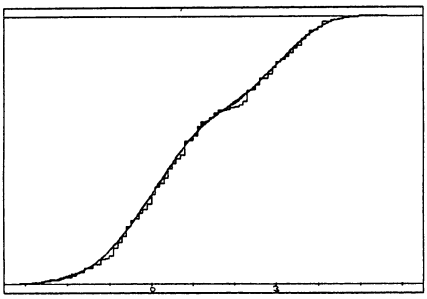


Figure 3. Empirical distribution function for a sample of size 200 from a mixture of two normal laws and two smooth estimators based on k_1 and k_2.

In figure 4, the distribution of cholesterol rates have been estimated for two groups (men and women). This example shows the usefulness of smooth representations.

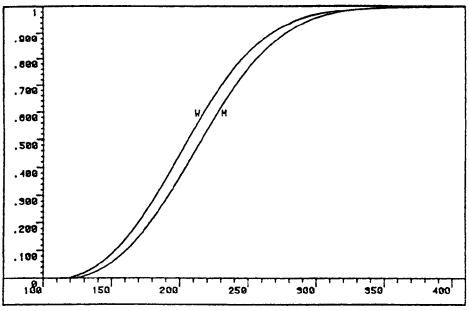


Figure 4. Estimation with k_1 and h_n=30 of the cholesterol rates in mg/100ml. 434 women and 510 men.

The results included in this section are essentially empirical. Some additional wotk is needed to control the effect of the degree p. In particular in order to choose, in practice, the degree p, some cross-validation procedure could be introduced, as done in density estimation by Hall and Marron (1987). See the following section.

4. BANDWIDTH CHOICE

In general, in nonparametric estimation it is commonly recognized that the choice of the smoothing parameter is more important that the choice of weights. Sarda (1990) introduces a cross-validation procedure which leads to a data-based bandwidth. In the following section, we assume that v is a weight function such that:

(H.1) v is bounded and supported on a compact set.

Moreover we make the following assumptions

(H.2) $h_n \in H_n = [C_1 n^{-a}, C_2 n^{-b}]$, $1/4<b\leq a<1/2$,

(H.3) k is absolutely continuous,

(H.4) $\int_{-\infty}^{\infty} xK(x)dx=0$ and $\int_{-\infty}^{\infty} x^2K(x)dx<\infty$,

(H.5) F is twice differentiable and
F and $|f'|$ are bounded from below on the support of v.

A common measure of accuracy for an estimator is the so-called Mean Integrated Squared Error, MISE(h_n), defined by

$$\mathrm{MISE}(h_n) = E\int_{-\infty}^{\infty} (\hat{F}(x)-F(x))^2 v(x)dF(x).$$

Then we have the following expression for MISE(h_n)

$$\mathrm{MISE}(h_n) = A_1 n^{-1} - A_2 h_n n^{-1} + A_3 h_n^4 + O(h_n^2/n), \tag{4.1}$$

where A_1, A_2, and A_3 are positive constants derived from (3.1) and (3.2). As is usual the case in smoothing parameter selection, it is easy to see that for h_n too small the two first terms (variance) in (4.1) will dominate the expression, while for big h_n the third term (bias) will dominate it. The balance between the two components occurs for

$$h_0 = Cn^{-1/3}.$$

Since h_0 cannot be computed in practice, one way to "estimate" it is to take an empirical approximation of MISE(h_n) and choose the bandwidth which minimizes it. Classical procedures rests on cross-validation (see Härdle and Marron (1985) for regression, Bowman (1984) and Marron (1987) for density and Sarda and Vieu (1990) for hazard estimations).

Let $\hat{F}_j$ be the leave-one-out estimator defined by

$$\hat{F}_j(x) = (n-1)^{-1}\sum_{i\neq j} k\left(\frac{x-X_i}{h_n}\right).$$

Then the so-called cross-validation criterion is given by

$$\mathrm{CV}(h_n) = n^{-1}\sum_{j=1}^{n}\left[\hat{F}_j(X_j)-F_n(X_j)\right]^2 v(X_j),$$

and the cross-validated bandwidth selected by this criterion is the bandwidth $\hat{h}_n$ minimizing CV(h_n) (see Sarda (1990) for motivation). We then have the following asymptotic optimality result. To this end let MISE'(h_n) be defined by

$$\mathrm{MISE}'(h_n) = \mathrm{MISE}(h_n) - A_1 n^{-1}.$$

THEOREM Under assumptions (H.1)-(H.5), if $\hat{h}_n$ minimizes CV(h_n) over H_n, then

$$\lim_{n\to\infty}\left[\frac{\mathrm{MISE}'(\hat{h}_n)}{\inf_{h\in H_n}\mathrm{MISE}'(\hat{h}_n)}\right]=1 \qquad \text{a.s.} \tag{4.2}$$

The proof of the theorem is in Sarda (1990). It is based on the same arguments as those usedin Härdle and Marron (1985) for regression estimation.

This theorem can be extend to dependent variables using the same techniques as in Hart and Vieu (1990). A local choice of a bandwidth is also possible (see Mielniczuk et al. (1989) in connection with density estimation).

REFERENCES

Azzalini, A. (1981) "A note on the estimation of a distribution function and quantiles by a kernel method", Biometrika 68, 1, 326-328.

Bowman, A. (1984) "An alternative method of cross-validation for the smoothing of density estimates", Biometrika 71, 352-360.

Cleveland, W.S. (1979) "Robust locally weighted regression and smoothing scatterplots", J. Amer. Statist. Assoc. 74, 829-836.

Epanechnikov, V.A. (1969) "Nonparametric estimation of a multivariate probability density", Theory Probab. Appl. 14, 153-158.

Gasser , T. and Müller, H.G. (1970) " Kernel estimation of regression functions", in T. Gasser and M. Rosenblatt (eds.), Smoothing techniques for curve estimation, Lecture Notes in Mathematics 757, Springer Verlag, 23-68.

Hall, P. and Marron, J.S. (1987) "Choice of kernel order in density estimation" Ann. Statist. 16, 161-173.

Härdle W. and Marron, J.S. (1985) "Optimal bandwidth selection in nonparametric regression function estimation", Ann. Statist. 13, 1465-1481.

Hart, J. and Vieu, P. (1990) "Data-driven bandwidth choice for density estimates based on dpenedent data", Ann. Statist. (to appear).

Lejeune, M. (1984) "Optimization in nonparametric regression", Proceedings Compstat 84, Physica Verlag, 421-426.

Lejeune, M. and Sarda, P. (1989) "Smooth estimators of distribution and density functions", Preprint.

Marron, J.S. (1987) "A comparison of cross-validation techniques in density estimation", Ann. Statist. 15, 152-162.

Mielniczuk, J., Sarda, P. and Vieu, P. (1989) "Local data-driven bandwidth choice for density estimation", J. Statist. Plan. Inference 23, 53-69.

Müller, H.G. (1987) "Weighted local regression and kernel methods for nonparametric curve fitting", J. Amer. Statist. Assoc. 82, 231-238.

Nadaraya, E.A. (1964) "Some new estimates for distribution functions", Theory Probab. Appl. 497-500.

Reiss, R.D. (1981) "Nonparametric estimation of smooth distribution functions", Scand. J. Statist. 8, 116-119.

Sarda, P. (1990) "Smoothing parameter selection for smooth distribution functions", Preprint.

Sarda, P. and Vieu, P. (1990) "Smoothing parameter selection in hazard estimation", preprint.

Winter, B.B. (1973) "Strong uniform consistency of integrals of density estimators", Canad. J. Statist. 1, 247-253.
Winter, B.B. (1979) "Convergence rate of perturbed empirical distribution functions", J. Appl.Probab. 16, 163-176.
Yamoto, H. (1973) "Uniform convergence of an estimator of distribution functions", Bull. Math. Statist. 15, 69-78.

SMOOTHING TECHNIQUES IN TIME SERIES ANALYSIS

P. VIEU
Laboratoire de Statistique et Probabilités
U.A. C.N.R.S. 745
Université P. Sabatier
118 route de Narbonne
31062 Toulouse Cedex
FRANCE

ABSTRACT. In this article, the problem of estimating the autoregressive functions of certain processes by the nonparametric kernel method is investigated. Also, consistency properties are briefly presented under mixing conditions. It is indicated how one has to choose the smoothing parameter in practical situations, and two improvements on the usual cross-validation procedure are proposed. Finally, applications to the prediction of some economic data are presented.

1. Introduction

Let $\{Z(i),\ i\in N\}$ be a real-valued process. Important tools by which we obtain information as to how future realizations of the process depend on past values are the autoregressive functions; they are defined, for positive integers k and s and for any $i\geq k$, by

$$R_{k,s}(u) = E(Z(i+s)|(Z(i),\ldots,Z(i-k+1)=u)). \tag{1.1}$$

In the following, it will be assumed that such functions exist. Note that this assumption is satisfied as long as the process is stationary.

Such a function can be seen as a particular case of a regression function of some bivariate process $\{(X(i),Y(i)),i\in N\}$, where the $X(i)$'s are real-valued random variables and the $Y(i)$'s are q-dimensional random vectors. Let

$$m(u) = E(Y(i)||X(i) = u). \tag{1.2}$$

By setting q=k and

$$Y(i)=Z(i+s),\ X(i) = (Z(i),\ldots,Z(i-k+1)), \tag{1.3}$$

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 271–283.
© 1991 *Kluwer Academic Publishers.*

it is seen that the autoregressive functions defined in (1.1) are particular cases of functions of the form (1.2). Therefore, it is natural to estimate the autoregressive functions of the process $\{Z(i), i \in N\}$ by smoothing techniques which are usually introduced in the setting of regression estimation. Most results in this setting are stated for independent pairs $(X(i), Y(i))$, and this assumption cannot be realized for the variables defined in (1.3). It is therefore necessary to study how the well-known smoothing regression techniques behave when the sample consists of dependent data.

A usual model for dependence in such problems is provided by some mixing condition. A short discussion about mixing dependence structures is presented in Section 2. Then, using existing results obtained by Collomb (1984), Section 3 presents a short survey of consistency results of the kernel autoregressive function estimates, constructed by the familiar Watson-Nadaraya method. Next, the important problem of bandwidth selection is discussed in Section 4, and several versions of the usual cross-validation rule are presented. In Section 5, the nonparametric kernel autoregressive estimate is applied to some real economic data.

2. The Mixing Conditions

It is not the purpose of this section to make an exhaustive survey of the various different mixing conditions that can be found in the literataure. The reader can find comprehensive discussions on this subject in Bradley (1985), Doukhan and Léon (1989) and in Chapter II of Györfi et al (1989). We will only consider two of these mixing conditions:

The φ-mixing condition introduced by Ibragimov (1962);
The α-mixing condition introduced by Rosenblatt (1956).

Collomb (1984) and other authors have studied nonparametric autoregressive estimates of a φ-mixing process. The φ-mixing condition is one of the more restrictive modes of mixing. Nonetheless, it was the first to be used mainly because several powerful probabilisitic results, such as a moment inequality (Billingsley, 1968, Lemma 1), and an exponential Berstein-type inequality (Collomb, 1984, Lemma 1) were available. Note that, although φ-mixing is very restrictive, compared to α- mixing, for example, it is nevertheless quite satisfactory to model Markovian processes and autoregressive processes (see Sections III.4.3 and III.4.4 in Györfi et al (1989) and the references cited therein).

However, as shown by Ibragimov and Linnick (1971), the φ-mixing condition is too restrictive to model dependence of Gaussian processes. Hence it is of interest to also study the case of α-mixing process. For this purpose, several moment inequalities have been established (see Györfi et al (1989), Section II.2.2.3). Bernstein-type inequalities have been established by Roussas and Ioannides (1987). It is to be noted that Carbon (1983) first gave

such a result but his inequality was considerably less powerful than the one proved by Roussas and Ioannides (1987).

3. Consistency of Nonparametric Autoregression Estimate

In the setting of the estimation of a regression function m, as defined by (1.2), on the basis of independent pairs (X(i),Y(i)), the Watson-Nadaraya kernel estimate has been studied by a number of authors (see, e.g., Collomb (1981, 1985a) and Härdle (1990) for surveys). One of the more interesting results can be stated as follows. By a suitable choice of h and K, and under the assumption that m is t-times continuously differentiable, the Watson-Nadaraya estimate is consistent with a rate of convergence:

$$n^{-t/(2t+q)} \text{ in the setting of } L_p \text{ norm with } 1\le p<\infty, \tag{3.1a}$$

$$(n/\log n)^{-t/(2t+q)} \text{ in the setting of } L_\infty \text{ norm.} \tag{3.1b}$$

Accoring to Stone (1982), these rates are optimal in some minimax sense (see also Hall, 1989b).

Using expressions (1.1)-(1.3) and the relevant discussion presented in connection with them, we are led to the definition of an estimate of the autoregressive function given by (1.1). To this end, fix k and s and set $R_{k,s} = R$ for simplicity.

Kernel estimates have the following form:

$$\hat{R}_h(u) = \frac{\sum_{i=k}^{n-s} Z(i+s)K(h^{-1}(Z(i),\dots,Z(i-k+1))-u)}{\sum_{i=k}^{n-s} K(h^{-1}(Z(i),\dots,Z(i-k+1))-u)}, \tag{3.2}$$

where h (h=h(n)) is the bandwidth and K is a kernel defined on $\mathbb{R}^k$. The purpose of this section is to present the main known results about consistency properties of such estimates. More extensive discussions may be found in Roussas (1990) and Györfi et al. (1989).

Collomb (1984) initiated the study of this field by showing uniform complete convergence of the estimate in (3.2) under the φ-mixing condition (see also Sarda and Vieu (1986) for similar results under weaker conditions). Then Collomb and Härdle (1986), still working under φ-mixing, improved the result by Collomb (1984) by obtaining a rate of convergence. These results have been extended to several other dependence structures (see, e.g., Peligrad (1987), Robinson (1983), Bosq and Delecroix (1985), Collomb (1985b), and Yakowitz (1987) for k-NN estimate). Finally, rates of convergence for α-mixing processes are given by Roussas (1990) and Györfi et al (1989). Consistency results, in a Markovian setting but without mixing conditions, can be found in Yakowitz (1989).

4. Automatic Bandwidth Selection

4.1 THE QUADRATIC MEASURE OF ERRORS

Surely the main task in using in practice nonparametric methods is the choice of the smoothing parameter or bandwidth, denoted by h here. One would choose an estimate from among several available estimates on the basis of a measure of error. Measures of errors which have been thoroughly studied are quadratic errors, mainly because of their mathematical simplicitly, as they compare to errors based on either the L_1 norm or the L_∞ norm.

In this paper, we wish to select a bandwidth h which minimizes the following Average Squared Error

$$ASE(h) = n^{-1}\Sigma_{i=k}^{n-s} (R-\hat{R}_h)^2 \; w(Z_{(i)},\ldots,Z(i-k+1)),$$

where R and $\hat{R}_h$ are given by (1.1) and (3.2), and w is a nonnegative function.

One may also choose other versions of quadratic errors, such as Integrated Squared Error (ISE) or Mean Integrated Squared Error (MISE), as well as measures of errors based on either the L_1 norm or the L_∞ norm. The question which arizes is as to how estimates so derived compare among themselves. According to Härdle and Vieu (1990), ASE, ISE and MISE estimates are asymptotically equivalent. See, also Marron and Härldle (1986) for a larger class of estimates but within the framework of independent identically distributed random (i.i.d.) variables (r.v.'s). The comparison of the previous estimates with those based on the L_1 or L_∞ criteria is a much harder task even for i.i.d. r.v.'s . In this latter framework, simulations carried out by Hall and Wand (1988a, 1988b) indicate that there is no much difference between the bandwidth which minimizes the ASE criterion and the bandwidth which minimizes an L_1 criterion. The first reference on bandwidth selection, in the context of L_1 norm, is Devroye (1989).

4.2 THE CROSS-VALIDATION CRITERION

In regression estimation from independent data, a popular method of choosing a bandwidth which minimizes the ASE is cross-validation. The idea of using cross-validation goes back to Collomb (1976), but the first theoretical results are those of Hall (1984) and Härdle and Marron (1985).

In the present problem of estimating the autoregressive function R, the bandwidth is selected so as to minimize the quantity

$$C(h) = \frac{1}{n}\,\Sigma\Big[Z(i+s)-\hat{R}_h^i(Z(i),\ldots,Z(i-k+1)\Big]^2 w(Z(i),\ldots,Z(i-k+1)), \quad (4.1)$$

where $\hat{R}_h^i$ is the leave-out version of the estimate $\hat{R}_h$ defined by

$$\hat{R}_h^i(u) = \frac{\sum_{j \neq i} Z(j+s)\, K(h^{-1}(Z(j),\ldots,Z(j-k+1))-u)}{\sum_{j \neq i} K(h^{-1}(Z(j),\ldots,Z(j-k+1))-u)}. \qquad (4.2)$$

Let $\hat{h}$ be the cross-validated bandwidth; i.e., $\hat{h}$ = argminC(h), and let $\bar{h}$ be the theoretically best bandwidth; i.e., $\bar{h}$ = argminASE(h).

Under suitable assumptions, which are roughly the same as those used in the i.i.d. case by Härdle and Marron (1985), the following result is shown in Härdle and Vieu (1990), under α-mixing:

$$ASE(\hat{h})/ASE(\bar{h}) \xrightarrow{a.s.} 1. \qquad (4.3)$$

This asymptotically optimal property is of great practical interest since it allows the automatic selection of an optimal bandwidth. In fact, (4.3) is established in a more general stochastic convergence mode.

4.3 A TIME-SMOOTHED VERSION OF CROSS-VALIDATION

The underlying idea in the criterion C is how well our estimate would predict Z(i+s), if we, actually, did not have the data at time i. This leads to the definition of the quantity in (4.2). Under dependence, however, it makes sense to suppress the data not only at time i, but those data as well which are close to i in terms of time. This remark motivates the introduction of the following version of cross-validation. To this end, define the quantity C as in (4.1), but employ $\hat{R}_h^i$ given below

$$\hat{R}_h^i(u) = \frac{\sum_{|j-i|>\ell(n)} Z(j+s)K(h^{-1}(Z(j),\ldots,Z(j-k+1))-u)}{\sum_{|j-i|>\ell(n)} K(h^{-1}(Z(j),\ldots,Z(j-k+1))-u)}. \qquad (4.4)$$

Roughly speaking, the significance of the parameter $\ell(n)$ above is that it indicates the distance in time beyond which two realizations of the process could be treated as if they were independent. Härdle and Vieu (1990) have established the property in (4.3) for the above defined version of cross-validation.

It is rather obvious that suppressing more than one data point, will not cause a problem for large sample size n, as is the case in the examples in Härdle and Vieu (1990), and also in Section 5 below. For small sample sizes, however, this poses a potential problem. For this reason, we propose to use a "time-smoothed" version of cross-validation, defining the criterion C as in (4.1), but with

$$\hat{R}_h^i(u) = \frac{\sum_j Z(j+s)K(h^{-1}(Z(j),\ldots,Z(j-k+1))-u)g(i-j)}{\sum_j K(h^{-1}(Z(j),\ldots,Z(j-k+1))-u)g(i-j)} . \quad (4.5)$$

The function g is required to satsify the following properties:

$$g(0)=0,\ g(t)=1 \text{ if } |t|>\ell(n), \text{ and } 0\le g(t)\le 1 \text{ for all } t. \quad (4.6)$$

This latest version of cross-validation contains as special cases the two previously presented versions. The function g plays a role similar to that of a kernel function K, in that K classifies data according to their closeness in space, whereas g classifies data according to their closeness in time.

A close examination of the proof in Härdle and Vieu (1990) reveals that property (4.3) holds here as well. That is, one has

Theorem 1. Under the same conditions as those in Härdle and Vieu (1990), the bandwidth which minimizes the criterion C, defined by (4.1), (4.5) and (4.6), is asymptotically optimal with respect to ASE in the same sense of (4.3).

In the present context, the problem is the fact that this new method is not completely automatic since one would have to choose, in practice, the function g, and what is perhaps more crucial, the parameter $\ell(n)$. The real and simulated examples in Hart and Vieu (1990), and Härdle and Vieu (1990), and also the recent paper by Chu and Marron (1990) illustrate the role of the parameter $\ell(n)$.

4.4 A LOCAL ADAPTIVE VERSION OF CROSS-VALIDATION

Utilizing the ideas described in Vieu (1991) for regression estimation from independent data, it is possible to choose a bandwidth, depending on the point at which the estimation is made. This means that, instead of looking for the minimization of some global measure of errors, such as ASE, ISE or MISE, we will search, at each S, for a bandwidth, depending on S, which minimizes

$$E_S(h) = E(R(S) - \hat{R}_h(S))^2 .$$

This can be done by choosing the bandwidth that minimizes the following version of cross-validation

$$C_S(h) = \frac{1}{n} \Sigma\left[Z(i+s)-\hat{R}_h^i(Z(i),\ldots,Z(i-k+1)\right]^2 w_{n,S}(Z(i),\ldots,Z(i-k+1)),$$

where the leave out estimate is defined as in (4.2), and where the weight functions are chosen to be concentrating around S. Examples of weight functions are given in Vieu (1990) along with practical guidelines as to how one would choose them. Furthermore, by

combining the proofs in Vieu (1990) and in Härdle and Vieu (1990), it is not difficult to establish the following result.

Theorem 2. Under the conditions of Härdle and Vieu (1990) and Vieu (1990), we have

$$\sup_S E_S(\hat{h}_S)/E_S(\bar{h}_S) \xrightarrow{a.s.} 1,$$

where $\hat{h}_S$ (respectively, $\bar{h}_S$) is the minimizer of C_S (respectively, of E_S).

4.5 OTHER BANDWIDTH SELECTION RULES

It appears that the cross-validation technique is the only one which has been studied in connection with the problem of choosing bandwidths for dependent data. In this section, we are going to discuss briefly other techniques in the context of regression estimation. Several other criteria have been proposed (see, e.g., Rice (1984)) in order to generate asymptotically optimal bandwidths. These criteria, however, are more or less equivalent to cross-validation. For details, see Härdle et al (1988).

An alternative route for bandwidth selection would be to compute directly the minimizer of ASE and to estimate it directly. If m is t-times continuously differentiable, it is well known that the theoretically optimal bandwidth has the following form

$$\bar{h} = \text{argmin}ASE(h) = Cn^{-1/(2t+k)} .$$

The constant C depends, among other things, on the function m through its partial derivatives. Therefore, estimating directly this optimal bandwidth requires an arbitrary choice of t and the estimation of the derivatives of m. This estimation creates a new smoothing parameter selection problem. However, despite these problems, in the setting of density estimation, these plug-in techniques may be competitive (see, e.g., Scott and Terel, 1987). Unfortunately, until now, this was not possible in a regression setting, because the shape of the Watson-Nadaraya estimate makes the calculus of the derivatives very difficult, due to the presence of a random quantity in the denominator. The Watson-Nadaraya approach does provide a way of estimating derivatives, with the resulting estimates enjoying good theoretical properties, as shown in Sarda and Vieu (1988). These estimates, however, are unusable in practice.

In recent work, Mack and Müller (1989a, 1989b) propose a new form of kernel estimates which avail themselves in derivative estimation. These estimates may facilitate exploration of the route proposed above.

5. An Example

In this section, we show how the kernel autoregressive estimate works on a data set drawn from economics. These data were kindly provided to me by Professor Wolfgang Härdle. The data are daily interest rates of Deutsche Mark in the Eurodollar market from March 1981 to March 1986. The sample size is 950. These data are presented in Figure 1 as a plot of Z(i+1) (vertical axis) versus Z(i) (horizontal axis). In fact, here we are estimating the autoregressive function of this time series as is defined in (1.1) with s=k=1.

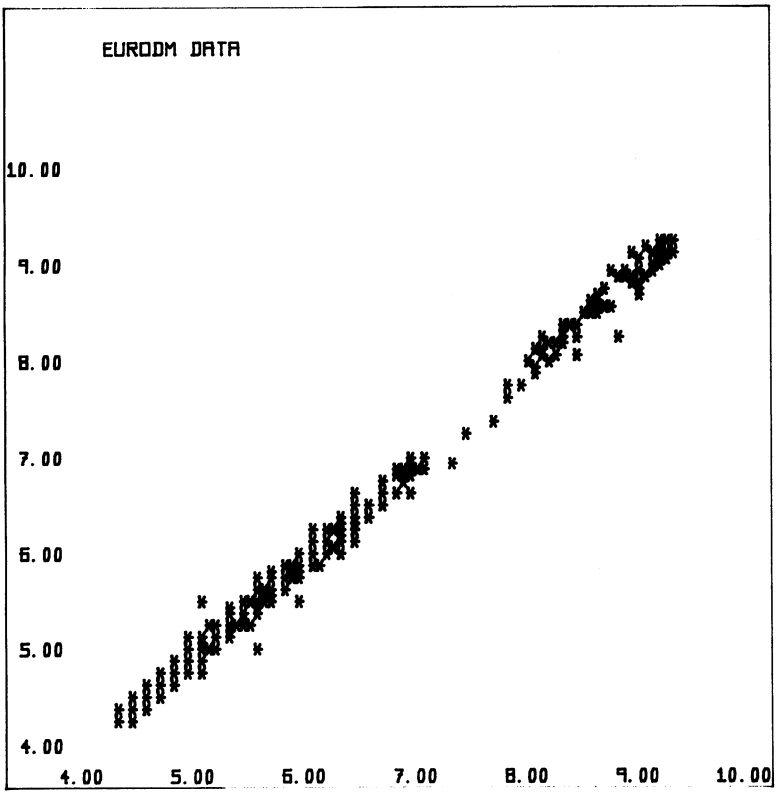


Figure 1. EURODM data: Y(i) = Z(i) = "rate today"
X(i) = Z(i-1) = "rate yesterday"

We applied a cross-validation algorithm, i.e., we minimized C defined by (4.1) and (4.4). The sample size is large enough so as not to have to introduce "time-smoothed" cross-validation, as was done in (4.5). As a weight function w, we used the indicator function of the set [a,b]. Several values for a and b, deleting from 0 percent up to 55 percent of the data, were tested. Also, several values of $\ell(n)$ were tried, from $\ell(n)=0$ up to $\ell(n)=10$. The selected bandwidth was always roughly the same. Several other examples were treated, and the same conclusion was reached. However, in some cases, the choice of a w deleting 0 percent of the data, did cause some problems akin to the well known boundary effects. The interested reader may obtain from the author plots showing the small effects of w and $\ell(n)$. In all cases, the selected bandwidth was 0.41. This choice led to the estimate presented in Figure 2.

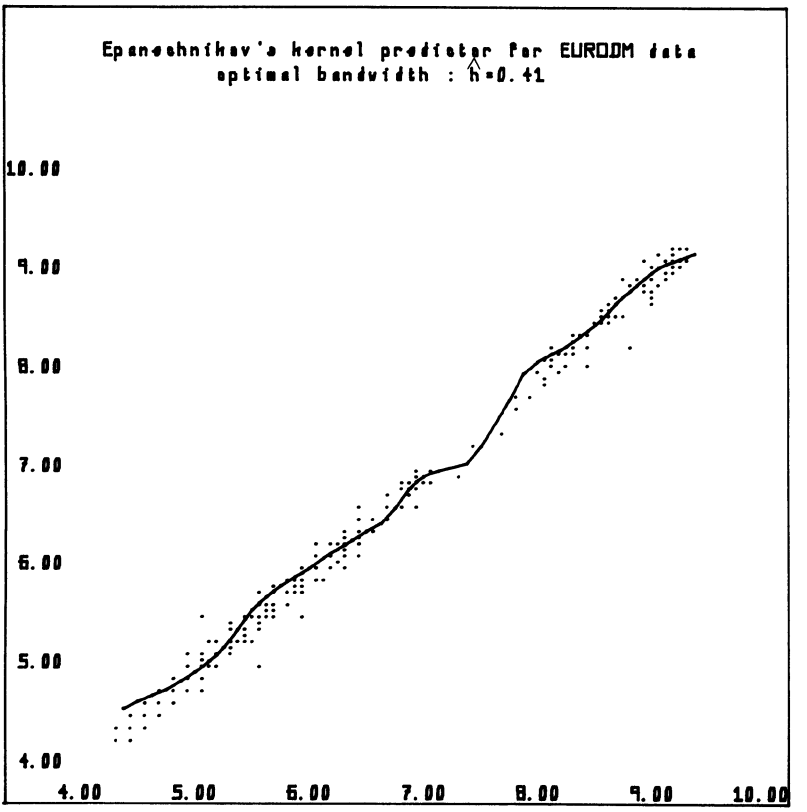


FIGURE 2. EURODM data: optimal kernel predictor, $\hat{h}=0.41$.

Even in this example (which was deliberately chosen because it looked very linear!), the nonparametric procedure is of interest. First, Figure 2 reveals two structural changes near the points 5.5 and 7. This may encourage an econometrician to chose as a parametric model for further study a blockwise linear relationship in three blocks. It may also be the case that the econometrician absolutely wishes to have a linear relation. In such a case, our nonparametric approach may be used to produce as a linear regression estimate the line which is "closest" to the above nonparametric estimate. Actually, such a method, which uses nonparametric estimates for linear regression estimation, was introduced by Cristobàl et al. (1987), and has also been discussed in Gonzalez and Vilàr (1987) in a time series analysis setting.

6. Closing Remarks

The discussion in the previous sections has provided a way of choosing the bandwidth under dependence. It may then be appropriate to investigate the somewhat less important problem of choosing the kernel. However, this problem has not been solved even under independence, although considerable work has been done towards this end (see, e.g., Marron and Nolan (1989) or Wand and Schucany (1990)).

Another much more important undertaking is that of investigating problems arizing from the multivariate function R. These problems may be attributed to difficulty in interpretation and representation, and poor rate of convergence for high values of p. Actually, this rate may be bad even in the optimal case. In the present case, the dimensionality problem may come from two sources; namely, from large values of q, when Z(i) is q-dimensional, or from large values of the integer k. It would be, of course, meaningless to propose solutions here, since these problems have not been resolved even in the context of independence.

However, two techniques seem to be promising: Generalized Additive Models (GAM) (Stone 1985 and Buja et al., 1989) and Projection Pursuit Regression (PPR) (Friedmann and Stueltzes (1981) and Huber (1985)). Few theoretical results are known concerning these models. In fact, there are only two recent papers (Stone (1986) for GAM, and Hall (1989a) for PPR), where consistency results are established. It is hoped that these techniques will be developed further in the near future and will also be applied to time series.

In closing, it should be mentioned that the theory developed in this paper applies to other functional estimation problems as well. Thus, for density estimation under dependence, see Chapters IV and V.3 in Györfi et al (1989) for consistency results, and Hart and Vieu (1990) for optimal bandwidth selection. For hazard rate estimation, again under dependence, the reader is referred to Sarda and Vieu (1989) and Roussas (1989) for consistency results, and Sarda and Vieu (1990) for optimal bandwidth selection.

References

Billingsley, P. (1968). 'Convergence of probability measures'. Wiley and Sons, New York.
Bosq, D. and Delecroix, M. (1985). 'Nonparametric prediction of Hilbert space-valued'. Stochastic Process. Appl. **19**, 271-280.
Bradley, R. (1985). 'Some remarks on strong mixing conditions'. VNU Science Press, Utrecht.
Buja, A., Hastie, T. and Tibshirani, R. (1989). 'Linear smoothers'. Ann. Statist. **17**, 453-455.
Carbon, M. (1983). 'Inégalité de Bernstein pour les processus mélangeants'. C.R.A.S., 297, I, 303-306.
Chu, C. and Marron, J. (1990). 'Comparison of two bandwidth selectors with dependent errors'. Preprint.
Collomb, G. (1976). 'Estimation nonparamétrique par la méthode du noyau'. Thèse, U.P.S. Toulouse.
Collomb, G. (1981). 'Estimation nonparamétrique de la régression'. Inter. Statist. Rev., **49**, 75-93.
Collomb, G. (1984). 'Propriétés de convergence presque complète du prédicteur à noyau. A.f.W.u.v.G. **66**, 441-460.
Collomb, G. (1985a). 'Nonparametric regression: an up to date bibliography. Statistics **2**, 309-324.
Collomb, G. (1985b). 'Nonparametric time series analysis and prediction'. Statistics **2**, 297-307.
Collomb, G. and Härdle, W. (1986). 'Strong uniform convergence rates'. Stochastic Process. Appl. **23**, 77-89.
Cristobàl, J., Faraldo, F. and Gonzalez, W. (1987). 'A class of linear regression'. Ann. Statist., **15**, 603-609.
Devroye, L. (1989). 'An L1 asymptotically optimal kernel estimate'. Preprint.
Doukhan, P. and Léon, P. (1988). Quelques notions de mélange'. Preprint. Univ. Paris Sud.
Friedman, J. and Stuetzle, W. (1981). 'Projection Pursuit Regression'. J. Amer. Statist. Assoc., **76**, 817-823.
Gonzalez, W. (1987). 'A class of nonparametrically constructed parameters'. Math. Statist. Probab. Theory B, 85-95.
Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989). 'Nonparametric curve estimation'. Lect. Notes in Stat., 60.
Hall, P. (1984). 'Asymptotic properties of integrated squared errors'. Z.f.W.u.v.G., **67**, 175-196.
Hall, P. (1989a). 'On projection pursuit regression'. Ann. Statist., **17**(2), 573-588.
Hall, P. (1989b). 'On convergence rates in nonparametric problems'. Inter. Statist. Rev., **57**, 45-58.
Hall, P. and Wand, M. (1988a) 'On the minimization of absolute distance'. Statist. Probab. Lett., **6**, 311-314.
Hall, P. and Wand, M. (1988b). 'Minimizing L1 distance in nonparametric density estimation'. Preprint.
Härdle, W. (1990). 'Applied nonparametric regression'. Book in press.

Härdle, W., Hall, P. and Marron, J.S. (1988). 'How far are optimal smoothing parameters'. J. Amer. Statist. Assoc., **83**, 86-101.
Härdle, W. and Marron, J.S. (1985). 'Optimal bandwidth selection'. Ann. Statist., **13**(4), 1465-1487.
Härdle, W. and Vieu, P. (1990). 'Asymptotically optimal data-driven bandwidths for regression'. Preprint.
Hart, J. and Vieu, P. (1990). 'Data-driven bandwidth choice for density estimates'. Ann. Statist., **18**(2), 873-890.
Huber, P. (1985). 'Projection Pursuit'. Ann. Statist., **13**, 415-525.
Ibragimov, I. (1962). 'Some limit theorems for stationary processes'. Theory Probab. Appl., **7**, 349-382.
Ibragimov, I.A. and Linnik, Y.V. (1971). 'Independent and stationary sequence'. Walters, Nordhoff, Gröningen.
Mack, Y. and Müller, H-G. (1989a). Convolution type estimators'. Statist. Probab. Lett. **7**, 229-239.
Mack, Y. and Müller, H-G. (1989b). 'Derivative estimation in non-parametric regression'. Sankhyā, A, **51**, 59-72.
Marron, J. and Nolan, D. (1989). 'Canonical kernels for density estimation'. Statist. Probab. Lett., **7**, 195-199.
Marron, J. and Härdle, W. (1986). 'Random approximation to some measures of accuracy'. J. Multivariate Anal., **20**, 91-113.
Peligrad, M. (1988). 'Properties of uniform consistency of the kernel estimator'. Preprint.
Rice, J. (1984). 'Bandwidth choice for nonparametric regression'. Ann. Statist., **33**, 1215-1230.
Robinson, P. (1983). 'Nonparametric estimators for time series'. J.T.S.A., **4**, 185-207.
Rosenblatt, M. (1956). 'A central limit theorem and strong mixing'. Proc. Nat. Acad. Sci. U.S.A., **42**, 43-47.
Roussas, G. and Ioannides, D. (1987). 'Moment inequalities for mixing sequences.' Stoch. Anal. and Appl., 5(1), 61-120.
Roussas, G. (1989). 'Hazard rate estimation under dependence.' J. Statist. Plan. Inference, **22**, 81-94.
Roussas, G. (1990). 'Nonparametric regression estimation under mixing conditions.' Stoch. Process. Appl., **36**, 107-116.
Sarda, P. and Vieu, P. (1986). 'Prévision nonparamétrique d'un processus markovien. Cahiers du CERO, **28**, 203-209.
Sarda, P. and Vieu, P. (1988). Vitesse de convergence d'estimateurs à noyau. C.R.A.S., Ser 1, **306**, 83-86.
Sarda, P. and Vieu, P. (1989). 'Empirical distribution function for mixing variables.' Statistics **20**, 559-571.
Sarda, P. and Vieu, P. (1990). 'Estimation non paramétrique de la fonction de hasard.' Cahiers du CERO, **31**, 241-256.
Scot, D. and Terrel, G. *1987). 'Biased and unbiased cross-validation.' J. Amer. Statist. Assoc., **82**, 1131-1146.
Stone, C. (1977). 'Consistent nonparametric regression.' Ann. Statist., **5**, 595-645.
Stone, C. (1982). 'Optimal global rates.' Ann. Statist., **10**, 1040-1053.

Stone, C. (1985). 'Additive regression.' Ann. Statist., **13**, 689-705.
Stone, C. (1986). 'The dimensionality reduction principle for generalized additive model.' Ann. Statist., **14**, 590-606.
Vieu, P. (1991). 'Nonparametric regression: local optimal bandwidth choice.' J.R.S.S. Ser. B, 53(1).
Wand and Schucany, (1990). 'Gaussian based kernels.' Canadian J. Statist., to appear.
Yakowitz, S. (1985). 'Nearest-neighbour methods for time series analysis.' J.T.S.A. **8**(2), 235-247.
Yakowitz, S. (1989). 'Nonparametric density and regression.' J. Multivariate Anal., **30**, 124-136.

IV. Regression Models

NONPARAMETRIC INFERENCE IN HETEROSKEDASTIC REGRESSION

M. G. AKRITAS
Pennsylvania State University
Department of Statistics
University Park, PA 16802
U.S.A.

ABSTRACT. We consider the problem of making inferences about the parameters in a heteroskedastic regression model. The model assumes a general (not necessarily symmetric) error distribution and a parametric model for the error variance. We present a survey of some recent developments in making inference for this model. The methods to be discussed use the ranks of weighted observations. In particular, weighted R-estimators and weighted aligned rank tests will be presented. It is seen that there is no loss in the asymptotic efficiency of these procedures due to estimating the unknown weights. Computational aspects of the procedures are discussed. Computational considerations suggest a new class of estimates that are termed MR-estimates.

1. INTRODUCTION

Let $(y_i, x_i, \varepsilon_i)$, $i=1,\ldots,n$, be independent identically distributed random vectors such that ε_i is independent from the p-dimensional vector x_i, ε_i and y_i are univariate, and

$$y_i = x_i'\beta + \sigma_i\varepsilon_i, \quad i=1,\ldots,n. \tag{1}$$

Here β is an unknown p-dimensional parameter, the unobservable errors ε_i have marginal distribution function F with zero mean and finite variance and $\{\sigma_i\}$ are scaling constants that express possible heteroskedasticity. It will be assumed that the scaling constants are of the form

$$\log(\sigma_i) = \theta' z_i \tag{2}$$

where θ is a rx1 vector of unknown coefficients and z_i are rx1 random

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 287–296.
© 1991 Kluwer Academic Publishers.

vectors independent of the errors ε_i. This is one of the common ways for modeling the the variance function (Carroll and Ruppert (1988)).

Generalized (weighted) least squares (LS) is probably the most common method for analyzing model(1). However the efficiency of the generalized LS can decrease dramatically away from the normal distribution and is very sensitive to outliers. The robust estimators of Carroll and Ruppert (1982) remedy these drawbacks but require a symmetric distribution. The generalized rank estimators (presented in Section 2) also remedy the drawbacks of the generalized LS estimators, but do so without imposing the assumption of a symmetric error distribution. The computation of the generalized rank estimators is briefly discussed in Section 4. Computational considerations suggest a new class of estimators for heteroscedastic regression that are termed MR-estimators.

Parameter estimation is the first step in the analysis of the linear model. In addition interest lies in testing hypotheses about subsets of the parameter vector. A class of test statistics for the homoscedastic case that maintain good efficiency away from the target distribution and are easy to compute is the class of aligned rank statistics(Puri and Sen (1971, 1985)). The theory of aligned rank tests for the present hetero-heteroskedastic model is presented in Section 3.

Another issue that has received a lot of attention is the estimation of the variance function. An excellent account of this work can be found in Carroll and Ruppert(1988). This paper will not consider estimation of the variance function; instead we will use one of the methods described in the source quoted above and will proceed with the construction of the weighted rank procedures. A main result conserning these procedures is that there is no loss in asymptotic efficiency due to estimating the unknown weights.

The symbol $\|\cdot\|$ will be used to denote both Euclidean norm and the sup -norm. It should be clear from the context which is the case. The proofs for the results in Sections 2 and 3 can be found in Akritas (1990b) and Akritas and Albers (1990), respectively. However, the assumptions under which the results are obtained are given in the Appendix for convenient reference.

2. WEIGHTED RANK ESTIMATES

In this section we present the methodology and results that lead to the construction and asymptotic properties of the weighted R-estimators.

The methodology rests on an asymptotic linearity result of the rank statistic that is based on the weighted observations (Lemma 2.3). This generalizes the corresponding result of Jureckova (1971) (stated in her Theorem 3.1) to the present heteroskedastic case. However the method for

proving Lemma 2.3 is different from that used in Jureckova's Theorem3.1. The present method uses a similar expansion of the M-statistic with a corresponding ψ- function and then relies on the theory of the empirical distribution function to show the equivalence of the two expressions. The way that the linearity result is used to establish the asymptotic normality of the weighted rank estimates also differs from the approach used in Jureckova's paper. The main idea here is borrowed from Carroll and Ruppert (1982) which uses methods from Jureckova (1977). Though the relations between weighted R-estimates and weighted M-estimates are not pursued to the extend that Jureckova (1977) does, a by-product of the present method is the symptotic equivalence of weighted R-estimates and weighted M-estimates provided the ψ-function is suitably chosen.

Consider model (1) and let $X = (x_{ij})$ denote the matrix with rows x_i'. Further let Σ denote the nxn diagonal matrix whose i-th diagonal element is σ_i. In this notation, model (1) can be written as

$$\Sigma^{-1}Y = \Sigma^{-1}X\beta + \varepsilon = \underset{\sim}{1}\mu + D\beta + \varepsilon \tag{3}$$

where $Y=(y_1,\ldots,y_n)$, $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_n)$, $D=(I-n^{-1}\underset{\sim}{1}\underset{\sim}{1}')(\Sigma^{-1}X)$ and $\mu=n^{-1}\underset{\sim}{1}'\Sigma^{-1}X\beta$. Let φ be a nondecreasing function defined on (0,1) such that

$$\int_0^1\varphi(u)du = 0 \tag{4}$$

and set $a(i)=\varphi(i/(n+1))$. Also let $D=(d_{ij})$ have rows d_i'. If the σ_i were known, rank estimates for β would be obtained by solving

$$S(\beta) = \Sigma_{i=1}^n d_i a(R((y_i-x_i'\beta)/\sigma_i)) \doteq 0, \tag{5}$$

where $R((y_i-x_i'\beta)/\sigma_i)$ is the rank of $(y_i-x_i'\beta)/\sigma_i$ among $(y_k-x_k'\beta)/\sigma_k$, k = 1, ... ,n. Our convention will be that the solution to (5) (and to all equations with the symbol $\doteq$) has the property that the norm of the left hand side evaluated at the solution is less than twice its infimum over all β. This suggests that β be estimated by solving

$$\hat{S}(\beta) = \Sigma_{i=1}^n \hat{d}_i a(\hat{R}((y_i-x_i'\beta)/\hat{\sigma}_i)) \doteq 0, \tag{6}$$

where $\hat{D} = (\hat{d}_{ij})$ is defined as the matrix D except that $\hat{\Sigma}=\mathrm{diag}(\hat{\sigma}_1,\ldots,\hat{\sigma}_n)$ is used instead of Σ, $\hat{R}((y_i-x_i'\beta)/\hat{\sigma}_i)$ is the rank of $(y_i-x_i'\beta)/\hat{\sigma}_i$ among $(y_k-x_k'\beta)/\hat{\sigma}_k$, k=1,...,n, and $\hat{\sigma}_i= \exp(\hat{\theta}'z_i)$. In what follows we will let $\hat{\beta}_{opt}$ stand for the unknown solution to (5) and $\hat{\beta}$ stand for the solution

to (6). The main result will be that $n^{1/2}(\hat{\beta}-\hat{\beta}_{opt}) \to 0$, in probability.

Before we can present the steps leading to the asymptotic equivalence of $\hat{\beta}$ and $\hat{\beta}_{opt}$, we need additional notation. Let $\Delta=(\Delta_1,\Delta_2)$, where $\Delta_1 \in \mathbb{R}^p$ and $\Delta_2 \in \mathbb{R}^r$, and define

$$F_n(x;\Delta) = n^{-1}\Sigma_{i=1}^n I((y_i-x_i'\beta-\Delta_1'x_i)\sigma_i^{-1} \le x\alpha_{1i}(\Delta)^{-1}) \tag{7}$$

where $\alpha_{1i}(\Delta)=\exp(-\Delta_2'z_i)$. Clearly, $F_n(x;0)=F_n(x)$ is the empirical distribution function of the errors ε_i. Also note that at

$$\hat{\Delta} = (\hat{\beta}-\beta,\ \hat{\theta}-\theta) = (\hat{\Delta}_1,\hat{\Delta}_2) \tag{8}$$

$F_n(x;\hat{\Delta}) = n^{-1}\Sigma_{i=1}^n I(y_i-x_i'\hat{\beta} \le x\hat{\sigma}_i)$ is the empirical distribution function of the estimated residuals $\hat{\varepsilon}_i = (y_i-x_i'\hat{\beta})/\hat{\sigma}_i$. Finally let

$$U_n(\Delta;G) = n^{-1/2}\ \Sigma_{i=1}^n\ [d_i + \alpha_{2i}(\Delta)]\psi_G(g_i(\varepsilon_i,\Delta)) \tag{9}$$

where $\alpha_{2i}(\Delta) = x_i\sigma_i^{-1}(1-\alpha_{1i}(\Delta)) - n^{-1}\Sigma_k x_k\sigma_k^{-1}(1-\alpha_{1k}(\Delta))$, $g_i(\varepsilon_i,\Delta) = (\varepsilon_i - \Delta_1'x_i\sigma_i^{-1})\alpha_{1i}(\Delta)$ and, for any distribution function G, $\psi_G(x)=\varphi(G(x))$. Note that $g_i(\varepsilon_i,\hat{\Delta})= \hat{\varepsilon}_i$, $\alpha_{2i}(\hat{\Delta})=\hat{d}_i-d_i$, $g_i(\varepsilon_i,0)=\varepsilon_i$, and $\alpha_{2i}(0)=0$. In what follows we will also use the notation

$$U_n(\Delta) = U_n(\Delta;F_n(\cdot;\Delta)). \tag{10}$$

Thus relationship (6) (with $\hat{\beta}$ instead of β) can be rewritten as

$$U_n(\hat{\Delta}) \doteq 0. \tag{11}$$

We can now outline the steps leading to the main result. The first result pertains to the M-estimator corresponding to ψ_F.

LEMMA 2.1. Under the assumptions given in the Appendix,

$$\sup\{\|U_n(\Delta;F) - U_n(0;F) + E(\psi_F')St_1\|;$$

$$\Delta' = n^{-1/2}t' = n^{-1/2}(t_1',t_2'),\ \|t\|\le M\} \to 0,$$

in probability, where S is defined in assumption A1 in the Appendix.

REMARK 2.1. The result in Lemma 2.1 is similar to the result in relation (7.16) in Carroll and Ruppert (1982). Note that in the present context ψ_F is not necessarily odd and F is not necessarily symmetric as

Carroll and Ruppert(1982) require. However $\int\psi_F dF = 0$, by (4), and this is really what is needed for their proof to go through.

LEMMA 2.2. Under the assumptions given in the Appendix,

$$\sup\{\|U_n(\Delta) - U_n(0) + E(\psi_F')St_1\|;$$

$$\Delta' = n^{-1/2}t' = n^{-1/2}(t_1', t_2'),\ \|t\|\le M\} \to 0,$$

in probability.

The result in Lemma 2.2 generalizes the asymptotic linearity result of Jureckova (1971,Theorem 3.1) to the present heteroskedastic case. The proof of Lemma2.2 makes use of the result in Lemma 2.1 and the fact that $\sup\{\|U_n(\Delta)-U_n(\Delta;F)\|;\ \Delta=n^{-1/2}t,\ \|t\|\le M\} \to 0$ (Akritas (1990b)). Lemma 2.2 and arguments given in the reference just cited show the following.

THEOREM 2.1. Under the assumptions given in the Appendix, $n^{1/2}(\hat{\beta}-\hat{\beta}_{opt}) \to 0$, in probability.

3. WEIGHTED ALIGNED RANK TESTS

Consider the model given by (1), (2) and rewrite (1) as

$$y_i = \beta' x_i + \sigma_i \varepsilon_i = \beta_1' x_{1i} + \beta_2' x_{2i} + \sigma_i \varepsilon_i \tag{12}$$

where β_1 is q-dimensional, and assume we want to test

$$H_o:\ \beta_2 = 0.$$

A class of test procedures in the homoscedastic case that maintains good efficiency properties away from the target distribution and is easy to compute is the class of aligned rank tests.Aligned rank tests were first proposed by Hodges and Lehmann (1962) in a two-way layout context. Puri and Sen (1971) derive aligned rank tests for a number of factorial designs. See also Puri and Sen (1985). However, this theory has not been extended to any design with heteroskedastic errors. The theory for aligned rank tests for regression is developed by Chiang and Puri (1984); the same problen was also considered by Adichie (1978) but he imposed restrictive assumptions on the design matrix. An altermative approach to the asymptotic theory of aligned rank tests is given in Akritas (1989a) with assumptions on the design matrix similar to that of Chiang and Puri (1984). The extension of the theory of aligned rank tests to the present heteroskedastic model, that will be presented in this section, uses the latter approach.

Using the notation of relation (3) write model (12) in matrix form as

$$\Sigma^{-1}Y = \Sigma^{-1}X_1\beta_1 + \Sigma^{-1}X_2\beta_2 + \varepsilon$$

where $X = (X_1, X_2)$ with X_1 being nxq. Note that the model for the scaled (y_i, x_i) is homoscedastic. Thus if $\hat{\beta}_1$ is a $n^{1/2}$-consistent estimate under H_o, the aligned rank test for H_o is based on the convergence

$$\tilde{W}' D_2 (D_2' D_2)^{-1} D_2' \tilde{W} \to \frac{1}{12} \chi^2_{p-q} \tag{13}$$

where $\tilde{W}$ is n^{-1} times the vector of ranks of the $(y_i - \hat{\beta}_1' x_{1i})/\sigma_i$, and $D_2 = (I-P)\Sigma^{-1}X_2$, where P is the matrix that projects on the space spanned by the columns of $D_1 = \Sigma^{-1}X_1$ and the unit vector $\underset{\sim}{1} = (1,\ldots,1)'$. However, since the σ_i are unknown (and thus $\tilde{W}$, $D = (D_1, D_2)$ are also unknown), (13) can not be used to test H_o. Akritas and Albers (1990) show that

$$W^{**\prime} \hat{D}_2 (\hat{D}_2' \hat{D}_2)^{-1} \hat{D}_2' W^{**} \to \frac{1}{12} \chi^2_{p-q}, \tag{14}$$

where W^{**} is n^{-1}times the vector of ranks of the aligned and scaled data $W_i = (y_i - \hat{\beta}_1' x_{1i})/\hat{\sigma}_i$, $\hat{\sigma}_i = \exp(\hat{\theta}' z_i)$, and $\hat{D}_2 = (I-\hat{P})\hat{\Sigma}^{-1}X_2$, $\hat{\Sigma}$ being the diagonal matrix with diagonal elements $\hat{\sigma}_i$ and $\hat{P}$ the matrix that projects on the space spanned by the columns f $\hat{D}_1 = \hat{\Sigma}^{-1}X_1$ and the unit vector $\underset{\sim}{1}$.

To outline the mehtod for proving (14) let $F_n(x;\Delta)$, $\hat{\Delta}$ be given by (7) and (8), respectively, except that now β $(\hat{\beta})$ is replaced by β_1 $(\hat{\beta}_1)$, and x_i is replaced by x_{1i} (so that Δ_1 is now q-dimensional). It follows that $F_n(\cdot;\hat{\Delta})$ is the empirical distribution function of the W_i's and thus any statistic based on the ranks of the W_i's can be expressed in terms of the $F_n(W_i;\hat{\Delta})$'s. The transformation $W_i \mapsto F_n(W_i;\hat{\Delta}) = W_i^{**}$ will be refered to as the *rank transformation,* and the "data" W_i^{**}, i=1,...,n, will be refered to as the *rank transformed data.* The asymptotic version of the rank transformation is, under H_o, $W_i \mapsto F(W_i) = W_i^*$, where F is the unknown distribution fuction of the errors ε_i. The "data" W_i^*, i=1,...,n, will be called the *transformed data.*

The present method for proving (14) first constructs a quadratic form based on the transformed data which has asymptotically, under H_o, the chi-squared distribution shown on the right hand side of (14). This is achieved by studying the model for the transformed data under H_o. The transformed data are then replaced by the rank transformed data in the quadratic form. This yields the test statistic on the left hand side of (14). By showing that the quadratic form on the left hand side of (14) is symptotically equivalent to the same quadratic form but with W^{**} replaced by W^* relation (14) is established. For details see Akritas and Albers (1990). It should be mentioned that the idea of constructing a test statistic by first studying the model for the asymptotic version of the rank transformation has been fruitfully applied in a number of contexts (see Akritas (1990), (1989a,b,c), (1988)).

4. COMPUTATIONAL ASPECTS AND A NEW CLASS OF ESTIMATORS

The weighted R- estimators of relation (6) can be computed by the same methods used for the unweighted rank estimators. The major method rests on direct minimization, using numerical search techniques, of an appropriate convex surface proposed by Jaeckel (1972). A survey of algorithms for the computation of rank estimates is provided in Aubuchon and Hettmansperger (1984). See also Osborne (1982, 1985).

However these algorithms may require an unrealistic amount of time to locate the minimum of the dispersion surface. Linearized versions of R-estimators have also been developed. See Kraft and van Eeden(1970,1972a, 1972b), and McKean and Hettmansperger (1978). However these require a starting value along with a scale estimate, and the effect of this on small sample efficiency is not well understood.

In contrast, the algorithms for obtaining M-estimates for the regression parameters have good convergence properties (Huber (1981, pp. 179-192), and references therein). In addition, M-estimators can be obtained through SAS using iteratively reweighted least squares and the nonlinear regression procedure NLIN (see also Street,Carroll and Ruppert (1988)).

It would be desirable to have a robust estimator for the parameters of the heteroskedastic regression model that combine the computational addvantages of the weighted M-estimator and general validity of the weighted R-estimator. The asymptotic equivalence of the M- and R- estimates when the M-estimator uses $\psi(x) = \varphi(F(x))$, where φ is the sccore function for the R-estimator, sugests the estimator that is obtained by solving the M-estimating equation with $\psi(x) = \varphi(\hat{F}(x))$. Here $\hat{F}$ is an estimate of the error distribution function F obtained from the residuals of a weighted LS fit. (A strictly increasing version of the e.d.f. can easily be

constructed as in Billingsley (1968, p. 104)). This could be iterated in order to reduce the effect of the initial LS fit. One iteration would probably suffice. Such an estimator can be called an MR-estimator (see Remark 4.1). Clearly, algorithms for obtaining M-estimators can be used for obtaining MR-estimators.

REMARK 4.1. Consider the following algorithm: 1) Let $\hat{F}$ be the above estimate of the error distribution function obtained from the residuals of a weighted LS fit. 2) Obtain $\hat{\beta}$, the weighted M-estimator corresponding to $\psi(x) = \varphi(\hat{F}(x))$ and update $\hat{F}$. 3) Update $\hat{\beta}$ by obtaining the weighted M-estimator corresponding to the updated $\varphi(\hat{F}(x))$. 4) Iterate until convergence. At convergence, the resulting estimator is the weighted R-estimator. The R in the name MR-estimator is due to the fact that it consists of the first two steps of the algorithm that gives the R-estimator.

With a somewhat messier notation and an extention of the arguments leading to Theorem 2.1 it can be shown that the MR-estimator is asymptotically equivalent to the corresponding M- and R-estimators, and that it does not require the assumption of a symmetric error distribution.

It should be stressed that MR-estimators are only usefull in heteroscedastic regression. Indeed, in the homoscedastic case M-estimators and their standard errors are valid for both symmetric and asymmetric error distributions (cf. Carroll and Welsh (1988)).

APPENDIX : Assumptions

The proofs in Section 2 are derived under the following assumptions.

A1. The matrix D is of full rank and $n^{-1}D'D \to S$ almost surely, where S is some positive definite matrix.

A2. $\sup\{n^{-1/2}[\|x_i\|+\|z_i\|+\|\varepsilon_i z_i\|];\ i=1,\ldots,n\} \to 0$, almost surely.

A3. $E(\|z_i\|^2) < \infty$, and $E(\|x_{ij}z_i\|^2) < \infty$ for all $j=1,\ldots,p$.

A4. F has zero mean, finite variance and a uniformly continuous density f which is bounded.

A5. φ is nondecreasing, bounded, with bounded continuous derivative.

A6. The σ_i are bounded away from zero.

The proofs in Section 3 are derived under the following assumptions. (Note that now the matrix D is different from that of Section 2).

B1. The matrix D is of full rank and (i) $n^{-1}D'D$ converges to some positive definite matrix almost surely, (ii) the largest diagonal element of $D(D'D)^{-1}D'$ goes to zero almost surely.

B2. $n^{-1}\Sigma_{i=1}^{n}[\|x_i\| + (x_{ij_1}x_{ij_2})^2] = O_p(1)$,

B3. $\sup\{n^{-1/2}\|x_i\|;\ i=1,\ldots,n\} \to 0$, almost surely.

B4. F has zero mean, finite variance, and its dencity is bounded and has a bounde derivative.

B5. The distribution of the random vector z_i has bouned support.

REMARK A.1. Relation (2) implies that assumption B5 is equivalent to the assumption that the σ_i are bounded away from zero and infinity. Thus it is only slightly stronger than assumption A6 which is common in the theory of the heteroskedastic linear model.

REFERENCES

Adichie, J. N. (1978) 'Rank tests of subhypotheses in the general linear regression. *Ann. Statist., 6, 1012-1026.*

Akritas, M. G. (1988) 'Rank transform statistics with censored data', *Statist. Probab. Lett.* Tentatively accepted.

Akritas, M. G. (1989a) 'An alternative derivation of aligned rank tests for regression', *J. Statist. Plan. Inference.* In press.

Akritas, M. G. (1989b) 'Limitations of the rank transform procedure: A study of repeated measures designs. Part I', *J. Amer. Statist. Assoc.* In press.

Akritas, M. G. (1989c) 'Aligned rank tests for regression with censored data', Submitted for publication.

Akritas, M. G. (1990a) 'The rank transform method in some two-factor designs', *J. Amer. Statist. Assoc., 85, 73-78.*

Akritas, M. G. (1990b) 'Nonparametric estimation in heteroscedastic regression', Submitted for publication.

Akritas, M. G. and Albers, W. (1990) 'Aligned rank tests for the linear model with heteroscedastic errors', manuscript in preparation.

Aubuchon, J. C. and Hettmansperger, T.P.(1984) 'On the use of rank tests and estimates in the linear model', in Krishnaiah, P.R. and Sen, P.K. (Eds), *Handbook of Statistics, 4, 259-274.*

Billingsley, P. (1968) *Convergence of Probability Measures,* Jonh Wiley & Sons, New York.

Carroll, R.J. and Ruppert, D. (1982) 'Robust estimation in heteroscedastic linear models', *Ann. Statist., 10, 429-441.*

Carroll, R. J. and Ruppert, D. (1988) *Transformation and Weighting in Regression,* Chapman and Hall, New York.

Carroll, R. J. and Welsh, A. H. (1988) 'A note on asymmetry and robustness in linear regression', *Amer. Statist., 42, 285-287.*

Chiang, C.-Y. and Puri, M. L. (1984) 'Rank procedures for testing sub-

hypotheses in linear regression. *Ann. Inst. Statist. Math., 36, Part A, 35-50.*

Hajek, J. and Sidak, Z. (1967) *Theory of Rank Tests,* Academic Press,N.Y.

Huber, P. J. (1981) *Robust Statistics,* John Wiley & Sons, New York.

Jaeckel, L. A. (1972) 'Estimation regression coeffficients by minimizing the dispersion of the residuals. *Ann. Math. Statist., 43, 1449-1458.*

Jureckova, J. (1971) 'Nonparametric estimate of regression coefficients' *Ann. Math. Statist., 42, 1328-1338.*

Jureckova, J.(1977) 'Asymptotic relations of M-estimates and R-estimates in linear regression model', *Ann. Statist., 5, 464-472.*

Kraft, C. H. and van Eeden, C.(1970) 'Efficient linearized estimates based on ranks', in M. L. Puri (Ed), *Nonparametric Techiques in Statistical Inference,* Cambridge University Press, 267-273.

Kraft, C. H. and van Eeden, C. (1972a) 'Relative efficiencies of quick and efficient methods of computing estimates from rank tests' *J. Amer Statist. Assoc., 67, 199-202.*

Kraft, C. H. and van Eeden, C. (1972b) 'Linearized rank estimates and signed-rank estimates for the general linear hypothesis', *Ann. Math. Statist., 43, 42-57.*

McKean, J. W. and Hettmansperger, T. P. (1978) 'A robust analysis of the general linear model based on one-step R-estimates', *Biometrika 65, 571-579.*

Osborne, M. R. (1982) 'A finite algorithm for the rank regression problem', in Gani, J., and Hannan, E. J. (Eds), *Essays in Statistical Science, 241-252, J. App. Probability,* special volume *19A.*

Osborne, M.R. (1985) *Finite Algorithms in Optimization and Data Analysis* J. Wiley & Sons, New York.

Puri, M. L. and Sen, P. K. (1971) *Nonparametric Methods in Multivariate Analysis,* J. Wiley & Sons, New York.

Puri, M. L. and Sen, P.K. (1985) *Nonparametric Methods in General Linear Models,* J. Wiley & Sons, New York.

Street, J.O., Carroll, R. J. and Ruppert, D. (1988) 'A note on computing Robust Regression estimates via iteratively reweighted least squares' *Amer. Statist., 42, 152-154.*

BOUNDED INFLUENCE REGRESSION IN THE PRESENCE OF HETEROSKEDASTICITY OF UNKNOWN FORM

Miguel A. DELGADO
Department of Economics
Indiana University
Bloomington, IN 47405 USA

ABSTRACT. In a regression model with conditional heteroskedasticity of unknown form, we propose a general class of M-estimators scaled by nonparametric estimates of the conditional standard deviations of the dependent variable. We give regularity conditions under which these estimators are asymptotically equivalent to M-estimators scaled by the true conditional standard deviations. The practical performance of these estimators is investigated through a Monte Carlo experiment.

1. Introduction

Cross-sectional data are typically heteroskedastic, often containing aberrant observations or gross errors. Under these circumstances, the least squares estimation method produces inefficient estimates. It is difficult to determine the source of inefficiency. As Huber (1973) pointed out:

> 'In the regression case, uncontrollable inhomogeneity of variance among the disturbances and genuinely long tailed error distributions have almost indistinguishable effects both impairing the efficiency of the estimates'.

Hence, it seems worthwhile to correct for heteroskedasticity using robust regression analysis. Carroll and Ruppert (1982) considered M-estimators scaled differently for each observation. The scale estimators were obtained under prior information on the functional form of the heteroskedasticity. Simulations reported by Carroll and Ruppert (1982) are encouraging. These estimators are expected to be nonrobust when the parameterization of the heteroskedasticity is incorrect. Asymptotically efficient estimators in the presence of unknown heteroskedasticity have been obtained by Carroll (1982), Robinson (1987) and Delgado (1989). However, these estimators have an unbounded influence function to residuals and leverage.

In this paper we establish the asymptotic properties of M-estimators with bounded influence function in the presence of heteroskedasticity of unknown nature. The conditional scale estimates are consistent nonparametric estimates of the conditional standard deviation of the dependent variable. The rest of the paper is organized as follows. In the next section we discuss a general class of optimal M-estimators which scale by the true conditional standard deviation of the dependent variable. In section 3 we present our M-estimators.

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 297–313.
© 1991 *Kluwer Academic Publishers.*

In section 4 we give conditions under which these M-estimators are adaptive. In section 5 we report the results of a Monte Carlo experiment and in section 6 we sumarize our conclusions. Proofs are confined to the appendix.

2. Optimal Bounded Regression in the Presence of Heteroskedasticity

We consider the usual regression model. We have independent observations $\{(Y_i, X_i), 1\leq i\leq n\}$ from a $\mathbb{R}\times^p\mathbb{R}$ random variable (Y, X) where,

$$Y_i = X_i'\theta^o + \varepsilon_i\ \sigma(X_i);\ i\geq 1 \tag{1.2}$$

and θ^o is a px1 vector of unknown parameters, $\sigma(.)$ is an unknown function and ε_i are, conditionally on X_i, symmetric about zero with $Var(\varepsilon_i|X_i)= 1$. The class of weighted least squares estimators is defined as,

$$\hat{\theta}_n(a)= \left\{\Sigma_i\ X_iX_i'\ a_i^{-2}\right\}^{-1} \Sigma_i\ X_iY_i\ a_i^{-2} \tag{2.2}$$

where $\{a_i, 1\leq i\leq n\}$ are suitable weights. Under regularity conditions, $n^{1/2}[\hat{\theta}_n(a)- \theta^o]$'s limiting covariance matrix has lower Gauss-Markov bound $\Phi_o= \{E[XX'\sigma^{-2}(X)]\}^{-1}$, and it is achieved by the unfeasible generalized least squares estimator $\hat{\theta}_n(\sigma)$, where $\sigma_i= \sigma(X_i)$. When the errors are conditionally normally distributed, Φ_o is the Cramer-Rao bound, but the normal model is never exactly true. In the presence of departures from the normality hypothesis, $\hat{\theta}_n(\sigma)$ may be very inefficient. On the other hand outlying X_i observations can adversely affect estimators such as (2.2). Therefore, it seems reasonable to consider estimators which bound the influence of the data.

Maronna and Yohai (1981) proved the asymptotic properties of a very general class of M-estimators of location and scale, implicitly defined as the simultaneous solution to,

$$\Sigma_i\ \phi(X_i, (Y_i- X_i'\theta)/\sigma)\ X_i = 0 \quad \text{and} \quad \Sigma_i\ \chi(X_i, (Y_i- X_i'\theta)/\sigma) = 0,$$

where $E[\phi(X, (Y- X'\theta^o)/\sigma)\ X|X]= 0$ and $E[\chi(X, (Y- X'\theta^o)/\sigma)|X]= 0$. All known proposals for $\phi(.,.)$ (see Hampel et. al 1986) may be written in the form,

$$\phi(X, r)= w(X)\ \psi(r\ v(X)) \tag{2.4}$$

for appropriate functions ψ: $\mathbb{R} \rightarrow \mathbb{R}$ and weight functions w: $\mathbb{R}^p \rightarrow \mathbb{R}^+$ and v: $\mathbb{R}^p \rightarrow \mathbb{R}^+$. The function $\psi(.)$ bounds the influence of residuals and w(.) and v(.), the influence of leverage. Relles (1968) and Huber (1973), uses w(X)=1, v(X)= 1. There are a large number of ψ-function proposals. A popular proposal is the Huber's ψ-function, i.e. $\psi(u)= u \min\{1, c/|u|\}$, where $c> 0$ is an appropriate chosen constant. In these cases, the corresponding M-estimators have bounded residual influence but the influence of leverage is unbounded. For a discussion on different choices of w(.) and v(.), see Hampel et. al (1986). Under conditional heteroskedasticity, estimators as (2.3) are not scale invariant. Carroll and Ruppert (1982) named $\check{\theta}(\sigma)$, the optimal M-estimator (OME) under heteroskedasticity, where

$$Q_n\{\check{\theta}_n(a),\ a\}= 0, \tag{2.5}$$

and $Q_n\{\theta,\ a\}= \sum_i \phi(X_i,\ (Y_i- X_i'\theta)/a_i)\ X_i/\ a_i$.

The conditions for asymptotic normality of the OME are very similar to those given in Maronna and Yohai (1981). Let us introduce the following notation, $\lambda(\Delta)= E\{X\ \sigma(X)^{-1}\ \phi(X,\ \varepsilon- \Delta'X\ \sigma(X)^{-1})\}$, $L(X)= \sup_U|\phi(X,\ U)|\ \|X\|$ and $H(X)= \sup_U|\phi'(X,\ U)|$ where $\phi'(X,\ U)= \partial\phi(X,\ U)/\partial U$. We assume,

N1.- For each X, $\phi(X,\ .)$ is odd, uniformly continuous, nondecreasing, $\phi(X,\ U)>$ 0 for $U>$ 0 and the conditional distribution of ε is symmetric about zero.

N2.- $\Pr\{\sigma(X)\leq \delta\}= 0$ for some $\delta>$ 0.

N3.- $\lambda(\Delta)$ has a nonsingular derivative at 0, $- V(\sigma)$ say, (that is $|\lambda(\Delta)- \lambda(0)+ V(\sigma)\ \Delta|= o(\|\Delta\|)$).

N4.- $E\{H(X)\ \|X\|^2\ L(X)\}< \infty$.

N5.- $E\{H(X)\ \|X\|^2\}< \infty$.

N6.- $E\|X\|< \infty$.

N7.- $E\{L(X)^2\}< \infty$.

Condition N1 implies that $E\{\phi[X_i,\ \varepsilon_i S(X_i)]|X_i\}= 0$ for any function $S(X_i)$ depending on X_i $i\geq 1$. Condition N2 is required by Robinson (1987) and Delgado (1989) for the adaptation proof of the semiparametric weighted least squares estimator (SWLSE). Alternatively, N2 may be removed by multiplying $\|X\|$ by $\sigma(X)^{-1}$ in N4-N7. We have found convenient to set out the conditions in this way in order to make comparisons with conditions needed in the next section. Conditions N3-N7 are required by Maronna and Yohai (1981). N3 and N7 guarantee that the asymptotic covariance matrix of the OME is positive definite (p.d). Note that for bounded influence estimators, as those with ϕ-functions as (2.4), L(X) is bounded. However, for classical M-estimators (i.e. $w(X)=v(X)=$ 1), N4 implies that $E\|X\|^3< \infty$ which seems a quite strong requirement. A normality proof in this case is possible, assuming $E\|X\|^2< \infty$, by using the results in Yohai and Maronna (1979). When, in (2.4), $\|X\|\ w(X)$, $\psi(.)$ and $\psi'(.)$ are bounded, N4-N7 holds if $E\|X\|^2< \infty$.

Theorem 1.- If N1-N7 hold,

$$n^{1/2}[\check{\theta}_n(\sigma)- \theta^o] \xrightarrow{d} N(0,\ \Phi_1)$$

where, $\Phi_1= V(\sigma)^{-1}\ E[\phi(X,\ \varepsilon)^2\ XX'/\sigma^2(X)]\ V(\sigma)^{-1}$.

Proof.- See Appendix.

The purpose of this paper is to obtain estimators first order asymptotically equivalent to $\check{\theta}_n(\sigma)$. We propose to estimate $\sigma(X_i)$ by nonparametric regression.

3.- Bounded Influence Function Semiparametric Estimators

When the functional form of $\sigma(X_i)$ is known and a preliminary root-n-consistent estimator of θ^o, $\tilde{\theta}_n$ say, is available, consistent estimators of

$\sigma^2(X_i)= [\sigma(X_i)]^2$ are obtained by regressing $(Y_i- X_i' \tilde{\theta}_n)^2$ against the known design of $\sigma^2(X_i)$. Under regularity conditions, the corresponding weighted least squares estimators (WLSE) of θ^o are as first order efficient as $\hat{\theta}_n(\sigma)$.

Carroll and Ruppert (1982) suggested the use the of consistent parametric weights to construct robust estimators such as those defined in (2.5) but using $w(X)=v(X)=1$ and X fixed. They proved that these estimators are, under regularity conditions, asymptotically equivalent to $\check{\theta}_n(\sigma)$. These estimators are nonrobust with respect to the assumed parameterization of $\sigma^2(X)$.

Rose (1978) proposed several nonparametric estimators of $\sigma^2(X_i)$. One of them is obtained by performing a nonparametric regression of $(Y_i- X_i' \tilde{\theta}_n)^2$ against the regressors on which $\sigma^2(X_i)$ is known to depend. Carroll (1982) and Robinson (1987) proved, under different regularity conditions, that the corresponding SWLSE asymptotically achieves the Gauss-Markov bound. Carroll (1982) used kernel regression while Robinson (1987) used k nearest neighbors (k-nn) regression.

In this paper we follow Robinson's approach. If $\sigma(X)$ is known to depend on a subset dx1 ($d\leq p$) vector $\mathring{X}_i= (\mathring{X}_{i1},\ldots,\mathring{X}_{di})'$ of nondegenerate elements of X_i, and given a positive integer $k= k(n)$, the sequence of k-nn nonparametric weights $\{\omega_{ij}(k),\ i,j=1,\ldots,n\}$ is defined by

$$\omega_{ij}(k)= 1(i\neq j)\ r_{ij}^{-1} \sum_{\tau=p_{ij}}^{p_{ij}+r_{ij}-1} c_\tau(k) \qquad 1\leq i,j\leq n \tag{3.1}$$

where for $1\leq i\leq n$, $c_i(k)> 0$, $1\leq i\leq k$; $c_i(k)=0$ $i> k$; $\sum_{i=1}^{k} c_i(k)= 1$, and

$$p_{ij}= 1+ \sum_t 1(\rho_{it}< \rho_{ij}); \qquad r_{ij}= 1+ \sum_t 1(\rho_{it}= \rho_{ij}) \quad i\neq j$$

where the sums are over $1\leq t\leq n$, $t\neq i,j$ and $1(.)$ is the indicator function.

$$\rho_{ij}= \Sigma_{m=1}^{d} (\mathring{X}_{im}- \mathring{X}_{jm})^2/s_m^2, \qquad s_m^2= (n-1)^{-1}\Sigma_i\{\mathring{X}_{im}- [n^{-1}\Sigma_j \mathring{X}_{jm}]\}^2.$$

The uniform weights $c_i(k)= k^{-1}$, $i=1,\ldots,k$, satisfy these conditions. Other weights satisfying these conditions can be found in Stone (1977). These k-nn weights do not use the own observation. This sample splitting is not required for the consistency of the k-nn weights but it is technically convenient in semiparametric estimation. This sample splitting technique has also been employed by Robinson (1987) and Delgado (1989). Given a preliminary root-n-consistent estimator of θ^o, $\tilde{\theta}_n$ say, a consistent estimate of σ_i^2 is,

$$\hat{\sigma}_i^2= \sum_j (Y_j- X_j'\tilde{\theta}_n)^2\ \omega_{ij}(k). \tag{3.2}$$

We suggest estimating θ^o by the semiparametric weighted M-estimator (SWME), $\check{\theta}_n(\hat{\sigma})$ where $\hat{\sigma}_i= [\hat{\sigma}_i^2]^{1/2}$. Unlike $\hat{\theta}_n(\sigma)$ and $\hat{\theta}_n(\hat{\sigma})$, $\check{\theta}_n(\hat{\sigma})$ has bounded influence function.

One would expect that the higher order efficiency of $\check{\theta}_n(\hat{\sigma})$ will improve by using an iterative procedure (i.e. computing new $\hat{\sigma}_i$ at each iteration). A full iterated SWLSE is obtained at once by using the following pure nonparametric estimator of $\sigma^2(X_i)$,

$$\check{\sigma}_i^2= \sum_j Y_j^2\ \omega_{ij}(k) - \{\sum_j Y_j\ \omega_{ij}(k)\}^2, \tag{3.3}$$

Delgado (1989) proved, under Robinson (1987) regularity conditions, that $n^{1/2}[\hat{\theta}_n(\check{\sigma})- \hat{\theta}_n(\sigma)]= o_p(1)$. We propose to use also $\check{\theta}_n(\check{\sigma})$ as a possible alternative to $\check{\theta}_n(\hat{\sigma})$ where $\check{\sigma}_i= [\check{\sigma}_i^2]^{1/2}$.

4.- Asymptotic Theory

Robinson (1987) noted that it is technically convenient to relate the moment conditions on (Y, X) and the rate of convergence of k. We assume that,

K1.- $\overline{\lim}\ \max_i c_i(k)\ k < \infty$ as $n \to \infty$.

K2.-$k^{-\upsilon/4}n \longrightarrow 0$ and $n^{-1}k \longrightarrow 0$ as $n \longrightarrow \infty$ for $\upsilon > 4$.

Besides N1-N3, we need stronger moment conditions in N4-N7. In particular for the υ in K2 we need,

R1.- (a) $E|H(X)\|X\|^2\ L(X)|^{\upsilon/(\upsilon-2)} < \infty$ and (b) $E|H(X)\|X\|\ L(X)|\varepsilon\ \sigma(X)||^{\upsilon/(\upsilon-2)} < \infty$.

R2.- (a) $E|H(X)\ \|X\|^2|^{\upsilon/(\upsilon-2)} < \infty$ and (b) $E|H(X)\ \|X\|\ |\varepsilon\ \sigma(X)||^{\upsilon/(\upsilon-2)} < \infty$.

R3.- $E\|X\|^{2\upsilon/(\upsilon-2)} < \infty$.

R4.- $E|L(X)|^{2\upsilon/(\upsilon-2)} < \infty$.

R5.- $E|\varepsilon\ \sigma(X)|^{\upsilon} < \infty$.

The difference between assumptions N and R are similar as in the weighted least squares estimator case. The asymptotic normality of $\hat{\theta}_n(\sigma)$ needs that $E\|X\|^2 < \infty$ and the SWLSE of Robinson (1987) needs $E\|X\|^{2\upsilon/(\upsilon-2)} < \infty$. This last condition implies R1-R4 when, in (2.4), $w(X)\ \|X\|$, $\psi(.)$ and $\psi'(.)$ are bounded. R3 and R5 are needed in Robinson (1987), Newey (1987) and Delgado (1989).

Theorem 2.- If N1-N3, K1, K2 and R1-R5 hold:

(a) $$n^{1/2}[\check{\theta}_n(\hat{\sigma})- \theta^o] \xrightarrow{d} N(0,\ \Phi_1),$$

(b) $$n^{1/2}[\check{\theta}_n(\check{\sigma})- \theta^o] \xrightarrow{d} N(0,\ \Phi_1).$$

Proof.- See Appendix.

Maronna and Yohai (1981) recommended to construct interval estimators, estimating the asymptotic variance by its natural sample analog. In our case, Φ_1 is estimated by,

$$\hat{\Phi}_1 = \hat{V}_n(\hat{\sigma})^{-1}\ n^{-1}\textstyle\sum_i X_iX_i'\ \hat{\sigma}_i^{-2}\ \phi^2\left(X_i,\ [Y_i- X_i'\check{\theta}_n(\hat{\sigma})]\ \hat{\sigma}_i^{-1}\right)\ \hat{V}_n(\hat{\sigma})^{-1} \tag{4.1}$$

where $\hat{V}_n(\hat{\sigma})= n^{-1}\sum_i X_iX_i'\ \hat{\sigma}_i^{-2}\ \phi'\left(X_i,\ [Y_i- X_i'\check{\theta}_n(\hat{\sigma})]\ \hat{\sigma}_i^{-1}\right)$.

We can substitute $\hat{\sigma}_i$ by $\check{\sigma}_i$ in (4.1).

5.- Monte Carlo

The experiments follow the model

$$Y_i= \theta_1^o + \theta_2^o\ X_i + \varepsilon_i\sigma(X_i),\ i\geq 1 \tag{5.1}$$

with $\theta_1^0 = \theta_2^0 = 1$. and X_i~iid Uniform(0, 2). The ε_i's were generated iid and independent of X_i as follows,

CONTAMINATED.- ε_i~ iid $\{.9\ N(0,\ 1) + .1\ N(0,\ 9)\}/\sqrt{1.8}$.

NORMAL.- ε_i~iid N(0, 1).

Note that Var(ε)=1 in all models. The residual conditional variances are constructed according to the models,

$$\text{Model 1.-}\ \sigma(X_i)= \exp(\gamma\ X_i)$$

$$\text{Model 2.-}\ \sigma(X_i)= |1.+ X|^{\delta}.$$

We have only used uniform weights; i.e. in (3.1) $c_i= 1/k$ for all $i\leq k$. We report two choices of k, $\hat{\theta}_n(\hat{\sigma})^1$, $\hat{\theta}_n(\check{\sigma})^1$, $\check{\theta}_n(\hat{\sigma})^1$, $\check{\theta}_n(\check{\sigma})^1$ are computed with k= $[n^{1/2}]$, while $\hat{\theta}_n(\hat{\sigma})^2$, $\hat{\theta}_n(\check{\sigma})^2$, $\check{\theta}_n(\hat{\sigma})^2$, $\check{\theta}_n(\check{\sigma})^2$ are computed with k= $[n^{2/3}]$.

We compare the unfeasible estimator $\hat{\theta}_n(\sigma)$, the OLSE $\hat{\theta}_n(1)$, the SWLSE's $\hat{\theta}_n(\hat{\sigma})$ and $\hat{\theta}_n(\check{\sigma})$, the unfeasible OME $\check{\theta}_n(\sigma)$, the M-estimator with fixed scale $\check{\theta}_n(1)$ and the SWME's $\check{\theta}_n(\hat{\sigma})$ and $\check{\theta}_n(\check{\sigma})$. In order to save space we only report results for the slope coefficient. We have only considered the classical Huber's estimator, i.e. $\psi(U)= U \min\{1,\ c/|U|\}$ in (2.4), where we choose c= 1.345. The robust estimators were computed using reweighted least squares. This c produces, under normality, $\check{\theta}_n(\sigma)$ 95% as efficient as $\hat{\theta}_n(\sigma)$.We also report results for the least absolute deviation estimator (LADE) without scaling. This estimator is used as the starting point in the reweighted procedure for the Huber's estimators and for the estimation of the residuals in order to compute $\hat{\sigma}_i$.

The tables show the bias (BIAS), variance (VAR) and the relative efficiency (EFF) of the different estimators for sample sizes of n= 30, n= 100 and n= 500 with 10,000, 5,000 and 1,000 replications respectively. The efficiency is the ratio of the mean square error (MSE) of the estimator with respect to MSE of $\hat{\theta}_n(\sigma)$. We report results for the different disturbances and the heteroskedasticity models 1 and 2 with different parameter values for δ and γ which produce different degrees of heteroskedasticity. All the programs were written in FORTRAN-77 double precision and NAG-13 routines were used to generate the variates. The programs were run on the Indiana University VAXes.

The simulations strongly support the applicability of our theorem. Through the experiments, the SWME's are always more efficient than the SWLSE's. We observe important gains in efficiency of the semiparametric estimates as the sample size increases while the EFF of the others estimators is not significantly affected. It is obviously due to the fact that the nonparametric estimates of the residual variances become more accurate as the sample size increases. This also happens when working with parametric weights. Therefore, for the smallest sample size (n= 30) and when the heteroskedasticity is mild, the SWME's are, sometimes, more inefficient than the LADE. However, when n= 100 or 500, the SWME's EFF is typically greater than one under departures from normality. Samples sizes of 30 or 100 are small in a cross-section context where samples of several thousands of observations are common. In the normal case, the SWME's EFF is quite close to their asymptotic values when n=500 and always is greater than the SWLSE's EFF. It is observed, in general, that for similar degrees of

heteroskedasticity, the semiparametric estimators based on $\hat{\sigma}_i$ behaves slightly better than those based on $\check{\sigma}_i$ when the conditional variances are small, while the latter perform better than the former when the conditional variances are larger. The choice of k does not seem to affect significantly the results. We proceed to discuss the tables in some detail.

Table 1 shows results for CONTAMINATED errors. Note that the contamination is quite moderate. Simulations using this distribution have been also reported by Huber (1973) and Carroll and Ruppert (1982). We have considered MODEL 1 and MODEL 2 with γ= 0, -1, -2 and δ= 1,3 5. So, in MODEL 1 we have small variances while in MODEL 2 the variances are larger. The severity of the heteroskedasticity is comparable in the two models when γ= -1, -2 and δ= 3, 5. When n=500 the semiparametric estimators efficiencies are fairly close to their asymptotic values. However, when n= 30, the semiparametric estimators are more inefficient than the LADE and OLSE, in some cases, especially when the heteroskedasticity is mild. When the heteroskedasticity is heavy, i.e. γ= -2 and δ= 5, the SWLSE's are more efficient than than the OLSE for n= 30, but they are more inefficient than the LADE in MODEL 1 (when γ= -2). As expected, the SWME's are always more efficient than the SWLSE's. However, the SWME's are generally more inefficient than the LADE for n=30 in MODEL 1. In MODEL 2, with n= 30, the SWME's appear to be more inefficient than the LADE only when δ= 0, 1 and k= $[n^{1/2}]$. When n=100, the SWLSE's are always more efficient than the OLSE but the SWLSE's are still more inefficient than the LADE in MODEL 1 and MODEL 2 for δ= 1,3. For n= 100, the SWME's are always more efficient than LADE, OLSE and SWLSE's. They, some times, appear to be more efficient than $\hat{\theta}_n$ (e.g. γ=0, -1, δ= 3), as the asymptotic theory predicts. This prediction is fully supported when n= 500. In this case, the SWME's EFF is always greater than 1 and the SWLSE's EFF is closer to 1, but they still behaves worse than the LADE when γ and δ are small. We have tried other distributions (e.g. Student and Laplace) with similar results.

In Table 2 we report results for NORMAL errors with MODEL 1 (γ= -2) and MODEL 2 (δ= 5). When n= 30, the LADE is more efficient than $\check{\theta}_n(\check{\sigma})$ and $\hat{\theta}_n(\check{\sigma})$ in MODEL 1. In the other cases, the SWME's are more efficient than LADE. As the sample size increases, the SWME's EFF approaches their asymptotic value (.95). Note that for n=500, the SWME's EFF are closer to their asymptotic value than the SWLSE's EFF. In fact the SWME's appear to be more efficient than the SWLSE's in all cases reported, though asymptotically it is 5% more inefficient. this fact is of practical importance. In Tables 3 and 4 we report results for LAPLACE and T_4 errors. The OME's EFF is smaller, in these cases, than in the CONTAMINATED case and, therefore, the SWME's when n= 500 are less spectacular. However, the differences in EFF between the SWME's and SWLSE's are larger than in the CONTAMINATED case. For n= 30, the SMWE's typically perform better than the LADE and always perform better than the SWLSE's.

6. Conclusions

We have seen that the introduction of robust methods in the estimation of semiparametric models is of practical relevance. The Monte Carlo reported shows that our method can be widely used, without great losses in efficiency when the data are close to normal but heteroskedastic.

APPENDIX

Proof of Theorem 1: In proving the theorem we do not need to assume that $\check{\theta}_n(\sigma)$ uniquely solves $Q_n(\theta, \sigma)= 0$ but rather that $\|Q_n(\check{\theta}_n(\sigma), \sigma)\|$ is majorized by twice the infimum of $\|Q_n(\theta, \sigma)\|$ over θ. Let define,

$$U_n(\Delta, a)= n^{-1/2} \sum_i \phi\left(X_i, (\varepsilon_i - n^{-1/2} \Delta' X_i \sigma_i^{-1})\sigma_i a_i^{-1}\right) X_i a_i^{-1}. \tag{a.1}$$

Since $Q_n(\check{\theta}_n(\sigma), \sigma)= U_n(n^{1/2}(\check{\theta}_n(\sigma)- \theta^o), \sigma)= 0$,

$$\|U_n(n^{1/2}(\check{\theta}_n(\sigma)- \theta^o), \sigma)\| \le 2 \|U_n(\Delta^*, \sigma)\|, \tag{a.2}$$

where

$$\Delta^*= - V(\sigma)^{-1} U_n(0, \sigma)= O_p(1), \tag{a.3}$$

by N3 and since by Chevyshev's inequality $U_n(0, \sigma)= O_p\{[E\|U(0, \sigma)\|^2]^{1/2}\} = O_p(1)$ by N1 and N7. Then noting that, by the Lindeberg-Levy central limit theorem (clt), $\Delta^* \xrightarrow{d} N(0, \Phi_1)$, the theorem follows from (a.2) after establishing that,

$$\|U_n(\Delta^*, \sigma)\|= o_p(1), \tag{a.4}$$

$$n^{1/2}(\check{\theta}_n(\sigma)- \theta^o)= O_p(1). \tag{a.5}$$

We conclude (a.4) from (a.3) and

$$\sup_{\|\Delta\|\le M} \|U_n(\Delta, \sigma) - U_n(0, \sigma) + V(\sigma) \Delta\| = o_p(1). \tag{a.6}$$

(a.6) follows from N3 and,

$$\sup_{\|\Delta\|\le M} \|U_n(\Delta, \sigma) - U_n(0, \sigma) - E\{U_n(\Delta, \sigma)\}\| = o_p(1). \tag{a.7}$$

In order to prove (a.7), take for convenience M=1. For $0< \delta< 1$ define,

$$S_n(\Delta, \sigma)= \sup_{\|\Delta^1- \Delta\|\le \delta} \|U_n(\Delta^1, \sigma) - U_n(\Delta, \sigma)\|. \tag{a.8}$$

Since for each fixed δ we can cover the ball of radius 1 in $\mathbb{R}^p$ by a finite number of balls of radius δ, we conclude (a.7) from,

$$U_n(\Delta, \sigma) - U_n(0, \sigma) - E\left\{U_n(\Delta, \sigma) - U_n(0, \sigma)\right\}= o_p(1), \tag{a.9}$$

for each fixed Δ, and for all $0< \delta< 1$, all n and all $\|\Delta\|\le 1$,

$$E\left\{S_n(\Delta, \sigma)\right\}\le \delta\, O(1), \tag{a.10}$$

$$S_n(\Delta, \sigma)- E\left\{S_n(\Delta, \sigma)\right\}= o_p(1), \tag{a.11}$$

(a.9) follows from,

$$\mathrm{Var}\left\{U_n(\Delta,\ \sigma)-\ U_n(0,\ \sigma)\right\}= n^{-1}\ \sum_i\ \mathrm{Var}\left\|X_i\sigma_i^{-1}\ \phi\left(X_i,\ (\varepsilon_i-\ n^{-1/2}\ \Delta' X_i\ \sigma_i^{-1})\right)\right\|$$

$$\leq n^{-1}\ \sum_i\ E\left\|X_i\sigma_i^{-1}\ \phi\left(X_i,\ (\varepsilon_i-\ n^{-1/2}\ \Delta' X_i\ \sigma_i^{-1})\right)\right\|^2$$

$$\leq n^{-1/2}\ E\left\{\sigma(X)^{-2}\ \|X\|^2\ H(X)\ L(X)\right\}= O(n^{-1/2}),$$

applying a mean value theorem (mvt) argument, N2 and N4. (a.10) follows from

$$E\left\{S_n(\Delta,\ \sigma)\right\}\leq\ \delta\ E\left\{\sigma(X)^{-2}\ \|X\|^2\ H(X)\right\}\ \leq\ \delta\ O(1),$$

by N2 and N5. (a.11) is proved using similar arguments as in (a.9), in particular,

$$\mathrm{Var}\left\{S_n(\Delta,\ \sigma)\right\}\leq\ E\left\{S_n^2(\Delta,\ \sigma)\right\}\leq\ \delta\ n^{-1/2}\ E\left\{\sigma(X)^{-2}\ \|X\|^2\ H(X)\ L(X)\right\}= O(n^{-1/2}).$$

(a.4) holds if for each $\eta>\ 0$, $\tau>\ 0$ and M, there exists a M_1 satisfying,

$$\Pr\left\{\inf_{\|\Delta\|\leq\ M_1}\ \|U_n(\Delta,\ \sigma)\|\ >\ \eta\right\}\ >\ 1-\ \tau, \qquad \text{(a.12)}$$

and (a.12) follows from(a.6) using same arguments as Jurečkovà's (1977) proof of her Lemma 5.2.

Proof of Theorem 2 (a): Let introduce the following notation,

$$\tilde{\sigma}_i^2= \sum_j(Y_j-\ X_j'\tilde{\theta}_n)^2\omega_{ij}\ \text{ and }\ \bar{\sigma}_i^2=\sum_j\ \sigma_j^2\ \omega_{ij}$$

where $\omega_{ij}= \omega_{ij}(k)$. We need the following Lemmas, proved in Robinson (1987),

Lemma 1.- Let f(.) be a Borel function such that $E|f(x)|^p< \infty$,

$$E\{\sum_i|f(X_i)-\ f(X_1)|^p\omega_{1i}\}= o(1).$$

Lemma 2.- $\Pr\{\min_i\ \sigma_i\leq\ \delta\}= 0$ all n and some $\delta>\ 0$.

Lemma 3.- $\Pr\{\bar{\sigma}_1\leq\ \delta\}= 0$ all n and some $\delta>\ 0$.

Lemma 4.- $\Pr\{\min_i\ \bar{\sigma}_i\leq\ \delta\}= 0$ all n and some $\delta>\ 0$.

Lemma 5.- $\{\min_i\ \tilde{\sigma}_i^2\}^{-1}= O_p(1)$.

Lemma 6.- $\{\min_i\ \hat{\sigma}_i^2\}^{-1}= O_p(1)$.

Lemma 7.- $\sum_i|\hat{\sigma}_i^2-\ \tilde{\sigma}_i^2|^2= O_p(k^{-1})$.

Lemma 8.- $E\{|\tilde{\sigma}_1^2-\ \bar{\sigma}_1^2|^{\upsilon/2}\}= O(k^{-\upsilon/4})$.

Lemma 9.- $\max_i|\hat{\sigma}_i^2-\ \tilde{\sigma}_i^2|= O_p(k^{-1/2})$.

Lemma 10.- $\max_i |\tilde{\sigma}_i^2 - \bar{\sigma}_i^2| = O_p(n\, k^{-\upsilon/4})$.

We follow the same strategy of proof as in Theorem 1. We first prove that,

$$\sup_{\|\Delta\| \le M} \|U_n(\Delta, \hat{\sigma}) - U_n(0, \hat{\sigma}) + V(\sigma)\, \Delta\| = o_p(1). \tag{b.1}$$

We conclude (b.1) from (a.6) and,

$$\sup_{\|\Delta\| \le M} \|U_n(\Delta, \hat{\sigma}) - U_n(0, \hat{\sigma}) - \{U_n(\Delta, \sigma) - U_n(0, \sigma)\}\| = o_p(1). \tag{b.2}$$

(b.2) follows from,

$$\sup_{\|\Delta\| \le M} \|n^{-1/2} \textstyle\sum_i X_i\, (\hat{\sigma}_i^{-1} - \tilde{\sigma}_i^{-1})\, R_i(\Delta, \hat{\sigma})\| = o_p(1), \tag{b.3}$$

$$\sup_{\|\Delta\| \le M} \|n^{-1/2} \textstyle\sum_i X_i\, (\tilde{\sigma}_i^{-1} - \bar{\sigma}_i^{-1})\, R_i(\Delta, \hat{\sigma})\| = o_p(1), \tag{b.4}$$

$$\sup_{\|\Delta\| \le M} \|n^{-1/2} \textstyle\sum_i X_i\, (\bar{\sigma}_i^{-1} - \sigma_i^{-1})\, R_i(\Delta, \hat{\sigma})\| = o_p(1), \tag{b.5}$$

$$\sup_{\|\Delta\| \le M} \|n^{-1/2} \textstyle\sum_i X_i\, \sigma_i^{-1}\, [R_i(\Delta, \hat{\sigma}) - R_i(\Delta, \tilde{\sigma})]\| = o_p(1), \tag{b.6}$$

$$\sup_{\|\Delta\| \le M} \|n^{-1/2} \textstyle\sum_i X_i\, \sigma_i^{-1}\, [R_i(\Delta, \tilde{\sigma}) - R_i(\Delta, \bar{\sigma})]\| = o_p(1), \tag{b.7}$$

$$\sup_{\|\Delta\| \le M} \|n^{-1/2} \textstyle\sum_i X_i\, \sigma_i^{-1}\, [R_i(\Delta, \sigma) - R_i(\Delta, \sigma)]\| = o_p(1), \tag{b.6}$$

where $R_i(\Delta, a) = \phi\left(X_i, (\varepsilon_i - n^{-1/2}\, \Delta' X_i\, \sigma_i^{-1})\sigma_i a_i^{-1}\right) - \phi\left(X_i, \varepsilon_i \sigma_i a_i^{-1}\right)$.
Using a mvt argument, the left hand side of (b.3) is bounded by,

$$K\, n^{-1} \textstyle\sum_i \|X_i\|^2\, H(X_i)\, \hat{\sigma}_i^{-1}\, |\hat{\sigma}_i^{-1} - \tilde{\sigma}_i^{-1}| \le$$

$$K \max_i |\hat{\sigma}_i^2 - \tilde{\sigma}_i^2| \left\{\min_i \hat{\sigma}_i^2\, \min_i \tilde{\sigma}_i\, (\min_i \hat{\sigma}_i + \min_i \tilde{\sigma}_i)\right\}^{-1} n^{-1} \textstyle\sum_i \|X_i\|^2\, H(X_i) = O_p(k^{-1/2})$$

by Lemmas 5, 6 and 9 and R1 (where, henceforth, K is a generic constant). The left hand side of (b.4) is bounded by,

$$K\, n^{-1} \textstyle\sum_i \|X_i\|^2\, H(X_i)\, \hat{\sigma}_i^{-1}\, |\tilde{\sigma}_i^{-1} - \bar{\sigma}_i^{-1}| \le$$

$$K \max_i |\tilde{\sigma}_i^2 - \bar{\sigma}_i^2| \left\{\min_i \hat{\sigma}_i\, \min_i \tilde{\sigma}_i\, \min_i \bar{\sigma}_i\, (\min_i \tilde{\sigma}_i + \min_i \bar{\sigma}_i)\right\}^{-1} n^{-1} \textstyle\sum_i \|X_i\|^2\, H(X_i)$$

$$= O_p(n\, k^{-\upsilon/4}),$$

by Lemmas 4, 5, 6 and 10 and R1. The left hand side of (b.5) is bounded by,

$$\left\{\min_i \hat{\sigma}_i\, \min_i \sigma_i\, \min_i \bar{\sigma}_i\, (\min_i \sigma_i + \min_i \bar{\sigma}_i)\right\}^{-1} K\, n^{-1} \textstyle\sum_i \|X_i\|^2\, H(X_i)\, |\bar{\sigma}_i^2 - \sigma_i^2| = o_p(1),$$

by Lemmas 2,3 and 6 and Markov's inequality, noting that, by Holder's inequality,

$$E\left\{\|X_1\|^2\, H(X_1)\,|\bar{\sigma}_1^{-2} - \sigma_1^2|\right\} \le \left\{E|\,\|X_1\|^2\, H(X_1)\,|^{\upsilon/(\upsilon-2)}\right\}^{(\upsilon-2)/\upsilon} \left\{E|\bar{\sigma}_1^{-2} - \sigma_1^2|^{\upsilon/2}\right\}^{2/\upsilon} = o(1)$$

by Lemma 1 and R1.(b.6), (b.7) and (b.8) follow by using same arguments as in (b.3), (b.4) and (b.5), after bounding the former expressions using a mvt argument. Next we prove that,

$$\sup_{\|\Delta\| \le M} \|U_n(\Delta, \hat{\sigma}) - U_n(0, \sigma) + V(\sigma)\,\Delta\| = o_p(1). \tag{b.9}$$

We conclude (b.9) from (b.1) and,

$$\|U_n(0, \hat{\sigma}) - U_n(0, \sigma)\| = o_p(1). \tag{b.10}$$

(b.10) follows from,

$$\|n^{-1/2} \textstyle\sum_i X_i\, (\bar{\sigma}_i^{-1} - \sigma_i^{-1})\, \phi(X_i, \varepsilon_i)\| = o_p(1), \tag{b.11}$$

$$\|n^{-1/2} \textstyle\sum_i X_i\, \bar{\sigma}_i^{-1}\, \{\phi(X_i, \varepsilon_i\sigma_i\bar{\sigma}_i^{-1}) - \phi(X_i, \varepsilon_i)\}\| = o_p(1), \tag{b.12}$$

$$\|n^{-1/2} \textstyle\sum_i X_i\, (\hat{\sigma}_i^{-1} - \tilde{\sigma}_i^{-1})\, \phi(X_i, \varepsilon_i\sigma_i\bar{\sigma}_i^{-1})\| = o_p(1), \tag{b.13}$$

$$\|n^{-1/2} \textstyle\sum_i X_i\, (\tilde{\sigma}_i^{-1} - \bar{\sigma}_i^{-1})\, \phi(X_i, \varepsilon_i\sigma_i\bar{\sigma}_i^{-1})\| = o_p(1), \tag{b.14}$$

$$\|n^{-1/2} \textstyle\sum_i X_i\, \hat{\sigma}_i^{-1}\, \{\phi(X_i, \varepsilon_i\sigma_i\hat{\sigma}_i^{-1}) - \phi(X_i, \varepsilon_i\sigma_i\bar{\sigma}_i^{-1})\}\| = o_p(1). \tag{b.15}$$

Using the fact that $\phi(X,.)$ is odd and ε is conditionally symmetric about zero, (b.11) follows from Chevyshev's inequality since,

$$E\|n^{-1/2} \textstyle\sum_i X_i\, (\bar{\sigma}_i^{-1} - \sigma_i^{-1})\, \phi(X_i, \varepsilon_i)\|^2 = E\{\|X_1\|^2\, |\bar{\sigma}_1^{-1} - \sigma_1^{-1}|^2\, \phi(X_1, \varepsilon_1)^2\|\}$$
$$\le K\, E\{L(X_1)^2\, |\bar{\sigma}_1^{-2} - \sigma_1^2|\},$$

by N2, Lemma 3 and 4; and by Holder's inequality,

$$E\{L(X_1)^2\, |\bar{\sigma}_1^{-2} - \sigma_1^2|\} \le \{E|L(X_1)|^{2\upsilon/(\upsilon-2)}\}^{(\upsilon-2)/\upsilon}\, E\{|\bar{\sigma}_1^{-2} - \sigma_1^2|^{\upsilon/2}\}^{2/\upsilon} = o_p(1),$$

by Lemma 1 and R4.We conclude (b.12) from Chevyshev's inequality, since, by triangle inequality, Lemma 3 and 4 and Holder's inequality,

$$E\|n^{-1/2} \textstyle\sum_i X_i\, \bar{\sigma}_i^{-1}\, \{\phi(X_i, \varepsilon_i\sigma_i\bar{\sigma}_i^{-1}) - \phi(X_i, \varepsilon_i)\}\|^2 \le$$

$$E|\,\|X_1\|^2\, \bar{\sigma}_1^{-2}\, \{\phi(X_1, \varepsilon_1\sigma_1\bar{\sigma}_1^{-1}) - \phi(X_1, \varepsilon_1)\}^2| =$$

$$E|\,\|X_1\|^2\, \bar{\sigma}_1^{-2}(\bar{\sigma}_1^{-1} - \sigma_1^{-1})\, \varepsilon_1\sigma_1\{\phi(X_1, \varepsilon_1\sigma_1\bar{\sigma}_1^{-1}) - \phi(X_1, \varepsilon_1)\} \times$$
$$\int_0^1 \phi(X_1, \varepsilon_1\sigma_1(\tau\, \bar{\sigma}_1^{-1} + (1-\tau)\, \sigma_1^{-1}))\, d\tau \le$$

$$E|\,\|X_1\|\, |\varepsilon_1\sigma_1|\, |\bar{\sigma}_1 - \sigma_1|\, H(X_1)\, L(X_1)| \le$$

$$\left\{E|\,\|X_1\|\, |\varepsilon_1\sigma_1|\, H(X_1)\, L(X_1)|^{\upsilon/(\upsilon-2)}\right\}^{(\upsilon-2)/\upsilon} \left\{E|\bar{\sigma}_1^{-2} - \sigma_1^2|^{\upsilon/4}\right\}^{2/\upsilon} = o(1)$$

by Lemma 1. Now note that the left hand side of (b.13) is bounded by,

$$\left\{\min_i\hat{\sigma}_i \ \min_i\tilde{\sigma}_i \ (\min_i\tilde{\sigma}_i + \min_i\hat{\sigma}_i)\right\}^{-1} \left\{\sum_i |\hat{\sigma}_i^2 - \tilde{\sigma}_i^2|^2\right\}^{1/2}\left\{n^{-1}\sum_i L(X_i)^2\right\}^{1/2} = o_p(1)$$

by lemmas 5, 6 and 7. The left hand side of (b.14) is bounded by,

$$\{\min_i \tilde{\sigma}_i \ \min_i \bar{\sigma}_i\}^{-1} \ n^{-1/2} \sum_i L(X_i) \ |\tilde{\sigma}_i - \bar{\sigma}_i| = o_p(1),$$

by Lemmas 4 and 5 and Chevyshev's inequality, since,

$$E|n^{-1/2} \sum_i L(X_i) \ |\tilde{\sigma}_i - \bar{\sigma}_i||^2 \le C_1 + C_2$$

where, by Hölder's inequality,

$$C_1 = E\{L(X_1)^2 \ |\tilde{\sigma}_1 - \bar{\sigma}_1|^2\} \le \left\{E|L(X_1)|^{2(\upsilon-2)/\upsilon}\right\}^{\upsilon/(\upsilon-2)}\left\{E|\tilde{\sigma}_1^2 - \bar{\sigma}_1^2|^{\upsilon/2}\right\}^{2/\upsilon} = O(k^{-1/2})$$

and by Cauchy's and Hölder's inequalities,

$$C_2 = E\left\{ n^{-1}\sum_{i\neq j}\sum L(X_i) \ |\tilde{\sigma}_i - \bar{\sigma}_i| \ L(X_j) \ |\tilde{\sigma}_j - \bar{\sigma}_j|\right\} \le E\left[n^{-1} \sum_{i\neq j}\sum L(X_i)^2 \ |\tilde{\sigma}_j - \bar{\sigma}_j|^2\right]$$

$$\le \left\{E|L(X_1)|^{2\upsilon/(\upsilon-2)}\right\}^{(\upsilon-2)/\upsilon} \left\{n \ E|\tilde{\sigma}_1^2 - \bar{\sigma}_1^2|^{\upsilon/2}\right\}^{2/\upsilon} = O\left[(n \ k^{-\upsilon/4})^{2/\upsilon}\right]$$

by Lemma 8 and R4. Now note that,

$$\|n^{-1/2}\sum_i \phi(X_i, \ \varepsilon_i\sigma_i\hat{\sigma}_i^{-1}) - \phi(X_i, \ \varepsilon_i\sigma_i\bar{\sigma}_i^{-1})\| =$$

$$\|n^{-1/2}\sum_i \varepsilon_i\sigma_i \ (\hat{\sigma}_i^{-1} - \bar{\sigma}_i^{-1}) \int_0^1 \phi'(X_i, \ \varepsilon_i\sigma_i[\hat{\sigma}_i^{-1}\tau + \bar{\sigma}_i^{-1}(1-\tau)]) \ d\tau\|$$

$$\le K \ n^{-1/2}\sum_i\|X_i\| \ H(X_i) \ |\varepsilon_i\sigma_i| \ |\hat{\sigma}_i^{-1} - \tilde{\sigma}_i^{-1}| + K \ n^{-1/2}\sum_i\|X_i\| \ H(X_i) \ |\varepsilon_i\sigma_i| \ |\hat{\sigma}_i^{-1} - \tilde{\sigma}_i^{-1}|$$

$$= o_p(1),$$

using arguments in (b.13) and (b.14). Then by (b.9), $U(\Delta^*, \hat{\sigma}) = o_p(1)$ and

$$0 = \|U_n[n^{1/2}(\check{\theta}_n(\hat{\sigma}) - \theta^o), \ \hat{\sigma}]\| \le 2 \ \|U_n(\Delta^*, \ \sigma)\| = o_p(1), \tag{b.18}$$

and given (b.9), $n^{1/2}(\check{\theta}_n(\hat{\sigma}) - \theta^o) = O_p(1)$, since for each $\eta > 0$, $\tau > 0$ and M, there exists a M_1 satisfying,

$$\Pr\left\{\inf_{\|\Delta\| \le M_1} \|U_n(\Delta, \ \hat{\sigma})\| > \eta\right\} > 1 - \tau,$$

using same arguments as Jurečková's (1977) proof of her Lemma 5.2. By (b.18),

$$n^{1/2}(\check{\theta}_n(\hat{\sigma}) - \theta^o) = V(\sigma)^{-1} \ U_n(0, \ \sigma) + o_p(1) \longrightarrow_d N(0, \ \Phi_1),$$

by the clt.

Proof of Theorem 2. (b): Following arguments given in the proof to Theorem 2, it suffices to prove that,

$$\sup_{\|\Delta\| \le M} \|U_n(\Delta, \ \check{\sigma}) - U_n(0, \ \check{\sigma}) - \{U_n(\Delta, \ \sigma) - U_n(0, \ \sigma)\}\| = o_p(1), \tag{b.19}$$

and

$$\|U_n(\Delta, \check{\sigma}) - U_n(0, \sigma)\| = o_p(1). \tag{b.20}$$

(b.19) and (b.20) follow in a similar manner as (b.2) and (b.10) but using the following Lemmas proved in Delgado (1989),

Lemma 12.- $\Pr\{\min_i \mathring{\sigma}_i \leq \delta\} = 0$ all n and some $\delta > 0$.

Lemma 13.- $\{\min \check{\sigma}^2\}^{-1} = O(1)$.

Lemma 14.- $E\{|\check{\sigma}_i^2 - \mathring{\sigma}_i^2|^{\upsilon/2}\} = O(k^{-\upsilon/4})$.

Lemma 15.- $\max_i |\check{\sigma}_i^2 - \mathring{\sigma}_i^2| = O_p(k^{-1/2})$.

where $\mathring{\sigma}_i^2 = \sum_j E(Y_j^2|X_j)\, \omega_{ij} - \{\sum_j X_j' \theta^o \omega_{ij}\}^2$.

Acknowledgements: This work was started at the London School of Economics as part of my doctoral dissertation. I thank Peter Robinson for encouraging me to investigate this problem. Financial support from the Economic and Social Research Council (ESRC) reference numbers: B00232156 and R000231441 and from a Indiana University Faculty Grant is gratefully acknowledged.

References

Carroll, R.J. (1982), "Adapting for heteroscedasticity in linear models", Annals of Statistics 10, 1244-1233.

Carroll, R.J. and D. Ruppert (1982), "Robust estimation in heteroscedastic linear models", Annals of Statistics 10,429-441.

Delgado, M.A. (1989), "An alternative asymptotically efficient semiparametric generalised least squares estimator", Unpublished.

Hampel, F.R., E.M. Rouchetti, P.J. Rousseew and W.A. Stahel (1986), Robust Statistics, Wiley: New York.

Huber, P.J. (1973), "Robust regression: asymptotics, conjectures and Monte Carlo", Annals of Statistics 1, 799-821.

Jureckova, J. (1977), "Asymptotic relations of M-estimates and R-estimates in linear regression model", Annals of Statistics 5,464-472.

Maronna, R.A. and V.J. Yohai, "Asymptotic behaviour of general M-estimates for regression and scale with random carriers", Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 58, 7-20

Relles, D. (1968), Robust Regression by Modified Least Squares, Ph.D. Thesis, Yale University.

Robinson, P.M. (1987), "Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form", Econometrica 55, 875-891.

Rose, R.L. (1978), Nonparametric Estimation of Weights in Least-Squares Regression Analysis, Ph.D. Thesis, University of California at Davis.

Stone, C.J. (1977), "Consistent nonparametric regression", (with discussion) Annals of Statistics 5,595-645.

Yohai, V.J. and R.A. Maronna (1979), "Asymptotic behavior of M-estimators for the linear model", Annals of Statistics 7, 258-268.

TABLE 1

CONTAMINATED ERRORS

MODEL 1 (γ= 0.)

	N= 30			N= 100			N=500		
	BIAS	VAR	EFF	BIAS	VAR	EFF	BIAS	VAR	EFF
$\hat{\theta}_n(\sigma)$	-5E-5	.1028	1.0000	-.0045	.0317	1.0000	-6E-4	.0062	1.0000
$\hat{\theta}_n(1)$	-5E-5	.1028	1.0000	-.0045	.0317	1.0000	-6E-4	.0062	1.0000
LADE	-.0027	.1007	1.0204	.0022	.0282	1.1235	-.0016	.0059	1.0528
$\hat{\theta}_n(\hat{\sigma})^1$	-.0015	.3541	.2903	-.0043	.0569	.5564	6E-4	.0078	.7965
$\hat{\theta}_n(\hat{\sigma})^2$	.0031	.3073	.3348	-6E-4	.0488	.6483	-4E-4	.0069	.8961
$\hat{\theta}_n(\check{\sigma})^1$	.0026	.3703	.2776	-.0032	.0585	.5415	8E-4	.0079	.7908
$\hat{\theta}_n(\check{\sigma})^2$	.0132	.2881	.3566	-.0025	.0477	.6639	-3E-4	.0069	.9009
$\check{\theta}_n(\sigma)$	-9E-4	.0766	1.3416	-.0025	.0220	1.4392	-8E-4	.0043	1.4327
$\check{\theta}_n(1)$	-9E-4	.0766	1.3416	-.0025	.0220	1.4392	-8E-4	.0043	1.4327
$\check{\theta}_n(\hat{\sigma})^1$	2E-4	.1488	.6909	-.006	.030	1.0429	-7E-4	.0050	1.2353
$\check{\theta}_n(\hat{\sigma})^2$	3E-4	.1136	.9053	-.003	.027	1.1471	-.0012	.0049	1.2697
$\check{\theta}_n(\check{\sigma})^1$	6E-4	.1370	.7501	-.004	.029	1.0752	-7E-4	.0050	1.2413
$\check{\theta}_n(\check{\sigma})^2$	.0135	.1085	.9464	-.001	.027	1.177	-.0011	.0048	1.2757

MODEL 1 (γ= -1.)

	N= 30			N= 100			N=500		
	BIAS	VAR	EFF	BIAS	VAR	EFF	BIAS	VAR	EFF
$\hat{\theta}_n(\sigma)$	-2E-5	.0146	1.0000	4E-4	.0046	1.0000	-6E-4	88E-5	1.0000
$\hat{\theta}_n(1)$	.0026	.0339	.4310	-.0031	.0096	.4816	-7E-7	.0020	.4384
LADE	-5E-4	.0193	.7594	-5E-4	.0053	.8824	-9E-4	.0010	.8064
$\hat{\theta}_n(\hat{\sigma})^1$	-5E-4	.0529	.2767	.0014	.0086	.5391	-4E-4	.0011	.7808
$\hat{\theta}_n(\hat{\sigma})^2$	.0017	.0430	.3405	-2E-5	.0073	.6342	-7E-4	99E-4	.8942
$\hat{\theta}_n(\check{\sigma})^1$	.0044	.0578	.2533	.0014	.0091	.5129	-2E-4	.0011	.7785
$\hat{\theta}_n(\check{\sigma})^2$	.0094	.0430	.3402	.0010	.0069	.6669	-7E-4	97E-4	.9075
$\check{\theta}_n(\sigma)$	-.0010	.0109	1.3354	-3E-4	.0032	1.4432	-9E-4	62E-5	1.4167
$\check{\theta}_n(1)$	.0016	.0257	.5697	-.0026	.0072	.6453	-9E-4	.0014	.6098
$\check{\theta}_n(\hat{\sigma})^1$	-3E-4	.0217	.6738	-6E-4	.0046	1.0182	-.0013	71E-5	1.2382
$\check{\theta}_n(\hat{\sigma})^2$	5E-4	.0176	.8317	-3E-4	.0042	1.1051	-.0012	69E-5	1.2695
$\check{\theta}_n(\check{\sigma})^1$	.0044	.0210	.6959	3E-4	.0045	1.0351	-.0012	71E-5	1.2381
$\check{\theta}_n(\check{\sigma})^2$	.0087	.0187	.7822	4E-4	.0043	1.0675	-.0011	70E-5	1.2574

TABLE 1 (Cont.)

MODEL 2 (δ= 3.)

	N= 30			N= 100			N=500		
	BIAS	VAR	EFF	BIAS	VAR	EFF	BIAS	VAR	EFF
$\hat{\theta}_n(\sigma)$	-.0015	4.8231	1.0000	-.0361	1.3672	1.0000	.0197	.2781	1.0000
$\hat{\theta}_n(1)$	-.0653	23.390	.2061	-.0348	6.9974	.1975	-.0172	1.3980	.1992
LADE	-.0455	9.0602	.5322	.0054	2.4612	.5560	-.0051	.4824	.5772
$\hat{\theta}_n(\hat{\sigma})^1$	-.0397	19.105	.2524	-.0245	2.6323	.5197	.0322	.3846	.7222
$\hat{\theta}_n(\hat{\sigma})^2$	-.0320	16.698	.2888	-.0268	2.2223	.6156	.0226	.3381	.8226
$\hat{\theta}_n(\check{\sigma})^1$	-.0519	21.121	.2283	-.0249	2.7676	.4943	.0304	.3895	.7132
$\hat{\theta}_n(\check{\sigma})^2$	-.0290	16.021	.3010	-.0276	2.2027	.6211	.0220	.3375	.8238
$\check{\theta}_n(\sigma)$	.0139	3.6326	1.3276	-.0236	1.0115	1.3522	.0219	.1975	1.4067
$\check{\theta}_n(1)$	-.0487	8.7229	.5527	.0043	2.4801	.5517	-5E-4	.4954	.5622
$\check{\theta}_n(\hat{\sigma})^1$	-2E-4	7.7019	.6262	-.0132	1.4622	.9357	.0354	.2500	1.1084
$\check{\theta}_n(\hat{\sigma})^2$	-.0290	6.9172	.6971	-.0159	1.3926	.9825	.0250	.2406	1.1543
$\check{\theta}_n(\check{\sigma})^1$	.0117	7.3163	.6592	-.0082	1.4371	.9522	.0344	.2487	1.1144
$\check{\theta}_n(\check{\sigma})^2$	-.0063	6.5064	.7413	-.0163	1.3660	1.0016	.0245	.2401	1.1573

MODEL 2 (δ= 5.)

	N= 30			N= 100			N=500		
	BIAS	VAR	EFF	BIAS	VAR	EFF	BIAS	VAR	EFF
$\hat{\theta}_n(\sigma)$	.0194	39.714	1.0000	-.1011	9.8887	1.0000	.0592	1.8264	1.0000
$\hat{\theta}_n(1)$	-.5159	1461.4	.0271	-.2760	442.68	.0223	.1996	88.534	.0207
LADE	-.2619	258.41	.1536	.0603	56.269	.1759	.0139	9.4482	.1936
$\hat{\theta}_n(\hat{\sigma})^1$	-.1122	206.76	.1920	-.0536	21.595	.4583	.0848	2.6485	.6890
$\hat{\theta}_n(\hat{\sigma})^2$	-.1928	231.36	.1716	-.1033	19.826	.4990	.0598	2.3742	.7695
$\hat{\theta}_n(\check{\sigma})^1$	-.1644	216.73	.1832	-.0530	21.978	.4503	.0783	2.6674	.6844
$\hat{\theta}_n(\check{\sigma})^2$	-.2036	211.56	.1876	-.1026	19.534	.5065	.0569	2.3597	.7744
$\check{\theta}_n(\sigma)$	.0720	30.293	1.3107	-.0727	7.3305	1.3494	.0658	1.2893	1.4144
$\check{\theta}_n(1)$	-.2749	256.30	.1549	.0464	57.008	.1736	.0047	9.9333	.1842
$\check{\theta}_n(\hat{\sigma})^1$	.0353	82.535	.4812	-.0227	11.716	.8448	.1010	1.7162	1.0599
$\check{\theta}_n(\hat{\sigma})^2$	-.1151	99.400	.3995	-.0634	12.559	.7879	.0737	1.6991	1.0735
$\check{\theta}_n(\check{\sigma})^1$	.0584	75.235	.5276	-.0182	11.026	.8978	.0982	1.6752	1.0861
$\check{\theta}_n(\check{\sigma})^2$	-.0652	89.291	.4447	-.0620	12.217	.8100	.0719	1.6848	1.0827

TABLE 1 (Cont.)

MODEL 1 (γ= -2.)

	N= 30			N= 100			N=500		
	BIAS	VAR	EFF	BIAS	VAR	EFF	BIAS	VAR	EFF
$\hat{\theta}_n(\sigma)$	-3E-6	.0016	1.0000	6E-4	46E-5	1.0000	-1E-4	95E-6	1.0000
$\hat{\theta}_n(1)$	.0028	.0227	.0701	-.0028	.0063	.0725	-4E-4	.0013	.0697
LADE	3E-4	.0047	.3359	-5E-5	.0012	.3791	-5E-4	23E-5	.4103
$\hat{\theta}_n(\hat{\sigma})^1$	-3E-4	.0065	.2449	.0010	98E-5	.4681	-3E-5	12E-5	.7513
$\hat{\theta}_n(\hat{\sigma})^2$	6E-4	.0056	.2826	1E-4	78E-5	.5868	-1E-5	11E-5	.8546
$\hat{\theta}_n(\check{\sigma})^1$	.0041	.0172	.0931	9E-4	.0014	.3355	-3E-4	13E-5	.7229
$\hat{\theta}_n(\check{\sigma})^2$	.0107	.0192	.0829	6E-4	.0016	.2841	-4E-4	15E-5	.6048
$\check{\theta}_n(\sigma)$	-3E-4	.0012	1.3003	3E-4	32E-5	1.3991	-2E-4	68E-6	1.4010
$\check{\theta}_n(1)$	.0021	.0174	.0920	-.0022	.0047	.0974	-6E-4	97E-5	.0977
$\check{\theta}_n(\hat{\sigma})^1$	-3E-4	.0027	.599	4E-4	50E-5	.9189	-3E-4	79E-6	1.1982
$\check{\theta}_n(\hat{\sigma})^2$	5E-5	.0025	.6466	2E-4	46E-5	1.0006	-3E-4	77E-6	1.2366
$\check{\theta}_n(\check{\sigma})^1$	.0028	.0051	.3122	4E-4	74E-4	.6718	-5E-4	87E-6	1.0921
$\check{\theta}_n(\check{\sigma})^2$	.0084	.0068	.2326	5E-4	.0010	.4450	-6E-4	11E-5	.8253

MODEL 2 (δ= 1.)

	N= 30			N= 100			N=500		
	BIAS	VAR	EFF	BIAS	VAR	EFF	BIAS	VAR	EFF
$\hat{\theta}_n(\sigma)$	-4E-5	.3627	1.0000	-.0067	.1061	1.0000	.0015	.0222	1.0000
$\hat{\theta}_n(1)$	-.0057	.4716	.7691	-.0047	.1399	.7588	3E-4	.0283	.7836
LADE	-.0079	.3865	.9383	.0024	.1123	.9455	-.0022	.0227	.9779
$\hat{\theta}_n(\hat{\sigma})^1$	-.0053	1.2806	.2832	-.0056	.2016	.5269	.0042	.0286	.7750
$\hat{\theta}_n(\hat{\sigma})^2$	.0018	1.1202	.3237	-.0037	.1692	.6278	.0019	.0254	.8714
$\hat{\theta}_n(\check{\sigma})^1$	-.0061	1.3736	.2640	-.0048	.2065	.5143	.0042	.0289	.7661
$\hat{\theta}_n(\check{\sigma})^2$	.0089	1.0821	.3352	-.0025	.1670	.6358	.0019	.0254	.8738
$\check{\theta}_n(\sigma)$	5E-4	.2715	1.3357	-.0042	.0778	1.3637	-.0015	.0155	1.4262
$\check{\theta}_n(1)$	-.0070	.3113	1.1649	-.0010	.0914	1.1616	-.0013	.0189	1.1709
$\check{\theta}_n(\hat{\sigma})^1$	-4E-4	.5389	.6731	-.0040	.1101	.9640	.0031	.0186	1.1962
$\check{\theta}_n(\hat{\sigma})^2$	-.0036	.4175	.8688	-.0027	.0979	1.0842	-.0013	.0179	1.2353
$\check{\theta}_n(\check{\sigma})^1$	.0064	.5006	.7244	-.0037	.1072	.9902	.0030	.0185	1.1962
$\check{\theta}_n(\check{\sigma})^2$	.0087	.3956	.9166	-.0018	.0960	1.1063	-.0013	.0179	1.2353

TABLE 2

NORMAL ERRORS

MODEL 1 (γ= -2.)

	N= 30			N= 100			N=500		
	BIAS	VAR	EFF	BIAS	VAR	EFF	BIAS	VAR	EFF
$\hat{\theta}_n(\sigma)$	-2E-4	.0016	1.0000	-2E-4	46E-5	1.0000	-2E-4	92E-6	1.0000
$\hat{\theta}_n(1)$	.0028	.0448	.0707	-7E-4	.0069	.0675	-7E-4	.0013	.0681
LADE	1E-4	.0072	.2232	-7E-5	.0019	.2458	-6E-4	35E-5	.2581
$\hat{\theta}_n(\hat{\sigma})^1$	1E-4	.0049	.3277	-7E-5	64E-5	.7198	-1E-4	10E-5	.9057
$\hat{\theta}_n(\hat{\sigma})^2$	8E-4	.0039	.4121	6E-4	56E-5	.8225	-3E-4	97E-6	.9500
$\hat{\theta}_n(\check{\sigma})^1$	.0048	.0118	.1369	5E-4	.0010	.4638	-4E-4	11E-5	.8439
$\hat{\theta}_n(\check{\sigma})^2$	.0117	.0122	.1311	8E-4	.0013	.3543	-6E-4	14E-5	.6584
$\check{\theta}_n(\sigma)$	-3E-4	.0017	.9637	1E-4	48E-5	.9653	-3E-4	98E-6	.9443
$\check{\theta}_n(1)$	.0025	.0217	.0746	-6E-4	.0064	.0725	-8E-4	.0013	.0733
$\check{\theta}_n(\hat{\sigma})^1$	1E-4	.0031	.5201	-1E-4	58E-5	.8042	-3E-4	10E-5	.9087
$\check{\theta}_n(\hat{\sigma})^2$	9E-5	.0029	.5729	2E-5	54E-5	.8647	-3E-4	97E-6	.9487
$\check{\theta}_n(\check{\sigma})^1$	.0034	.0060	.2692	2E-4	85E-5	.5465	-5E-4	11E-5	.8598
$\hat{\theta}_n(\check{\sigma})^2$	.0084	.0079	.2039	7E-4	.0012	.3959	-6E-4	14E-5	.6583

MODEL 2 (δ= 5.)

	N= 30			N= 100			N=500		
	BIAS	VAR	EFF	BIAS	VAR	EFF	BIAS	VAR	EFF
$\hat{\theta}_n(\sigma)$	.0176	39.134	1.0000	-.0972	9.8403	1.0000	.0637	1.7540	1.0000
$\hat{\theta}_n(1)$	-.5878	1490.7	.0262	-.2371	441.68	.0223	.2241	87.482	.0201
LADE	-.3699	394.42	.0992	.0495	86.638	.1137	.0062	14.855	.1183
$\hat{\theta}_n(\hat{\sigma})^1$	.0112	160.79	.2434	-.0311	15.677	.6282	.0754	2.0997	.8350
$\hat{\theta}_n(\hat{\sigma})^2$	-.0714	162.41	.2409	-.1191	15.872	.6200	.0624	2.0867	.8409
$\hat{\theta}_n(\check{\sigma})^1$	-.0039	158.70	.2466	-.0149	15.011	.6561	.0688	2.0572	.8526
$\hat{\theta}_n(\check{\sigma})^2$	-.0192	142.78	.2741	-.1137	15.431	.6377	.0586	2.0709	.8475
$\check{\theta}_n(\sigma)$	.0546	40.576	.9644	-.0918	10.283	.9570	.0711	1.8492	.9481
$\check{\theta}_n(1)$	-.3630	391.76	.0999	-.0363	87.009	.1132	7E-4	15.251	.1153
$\check{\theta}_n(\hat{\sigma})^1$	.0569	99.315	.3940	-.0354	13.701	.7188	.0859	2.0444	.8568
$\check{\theta}_n(\hat{\sigma})^2$	-.0367	117.45	.3331	-.1047	14.854	.6626	.0668	2.0589	.8520
$\check{\theta}_n(\check{\sigma})^1$	.0663	91.335	.4284	-.0169	12.966	.7596	.0795	2.0009	.8758
$\hat{\theta}_n(\check{\sigma})^2$	.0094	104.85	.3732	-.0990	14.465	.6804	.0636	2.0433	.8587

LINEAR REGRESSION WITH RANDOMLY RIGHT-CENSORED DATA USING PRIOR NONPARAMETRIC ESTIMATION

W. GONZALEZ-MANTEIGA and C. M. CADARSO-SUAREZ
Departamento de Estadistica e I.O.
Facultad de Matemáticas.
Universidad de Santiago de Compostela.
Spain.

ABSTRACT. A new non-iterative method for estimating the coefficients of a linear regression model from censored data is developed on the basis of prior non parametric estimation of the regression function, and is compared with others by extensive simulation. The results of its application to a set of real data are reported, together with those of a variant that appears to be more suitable under skewed censorship.

1. Introduction.

A linear regression model of fixed design can be expressed in the general form

$$T_i = A^t(z_i)\theta + \varepsilon_i \qquad (i=1,\dots,n;\ z_i\in C\subset\mathbb{R}^p\ ,\ C \text{ compact})$$

with [1.1]

$$E\varepsilon_i = 0 \qquad (i=1,\dots,n)$$

where $A\colon \mathbb{R}^p \longrightarrow \mathbb{R}^q$ is a general functional, the T_i are the dependent variables and the residuals ε_i do not in general have equal variances. Particular A's yield polynomial regression $(A^t(z)=(1,z,..,z^{q-1}))$. In this article we are interested in models which the T_i are the survival times or functions of survival times (typically the logarithm), in which case it is common for the observed T_i to be censored. In medicine, for example, a patient may still alive when the study is wound up, may have died of causes unrelated to the pathology being studied, or may have ceased to participate in the study for other reasons.

The general censorship model that is assumed in this article, and which is applicable to the above causes of censorship among others, is random right censorship, in which the observed data are $\{(z_i,\delta_i,Y_i)\}_{i=1}^n$, where $Y_i=\min\{T_i,C_i\}$ and $\delta_i=\mathbf{1}_{\{T_i\le C_i\}}$, the C_i being the censoring variables. The value of δ_i is 1 if T_i is not censored and 0 if it is.

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 315–328.
© 1991 *Kluwer Academic Publishers.*

If the distributions of the T_i and C_i are denoted respectively by:

$$F(t|z_i) = 1 - P(T>t|z_i) \text{ and } G(t|z_i) = 1 - P(C>t|z_i)$$

and if the T_i and C_i are assumed to be conditionally independent, then the distribution of the Y_i is

$$H(t|z_i) = 1 - [1 - F(t|z_i)][1 - G(t|z_i)] = 1 - P(Y>t|z_i)$$

and the "subdistributions" of the censored and uncensored data are respectively

$$\tilde{H}(t|z_i) = P(Y_i \le t, \delta_i=1) = P(Y \le t, \delta=1|z_i) = \\ = \int_{-\infty}^{t} (1 - G(s|z_i))dF(s|z_i)$$

and

$$\bar{\tilde{H}}(t|z_i) = P(Y_i \le t, \delta_i=0) = P(Y \le t, \delta=0|z_i) = \\ = \int_{-\infty}^{t} (1 - F(s|z_i))dG(s|z_i)$$

The cumulative hazard rate function is

$$\Lambda(t|z_i) = - Ln(1 - F(t|z_i)) = \int_{-\infty}^{t} \frac{dF(s|z_i)}{1 - F(s|z_i)} = \\ = \int_{-\infty}^{t} \frac{d\tilde{H}(s|z_i)}{1 - H(s|z_i)}$$

Our aim is to estimate the regression function

$$m(T|z_i) = A^t(z_i)\theta = E[T|z_i] = ET_i = \int t dF(t|z_i)$$

During the last 15 years, the following articles have proposed ways of estimating θ from right-censored data for the multiple regression model ($A^t(z) = (1,z)$).

A) Miller (1976). Miller's generalized least squares method for censored data estimates θ by iteratively minimizing the functional

$$\Psi_1(\theta) = \int \varepsilon^2 d\hat{F}_{km}(\varepsilon;\theta_0) \qquad [1.2]$$

where $\hat{F}_{km}(\varepsilon;\hat{\theta}_0)$ is the Kaplan-Meier (1958) estimator (hereinafter KM) constructed from the sample $\{(Y_i-A^t(z_i)\hat{\theta}_0, \delta_i)\}_{i=1}^n$, $\hat{\theta}_0$ being an initial estimate. For obtaining which it is recommended that least squares regression be performed for the uncensored subsample.

B) Buckley and James (1979). This method for fixed C_i is based on the behaviour of

$$Y_i' = \delta_i Y_i + (1-\delta_i)E[T_i|T_i>Y_i]$$

which satisfies

$$E[Y_i'|z_i] = A^t(z_i)\theta \qquad [1.3]$$

Given an initial estimate of θ (preferably obtained by least squares

regression for the sample $\{(z_i,Y_i)\}_{i=1}^{n}$ as uncensored data), then obtained by least squares for the dependent variables

$$\hat{Y}_i = \begin{cases} Y_i & \text{if } \delta_i = 1 \\ \delta_i Y_i + (1-\delta_i)\hat{E}[T_i|T_i>Y_i] & \text{if } \delta_i = 0 \end{cases}$$

This method too is iterative.

C) Koul, Susarla and Van Ryzin (1981a). This non-iterative method is based on the identity

$$E[\delta_i Y_i(1-G(Y_i|z_i))^{-1}|z_i] = A^t(z_i)\Theta \qquad [1.4]$$

A modified KM estimator $\hat{G}$ of $G(t|z_i)$ (assumed identical for all z_i) is obtained from the sample $\{(Y_i,\delta_i)\}$; and Θ is then estimated by least squares for the dependent variables

$$\hat{Y}_i = \delta_i Y_i(1-\hat{G}(Y_i))^{-1}\mathbf{1}_{\{T_i\le\tau_n\}},$$

where τ_n is a suitable sequence of numbers tending to infinity with n.

D) Koul, Susarla and Van Ryzin (1982). To generalize the Buckley-James method to random censoring variables, the variable

$$\begin{aligned} Y_i' &= \delta_i Y_i + (1-\delta_i)E[T_i|\delta_i=0] = \\ &= \delta_i Y_i + (1-\delta_i)E[T_i|\delta_i=0] = \\ &= \delta_i Y_i + (1-\delta_i)\frac{\int s(1-G(s|z_i))dF(s|z_i)}{\int(1-G(s|z_i))dF(s|z_i)} = \\ &= \delta_i Y_i + (1-\delta_i)\frac{\int s(1-G(s|z_i))dF(s-A^t(z_i)\Theta)}{\int(1-G(s|z_i))dF(s-A^t(z_i)\Theta)} \end{aligned}$$

is estimated by

$$\delta_i Y_i + (1-\delta_i)\frac{\int s(1-\hat{G}(s))d\hat{F}_{km}(s-A^t(z_i)\hat{\Theta}_0)}{\int((1-\hat{G}(s))d\hat{F}_{km}(s-A^t(z_i)\hat{\Theta}_0)}$$

where, as in method C above, $\hat{G}$ is a modified KM estimator of the distribution of the censoring C_i, which is assumed identical for all values of z_i.

As in the original Buckley-James method, calculation of $\hat{\Theta}_1$, is iterative, the recommended initial estimate now being the $\hat{\Theta}_0$ obtained as in method C above.

E) Koul, Susarla and Van Ryzin (1981b). In this weighted least squares method the weights are obtained using the same estimator $\hat{G}$ as in methods C and D above. Specifically, θ is estimated by minimizing

$$\Psi_2(\theta) = \sum_{i=1}^{n} \frac{\delta_i}{1 - \hat{G}(Y_i)} \left(Y_i - A^t(z_i)\theta\right)^2 \qquad [1.5]$$

Of the above methods, A, B and D suffer from computational complications because of their iterative nature, while C and E (and D) are severely limited by the assumption that the distributions of the censoring variables are independent. All five methods have in common their reducing to the least squares estimator if the data are all uncensored, in which case the functionals [1.2] and [1.5] become

$$\Psi_3(\theta) = \sum_{i=1}^{n} \left(Y_i - A^t(z_i)\theta\right)^2 \qquad [1.6]$$

while the variable averaged in [1.3] and [1.4] is T_i.

In this article we base a new non-iterative method for regression with randomly right-censored data on prior nonparametric estimation of the regression function $m(t|z)$ $(z \in C)$. In Section 2 below we describe the new method in detail, in Section 3 its performance in an extensive simulation study is compared with that of previous methods, and in Section 4 it is applied to a set of real data together with a variant that appears to be preferable under skewed censorship.

2. The new Method.

In that follows, $F(t|z_i)$ and $G(t|z_i)$ are assumed belong to the families

$$\left\{F(\circ|z)\right\}_{z \in C} \text{ and } \left\{G(\circ|z)\right\}_{z \in C}$$

which vary smoothly with $z \in C$. Of the few published studies on nonparametric estimation of regression functions (i.e. of conditional distributions) from censored data, the most interesting are those of Beran (1981) and Dabrowska (1987, 1989). Unfortunately, their estimators for F are based, via the identity

$$\hat{F}_n(t|z) = 1 - \exp[-\hat{\Lambda}_n(t|z)] \ ,$$

on nonparametric estimation of the conditional cumulative hazard rate

$$\hat{\Lambda}_n(t|z) = \int_{-\infty}^{t} \frac{d\hat{\tilde{H}}_n(s|z)}{1 - \hat{H}_n(s|z)} \quad , \quad z \in C$$

where

$$\hat{\tilde{H}}n(s|z) = \sum_{j=1}^{n} \mathbf{1}_{\{Yj\leq s,\,\delta j=1\}} Bnj(z)$$

and

$$\hat{H}n(s|z) = \sum_{j=1}^{n} \mathbf{1}_{\{Yj\leq s\}} Bnj(z)$$

are nonparametric estimates of $\tilde{H}$ and H respectively, $\{Bnj(z)\}$ being the sequence of weights defined by the particular nonparametric estimation method used (for an excellent review of nonparametric estimations methods, see Härdle (1990)). In consequence, the resulting nonparametric estimator of the regression function,

$$mn(t|z) = \int t d\hat{F}n(t|z) \qquad [2.1]$$

is not particularly practical.

A more workable class of estimators is obtained as follows. First, to prevent the possibility of zero denominators, the nonparametric estimator of the conditional cumulative hazard rate is changed from

$$\hat{\Lambda}n(t|z) = \sum_{j=1}^{n} \left[\frac{\mathbf{1}_{\{Yj\leq t,\,\delta j=1\}} Bnj(z)}{1 - \sum_{r=1}^{n} \mathbf{1}_{\{Yr\leq Yj\}} Bnr(z)} \right]$$

to

$$\hat{\Lambda}n^{1}(t|z) = \sum_{j=1}^{n} \left[\frac{\mathbf{1}_{\{Yj\leq t,\,\delta j=1\}} Bnj(z)}{1 - \sum_{r=1}^{n} \mathbf{1}_{\{Yr\leq Yj\}} Bnr(z) + Bnj(z)} \right],$$

which makes the estimator of 1-F(t|z)

$$1-\hat{F}n(t|z) = \exp\{-\hat{\Lambda}n^{1}(t|z)\} =$$

$$= \prod_{j=1}^{n} \exp\left\{ - \frac{\mathbf{1}_{\{Yj\leq t,\,\delta=1\}} Bnj(z)}{1 - \sum_{r=1}^{n} \mathbf{1}_{\{Yr<Yj\}} Bnr(z)} \right\}$$

This expression is now approximated by Taylor expansion to the first order, to give

$$1-\hat{F}n(t|z) =$$

$$= \begin{cases} \prod_{j=1}^{n} \left[\dfrac{1 - \sum_{r=1}^{n} 1_{\{Y_r \leq Y_j\}} B_{nr}(z)}{1 - \sum_{r=1}^{n} 1_{\{Y_r < Y_j\}} B_{nr}(z)} \right]^{1_{\{Y_j \leq t, \delta_j = 1\}}} & \text{if } t < Y_{(n)} \\ 0 & \text{if } t \geq Y_{(n)} \end{cases} \qquad [2.2]$$

The new estimator [2.2] generalizes the KM estimator, to which it reduces if the independent variable is constant ($B_{nj}(z)=1/n$ for all j); if the z_i do vary but there is no censorship, [2.2] gives the classical nonparametric estimator of the conditional distribution. What makes [2.2] useful is that the jump in $\hat{F}_n$ at each Y_j for which $\delta_j=1$ is $B_{nj}(z)/(1-\hat{G}_n((Y_j|z))$, where the denominator is obtained from [2.2] by substituting $\hat{G}_n$ for $\hat{F}_n$ and interchanging the roles of T and C; as a result, [2.1] becomes

$$m_n(t|z) = \sum_{j=1}^{n} \frac{Y_j \delta_j B_{nj}(z)}{1 - \hat{G}_n(Y_j|z)} \qquad [2.3]$$

In the absence of censorship ($\delta_i = 1$, $\hat{G}_n = 0$), [2.3] is the classical nonparametric regression function estimator (see Härdle (1990)).

Having obtained [2.3], we now define an estimator for Θ in [1.1] as the value minimizing the functional

$$\Psi_4(\Theta) = \int \left(m_n(t|z) - A^t(z)\Theta \right)^2 d\Omega_n(z) \qquad [2.4]$$

where $\Omega_n(z)$ is a weighting function. Estimators of this type for random design models given uncensored data have been extensively studied by Cristóbal, Faraldo and González-Manteiga (1987) and Stute and González-Manteiga (1990) for nonparametric estimation by the kernel and k nearest neighbour methods respectively. In the present context, optimization of [2.4] using the empirical weighting function on the z_i as Ω_n leads to the class of estimators

$$\hat{\Theta}_n = \left[\sum_{j=1}^{n} A(z_j) A^t(z_j) \right]^{-1} \left[\sum_{i=1}^{n} \sum_{j=1}^{n} A(z_i) \frac{B_{nj}(z_i) Y_j \delta_j}{1-\hat{G}_n(Y_j|z_i)} \right] \qquad [2.5]$$

These estimators are calculated non-iteratively, and do not need the strong assumptions concerning the censoring variables that are made in methods B-E of Section 1 above. Like methods A-E, they reduce to least squares for uncensored data, since [2.4] becomes [1.6] when Ω_n is the empirical weighting function and

$$m_n(t|z) = \sum_{i=1}^{n} 1_{\{Y_i\}}(z).$$

We are currently investigating the asymptotic distribution of the $\hat{\Theta}_n$. Their asymptotic covariance is expected to be

$$\sum = \frac{1}{n} B^{-1} \int A(z)g(z)A^t(z)dz\, B^{-1}$$

where

$$g(z) = \int \frac{1}{(1-H(u|z))^2} \left(\int_u^{\infty} (1-F(t|z))dt \right)^2 d\tilde{H}(u|z)$$

and

$$B = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} A(z_i)A^t(z_i) .$$

Note that if [1.1] is homoscedastic and there is no censorship, then $g(z) = \sigma^2$ and

$$\sum \simeq \frac{\sigma^2}{n} B^{-1} ,$$

the asymptotic covariance of the least squares estimator, which suggest that the new estimators may well be asymptotically efficient. In view of this, the simulation study described in the next section concentrates on the small sample behaviour of the contending methods.

3. Simulation results.

In this section we describe a simulation study in which the small sample behaviours of the new estimators and methods A, B and D were compared; methods C and E were excluded because their behaviour was specially worst than in the methods A, B and D in our previous simulations. The simulated model was

$$Y_i = \min \{T_i, C_i\} \qquad (i=1,\ldots,n)$$

where

$$T_i = \Theta_1 + \Theta_2 z_{1i} + \Theta_3 z_{2i} + \varepsilon_i$$

and $\qquad (i=1,\ldots,n) \qquad [3.1]$

$$C_i = \Theta_1 + \Theta_2 z_{1i} + \Theta_3 z_{2i} + \tilde{\varepsilon}_i$$

with

$$\Theta_1 = 2,\ \Theta_2 = 0.01 \text{ and } \Theta_3 = 0.5 ,$$

the ε_i and $\tilde{\varepsilon}_i$ being independent normally distributed random variables with variance σ^2 and mean 0 (for the ε_i) or Δ (for the $\tilde{\varepsilon}_i$). T and C were thus conditionally independent for each value of (z_{i1}, z_{i2}). For each of two sample sizes (n = 25 and n = 50), three variances (σ^2 = 0.25, σ^2 = 1 and σ^2 = 4) and three degrees of censorship DC (16%, 33% and 50%, DC being defined as $P(C_i<T_i) = P(\tilde{\varepsilon}_i<\varepsilon_i) = 1-\Phi(\Delta/\sigma\sqrt{2})$), N =

1000 samples corresponding to the model [3.1] were generated using the IMSL routine GGNML. For uniformity, (z_{1i}, z_{2i}) cycled through all points of $[0,1]\times[0,1]$ on the lattice $\{(k/l, s/m) \mid 0<k<l,\ 0<s<m,\ lm=n\}$ for $(l,m) = (5,5)$ or $(5,10)$.

In the particular new estimator that was applied to the simulated data, the underlying nonparametric estimation was performed by the kernel method with Nadaraya-Watson weighting (Nadaraya (1964), Watson (1964)). Specifically,

$$B_{ni}(z) = \frac{K\left(\dfrac{z - z_i}{h_n}\right)}{\displaystyle\sum_{r=1}^{n} K\left(\dfrac{z - z_r}{h_n}\right)}$$

where $z_i = (z_{1i}, z_{2i})$ $(i=1,\ldots,n)$ and K is the uniform kernel on $[-1,1] \times [-1,1]$; good values for h_n were found among different values to be 0.6 and 0.8. Subsequently, the 25-point samples were reanalysed using values de h_n chosen to minimize the cross-validation functional

$$\Psi_5(h_n) = \sum_{i=1}^{n} \left(Y_i - A^t(z_i)\hat{\Theta}_{-i}\right)^2 \qquad [3.2]$$

where $\hat{\Theta}_{-i}$ is [2.5] constructed without the i-th sample point.

Checking differents τ_n's was also used to choose the τ_n for method D, to find the initial estimate.

The performance of the regression parameter estimators was measured by the estimated multidimensional mean squared error,

$$\widehat{M.S.E.} = \sum_{i=1}^{3} \left[\left(\frac{\sum_{j=1}^{n} \hat{\Theta}_i^j}{N} - \Theta_i \right)^2 + \frac{1}{N} \sum_{j=1}^{N} \left(\hat{\Theta}_i^j - \bar{\hat{\Theta}}_i \right)^2 \right]$$

where

$$\bar{\hat{\Theta}}_i = \frac{1}{N} \sum_{j=1}^{N} \hat{\Theta}_i^j .$$

The results are listed in Tables 1-6. Miller's method performed well, generally better than methods B and D, doubtless because, unlike the latter methods, it is well suited to the simulation model used, in which the C_i are random variables (method B assumes constant C_i) and their distributions depend on the z_i (method D assumes fixed distributions). However, the new estimator was in general better than all the others (regardless of how the window parameter h_n was chosen) except when the variance was small, in which case initial smoothing was negligible.

TABLE 1.

n = 25 16% DC	$\sigma^2 = 0.25$	$\sigma^2 = 1$	$\sigma^2 = 4$
A	0.38439	1.50538	5.99790
B	0.54969	1.75343	6.40318
D	0.59986	1.74335	6.35973
NEW METHOD			
hn = 0.6	0.32228	0.44387	0.88835
hn = 0.8	0.32746	0.35295	0.47388
h_{cv}	0.31545	0.48960	1.18481

TABLE 2.

n = 25 33% DC	$\sigma^2 = 0.25$	$\sigma^2 = 1$	$\sigma^2 = 4$
A	0.44490	1.73189	6.85069
B	0.93599	2.55834	8.63294
D	0.94390	2.44727	8.22965
NEW METHOD			
hn = 0.6	0.47866	0.64498	1.26601
hn = 0.8	0.33094	0.37043	0.53407
h_{cv}	0.39776	0.61217	1.39255

TABLE 3.

n = 25 50% DC	$\sigma^2 = 0.25$	$\sigma^2 = 1$	$\sigma^2 = 4$
A	0.66102	2.52652	9.96524
B	1.61695	3.92416	12.50964
D	2.19767	4.03832	12.42808
NEW METHOD			
hn = 0.6	0.78819	1.07605	1.97866
hn = 0.8	0.38559	0.46584	0.75246
h_{cv}	0.59714	0.83174	1.89210

TABLE 4.

n = 50 16% DC	$\sigma^2 = 0.25$	$\sigma^2 = 1$	$\sigma^2 = 4$
A	0.19119	0.74450	2.93471
B	0.23080	0.80980	3.07125
D	0.23689	0.79952	3.05348
NEW METHOD $h_n = 0.6$ $h_n = 0.8$	 0.27509 0.32978	 0.34762 0.34219	 0.62973 0.40438

TABLE 5.

n = 50 33% DC	$\sigma^2 = 0.25$	$\sigma^2 = 1$	$\sigma^2 = 4$
A	0.26036	0.99208	3.91293
B	0.29019	0.95984	3.57532
D	0.31455	0.97052	3.54208
NEW METHOD $h_n = 0.6$ $h_n = 0.8$	 0.37015 0.32428	 0.48934 0.34372	 0.90029 0.43818

TABLE 6.

n = 50 50% DC	$\sigma^2 = 0.25$	$\sigma^2 = 1$	$\sigma^2 = 4$
A	0.37819	1.43661	5.65224
B	0.41908	1.31204	4.70697
D	0.44982	1.31804	4.68591
NEW METHOD $h_n = 0.6$ $h_n = 0.8$	 0.63885 0.35532	 0.84188 0.40175	 1.53530 0.57637

4. Application to real data.

Tables 7-9 show the results of applying the new estimator of Section 3 and methods A, B, C to data for the survival times of patients who underwent surgery for kidney cancers (method D reduces to method B for these data). The base 10 logarithm of post-operative survival time (in days) was regressed in turn of age, tumour size and globular sedimentation rate (GSR) at the time of surgery. In applying the new method, the window h_n was obtained by minimizing [3.2] (h_{cv}).

TABLE 7.

$\log_{10}(T)$ / AGE	α	β	
MILLER	2.50842	0.00018	
BUCKLEY-JAMES	2.30882	0.00280	
KOUL-SUSARLA-VAN RYZIN	1.77723	0.00350	
NEW METHOD	1.84949 2.09544	0.00390 -0.00012	hcv = 19 h = 12
n = 82	MODEL $\equiv \mathrm{Log}_{10}(T) = \alpha + \beta\mathrm{AGE} + \varepsilon$		

TABLE 8.

$\log_{10}(T)$ / SIZE	α	β	
MILLER	3.01903	-0.05526	
BUCKLEY-JAMES	3.88509 3.88453	-0.15026 -0.15031	loop with near values
KOUL-SUSARLA-VAN RYZIN	1.72708	0.00238	
NEW METHOD	2.09518 2.03843	-0.00682 -0.00280	hcv = 6 h = 12
n = 73	MODEL $\equiv \mathrm{Log}_{10}(T) = \alpha + \beta\mathrm{SIZE} + \varepsilon$		

TABLE 9.

$\log_{10}(T)$ / GSR	α	β	
MILLER	3.29145 2.63831	-0.00232 -0.00124	loop with far values
BUCKLEY-JAMES	2.88494	-0.00503	
KOUL-SUSARLA-VAN RYZIN	1.69845	0.00553	
NEW METHOD	2.09518 2.03843	-0.00682 -0.00280	hcv = 79 h = 125
n = 70	MODEL $\equiv \mathrm{Log}_{10}(T) = \alpha + \beta\mathrm{GSR} + \varepsilon$		

The results of the various methods differed appreciably for both regression coefficients. For α, the new method systematically yielded values that were smaller than those afforced by methods A and B and larger than those of method C; this was clearly due to the censored

data in this data set generally being those with the largest values of T, which meant that $G \simeq 0$ globally in method C and locally in the new method. For β, different signs were obtained by different methods, except for β_{AGE}, all estimates of which were positive; since the values of β_{AGE} were very small, survival time appears to be practically independent of patient age. For both β_{SIZE} and β_{GSR}, the collaborating urologists predicted negative values. Negative values of β_{SIZE} were obtained by the new method and by methods A and B (the method B iteration algorithm oscillated between two values, though they were close to each other) but not by method C. Negative values of β_{GSR} were obtained by methods A, B and the new method, but not by method C or the new method. the method A have not found a solution because the sequence became trapped in a loop with two values too far apart.

On the whole, the new method behaved rather like method C, with positive β_{GSR} and low values of α. The reason for this similarity must be that the factor $\delta_j Y_j/[1-G_n(Y_j|z)]$ in the terms of the nonparametric pilot estimator [2.3] is very similar to the $\hat{Y}_i$ of method C (see Section 1), except that the distribution of the censoring variable is no longer fixed. Hence it would appear to be the form of this factor that is responsible for the pessimistically low estimates of α produced by both these methods in a situation in which the censored data are mostly those with large values of the dependent variable. To remedy this, we reprocessed the same data using a different nonparametric pilot estimator. Specifically, the offending factor in the terms of [2.3] was replaced by the Buckley-James-type expression $\delta_i Y_i + (1-\delta_i)\hat{E}[Y_i/T_i>Y_i]$, where

$$\hat{E}[Y_i/T_i>Y_i] = \hat{E}[Y_i/\delta_i=0] = \frac{\sum_{j=1}^{n} Y_j(1-\delta_j)B_{nj}(z_i)}{\sum_{j=1}^{n} (1-\delta_j)B_{nj}(z_i)} .$$

The $\{B_{nj}(z)\}$ were once again derived from the uniform kernel with a window parameter determined by minimization of the cross-validation functional [3.2]. The results (Table 10) and Figs. 1-3 show that the use of a Buckley-James-type dependent variable in the pilot estimation led to estimated regression parameters similar to -but more easily computed than- those given by the Buckley-James method itself.

TABLE 10.

$Log_{10}(T)$ / T. SIZE	2.84825	-0.05004	$h_{cv} = 3$
$Log_{10}(T)$ / AGE	2.38213	0.00000	$h_{cv} = 50$
$Log_{10}(T)$ / GSR	2.56469	-0.00209	$h_{cv} = 46$

FIGURE 1.

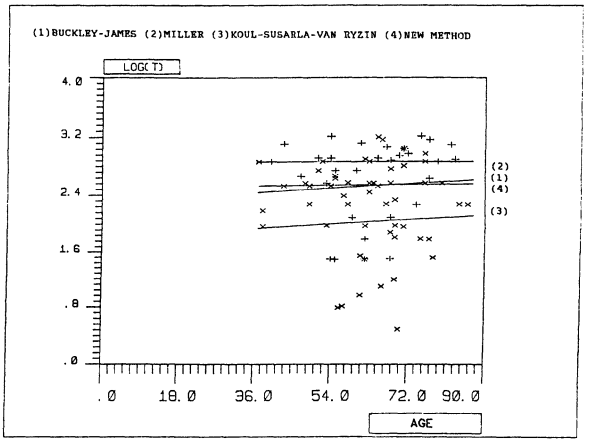


FIGURE 2.

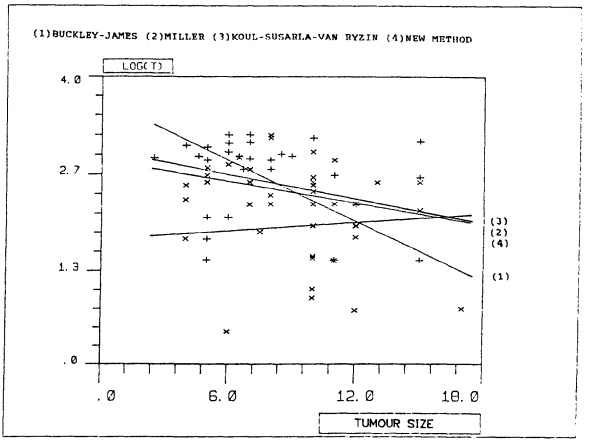
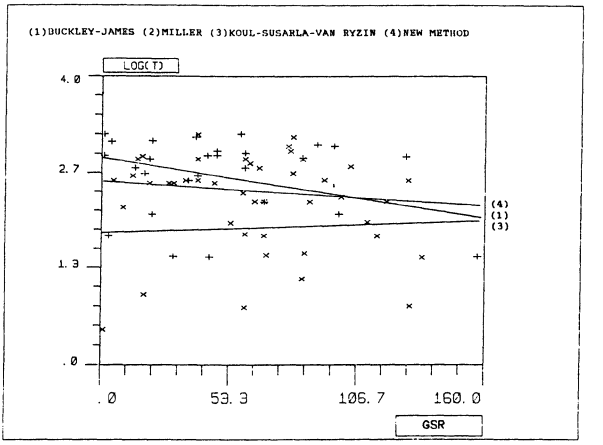


FIGURE 3.



Acknowledgements
We thank the Surgical Service of the Hospital General de Galicia (Santiago de Compostela, Spain) for providing the kidney tumour data used and Prof. H. Koul for mentioning two references that we were not acquainted with.
This work was supported by the Xunta de Galicia under Projet No. XUGA80505588.

5. References.

Beran, R. (1981) 'Nonparametric regression with randomly censored survival data', Technical Report, Univ. California, Berkeley.

Buckley, J. and James, I. (1979) 'Linear regression with censored data', Biometrika, Vol. 66, 429-436.

Cristóbal Cristóbal, J.A., Faraldo Roca, P. and González Manteiga, W. (1987) 'A class of linear regression parameter estimators constructed by nonparametric estimation', Ann. Statist., Vol. 15, 603-609.

Dabrowska, D.M. (1987) 'Nonparametric regression with censored survival data', Scand. J. Statist., Vol. 14, 181-197.

Dabrowska, D.M. (1989) 'Uniform consistency of the kernel conditional Kaplan-Meier estimate', Ann. Statist., Vol. 17, 1157-1167.

Härdle, W. (1990) Applied nonparametric regression, Cambridge University Press, to appear.

Kaplan, E.L. and Meier, P. (1958) 'Nonparametric estimation from incomplete observations', J. Amer. Statist. Assoc., Vol. 53, 457-481.

Koul, H., Susarla, V. and Van Ryzin, J. (1981a) 'Regression analysis with censored randomly right-censored data', Ann. Statist., Vol. 9, 1276-1288.

Koul, H., Susarla, V. and Van Ryzin, J. (1981b) 'Least squares regression with censored survival data', Topics in Applied Statistics (edited by Chaubey, V. and Dwivedi, T.), 151-165.

Koul, H., Susarla, V. and Van Ryzin, J. (1982), 'Multi-step estimator of regression coefficients in a linear model with censored survival data', I.M.S. Lecture notes series, Vol. 12, 85-100.

Miller, R.G. (1976) 'Least squares regression with censored data', Biometrika, Vol. 63, 449-464.

Miller, R.G. (1982) 'Regression with censored data', Biometrika, Vol. 69, 521-531.

Nadaraya, I.E. (1964) 'On estimating regression', Theory Probab. Appl., Vol. 9, 141-142.

Stute, W. and González Manteiga, W. (1990) 'Nearest neighbor smoothing in linear regression', J. Multiv. Analysis, Vol. 34.

Watson, G.S. (1964) 'Smooth regression analysis', Sankhyā Ser. A, Vol. 26, 359-372.

UNIVERSAL CONSISTENCIES OF A REGRESSION ESTIMATE FOR UNBOUNDED REGRESSION FUNCTIONS

L. GYÖRFI
Hungarian Academy of Sciences
Technical University of Budapest
H-1111 Budapest
Stoczek u. 2
Hungary

ABSTRACT The concepts of weak and strong universal consistencies are introduced. Based on equivalence theorems the universal consistencies of partitioning estimates are proved.

1. Introduction

(X,Y) is a pair of random variables taking values in $\mathbb{R}^d \times \mathbb{R}$. If $\mathbb{E}|Y|<\infty$ then the regression function m is defined by

$$m(x)=\mathbb{E}(Y/X=x).$$

For regression estimation problems we have to estimate m when a sequence $(X_1,Y_1),\ldots(X_n,Y_n)$ is given which is a sequence of i.i.d. copies of (X,Y). In the sequel μ denotes the probability measure of X.

Consider a regression estimate of the form

$$m_n(x) = \sum_{i=1}^{n} W_{ni}(x)Y_i, \tag{1.1}$$

where $W_{ni}(x)=W_{ni}(X_1,\ldots X_n,x)$ such that $W_{ni}(x)\geq 0$ and $\sum_i W_{ni}(x)=1$. Such weights $\{W_{ni}(x)\}$ are called *normal*.

The concept of universal consistency is the convergence of $\int|m_n(x)-m(x)|^p\mu(dx)$ under general conditions on the distribution of (X,Y). If $\lim_{n\to\infty} \mathbb{E}\left\{\int|m_n(x)-m(x)|^p\mu(dx)\right\} = 0$, then it is called *weak consistency*, if $\lim_{n\to\infty} \int|m_n(x)-m(x)|^p\mu(dx) = 0$ a.s., then it is called *strong consistency*. In the literature of nonparametric regression estimates there are usually two types of hard-to-verify conditions:

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 329–338.
© 1991 *Kluwer Academic Publishers.*

i. X has a density,

ii. m(x) is continuous or differentiable.

Universal consistency means the consistency without any unnecessary conditions on the underlying distribution. Obviously the necessary condition of consistency in L_p is that $\mathbb{E}|Y|^p<\infty$.

2. Weak universal consistency in L_p

Definition 1 m_n is called *weakly universally consistent* in L_p if

$$\lim_{n\to\infty} \mathbb{E}\left\{\int |m_n(x)-m(x)|^p \mu(dx)\right\} = 0 \tag{2.1}$$

for all Y with $E|Y|^p<\infty$ $(p\geq 1)$.

The problem of weak universal consistency was completely solved by Stone (1977), who gave necessary and sufficient conditions on normal weights for weak universal consistency in L_p. It is easy to prove a similar equivalence result:

Theorem 1 Assume that the weights are normal. Then the following statements are equivalent:

(i) m_n is weakly universally consistent in L_1.

(ii) m_n is weakly universally consistent in L_p for all $p\geq 1$.

(iii) m_n is weak consistent in L_1 for all bounded $|Y|$ and there is a $C>0$ such that for all $Y\in L_1$

$$\limsup_{n\to\infty} \mathbb{E}\left\{ \sum_{i=1}^{n} \int W_{ni}(x)\mu(dx)|Y_i|\right\} \leq C\, \mathbb{E}(|Y|). \tag{2.2}$$

The proof of Theorem 1 is a slight modification of the proof of Theorem 2 in the sequel, therefore we omit it.

Stone (1977) applied his result for proving the weak universal consistency in L_p of the k_n-nearest-neighbor (k_n-NN) estimate which can be formulated as follows: let $0<k_n<n$ be an integer such that

$$\lim_{n\to\infty} k_n = \infty, \quad \lim_{n\to\infty} k_n/n = 0, \tag{2.3}$$

and let $\rho_n(x)$ denote the distance of x and its k_n-th nearest neighbor from the samples $X_1, X_2, \ldots, X_n$. Then the k_n-NN regression estimate is defined by

$$m_n(x)=\frac{\sum_{i=1}^{n} I_{\{\|x-X_i\|\le\rho_n(x)\}}Y_i}{\sum_{i=1}^{n} I_{\{\|x-X_i\|\le\rho_n(x)\}}}. \tag{2.4}$$

Stone (1977) proved that (2.3) is the necessary and sufficient condition of the weak universal consistency in L_p of k_n-NN estimate (2.4) for any $p\ge1$.

Devroye and Wagner (1980), and independently Spiegelman and Sachs (1980) considered the kernel estimate: let h_n be real such that

$$\lim_{n\to\infty} h_n = 0, \quad \lim_{n\to\infty} nh_n^d = \infty. \tag{2.5}$$

Choose a nonnegative function K(x), which is called kernel, and put

$$K_n(x) = \frac{1}{h_n^d} K\left(\frac{x}{h_n}\right). \tag{2.6}$$

Then the kernel estimate is

$$m_n(x)=\begin{cases} \dfrac{\sum_{i=1}^{n} K_n(x-X_i)Y_i}{\sum_{i=1}^{n} K_n(x-X_i)}, & \text{if } \sum_{i=1}^{n} K_n(x-X_i)>0 \\ \dfrac{1}{n}\sum_{i=1}^{n} Y_i, & \text{otherwise.} \end{cases} \tag{2.7}$$

Under some conditions on kernel K, Devroye and Wagner (1980), and Spiegelman and Sachs (1980) proved that for (2.5) the kernel estimate (2.7) is weakly universally consistent in L_p for all $p\ge1$.

3. Strong universal consistency in L_p

Definition 2 m_n is called *strongly universally consistent* in L_p if

$$\lim_{n\to\infty} \int |m_n(x)-m(x)|^p\mu(dx) = 0 \tag{3.1}$$

a.s. for all Y with $E|Y|^p<\infty$ $(p\ge1)$.

Theorem 2 Assume that the weights are normal. Then the following statements are equivalent:

(i) m_n is strongly universally consistent in L_1.

(ii) m_n is strongly universally consistent in L_p for all $p\ge1$.

(iii) m_n is strong consistent in L_1 for all bounded $|Y|$ and there is a

C>0 such that for all $Y \in L_1$

$$\limsup_{n\to\infty} \sum_{i=1}^{n} \int W_{ni}(x)\mu(dx)|Y_i| \le C\, E(|Y|) \text{ a.s.} \tag{3.2}$$

Proof <u>(i) implies (iii)</u> The first half of (iii) is obvious. In order to prove the second half of (iii) for C=1, let $Y \in L_1$ be arbitrary and

$$\tilde{Y}=|Y|,$$

$$\tilde{m}(x)=E(\tilde{Y}/X),$$

$$\tilde{m}_n(x) = \sum_{i=1}^{n} W_{ni}(x)\tilde{Y}_i,$$

From condition (i) we have

$$\lim_{n\to\infty} \int|\tilde{m}_n(x)-\tilde{m}(x)|\mu(dx) = 0$$

a.s. therefore

$$\int\tilde{m}_n(x)\mu(dx)\to\int\tilde{m}(x)\mu(dx)$$

a.s., thus

$$\sum_{i=1}^{n} \int W_{ni}(x)\mu(dx)\tilde{Y}_i\to E(\tilde{Y}) \text{ a.s.}$$

Consequently

$$\sum_{i=1}^{n} \int W_{ni}(x)\mu(dx)|Y_i|\to E(|Y|) \text{ a.s.}$$

<u>(iii) implies (ii)</u> For an arbitrary L let

$$Y_{jL} = \begin{cases} Y_j & , \text{ if } |Y_j|\le L \\ L\,\mathrm{sign}(Y_j) & , \text{ otherwise} \end{cases} \tag{3.3}$$

and m_L and m_{nL} the functions m and m_n when $\{Y_j\}$ are replaced by $\{Y_{jL}\}$. Then by (iii) we have

$$\begin{aligned} &\limsup_{n\to\infty} \int|m_n(x)-m(x)|^p\mu(dx) \\ &\le 3^{p-1}\{\limsup_{n\to\infty} \int|m_n(x)-m_{nL}(x)|^p\mu(dx) \\ &+ \lim_{n\to\infty} \int|m_{nL}(x)-m_L(x)|^p\mu(dx) \\ &+ \int|m_L(x)-m(x)|^p\mu(dx)\}. \end{aligned} \tag{3.4}$$

By the first half of (iii)

$$\lim_{n\to\infty} \int |m_{nL}(x)-m_L(x)|^p \mu(dx)$$

$$\le (2L)^{p-1} \lim_{n\to\infty} \int |m_{nL}(x)-m_L(x)| \mu(dx)=0. \tag{3.5}$$

By Jensen's inequality

$$\int |m_L(x)-m(x)|^p \mu(dx)=E|E(Y_L/X)-E(Y/X)|^p \le E|Y-Y_L|^p. \tag{3.6}$$

Moreover $|Y-Y_L|^p \in L_1$, therefore by the second half of (iii)

$$\limsup_{n\to\infty} \int |m_n(x)-m_{nL}(x)|^p \mu(dx) \le \limsup_{n\to\infty} \sum_{i=1}^{n} \int W_{ni}(x)\mu(dx)|Y_i-Y_{iL}|^p$$

$$\le C\ (E|Y-Y_L|^p). \tag{3.7}$$

By (3.4)-(3.7) we have

$$\limsup_{n\to\infty} \int |m_n(x)-m(x)|^p \mu(dx) \le 3^{p-1}(C\ (E|Y-Y_L|)^p+E|Y-Y_L|^p)$$

a.s. Since L is arbitrary we are ready.

4. Partitioning estimate

Let $\mathcal{P}_n=\{A_{n1},A_{n2},\dots\}$ be a sequence of partitions of $\mathbb{R}^d$ and $A_n(x)$ be the set from $\mathcal{P}_n$ to which x belongs and let μ_n be the empirical measure for the sample $X_1,X_2,\dots X_n$. The partitioning estimate for this partition is

$$m_n(x)=\begin{cases} \dfrac{\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i\in A_n(x)\}} Y_i}{\mu_n(A_n(x))}, & \text{if } \mu_n(A_n(x))>0 \\ \frac{1}{n}\sum_{i=1}^{n} Y_i, & \text{otherwise.} \end{cases} \tag{4.1}$$

Devroye and Györfi (1985b) proved the following: assume that for each sphere S centered at the origin

$$\lim_{n\to\infty} \frac{1}{n} \#\{i; A_{ni}\cap S\neq\emptyset\} = 0 \tag{4.2}$$

and

$$\lim_{n\to\infty} \max_{\substack{i \\ A_{ni}\cap S\neq\emptyset}} \{ \sup_{x,y\in A_{ni}} \|x-y\| \} = 0. \tag{4.3}$$

If $|Y|\le L$, a.s. then for each $\varepsilon>0$ there is an n_0 such that for all $n>n_0$

$$\mathbb{P}(\int|m_n(x)-m(x)|\mu(dx)>\varepsilon)\le 2e^{-c\frac{\varepsilon^2}{L^2}n} \tag{4.4}$$

where c is a universal constant.

Theorem 3 Under (4.2) and (4.3) the partitioning estimate is weakly universally consistent for all $p\ge 1$.

Proof Let's verify (iii) of Theorem 1. If $|Y|\le L$, then (4.4) implies that $\int|m_n(x)-m(x)|\mu(dx)\to 0$ a.s. and in this case $\int|m_n(x)-m(x)|\mu(dx)$ is bounded by 2L, therefore $\mathbb{E}\left\{\int|m_n(x)-m(x)|\mu(dx)\right\}\to 0$, too. In order to prove the second half of (iii) with C=2L, we show that

$$\mathbb{E}\left\{\int W_{ni}(x)\mu(dx)|Y_i|\,|X_i,Y_i\right\}\le\frac{2}{n}|Y_i|, \tag{4.5}$$

(4.1) implies that

$$\int W_{ni}(x)\mu(dx) \le \frac{1}{n}+\frac{1}{n}\frac{\mu(A_n(X_i))}{\mu_n(A_n(X_i))},$$

therefore (4.5) is proved if

$$\mathbb{E}\left\{\frac{\mu(A_n(X_i))}{\mu_n(A_n(X_i))}|X_i,Y_i\right\}\le 1.$$

Apply the following simple inequality for the binomial (n-1,p) random variable Z:

$$\mathbb{E}\frac{1}{1+Z}=\sum_{m=0}^{n-1}\binom{n-1}{m}p^m(1-p)^{n-1-m}\frac{1}{1+m}=\frac{1}{np}\sum_{m=0}^{n-1}\binom{n}{m+1}p^{m+1}(1-p)^{n-(m+1)}$$
$$\le\frac{1}{np}.$$

Then

$$\mathbb{E}\left\{\frac{\mu(A_n(X_i))}{\mu_n(A_n(X_i))}|X_i,Y_i\right\}=\mathbb{E}\left\{\sum_j I_{\{X_i\in A_{nj}\}}\frac{\mu(A_{nj})}{\mu_n(A_{ni})}|X_i,Y_i\right\}$$

$$=\mathbb{E}\left\{\sum_j I_{\{X_i\in A_{nj}\}}n\mu(A_{nj})\frac{1}{1+\sum_{k\ne i}I_{\{X_k\in A_{nj}\}}}|X_i,Y_i\right\}$$

$$= \sum_j I_{\{X_i \in A_{nj}\}} n\mu(A_{nj}) \mathbb{E}\left\{ \frac{1}{1+\sum_{k\neq i} I_{\{X_k \in A_{nj}\}}} \right\}$$

$$\leq \sum_j I_{\{X_i \in A_{nj}\}} = 1.$$

Unfortunately we could prove the strong universal consistency only for a modified partitioning estimate:

$$m'_n(x) = \begin{cases} \dfrac{\frac{1}{n}\sum_{i=1}^n I_{\{X_i \in A_n(x)\}} Y_i}{\mu_n(A_n(x))}, & \text{if } \mu_n(A_n(x)) > \log n/n \\ \frac{1}{n}\sum_{i=1}^n Y_i, & \text{otherwise.} \end{cases} \tag{4.6}$$

Theorem 4 Assume (4.3) and that for each sphere S centered at the origin

$$\lim \frac{\log n}{n} \#\{i; A_{ni} \cap S \neq \emptyset\} = 0. \tag{4.7}$$

Then m'_n is strongly universally consistent in L_p for all $p \geq 1$.

Proof According to Theorem 2 we have to prove (iii). In order to prove its first half, let L be the bound of $|Y|$ and S a sphere centered at the origin. Introduce the notations

$$I_n = \{i;\ A_{ni} \cap S \neq \emptyset\}, \tag{4.8}$$

$$J_n = \{i;\ \mu_n(A_{ni}) \geq \log n/n\}, \tag{4.9}$$

$$L_n = \{i;\ \mu(A_{ni}) \geq c\log n/n\}, \tag{4.10}$$

where $c>1$. As we mentioned in the proof of Theorem 3, (4.2) and (4.3) imply that

$$\lim_{n\to\infty} \int |m(x)-m_n(x)|\mu(dx) = 0 \tag{4.11}$$

a.s. Moreover

$$\int |m_n(x)-m'_n(x)|\mu(dx) \leq 2L\mu\{x; m_n(x) \neq m'_n(x)\} \leq 2L(\mu(S^c)+\mu(A_n)), \tag{4.12}$$

where $A_n = \bigcup_{i \in I_n \cap J_n^c} A_{ni}$. Because of (4.11) and (4.12), the first half of (iii) is proved if we show that $\mu(A_n) \to 0$ a.s.

$$\mu(A_n) \le \sum_{i\in I_n\cap L_n^c} \mu(A_{ni}) + \sum_{i\in I_n\cap L_n\cap J_n^c} \mu(A_{ni})$$

$$\le c\frac{\log n}{n} \#\{i\in I_n\} + \sum_{i\in I_n\cap L_n} \mu(A_{ni}) I_{\{c\mu_n(A_{ni})<\mu(A_{ni})\}}. \tag{4.13}$$

The first term of the r.h.s. of (4.13) $\to 0$ by (4.7). For the second term we apply a lemma of Györfi and van der Meulen (1987): for each $c>1$ and set A

$$P\left(\frac{\mu(A)}{c} > \mu_n(A) \right) \le 2 \exp(-n\mu(A)(1-1/c)^2/4). \tag{4.14}$$

Thus for all $\varepsilon>0$

$$P\left[\sum_{i\in I_n\cap L_n} \mu(A_{ni}) I_{\{c\mu_n(A_{ni})<\mu(A_{ni})\}} > \varepsilon \right]$$

$$\le P\left[\sum_{i\in I_n\cap L_n} \mu(A_{ni}) I_{\{c\mu_n(A_{ni})<\mu(A_{ni})\}} > \varepsilon \sum_{i\in I_n\cap L_n} \mu(A_{ni}) \right]$$

$$\le \sum_{i\in I_n\cap L_n} P\left[I_{\{c\mu_n(A_{ni})<\mu(A_{ni})\}} > \varepsilon \right]$$

$$\le \sum_{i\in I_n\cap L_n} P\left[c\mu_n(A_{ni})<\mu(A_{ni}) \right]$$

$$\le \sum_{i\in I_n\cap L_n} 2 \exp(-n\mu(A_{ni})(1-1/c)^2/4)$$

$$\le 2 \sum_{i\in I_n} \exp\{-c\log n(1-1/c)^2/4\} = 2\#\{i; i\in I_n\} n^{-c(1-1/c)^2/4}. \tag{4.15}$$

Choose c such that

$$c(1-1/c)^2/4=4. \tag{4.16}$$

Then by (4.2) and (4.15) there is an n_o such that

$$P\left[\sum_{i\in I_n\cap L_n} \mu(A_{ni}) I_{\{c\mu_n(A_{ni})<\mu(A_{ni})\}} > \varepsilon \right] \le n^{-2}$$

for all $n>n_o$, therefore the second term of the r.h.s. of (4.13)$\to 0$ a.s. and and the first half of (iii) is proved. The second half of (iii) is proved if

$$\limsup_{n\to\infty} \{\max_j n\int W_{nj}(x)\mu(dx)\} \le C \text{ a.s.} \tag{4.17}$$

Let c be defined by (4.16) and

$$B_n = \bigcup_{i\in J_n\cap L_n} A_{ni}, \quad D_n = \bigcup_{i\in J_n} A_{ni}. \tag{4.18}$$

Then

$$nW_{nj}(x) = \begin{cases} \dfrac{I_{\{x\in A_n(X_j)\}}}{\mu_n(A_n(X_j))} & , \text{ if } \quad A_n(X_j)\subset D_n \\ 1 & , \text{ otherwise,} \end{cases}$$

therefore

$$n\int |W_{nj}(x)|\mu(dx) = \begin{cases} \dfrac{\mu(A_n(X_j))}{\mu_n(A_n(X_j))} & , \text{ if } \quad A_n(X_j)\subset D_n \\ 1 & , \text{ otherwise.} \end{cases}$$

Thus we are done with (4.17) for C=1+c if we show that

$$\sum_{n=1}^{\infty} P\left[\max_{\substack{1\le j\le n \\ A_n(X_j)\subset D_n}} \frac{\mu(A_n(X_j))}{\mu_n(A_n(X_j))} > c\right] < \infty. \tag{4.19}$$

Obviously

$$\left\{\frac{\mu(A_n(X_j))}{\mu_n(A_n(X_j))} > c,\ A_n(X_j)\subset D_n\right\} \subset \left\{A_n(X_j)\subset B_n\right\},$$

therefore by (4.14)

$$P\left[\max_{\substack{1\le j\le n \\ A_n(X_j)\subset D_n}} \frac{\mu(A_n(X_j))}{\mu_n(A_n(X_j))} > c\right] = P\left[\max_{\substack{1\le j\le n \\ A_n(X_j)\subset B_n}} \frac{\mu(A_n(X_j))}{\mu_n(A_n(X_j))} > c\right]$$

$$\le nP\left[A_n(X_1)\subset B_n, \frac{\mu(A_n(X_1))}{\mu_n(A_n(X_1))} > c\right] \le n\sum_{i\in L_n} P\left[X_1\in A_{ni}, \frac{\mu(A_{ni})}{\mu_n(A_{ni})} > c\right]$$

$$\le n\sum_{i\in L_n} \mu(A_{ni})P\left[c\mu_n(A_{ni})<\mu(A_{ni})\right]$$

$$\le 2n\sum_{i\in L_n} \mu(A_{ni})\exp(-n\mu(A_{ni})(1-1/c)^2/4)\le 2n^{-3}, \tag{4.20}$$

and Theorem 4 is proved.

Remark Unfortunately there is no known proof for the strong universal consistencies of the kernel and of the k_n-NN estimate. For the kernel estimate the first half of (iii) is proved by Devroye and Krzyzak

(1989), when they showed analog to (4.4) if Y is bounded and (2.5) is satisfied. For the k_n-NN there is no such result, it is only proved if X has a density (Devroye and Györfi (1985a), Theorem 10.5, and Zhao (1987)). If X is arbitrary and $k_n/\log n \to \infty$ then the first half of (iii) can be derived from the pointwise strong consistency of k_n-NN estimate for bounded Y (Devroye (1982)).

5. References

L. Devroye (1982) "Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates", *Z. Wahrscheinlichkeitstheorie*, **61**, pp. 467-481.

L. Devroye, L. Györfi (1985a) *Nonparametric Density Estimation: the L_1 View*, Wiley.

L. Devroye, L. Györfi (1985b) "Distribution free exponential bound for the L_1 error of partitioning estimates of a regression function" in *Probability and Statistical Decision Theory*, Proc. 4th Pannonian Symp. on Math. Stat., Bad Tatzmannsdorf, Austria, 1983, Konecny, J. Mogyoródi, W. Wertz, Eds., D. Reidel, Dordrecht.

L. Devroye, A. Krzyzak (1989) "An equivalence theorem for L_1 convergence of kernel regression estimate" *J. Statistical Planning and Inference*, **23**, pp. 71-82

L. Devroye, T. J. Wagner (1980) "Distribution-free consistency results in nonparametric discrimination and regression function estimation", *Annals of Statistics*, **8**, pp. 231-239.

L. Györfi, E. C. van der Meulen (1987) "Density-free convergence properties of various estimators of entropy" *Computational Statistics & Data Analysis*, *5*, pp.425-436.

C. Spiegelman, J. Sacks (1980) " Consistent window estimation in nonparametric regression", *Annals of Statistics*, **8**, pp. 240-246.

C. J. Stone (1977) "Consistent nonparametric regression" *Annals of Statistics*, **5**, pp. 1348-1360.

L. C. Zhao (1987) "Exponential bounds of mean error for the nearest neighbor estimates of regression functions" *J. Multivariate Analysis*, **21**, pp.168-178.

MINIMAX BAYES ESTIMATION, PENALIZED LIKELIHOOD METHODS, AND RESTRICTED MINIMAX ESTIMATION

Nancy E. Heckman
The University of British Columbia
Department of Statistics
Vancouver, B.C.
Canada

ABSTRACT. Suppose, based on n data points, one wishes to estimate an n–dimensional vector $\underset{\sim}{\theta}$ (e.g. one wishes to estimate $g(\underset{\sim}{t}_i)$, $\underset{\sim}{t}_i$ distinct and fixed, based on observations Y_i with $\mathrm{E}(Y_i|g) = g(\underset{\sim}{t}_i)$). The minimax restricted Bayes estimate of $\underset{\sim}{\theta}$ minimizes the maximum Bayes risk, $\mathrm{E}\|\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}\|^2$, where the minimum is taken over all affine estimators and the maximum is taken over a specified class of priors on $\underset{\sim}{\theta}$. The definition of the class of priors is motivated by an assumption of smoothness in the regression problem. The estimators are related to penalized likelihood and restricted minimax estimators.

1. Introduction.

The use of Bayes methods in estimating a regression function, g, from observations $(\underset{\sim}{t}_1, Y_1), \ldots, (\underset{\sim}{t}_n, Y_n)$, where $\mathrm{E}(Y_i|g) = g(\underset{\sim}{t}_i)$, is not new.

O'Hagan (1978) provides Bayes estimates of g using both proper and vague priors. Weerahandi and Zidek (1985) consider empirical Bayes estimators which may perform well over a large class of priors. They use a Taylor series expansion of g to justify a particular form of the prior. The form depends upon a few hyperparameters which can be estimated from the data.

A minimax Bayes estimate of $\underset{\sim}{g} = (g(\underset{\sim}{t}_1), \ldots, g(\underset{\sim}{t}_n))^t$ is one that minimizes the maximum Bayes risk under squared error loss. The maximum is taken over Ω, a large class of priors, and thus, hopefully, one avoids the problem of misspecification of the prior. Minimax Bayes estimates have been proposed in estimating an unknown parameter vector $\underset{\sim}{\theta}$: a specific class of priors has been studied by Leamer (1982), Polasek (1985), and DasGupta and Studden (1989), among others. Gupta and Hwang (1977) have applied the minimax Bayes principal to ranking and selection, where the procedure is called Γ-minimax.

The types of Ω's considered here are ones that would be appropriate for the regression problem, assuming that g is smooth. Consider the following example. For simplicity, suppose that g is a univariate function and one believes that g is not far from constant. Then one could define Ω to be all priors on g for which $\mathrm{E}(g(t_{i+1}) - g(t_i)) = 0$ and $\sum b_i^2 \operatorname{var}(g(t_{i+1}) - g(t_i)) \leq \rho$ for some specified ρ and non-zero weights b_i. This prior that places restrictions on the first divided differences of g is easily extended to one that places restrictions on higher order divided differences, and to priors on multivariate functions, through generalized divided differences, discussed in Section 4. The use of a kth divided difference prior may be thought of as a softening of the approximately linear model of Legostaeva and Shiryayev (1971) and Sacks

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 339–346.
© 1991 *Kluwer Academic Publishers.*

and Ylvisaker (1978). The approximately linear model estimate of $g(t)$ minimizes the maximum mean squared error where the maximum is taken over all g with $|g(t+\delta) - \sum_0^{k-1} g^{(i)}(t)\delta^i/i!| \leq C|\delta^k|$, for C specified by the user.

The general form of Ω is given in Section 2. Affine estimators of $\underset{\sim}{g}$ are considered, that is, estimators of the form $C\underset{\sim}{Y} + \underset{\sim}{c}$. The minimax Bayes estimator is calculated and it is shown that the estimator is Bayes.

In Section 3, the relationship between penalized likelihood and minimax Bayes estimates is discussed. Let g be in a Hilbert space, $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$, P_1 be a projection operator with k-dimensional nullspace, and let $\hat{g}$ minimize

$$\|\underset{\sim}{Y} - \underset{\sim}{g}\|^2 + \lambda\|P_1 g\|_{\mathcal{H}}^2.$$

Then, under certain conditions, there exist $\psi_j \in \mathcal{H}$ and $\hat{\beta}_j$ real such that $\hat{g} = \sum_1^n \hat{\beta}_j\psi_j$ and $\hat{g}$ is Bayes, assuming $g = \sum_1^n \beta_j\psi_j$, the ψ_j's non-random, the Y_j's normal, the β_j's normal and uncorrelated, the prior diffuse on $\beta_1, \ldots, \beta_k$, and $\mathrm{E}(\beta_i) = 0$, $i > k$. If one uses P_1 to define a generalized divided difference in a natural way, and defines Ω in terms of this generalized divided difference, then the resulting minimax Bayes estimate is also a linear combination of the ψ_j's. The estimate is Bayes under a prior which differs from the above only in the values of the variances of the β_j's. These facts are discussed in Section 3.

There is a correspondence between restricted minimax estimates and minimax Bayes estimates. Specifically, the restricted minimax estimator minimizes the maximum $\mathrm{E}\|\underset{\sim}{Y} - \underset{\sim}{g}\|^2$, where the expectation is taken with respect to the true function g and the maximum is taken over all g with $\|P_1 g\|_{\mathcal{H}}^2 \leq \rho$. For the Ω generated by P_1 (see Section 3), the minimax Bayes and the restricted minimax estimate are identical. Speckman (1985) and Nussbaum (1985) calculated the restricted minimax estimator when $\|P_1 g\|_{\mathcal{H}}^2 = \int \left(g^{(k)}(t)\right)^2 dt$. Thus, their estimator is minimax Bayes, and hence Bayes. Section 5 contains these results.

2. The Minimax Bayes Estimate.

Let θ_0 and $\underset{\sim}{Y}$ be n–vectors, ρ a non-negative real number and A an Lxn matrix, $L \leq n$, of rank $n - k$, $0 \leq k < n$. $\underset{\sim}{\theta}_0$, ρ, and A are specified and non-random, $\underset{\sim}{Y}$ is observed. Assume that $\underset{\sim}{\theta}$ and $\underset{\sim}{\epsilon}$ are random n-vectors and that

$$\underset{\sim}{Y} = \underset{\sim}{\theta} + \underset{\sim}{\epsilon}, \tag{1a}$$

$$\mathrm{E}(\underset{\sim}{\epsilon}) = \underset{\sim}{0}, \tag{1b}$$

$$\mathrm{cov}(\underset{\sim}{\epsilon}) = WW^t, \text{ diagonal and positive definite, } W \text{ invertible,} \tag{1c}$$

$$\left.\begin{array}{ll} \mathrm{E}(A\underset{\sim}{\theta}) = A\underset{\sim}{\theta}_0, & \text{trace } \mathrm{cov}(A\underset{\sim}{\theta}) \leq \rho, \\ \mathrm{cov}(\underset{\sim}{\theta}) \text{ defined,} & \mathrm{cov}(\underset{\sim}{\theta}, \underset{\sim}{\epsilon}) = 0. \end{array}\right\} \tag{1d}$$

Ω is the set of all priors on $\underset{\sim}{\theta}$ satisfying (1d).

Let the principal value decomposition of AW be PDQ_2^t, where P is $L \times n-k$, D is diagonal with $0 < D_{11}^2 \leq \cdots \leq D_{n-k,n-k}^2$, and $P^tP = I_{n-k} = Q_2^tQ_2$. Let Q_1 be $n \times k$, with $Q_1^tQ_2 = 0$ and $Q_1^tQ_1 = I_k$. If $k = 0$, then Q_1 is taken to be the zero matrix in all expressions.

Theorem 1.

$$\min_{C,\underset{\sim}{c}} \max_{\Omega} \mathrm{E}\|W^{-1}(C\underset{\sim}{Y} + \underset{\sim}{c} - \underset{\sim}{\theta})\|^2 = k + \sum_{i=1}^{n-k} s_{ii}/(s_{ii}+1) \tag{2}$$

where

$$s_{ii} = s_{ii}(\rho) = (\alpha/|D_{ii}| - 1)_+,$$

x_+ is the maximum of 0 and x, and α is chosen so that $\sum_1^{n-k} D_{ii}^2 s_{ii} = \rho$. The minimax estimator is

$$\hat{\underset{\sim}{\theta}} = WQ_1Q_1^tW^{-1}\underset{\sim}{Y} + WQ_2\hat{C}^*Q_2^tW^{-1}\underset{\sim}{Y} + WQ_2(I_{n-k} - \hat{C}^*)Q_2^tW^{-1}\underset{\sim}{\theta}_0 \tag{3}$$

where $\hat{C}^*$ is diagonal with

$$\hat{C}_{ii}^* = \hat{C}_{ii}^*(\rho) = s_{ii}/(s_{ii}+1) = (1 - |D_{ii}|/\alpha)_+.$$

Furthermore, this estimate is related to the following Bayes estimator. Suppose that $\underset{\sim}{\theta}$ and $\underset{\sim}{\epsilon}$ are independent and jointly normal. Let $\underset{\sim}{\theta}_1^* = Q_1^tW^{-1}\underset{\sim}{\theta}$ and $\underset{\sim}{\theta}_2^* = Q_2^tW^{-1}\underset{\sim}{\theta}$ be uncorrelated, $\mathrm{E}(\underset{\sim}{\theta}_2^*) = Q_2^tW^{-1}\underset{\sim}{\theta}_0$, and the covariance of $\underset{\sim}{\theta}_2^* = \mathrm{diag}(s_{11}, \ldots, s_{n-k,n-k})$. Then $\hat{C}^*Q_2^tW^{-1}\underset{\sim}{Y} + (I_{n-k} - \hat{C}^*)Q_2^tW^{-1}\underset{\sim}{\theta}_0$ is the Bayes estimate of $\underset{\sim}{\theta}_2^*$ under squared error loss, and $Q_1^tW^{-1}\underset{\sim}{Y}$ is the unbiased estimate of $\underset{\sim}{\theta}_1^*$. Combining these yields the estimate of $W^{-1}\underset{\sim}{\theta} = Q_1\underset{\sim}{\theta}_1^* + Q_2\underset{\sim}{\theta}_2^*$ given in (3). Thus, the minimax Bayes estimate of $\underset{\sim}{\theta}$ is the Bayes estimate with the above prior on $\underset{\sim}{\theta}_2^*$ and a diffuse prior on $\underset{\sim}{\theta}_1^*$.

Proof. The proof of the Theorem follows the argument in Heckman and Woodroofe (1989), for the case of linear estimators ($\underset{\sim}{c} = \underset{\sim}{0}$) and $\underset{\sim}{\theta}_0 = \underset{\sim}{0}$. Let $A_1 = AW$, $\underset{\sim}{Y}_1 = W^{-1}(\underset{\sim}{Y} - \underset{\sim}{\theta}_0)$, $\underset{\sim}{\theta}_1 = W^{-1}(\underset{\sim}{\theta} - \underset{\sim}{\theta}_0)$, $\underset{\sim}{\theta}_{01} = \underset{\sim}{0}$, and $\underset{\sim}{\epsilon}_1 = W^{-1}\underset{\sim}{\epsilon}$. Then (1a), (1b), and (1d) hold for the subscripted quantities, (1c) becomes $\mathrm{cov}(\underset{\sim}{\epsilon}_1) = I_n$, and we must now solve

$$\min_{C_1, \underset{\sim}{c}_1} \max_{\Omega_1} E\|C_1\underset{\sim}{Y}_1 + \underset{\sim}{c}_1 - \underset{\sim}{\theta}_1\|^2$$

where $C_1 = W^{-1}CW$, $\underset{\sim}{c}_1 = W^{-1}(C\underset{\sim}{\theta}_0 - \underset{\sim}{\theta}_0 + \underset{\sim}{c})$, and Ω_1 is the set of priors on $\underset{\sim}{\theta}_1$ with $\mathrm{E}(Q_2^t\underset{\sim}{\theta}_1) = \underset{\sim}{0}$ and trace $D^2\ \mathrm{cov}(Q_2^t\underset{\sim}{\theta}_1) \le \rho$. Let $P_1 = Q_1Q_1^t$ and $P_2 = Q_2Q_2^t$ so that $I_n = P_1 + P_2$. First, it will be shown that the minimizing C_1 and $\underset{\sim}{c}_1$ must satisfy $C_1P_1 = P_1$, $\underset{\sim}{c}_1 = \underset{\sim}{0}$, and $P_1C_1P_2 = 0$. Since $\mathrm{E}(P_2\underset{\sim}{\theta}_1) = \underset{\sim}{0}$,

$$\mathrm{E}\|C_1\underset{\sim}{Y}_1 + \underset{\sim}{c}_1 - \underset{\sim}{\theta}_1\|^2 = \mathrm{E}\|C_1(\underset{\sim}{Y}_1 - \underset{\sim}{\theta}_1)\|^2 + \mathrm{E}\|(C_1 - I_n)(\underset{\sim}{\theta}_1 - \mathrm{E}\underset{\sim}{\theta}_1)\|^2 + \|(C_1 - I_n)P_1\mathrm{E}\underset{\sim}{\theta}_1 + \underset{\sim}{c}_1\|^2.$$

Since $P_1\mathrm{E}(\underset{\sim}{\theta}_1)$ is unrestricted, its norm can be arbitrarily large. Thus for any fixed C_1 and $\underset{\sim}{c}_1$, the maximum over Ω_1 of $\mathrm{E}\|C_1\underset{\sim}{Y}_1 + \underset{\sim}{c}_1 - \underset{\sim}{\theta}_1\|^2$ will be infinite unless $(C_1 - I_n)P_1 = 0$. We also see that, to minimize the third term in the above, we should take $\underset{\sim}{c}_1 = \underset{\sim}{0}$.

The problem now reduces to one solved in Heckman and Woodroofe (1989). A sketch of the proof, including some heuristics, is given here.

Since $P_1\underset{\sim}{Y}_1$ and $P_2\underset{\sim}{Y}_1$ are uncorrelated, given $\underset{\sim}{\theta}_1$, and $C_1P_1 = P_1$ and $P_1P_2 = 0$,

$$\begin{aligned}\mathrm{E}\|C_1\underset{\sim}{Y}_1 - \underset{\sim}{\theta}_1\|^2 &= \mathrm{E}\|C_1P_1\underset{\sim}{Y}_1 - P_1\underset{\sim}{\theta}_1 + P_1C_1P_2\underset{\sim}{Y}_1 + P_2C_1P_2\underset{\sim}{Y}_1 - P_2\underset{\sim}{\theta}_1\|^2 \\ &= \mathrm{E}\|P_1(\underset{\sim}{Y}_1 - \underset{\sim}{\theta}_1)\|^2 + \mathrm{E}\|P_1C_1P_2\underset{\sim}{Y}_1\|^2 + \mathrm{E}\|P_2C_1P_2\underset{\sim}{Y}_1 - P_2\underset{\sim}{\theta}_1\|^2.\end{aligned}$$

The maximum of $\mathrm{E}\|P_1C_1P_2\underset{\sim}{Y}_1\|^2$ will always be strictly positive unless $P_1C_1P_2 = 0$. Therefore

$$\hat{\underset{\sim}{\theta}}_1 = C_1\underset{\sim}{Y}_1 = P_1\underset{\sim}{Y}_1 + P_2C_1P_2\underset{\sim}{Y}_1.$$

Let $C^* = Q_2^tC_1Q_2$, $\underset{\sim}{\theta}^* = Q_2^t\underset{\sim}{\theta}_1$, and $\underset{\sim}{Y}^* = Q_2^t\underset{\sim}{Y}_1$. Then, given $\underset{\sim}{\theta}^*$, $\underset{\sim}{Y}^*$ has mean $\underset{\sim}{\theta}^*$ and covariance I_{n-k}. The prior on $\underset{\sim}{\theta}^*$ is in Ω^*, the collection of all priors with $\mathrm{E}(\underset{\sim}{\theta}^*) = \underset{\sim}{0}$ and trace $D^2\ \mathrm{cov}(\underset{\sim}{\theta}^*) \le \rho$. Then (2) becomes

$$k + \min_{C_1} \max_{\Omega_1} \mathrm{E}\|P_2C_1P_2\underset{\sim}{Y}_1 - P_2\underset{\sim}{\theta}_1\|^2$$

$$= k + \min_{C^*} \max_{\Omega^*} \mathrm{E}\|C^* \underset{\sim}{Y}^* - \underset{\sim}{\theta}^*\|^2.$$

Let Σ^* denote the covariance of $\underset{\sim}{\theta}^*$. Then $\mathrm{E}\|C^* \underset{\sim}{Y}^* - \underset{\sim}{\theta}^*\|^2 = \text{trace } C^{*t} C^* + \text{trace } (C^* - I_{n-k})^t (C^* - I_{n-k})\Sigma^*$ is quadratic in C^* and linear in in Σ^*, and Σ^* lies in a convex compact set. It can be shown that, for any fixed Σ^*, the C^* that minimizes $\mathrm{E}\|C^* \underset{\sim}{Y}^* - \underset{\sim}{\theta}^*\|^2$ lies in a compact set, not depending on Σ^*. Thus we can use an argument involving the general minimax theorem (see e.g. Karlin, 1959, Theorem 1.5.1) to show that the min and max can be reversed.

For priors in Ω^*, $\mathrm{E}\|C^* \underset{\sim}{Y}^* - \underset{\sim}{\theta}^*\|^2$ depends only upon second moments, so we may assume without loss of generality that $\underset{\sim}{Y}^*$ and $\underset{\sim}{\theta}^*$ are jointly normal. For fixed Σ^*, the C^* that minimizes $\mathrm{E}\|C^* \underset{\sim}{Y} - \underset{\sim}{\theta}^*\|^2$ is found from the Bayes estimate:

$$C^* = \Sigma^*(\Sigma^* + I_{n-k})^{-1}.$$

Thus

$$\max_{\Sigma^*} \min_{C^*} \mathrm{E}\|C^* \underset{\sim}{Y}^* - \underset{\sim}{\theta}^*\|^2 = \max_{\Sigma^*} \text{trace} \Sigma^*(\Sigma^* + I_{n-k})^{-1},$$

where the maximum is taken over all nonnegative definite symmetric Σ^* with trace $D^2\Sigma^* \leq \rho$. A heuristic calculation of the maximizer is given here. Ignoring the non-negative definite constraint on Σ^*, assuming that trace $D^2\Sigma^* = \rho$, and using the method of Lagrange multipliers, we find that the minimizing Σ^* must satisfy

$$\lim_{\delta \to 0} \delta^{-1} \text{trace}\{(\Sigma^* + \delta R)(\Sigma^* + \delta R + I_{n-k})^{-1} - \Sigma^*(\Sigma^* + I_{n-k})^{-1}\} - \alpha^{-2} \text{trace } D^2 R$$

$$= \text{trace } (\Sigma^* + I_{n-k})^{-2} R - \alpha^{-2} \text{trace } D^2 R = 0$$

for all R. Therefore

$$\hat{\Sigma}^* = \alpha \text{ diag}(|D_{11}|^{-1}, \ldots, |D_{n-k,n-k}|^{-1}) - I_{n-k}.$$

To force $\hat{\Sigma}^*$ to be non-negative definite, change all negative entries of Σ^* to 0, yielding $\hat{\Sigma}^*_{ii} = s_{ii}$ and $\hat{C}^*$ diagonal with $\hat{C}^*_{ii} = (1 - |D_{ii}/\alpha)_+$.

Combining the calculations yields

$$W^{-1}\hat{\underset{\sim}{\theta}} = P_1 \underset{\sim}{Y}_1 + Q_2 \hat{C}^* Q_2^t \underset{\sim}{Y}_1 + W^{-1}\underset{\sim}{\theta}_0$$

$$= Q_1 Q_1^t W^{-1} \underset{\sim}{Y} + Q_2 \hat{C}^* Q_2^t W^{-1} \underset{\sim}{Y} - Q_1 Q_1^t \underset{\sim}{\theta}_0 - Q_2 \hat{C}^* Q_2^t W^{-1} \underset{\sim}{\theta}_0 + W^{-1}\underset{\sim}{\theta}_0.$$

Corollary. Suppose that (1) holds with $W = A = I_n$. Then

$$\min_{C, \underset{\sim}{c}} \max_{\Omega} \mathrm{E}\|C \underset{\sim}{Y} + \underset{\sim}{c} - \underset{\sim}{\theta}\|^2 = \frac{n\rho}{\rho + n}$$

and the minimax Bayes estimate of $\underset{\sim}{\theta}$ is

$$\hat{\underset{\sim}{\theta}} = \frac{\rho}{\rho + n} \underset{\sim}{Y} + \frac{n}{\rho + n} \underset{\sim}{\theta}_0,$$

the Bayes estimate of $\underset{\sim}{\theta}$, assuming normality, with $\mathrm{E}(\underset{\sim}{\theta}) = \underset{\sim}{\theta}_0$ and $\text{cov}(\underset{\sim}{\theta}) = \rho n^{-1} I_n$.

Theorem 2. Suppose that (1a) - (1c) hold and that, for given $\underset{\sim}{\theta}_0$ and ρ, Ω consists of all priors on $\underset{\sim}{\theta}$ such that $\text{cov}(\underset{\sim}{\theta})$ is defined, $\text{cov}(\underset{\sim}{\theta}, \underset{\sim}{\epsilon}) = 0$, and $\mathrm{E}\|A(\underset{\sim}{\theta} - \underset{\sim}{\theta}_0)\|^2 \leq \rho$. Then (2) and (3) hold.

Proof. The proof is similar to that of Theorem 1.

Comment. Often, one restricts oneself to linear estimators of $\underset{\sim}{\theta}$, that is estimates of the form $C\underset{\sim}{Y}$. In the minimax Bayes formulation, this restriction leads to computational difficulties, except in the case that $\underset{\sim}{\theta}_0 = \underset{\sim}{0}$ (and thus the affine minimax Bayes estimator is linear) and in a few other restrictive cases. Specifically, assume that (1) holds, with $W = I_n$ for simplicity. Following the argument and notation in the proof of Theorem 1, one can show that

$$\hat{C} = Q_1 Q_1^t + Q_2 \hat{C}^* Q_2^t$$

where

$$\hat{C}^* = (\hat{\Sigma}^* + \underset{\sim}{\theta}_0^* \underset{\sim}{\theta}_0^{*t})(I_{n-k} + \hat{\Sigma}^* + \underset{\sim}{\theta}_0^* \underset{\sim}{\theta}_0^{*t})^{-1},$$
$$\underset{\sim}{\theta}_0^* = Q_2^t \underset{\sim}{\theta}_0,$$

and $\hat{\Sigma}^*$ is the non-negative definite symmtric matrix which maximizes

$$\operatorname{trace}(\Sigma^* + \underset{\sim}{\theta}_0^* \underset{\sim}{\theta}_0^{*t})(I_{n-k} + \Sigma^* + \underset{\sim}{\theta}_0^* \underset{\sim}{\theta}_0^{*t})^{-1}$$

subject to trace $D^2 \Sigma^* \leq \rho$. Ignoring the non-negativity constraint and using the heuristic Lagrangian argument to calculate the maximizing Σ^*, one finds that

$$\hat{\Sigma}^* = \alpha \operatorname{diag}(|D_{11}|^{-1}, \ldots |D_{n-k,n-k}|^{-1}) - I_{n-k} - \underset{\sim}{\theta}_0^* \underset{\sim}{\theta}_0^{*t}.$$

If $\hat{\Sigma}^*$ is non-negative definite, we have found the solution.

If $\hat{\Sigma}^*$ is not non-negative definite and $\underset{\sim}{\theta}_0 \neq \underset{\sim}{0}$, it is not clear how to find the maximizing Σ^*, except in the simple case that $D = I_{n-k}$. Write $\underset{\sim}{\theta}_0^* \underset{\sim}{\theta}_0^{*t} = V \Gamma V^t$, with V orthonormal with first column equal to $\underset{\sim}{\theta}_0^* / \|\underset{\sim}{\theta}_0^*\|$, and Γ's elements 0 except $\Gamma_{11} = \|\underset{\sim}{\theta}_0^*\|^2$. Then the maximizing Σ^* is $V \hat{S} V^t$ where $\hat{S}$ is diagonal with ii^{th} entry equal to $(\alpha - 1 - \Gamma_{ii})_+$,

$$\hat{\underset{\sim}{\theta}} = Q_1 Q_1^t \underset{\sim}{Y} + Q_2 V B V^t Q_2^t \underset{\sim}{Y}$$

and

$$\min \max \mathrm{E} \|C\underset{\sim}{Y} - \underset{\sim}{\theta}\|^2 = k + \operatorname{trace} B,$$

where

$$B = (\hat{S} + \Gamma)(I_{n-k} + \hat{S} + \Gamma)^{-1}.$$

If $\rho > (n-k-1)\|\underset{\sim}{\theta}_0^*\|^2$ then

$$B = \frac{\rho + \|\underset{\sim}{\theta}_0^*\|^2}{n - k + \rho + \|\underset{\sim}{\theta}_0^*\|^2} I_{n-k}.$$

If $\rho \leq (n-k-1)\|\underset{\sim}{\theta}_0^*\|^2$ then

$$B_{11} = \frac{\|\underset{\sim}{\theta}_0^*\|^2}{1 + \|\underset{\sim}{\theta}_0^*\|^2},$$

and

$$B_{ii} = \frac{\rho}{n - k - 1 + \rho} \text{ for } i \geq 2.$$

3. Penalized Likelihood and Minimax Bayes Estimates.

Let $\mathcal{H}$ be a Hilbert space over the reals, L_i, $1 \le i \le n$, continuous linear functionals on $\mathcal{H}$, $\eta_i \in \mathcal{H}$ with $L_i(g) =< \eta_i, g >$ for all g, and $\underset{\sim}{L}(g) = (L_1(g), \cdots, L_n(g))^t$. Fix g_0 in $\mathcal{H}$ and consider

$$\min_{g \in \mathcal{H}} \|\underset{\sim}{Y} - \underset{\sim}{L}(g)\|^2 + \lambda \|P_1(g - g_0)\|^2_{\mathcal{H}} \tag{4}$$

where P_1 is a projection with k-dimensional nullspace, $k < n$, $\|\cdot\|$ denotes Euclidean norm and $\|\cdot\|_{\mathcal{H}}$ denotes the norm in $\mathcal{H}$. Let $\phi_1, \ldots \phi_k$ span the nullspace of P_1, $\xi_i = P_1(\eta_i)$, and $T_{ij} = L_i(\phi_j)$. If T is of full rank and if the ξ_i's are linearly independent then the unique minimizer, $\hat{g}$, exists and is of the form

$$g = \sum_1^k c_\nu \phi_\nu + \sum_1^n d_\nu \xi_\nu + g_0$$

where $(d_1, \ldots, d_n)T = \underset{\sim}{0}^t$. Thus $\hat{g} - g_0$ is in an n-dimensional subspace of $\mathcal{H}$. See, for instance, Wahba (1990).

A convenient basis for the solution is the Demmler-Reinsch basis (Demmler and Reinsch, 1975). Under the above conditions on T and P_1, there exists a basis $\psi_1, \ldots, \psi_n$ such that $\hat{g} - g_0 = \sum \hat{\beta}_i \psi_i$, $\psi_1, \ldots, \psi_k$ span the nullspace of P_1, the matrix Q with $Q_{ij} = L_i(\psi_j)$ is orthonormal, $< P_1\psi_i, P_1\psi_j > = \lambda_i \delta_{ij}$ where δ_{ij} is Kronecker's delta, and $0 = \lambda_1 = \cdots = \lambda_k < \lambda_{k+1} \le \cdots \le \lambda_n$. Writing $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $\underset{\sim}{Y}^* = \underset{\sim}{Y} - \underset{\sim}{L}(g_0)$, (4) becomes

$$\min_{\underset{\sim}{\beta}} \|\underset{\sim}{Y}^* - Q\underset{\sim}{\beta}\|^2 + \lambda \underset{\sim}{\beta}^t \Lambda \underset{\sim}{\beta}.$$

The solution is easily found to be

$$\hat{\underset{\sim}{\beta}} = (I_n + \lambda\Lambda)^{-1} Q^t \underset{\sim}{Y}^*.$$

Let $Q = [T : Q_2]$ and $\Lambda^* = \mathrm{diag}(\lambda_{k+1}, \ldots, \lambda_n)$. Then

$$\begin{aligned} \underset{\sim}{L}(\hat{g}) &= Q(I_n + \lambda\Lambda)^{-1} Q^t \underset{\sim}{Y}^* + \underset{\sim}{L}(g_0) \\ &= TT^t \underset{\sim}{Y} + Q_2(I_{n-k} + \lambda\Lambda^*)^{-1} Q_2^t \underset{\sim}{Y} + Q_2(I_{n-k} - (I_{n-k} + \lambda\Lambda^*)^{-1}) Q_2^t \underset{\sim}{L}(g_0). \end{aligned}$$

Furthermore, $\underset{\sim}{L}(\hat{g})$ is the Bayes estimate of $\underset{\sim}{L}(g)$ assuming that, given g, $\underset{\sim}{Y}$ is normal with mean $\underset{\sim}{L}(g)$ and covariance I_{n-k}, $g = \sum \beta_j \psi_j + g_0$, the β_j's are normal and uncorrelated, with a diffuse prior on $\beta_1, \ldots, \beta_k$, and, for $i > k$, $\mathrm{E}(\beta_i) = 0$ and $\mathrm{var}(\beta_i) = (\lambda\lambda_i)^{-1}$.

The penalized likelihood estimate is very similar to $\hat{\underset{\sim}{L}}_{MB}$, the minimax Bayes estimate found by applying Theorem 1 with $\underset{\sim}{\theta} = \underset{\sim}{L}(g)$, $A = \Lambda^{*1/2} Q_2^t$, $\underset{\sim}{\theta}_0 = \underset{\sim}{L}(g_0)$, and $W = I_n$:

$$\hat{\underset{\sim}{L}}_{MB} = TT^t \underset{\sim}{Y} + Q_2 \hat{C}^* Q_2^t \underset{\sim}{Y} + Q_2(I_{n-k} - \hat{C}^*) Q_2^t \underset{\sim}{L}(g_0)$$

where $\hat{C}^*_{ii} = (1 - \lambda_{k+i}^{1/2}/\alpha)_+$. $\hat{\underset{\sim}{L}}_{MB}$ is the Bayes estimate of $\underset{\sim}{L}(g)$ under the same prior as above, except that, for $i > k$, $\mathrm{var}(\beta_i) = (\alpha\lambda_i^{-1/2} - 1)_+$. Note that $\mathrm{var}(\beta_{k+1}) \ge \cdots \ge \mathrm{var}(\beta_n)$ for both the penalized likelihood and minimax Bayes estimates'priors.

The Demmler-Reinsch basis provides some insight into our prior. Consider $\mathcal{H}$ the Sobolev space of order k on $[0,1]$ (see, e.g. Wahba,1990), with $L_i(g) = g(t_i)$, and $(P_1 g)(t) = g(t) - \sum_0^{k-1} g^{(i)}(0) t^i / i!$. For $g - g_0 = \sum_1^n \beta_j \psi_j$,

$$\begin{aligned} \int (g^{(k)}(t) - g_0^{(k)}(t))^2 dt = \|P_1(g - g_0)\|^2_{\mathcal{H}} &= \underset{\sim}{\beta}^t \Lambda \underset{\sim}{\beta} \\ &= (Q\underset{\sim}{\beta})^t Q\Lambda Q^t (Q\underset{\sim}{\beta}) = (\underset{\sim}{g} - \underset{\sim}{g}_0)^t Q\Lambda Q^t (\underset{\sim}{g} - \underset{\sim}{g}_0) \\ &= \|A(\underset{\sim}{g} - \underset{\sim}{g}_0)\|^2. \end{aligned}$$

4. Generalized Divided Differences.

In minimax Bayes estimation of a regression function, it is natural to define a class of priors which restrict k^{th} order divided differences. If g is a function of a real variable, then the $n-k$ vector of kth divided differences of g, based on $t_1, \ldots, t_n$, is naturally defined and is equal to $A\underset{\sim}{g}$ for some A of full rank. One could then define Ω to be all priors on g with $\mathrm{E}(A\underset{\sim}{g}) = \underset{\sim}{0}$ and trace $\mathrm{cov}(BA) \leq \rho$, for B a diagonal matrix of positive weights. The matrix A has a property which indicates how one might define k^{th} divided differences of a function of more than one variable: $AT = 0$ for $T_{ij} = \phi_j(t_i)$, where $\phi_1, \ldots, \phi_k$ span the space of kth order polynomials.

Using the notation of Section 3, a generalized k^{th} divided difference associated with the linear functionals $L_1, \ldots, L_n$ and the projection operator P_1 with k- dimensional nullspace is any n-vector, $\underset{\sim}{q}$, such that $T^t\underset{\sim}{q} = \underset{\sim}{0}$, where $T_{ij} = L_i(\phi_j)$ and $\phi_1, \ldots, \phi_k$ span the null space of P_1. Suppose that T is of full rank k. Then there exists a generalized k^{th} divided difference matrix, A, of rank $n-k$ with $AT = 0$.

We can thus easily define k^{th} order divided differences for g, a function of d variables, and $L_i(g) = g(\underset{\sim}{t}_i)$, $i = 1, \ldots, n$, by considering P_1 with null space equal to all k^{th} order polynomials of d variables. For instance, if $k = 2$ and $d = 2$, we can take $\phi_1(x,y) \equiv 1$, $\phi_2(x,y) = x$, and $\phi_3(x,y) = y$. For technical details concerning the Hilbert space structure of the problem, see Wahba (1990) and references therein.

Of course for a given $P_1, L_1, \ldots, L_n$, there are an infinite number of choices of A. However, it is easily seen that Q_2^t, generated by the Demmler-Reinsch basis, as defined in Section 3, is a generalized divided difference matrix: for that basis, $T = Q_1$ and one immediately sees that $Q_2^t T = 0$. Therefore, the minimax Bayes estimate associated with the penalized likelihood estimate for $P_1, L_1, \ldots, L_n$ arises from an Ω defined in terms of generalized divided differences.

5. Restricted Minimax Estimates.

Let $\mathcal{H}$, P_1, and $L_1, \ldots, L_n$ be as in Section 3, and let Q, T, Q_2, and Λ^* be the matrices generated by the Demmler-Reinsch basis. Suppose that $\underset{\sim}{Y} = \underset{\sim}{L}(g) + \underset{\sim}{\epsilon}$ with $\mathrm{E}(\underset{\sim}{\epsilon}) = \underset{\sim}{0}$ and $\mathrm{cov}(\underset{\sim}{\epsilon}) = I_n$.

Theorem 3. For $g_0 \in \mathcal{H}$,

$$\min_{C, \underset{\sim}{c}} \max_{g \in \mathcal{H}: \|P_1(g-g_0)\|_{\mathcal{H}}^2 \leq \rho} \mathrm{E}\|C\underset{\sim}{Y} + \underset{\sim}{c} - \underset{\sim}{L}(g)\|^2$$

$$= k + \sum_1^{n-k} (1 - \Lambda_{ii}^{*1/2}/\alpha)_+$$

for some $\alpha = \alpha(\rho)$ and the restricted minimax estimate of $\underset{\sim}{L}(g)$ is equal to $\hat{\underset{\sim}{L}}_{MB}$, the minimax Bayes estimate of Section 3.

Proof. For the Demmler-Reinsch basis, $\psi_1, \ldots, \psi_n$, for $g - g_0 \in \mathcal{H}$, there exists $s = \sum \beta_j \psi_j$ such that $\underset{\sim}{L}(g - g_0) = \underset{\sim}{L}(s)$ and $\|P_1 s\|_{\mathcal{H}}^2 \leq \|P_1(g - g_0)\|_{\mathcal{H}}^2$. We can therefore restrict the problem to g in $\tilde{\mathcal{H}}$ of the form $g - g_0 = \sum \beta_j \psi_j$. The beginning of the proof is similar to that of Theorem 1: transform the problem to the quantities with subscripts 1 and show that $\underset{\sim}{c}_1 = \underset{\sim}{0}$. Then apply Lemma 3.1 of Speckman (1985).

Acknowledgements.

This research was supported under the Natural Sciences and Engineering Research Council of Canada grant 5-87969. The author also gratefully acknowledges partial financial support provided by the NATO Advanced Study Institute under the Directorship of Professor G. Roussas.

REFERENCES

DasGupta, A. and Studden, W.J. (1989) 'Frequentist behavior of robust Bayes estimates of normal means', Statistics and Decisions 7, 333–361.

Demmler, A. and Reinsch, C. (1975) 'Oscillation matrices with spline smoothing', Numerische Mathematik 24, 375–382.

Gupta, S. and Hwang, D. (1977) 'On some Γ–minimax subset selection and multiple comparison procedures', in S.S. Gupta and D.S. Moore (eds.), Proceedings Second Purdue Symposium, Academic Press, New York, pp. 139–155.

Heckman, N. and Woodroofe, M. (1989) 'Minimax Bayes estimation in nonparametric regression', manuscript.

Karlin, S. (1959) Mathematical Methods and Theory in Games Programming and Economics, Volume 1, Addison–Wesley, Massachusetts.

Leamer, E.E. (1982) 'Sets of posteriour means with bounded variance prior', Econometrica 50, 725–736.

Legostaeva, I.L. and Shiryayev, A.N. (1971) 'Minimax weights in a trend detection problem of a random process', Theory of Probability and its Applications 16, 344–349.

Nussbaum, M. (1985) 'Spline smoothing in regression models and asymptotic efficiency in L_2', Annals of Statistics 12, 984–997.

O'Hagan, A. (1978) 'Curve fitting and optimal design for prediction', Journal of the Royal Statistical Society, Series B 40, 1–42.

Polasek, W. (1985) 'Sensitivity analysis for general and hierarchical linear regression' in P.K. Goel and A. Zellner (eds.), Bayesian Inference and Decision Techniques with Applications, North Holland, Amsterdam.

Sacks, J. and Ylvisaker,D. (1978) 'Linear estimation for approximately linear models', Annals of Statistics 6, 1122–1137.

Speckman, Paul (1985) 'Spline smoothing and optimal rates of convergence in nonparametric regression models', Annals of Statistics 13, 970–983.

Wahba, G. (1990) Spline Models for Observational Data, CBMS–NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics.

Weerahandi, S. and Zidek, J.V. (1985) 'Smoothing locally smooth processes by Bayesian nonparametric methods', SIAM Institute for Mathematics and Society Technical Report 93.

ON EXPONENTIAL BOUNDS ON THE BAYES RISK OF THE NONPARAMETRIC CLASSIFICATION RULES

ADAM KRZYŻAK
Department of Computer Science
Concordia University
1455 De Maisonneuve Blvd. West
Montreal, Canada H3G 1M8

Abstract

We derive improved exponential bounds on Bayes probability of error for the nonrecursive and recursive kernel classification rule. We show, using the martingale device, that weak, strong and complete L_1 Bayes risk consistencies are equivalent. Consequently the conditions on the smoothing sequence $h_n \to 0$ and $nh_n^d \to \infty$ are necessary and sufficient for Bayes risk consistency of the nonrecursive kernel classification rule. For the recursive rule these conditions become $h_n \to 0$ and $\sum_{i=1}^n h_i^d \to \infty$. The rates of convergence of the kernel classification rules are also given.

1 Introduction

In nonparametric classification several problems have recently received special attention. One problem is to obtain Bayes risk consistency of classification rules under minimum assumptions on statistical properties of the data. Another is to obtain the rates of convergence under different restrictions on the underlying distributions. Finally, important results have been obtained in constructing automatic, data driven classification rules (Devroye [6]). Bayes risk consistency was obtained for all class densities by Van Ryzin [27], Wolverton and Wagner [29], Devroye and Györfi [7] and Greblicki [11]. Stone [24] was the first to show consistency of the k-NN rule for all class distributions. Subsequently similar results were obtained for the kernel classification rule by Devroye [4], Györfi [17], Krzyżak and Pawlak [20], Devroye and Wagner [3], Greblicki et al [12], for the recursive kernel rules by Krzyżak and Pawlak [19, 21] and for the k-NN rule by Györfi [16, 18]. The rates of convergence of nonparametric classification rules were studied by, among others, Stone [25, 26], Krzyżak [20] and Györfi [16, 18]. Devroye [5] proved that distribution-free rates do not exist.

In this paper we establish exponential bounds on the Bayes risk of the kernel classification rule. The bounds are optimal up to a factor. We also show that weak, strong and complete Bayes risk consistencies are equivalent for all class distributions. A similar result (without exponential bounds) for the recursive kernel rule was obtained by Greblicki and Pawlak [14]. In [14] equivalence was not established for the nonrecursive kernel rule because no optimal consistency conditions for that rule existed at that time. The L_1 equivalence

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 347–360.
© 1991 *Kluwer Academic Publishers.*

result for the kernel regression estimate of Devroye and Krzyżak [8] paved the way for the present paper.

In section 2 we give the equivalence convergence theorem for the kernel classification rule. In sections 3 and 4, respectively, we discuss the proofs of necessity and sufficiency. Section 5 deals with the global optimal rate of convergence of the kernel classification rules.

2 Kernel Classification Rule

Let (θ, X) be a pair of random variables taking values in the set $1, ..., M$, whose elements are called classes, and in R^d, respectively. The problem is to classify X, i.e. to decide on θ. Let us define *a posteriori* class probabilities

$$p_i(x) = P\{\theta = i | X = x\}, i = 1, ..., M, x \in R^d.$$

The Bayes classification rule

$$\Psi^*(X) = i \text{ if } p_i(X) > p_j(X), j < i, \text{ and } p_i(X) > p_j(X), j > i$$

minimizes the probability of error. The Bayes risk L^* is defined by

$$P\{\Psi^*(X) \neq \theta\} = \inf_{\Psi: R^d \to \{1,...,M\}} P\{\Psi(X) \neq \theta\}.$$

The local Bayes risk is equal to $P\{\Psi^*(X) \neq \theta \mid X = x\}$. Observe that $p_i(x) = E\{I_{\{\theta=i\}} \mid X = x\}$ may be viewed as a regression function of the indicator of the event $\{\theta = i\}$. Given the learning sequence $V_n = \{(\theta_1, X_1), ..., (\theta_n, X_n)\}$ of independent observations of the pair (θ, X), we construct an empirical classification rule Ψ_n, which classifies every $x \in R^d$ to any class maximizing

$$p_{in} = \frac{\sum_{j=1}^n I_{\{\theta_j=i\}} K\left(\frac{x-X_j}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)},$$

$i = 1, ..., M$, where K is a nonnegative Borel measurable kernel on R^d and $\{h_n\}$ is a sequence of positive numbers. The recursive version of Ψ_n is obtained by replacing $p_{in}(x)$ by

$$\frac{\sum_{j=1}^n I_{\{\theta_j=i\}} K\left(\frac{x-X_j}{h_j}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_j}\right)}.$$

The global performance of Ψ_n is measured by $L_n = P\{\Psi_n(X) \neq \theta \mid V_n\}$ and the local performance by $L_n(x) = P\{\Psi_n(x) \neq \theta \mid V_n\}$. A rule is said to be weakly, strongly, or completely Bayes risk consistent (BRC) if $L_n \to L^*$, in probability, almost surely, or completely, respectively, as $n \to \infty$.

Our main result is based on the exponential inequality for martingales. It enables proving the equivalence of all modes of Bayes risk consistency for all distributions of (X, θ).

We will say that a Borel kernel K is *regular* if

$$K(x) \geq cI_{S_{x,r}}$$

for some positive constants c and r, where $S_{x,r}$ is the ball of radius r centered at x, and if

$$\int \sup_{t \in S_{x,r}} K(t)dx < \infty. \tag{1}$$

Notice that (1) is satisfied if K is Riemann integrable. Examples of regular kernels include:

a) rectangular kernel

$$K(x) = \begin{cases} 1/2 & \text{for } \|x\| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

b) triangular kernel

$$K(x) = \begin{cases} 1 - \|x\| & \text{for } \|x\| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

c) Gaussian

$$K(x) = \frac{1}{\sqrt{2\pi}} exp\{(-1/2)\|x\|^2\}$$

d)

$$K(x) = \begin{cases} \frac{1}{2\pi}\left(\frac{sin(x/2)}{x/2}\right)^2 & \text{if } x \neq 0 \\ 1/2\pi & \text{if } x = 0. \end{cases}$$

We now present the main result.

Theorem 1. Let K be a regular kernel with compact support. Then the following statements are equivalent: (A) For all distributions of (θ, X) and every $\epsilon > 0$, and for both nonrecursive and recursive rules, there exists n_0 such that

$$P\{|L_n(X) - L^*(X)| > \epsilon\} \leq 4Me^{-\bar{c}n},$$

for all $n \geq n_0$. Constant $\bar{c}$ is equal to $\epsilon^2/(2 \cdot 4^4 M^2 \max\{\rho^2, 4(1+\rho)^2\})$ for nonrecursive rule. For recursive rule $\bar{c}$ is equal to nonrecursive $\bar{c}$ divided by γ (see part D below). ρ is a constant depending on K only. (B) For all distributions of (θ, X) and both rules

$$L_n \to L^* \text{ almost surely as n } \to \infty.$$

(C) For all distributions of (θ, X) and both rules

$$L_n \to L^* \text{ in probability as n } \to \infty.$$

(D) For nonrecursive rule we have

$$\lim_{n\to\infty} h_n = 0, \tag{2}$$

$$\lim_{n\to\infty} nh_n^d = \infty. \tag{3}$$

For recursive rule we assume that sequence h_n is regularly varying with $\gamma = \limsup \frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_i}{\min_{1\le j\le n} h_j}\right)^2 < \infty$, (2) holds and

$$\sum_{i=1}^{\infty} h_i^d = \infty. \tag{3'}$$

Statement (D) implies (A), (B) and (C) if K is merely regular.

Remark. The exponential bound in (A) is better than existing bounds for the Bayes risk and is optimal up to a constant. Sufficient conditions for BRC of the nonrecursive kernel rule have been established by Devroye and Wagner [2, 3], Krzyżak and Pawlak [20] and Krzyżak [22], but under stricter assumptions on the kernel and on h_n. Typical assumptions on the kernel require it to have compact support or a nonincreasing envelope. We merely require it to be Riemann integrable and bounded from zero at the origin. Strong, global Bayes risk consistency is usually inferred from local consistency through the Lebesgue dominated convergence theorem resulting in the following restriction on the smoothing sequence $nh^d/\log n \to \infty$ as n $\to \infty$. We deal with global risk directly and relax this restriction to $nh^d \to \infty$ as n $\to \infty$. Consequently, we obtain equivalence of all types of convergence. Finally, using the bound in A, we obtain better rates of convergence of the kernel classification rule than in [22] (see Theorem 2). For recursive rule we obtain in addition to [14] the exponential bound and we obtain global BRC directly.

3 Proof of Necessity

For the recursive rule the proof of necessity is given in [14]. First it is shown that (C) implies (2) and subsequently that (C) and (2) imply (3'). The necessity for the nonrecursive rule can be obtained in a same way and the proof is omitted.

4 Proof of Sufficiency

In this section we will consider nonrecursive rule only. The proof of sufficiency for recursive rule is similar and is omitted. In the proof of sufficiency we will need several lemmas of independent interest. The following technical lemma [8, Lemma 1] will be useful in establishing the bounds for the Bayes risk.

Lemma 1. If K is a regular kernel and μ is an arbitrary probability measure on R^d, then there exists a finite constant $\rho = \rho(K)$, depending only on K, such that

$$\sup_{y\in R^d, h>0} \int \frac{K_h(x-y)}{EK_h(x-X)}\mu(dx) \le \rho,$$

where $K_h(x) = K(x/h)$. Also, for any $\delta, \epsilon > 0$, there exists $h_0 > 0$ such that

$$\sup_{y\in R^d, h\le h_0} \int \frac{K_h(x-y)I_{\{\|x-y\|\ge\delta\}}}{EK_h(x-X)}\mu(dx) \le \epsilon.$$

The next lemma establishes the exponential inequality, which is crucial in the proof of sufficiency.

Lemma 2. Let K be a regular kernel. Then there exists a finite constant $\rho = \rho(K)$ depending on K only, such that for all $\epsilon > 0$ and for all $n > 0$,

$$P\{|\int |N_n(x)|\mu(dx) - E\int |N_n(x)|\mu(dx)| > \epsilon\} \leq 2\exp(-n\epsilon^2/32\rho^2)$$

and

$$P\{|\int |D_n(x)|\mu(dx) - E\int |D_n(x)|\mu(dx)| > \epsilon\} \leq 2\exp(-n\epsilon^2/32(1+\rho)^2),$$

where $N_n(x) = \sum_{j=1}^n \left(I_{\{\theta_j=i\}} - p_i(x)\right) K_h(x - X_j)/nEK_h(x - X)$ and $D_n(x) = \sum_{j=1}^n K_h(x - X_j)/nEK_h(x - X) - 1$.

Proof. We will apply exponential inequalities for martingale difference sequences (see Azuma [1] and Stout [23]). Define

$$T_j = E(\int |N_n| \mid V_j) - E(\int |N_n| \mid V_{j-1}).$$

Clearly

$$\sum_{j=1}^n T_j = E(\int |N_n| \mid V_n) - E(\int |N_n| \mid V_0) = \int |N_n| - E\int |N_n|,$$

where V_0 stands for the trivial sigma algebra (no data). Also $E(T_j \mid V_{j-1}) = 0$ so that T_j is a martingale difference sequence. The Azuma inequality may be applied if T_j are uniformly bounded for all j. Denoting

$$S_{j,k} = \frac{1}{n}\sum_{j\leq l\leq k} \frac{\left(I_{\{\theta_l=i\}} - p_i(x)\right) K_h(x - X_l)}{EK_h(x - X)},$$

we have by a simple argument

$$\begin{aligned}|T_j| &\leq \int |E(|S_{1,j-1} + S_{j,j} + S_{j+1,n}| \mid V_j) - E(|S_{1,j-1} + S_{j,j} + S_{j+1,n}| \mid V_{j-1})| \\ &\leq \int \sup_a |E|a + S_{j,j}| - |a + S_{j,j}|| \\ &\leq \int |S_{j,j} - ES_{j,j}| + \int E|S_{j,j} - ES_{j,j}|.\end{aligned}$$

Since

$$\int |S_{j,j}| = \int |\frac{1}{n}\frac{I_{\{\theta_j=i\}} - p_i(x))K_h(x - X_j)}{EK_h(x - X)}|\mu(dx),$$

the right hand side of the above equation is bounded above by

$$\frac{1}{n}\int \frac{K_h(x - X_j)}{EK_h(x - X)}\mu(dx) \leq \frac{1}{n}\sup_z \int \frac{K_h(x - z)}{EK_h(x - X)}\mu(dx) \leq \frac{\rho}{n},$$

where ρ is the constant from Lemma 1. Similarly we can show that $E\int |S_{j,j}| \leq \frac{\rho}{n}$. Hence

$$|T_j| \leq \frac{4\rho}{n}.$$

In what follows we will need the following inequality, due to Azuma [1, 23].

$$P\{|\sum_{i=1}^{n} T_i| > \epsilon\} \leq 2\exp(-\epsilon^2/(2\sum_{i=1}^{n} ||T_i||_\infty^2)),$$

where $||T||_\infty$ denotes the essential supremum of $|T|$ and T_i satisfy the conditions stated above. We have

$$P\{|\int |N_n| - E\int |N_n|| > \epsilon\} \leq 2e^{-n\epsilon^2/32\rho^2},$$

for all $\epsilon > 0$ and all $n > 0$.
The bound for $D_n(x)$ can be obtained similarly. We now define

$$S_{j,j} = \frac{1}{n}(K_h(x - X_j)/EK_h(x - X_1) - 1).$$

By repeating the argument for $N_n(x)$ with slight modifications we obtain for $D_n(x)$

$$\int |S_{j,j}| \leq \frac{1+\rho}{n} \text{ and } |T_j| \leq 4(1+\rho)/n.$$

Applying the exponential inequality we have,

$$P\{|\int |D_n| - E\int |D_n|| > \epsilon\} \leq 2\exp\left(-n\epsilon^2/32(1+\rho)^2\right),$$

for all $\epsilon > 0$ and all $n > 0$. The proof of Lemma 2 is complete.
The next lemma shows that $E\int |N_n|$ and $E\int |D_n|$ tend to zero as $n \to \infty$.

Lemma 3. Let K be a regular kernel, and let $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$. Then

$$E\int |N_n(x)|\mu(dx) \to 0, E\int |D_n(x)|\mu(dx) \to 0$$

as $n \to \infty$, for all distributions of (θ, X).

Proof. In the proof we will use the covering argument and the fact that continuous functions are dense in $L_1(\mu)$ [28]. Due to the last property there exists for any $\epsilon > 0$ a continuous function $m \in L_1(\mu)$ with compact support, such that $\int |p_i(x) - m(x)|\mu(dx) < \epsilon$. We have

$$E\int |N_n(x)|\mu(dx) = E\int |\sum_{j=1}^{n}(I_{\{\theta_j=i\}} - p_i(x))K_h(x - X_j)/nEK_h(x - X)|\mu(dx)$$

$$\leq E\int |\sum_{j=1}^{n}(I_{\{\theta_j=i\}} - p_i(X_j))K_h(x - X_j)/nEK_h(x - X)|\mu(dx)$$

$$+E\int |\sum_{j=1}^{n}(p_i(X_j) - m(X_j))K_h(x - X_j)/nEK_h(x - X)|\mu(dx)$$

$$+E\int |\sum_{j=1}^{n}(m(X_j) - m(x))K_h(x - X_j)/nEK_h(x - X)|\mu(dx)$$

$$+E\int |\sum_{j=1}^{n}(m(x) - p_i(x))K_h(x - X_j)/nEK_h(x - X)|\mu(dx)$$

$$= A_1 + A_2 + A_3 + A_4.$$

By the approximation property

$$A_4 \leq \int |m - p_i| < \epsilon.$$

A_3 can be split for arbitrary δ into the sum of two terms

$$A_{31} = E \int |\sum_{j=1}^{n} (m(X_j) - m(x)) K_h(x - X_j) I_{\{||x - X_j|| < \delta\}} / nEK_h(x - X)| \mu(dx)$$

and

$$A_{32} = E \int |\sum_{j=1}^{n} (m(X_j) - m(x)) K_h(x - X_j) I_{\{||x - X_j|| \geq \delta\}} / nEK_h(x - X)| \mu(dx).$$

Clearly,

$$A_{31} \leq \sup_x \sup_{y \in x + \delta h} |m(x) - m(y)| < \epsilon/2$$

for δ small enough.
We have for sufficiently small h (Lemma 1),

$$A_{32} \leq 2 \sup_x |m(x)| \sum_{j=1}^{n} \int E(K_h(x - X_j) I_{\{||x - X_j|| \geq \delta\}}) / nEK_h(x - X) \mu(dx) < \epsilon/2.$$

Furthermore,

$$A_2 \leq \int E(|p_i(X_1) - m(X_1)| K_h(x - X_1)) / EK_h(x - X) \mu(dx)$$

$$\leq \sup_y \int K_h(x - y) / EK_h(x - X) \mu(dx) \; E|p_i(X_1) - m(X_1)| \leq \rho\epsilon.$$

Term A_1 requires some extra work. We split the integral in A_1 into two integrals: one over a large compact set Ω and the other over the complement of Ω. Since the integrand in A_1 is bounded by some constant M, the first integral does not exceed $M\mu\{\Omega\} < \epsilon/2$ for Ω large enough. By the Cauchy-Schwartz inequality we have

$$E \int_\Omega E^{1/2} \left\{ \left\{ \frac{\sum_{j=1}^{n} (I_{\{\theta_j = i\}} - p_i(X_j)) K_h(x - X_j)}{nEK_h(x - X_1)} \right\}^2 \mid X_1, ..., X_n \right\} \mu(dx)$$

$$\leq ME \int_\Omega \left\{ \frac{\sum_{j=1}^{n} K_h^2(x - X_j)}{n^2 E^2 K_h(x - X_1)} \right\}^{1/2} \mu(dx)$$

$$\leq \frac{M}{\sqrt{n}} \int_\Omega \left\{ \frac{EK_h^2(x - X_1)}{E^2 K_h(x - X_1)} \right\}^{1/2} \mu(dx)$$

$$\leq \frac{M}{\sqrt{nh^d}} \left\{ \int_\Omega \frac{h^d EK_h^2(x - X_1)}{E^2 K_h(x - X_1)} \mu(dx) \right\}^{1/2}$$

(Jensen's inequality)

$$\leq \frac{M}{\sqrt{nh^d}} \left\{ \int_\Omega \frac{kh^d}{EK_h(x - X_1)} \mu(dx) \right\}^{1/2},$$

where k is an upper bound on K. The last integral is uniformly bounded for all $h \leq 1$. Indeed, notice that $EK_h(x - X_1) \geq c\mu(S_{x,rh})$, and also that Ω can be covered by $O(h^{-d})$ translates of $S_{x,rh/2}$. The integral over each of these translates is not larger than $O(h^d)$ and the whole integral is O(1). Therefore by (3) the integral over Ω is bounded above by $\epsilon/2$ for n large enough. Collecting these facts we see that

$$A_1 \leq \epsilon + o(1).$$

The $E \int |D_n(x)| \mu(dx)$ term can be handled similarly as the $E \int |N_n(x)| \mu(dx)$ term. The proof of Lemma 3 is complete.

We are now in a position to prove the sufficiency part of Theorem 1.

Proof of sufficiency. It suffices to show that (D) implies (A). Using the bound of Van Ryzin [27] (see also [4, 16]) relating the Bayes risk to conditional class distributions we have

$$\int |L_n(x) - L^*(x)| \mu(dx) \leq \sum_{i=1}^{M} \int |p_{in}(x) - p_i(x)| \mu(dx). \tag{4}$$

We will now derive exponential bounds for $\int |p_{in} - p_i|, i = 1, ..., M$. We can write

$$|p_{in}(x) - p_i(x)| = |\frac{N_n(x)}{D_n(x) + 1}|.$$

Now,

$$P\{\int |\frac{N_n(x)}{D_n(x) + 1}| \mu(dx) > \epsilon\}$$

$$= P\{\int |\frac{N_n(x)}{D_n(x) + 1}| \left(I_{\{S(x)\}} + I_{\{\bar{S}(x)\}} \right) \mu(dx) > \epsilon\}$$

$$\text{(where } S(x) = \{|D_n(x)| < 1/2\})$$

$$\leq P\{\int |N_n| d\mu + \sup_x |\frac{N_n(x)}{D_n(x) + 1}| P\{\bar{S}(X)|V_n\} > \epsilon\}$$

$$\leq P\{\int |N_n| > \epsilon/2\} + P\{P\{\bar{S}(X)|V_n\} > \epsilon/2\}.$$

Since the feature vector X is independent of the observations $V_n = \{\theta_1, X_1, ..., \theta_n, X_n\}$ we can apply Markov's inequality to the second probability in the last expression, and bound the whole expression from above by

$$P\{\int |N_n(x)| \mu(dx) > \epsilon/2\} + P\{\int |D_n(x)| \mu(dx) > \epsilon/4\}$$

$$\leq P\{|\int |N_n| - E\int |N_n|| > \epsilon/4\} + P\{E\int |N_n| > \epsilon/4\}$$

$$+ P\{|\int |D_n| - E\int |D_n|| > \epsilon/8\} + P\{E\int |D_n| > \epsilon/8\}. \tag{5}$$

Using (5) and the bounds from Lemma 2 and Lemma 3 we finally obtain for n sufficiently large,

$$P\{\int |p_{in}(x) - p_i(x)|\mu(dx) > \epsilon\}$$
$$\leq P\{|\int |N_n| - E\int |N_n|| > \epsilon/4\} + P\{E\int |N_n| > \epsilon/4\}$$
$$+P\{|\int |D_n| - E\int |D_n|| > \epsilon/8\} + P\{E\int |D_n| > \epsilon/8\}$$
$$\leq 4\exp(-n\epsilon^2/(2\cdot 4^4 \max\{\rho^2, 4(1+\rho)^2\}))$$

and hence by (4)

$$P\{\int |L_n(x) - L^*(x)|\mu(dx) > \epsilon\} \leq 4M\exp(-n\epsilon^2/(2\cdot 4^4 M^2 \max\{\rho^2, 4(1+\rho)^2\}))$$

for n large enough, so that $E\int |N_n| \leq \epsilon/4$ and $E\int |D_n| \leq \epsilon/8$. The proof is completed.

5 The Rates of Convergence

We can apply the results of Section 4 to obtain the rates of global risk convergence. The rates for the kernel classification rules were studied in [6, 22]. In [22] the rates of global convergence were obtained indirectly from the pointwise rates and were not optimal. The exponential inequality of Theorem 1 enables us to obtain better rates under weaker assumptions on the parameters of the estimate. The rate of convergence of the corresponding kernel regression estimate is optimal (see Remark 2). We say that $A_n = O(B_n)$ with probability 1 (completely) if for any sequence a_n convergent to zero, $a_n A_n/B_n \to 0$ as n $\to \infty$ with probability 1 ($\sum_{n=1}^{\infty} P\{a_n A_n/B_n > \epsilon\} < \infty$ for any $\epsilon > 0$). The following technical lemma will be useful in determining the effect of the kernel's tail on the rate of convergence. Define for any kernel K

$$K^*(t) = \sup_x\{K(x)I_{\{||x||>t\}}(x)\}$$

and

$$K^+(t) = \sup\{x : K^*(x) > t\}.$$

Lemma 4. Suppose that K is a regular kernel. Then for any bounded, Lipschitz function of order α that is the function satisfying

$$|f(x) - f(y)| \leq G(x)||x-y||^\alpha, 0 < \alpha \leq 1, G \in L_1(\mu), \tag{6}$$

we have for any $\epsilon > 0$,

$$\int_Q \frac{EK_h(x-X)|f(x)-f(X)|}{EK_h(x-X)}\mu(dx) \leq \int G(x)\mu(dx)[hK^+(\epsilon h^d)]^\alpha + c_2\epsilon$$

where Q is a compact subset of R^d, $c_2 = \{M(f) + \int |f(x)|\mu(dx)\}c_1/c$, $M(f) = \sup_{x\in Q} f(x)$ and c_1 is a constant depending upon K and d only.
The proof of Lemma 4 is omitted. Next we give the main result of this section.

Theorem 2. Suppose that K is a regular kernel and in addition

$$||x||^d K(x) \to 0 \text{ as } ||x|| \to \infty. \tag{7}$$

Assume also that the smoothing sequence satisfies (2) and $nh^d/logn \to \infty$ as n $\to \infty$. If $p_i(x), i = 1, ..., M$ are Lipschitz of order α, that is if

$$|p_i(x) - p_i(y)| \leq G(x)||x - y||^\alpha, 0 < \alpha \leq 1, G \in L1(\mu),$$

then for nonrecursive kernel rule

$$\int_Q |L_n(x) - L^*(x)|\mu(dx) = O\{\max\{(\frac{nh^d}{logn})^{-1/2}, b_n\}\} \tag{8}$$

with probability 1 (completely), where b_n is the smallest solution of the nonlinear equation

$$[hK^+(b_n h^d)]^\alpha = b_n \tag{9}$$

and Q is a compact subset of R^d. If, in addition, K has compact support and the smoothing sequence satisfies (2) and (3), then the rate in (8) can be improved and becomes

$$\int_Q |L_n(x) - L^*(x)|\mu(dx) = O\{\max\{(nh^d)^{-1/2}, b_n\}\} = O(n^{-\alpha/(2\alpha+d)}) \tag{10}$$

with probability 1 (completely).
For recursive rule with compact kernel and with regularly varying smoothing sequence satisfying (2) and (3′) we obtain

$$\int_Q |L_n(x) - L^*(x)|\mu(dx) = O\{\max\{(\sum_{i=1}^{n} h_i^d)^{-1/2}, b_n\}\}$$

with probability 1 (completely).

Remark. Note that no differentiability assumptions are imposed on p_i. Also, no restrictions are imposed on the distribution of X. The integrals in Theorem 2 can be extended to R^d provided that μ has compact support. Condition (9) expresses the influence of the kernel's tail on the rate of convergence. Condition (7) implies that $b_n \to 0$ as $n \to \infty$. If for some kernel, K^+ decreases at polynomial rate, that is

$$K^+(t) = t^{-\beta}, 0 < \beta < 1/d$$

then

$$b_n = h_n^{\alpha(1-d\beta)/(1+\alpha\beta)}.$$

In this case (8) can be replaced by

$$\int_Q |L_n(x) - L^*(x)|\mu(dx) = n^{-\alpha(1-d\beta)/(2\alpha+d(1-\alpha\beta))} \tag{11}$$

and the only restrictions on the smoothing sequence are (2), (3) and (3′). Thanks to the improved exponential bound of the previous section, a typical log n term is missing from the rate expressions in (10) and (11). Clearly the rates are not distribution-free,

but no distribution-free rates exist (as shown by Devroye [5]). The strong global rate in (10) matches the weak optimal rate for nonparametric regression obtained in [26]. The smoothness factor α cannot exceed 1. To obtain better rates than those in Theorem 2 we have to assume the existence of derivatives of p_i of order higher than 1. If the p_i are continuously differentiable up to order $p > 1$, then the rate in (10) is $O(n^{-p/(2p+d)})$ and approaches $n^{-1/2}$ as p increases without bound. The rates for nonrecursive and recursive rules are comparable.

Proof of Theorem 2. We will prove the theorem only for nonrecursive rule since the proof for recursive one is similar. Let us first consider the case of K with neither compact support nor a polynomial tail. Define the sequence $c_n = a_n \min\{(nh^d/logn)^{1/2}, b_n^{-1}\}$, where a_n is any sequence convergent to zero. Using (4) and (12) we have

$$P\{c_n \int_Q |L_n(x) - L^*(x)|\mu(dx) > \epsilon\} \leq P\{c_n \sum_{i=1}^{M} \int_Q |p_{in}(x) - p_i(x)|\mu(dx) > \epsilon\}$$

$$\leq \sum_{i=1}^{M} P\{c_n \int_Q |p_{in}(x) - p_i(x)|\mu(dx) > \epsilon/M\}$$

$$\leq \sum_{i=1}^{M} \Big\{ P\{c_n| \int_Q |N_n(x)|\mu(dx) - E \int_Q |N_n(x)|\mu(dx)| > \epsilon/4M\}$$

$$+P\{c_n E \int_Q |N_n(x)|\mu(dx) > \epsilon/4M\}$$

$$+P\{c_n| \int_Q |D_n(x)|\mu(dx) - E \int_Q |D_n(x)|\mu(dx)| > \epsilon/8M\}$$

$$+P\{c_n E \int_Q |D_n(x)|\mu(dx) > \epsilon/8M\}\Big\}. \tag{12}$$

Let us first consider the bias term $E \int_Q |N_n| d\mu$. We have,

$$E \int_Q N_n|\mu(dx) \leq \int_Q E|\sum_{j=1}^{n} (I_{\{\theta_j = i\}} - p_i(X_j)) K_h(x - X_j)/nEK_h(x - X_1)| d\mu$$

$$+ \int_Q E|\sum_{j=1}^{n} (p_i(X_j) - p_i(x)) K_h(x - X_j)/nEK_h(x - X_1)| d\mu$$

$$= I + II.$$

Using the same argument as for the term A_1 in the proof of Lemma 3 we obtain

$$I \leq \frac{2\sqrt{k}c_3}{\sqrt{nh^d}}$$

where c_3 is a covering constant. Hence

$$P\{c_n I > \epsilon/8M\} \leq P\{a_n > \epsilon\sqrt{logn}/16M\sqrt{k}c_3\} = 0$$

for n large enough. By Lemma 4 and the Lipschitz assumption we have,

$$II \leq \int_Q E|p_i(X_1) - p_i(x)|K_h(x - X_1)/EK_h(x - X_1)d\mu$$

$$\leq (\int_Q G(x)\mu(dx))[hK^+(b_n h^d)^\alpha + c_2 b_n = (\int_Q G(x)\mu(dx) + c_2)b_n = Wb_n.$$

Assumption (7) implies that $b_n \to 0$ as $n \to \infty$. Hence we obtain

$$P\{c_n II > \epsilon/8M\} \leq P\{a_n > \epsilon/8MW\} = 0$$

for n large enough. We therefore conclude that for any sequence a_n converging to zero the second (bias) term on the right side of (12) is equal to zero for n sufficiently large. Note that for the kernel with compact support, $b_n = h_n^\alpha$ and we have

$$a_n \min\{(nh^d)^{1/2}, h^{-\alpha}\}E\int_Q |N_n|d\mu \to 0$$

as $n \to \infty$ for any sequence a_n converging to zero. Let us now consider the first (variation) term on the right side of (12). Using the bound from Lemma 2 we have,

$$P\{c_n|\int_Q |N_n| - E\int_Q |N_n|| > \epsilon/4M\}$$

$$\leq P\{a_n(nh^d/logn)^{1/2}|\int_Q |N_n| - E\int_Q |N_n|| > \epsilon/4M\}$$

$$\leq 2exp(-logn\epsilon^2/2 \cdot 4^4 M^2 \rho^2 a_n^2 h^d). \tag{13}$$

The right side of the inequality above is summable, implying that the variation converges almost surely (completely) as $n \to \infty$ with rate at least $(logn/nh^d)^{1/2})$. In case of K with compact support or with a polynomial tail, the $logn$ term may be dropped from (13). In particular, for the kernel with compact support the optimal h satisfies $(nh^d)^{1/2} = h^{-\alpha}$, resulting in $h_{opt} = n^{-1/(2\alpha+d)}$ and $c_n = a_n n^{-\alpha/(2\alpha/+d)}$. Substituting this in the probability on the left side of (13) we obtain

$$P\{a_n n^{-\alpha/(2\alpha+d)}|\int_Q |N_n| - E\int_Q |N_n|| > \epsilon/4M\} \leq 2\exp(-n^{2\alpha/(2\alpha+d)}\epsilon^2/2 \cdot 4^4 M^2 \rho^2 a_n^2)$$

which is summable, implying that the variation converges almost surely (completely) as $n \to \infty$ with the rate at least $n^{-\alpha/(2\alpha+d)}$. The proof for D_n is similar; the proof of Theorem 2 is complete.

6 Conclusion

We have derived the exponential, distribution-free bounds for the Bayes risk for nonrecursive and recursive kernel classification rules. We have established the equivalence of all modes of the global convergence of the rules under optimal assumptions on the smoothing sequences. We have also derived the optimal global rates of convergence of the recursive and nonrecursive kernel regression estimates within the class of Lipschitz distributions.

The rates are optimal for nonparametric regression, but not for classification. It is intuitively clear that we do not have to estimate entire class conditional probabilities to obtain good rates for the Bayes risk. Therefore the question of optimal classification rates is still open. Another open problem involves the rates of convergence for the smooth class densities. We conjecture that good classification rates may be obtained for the kernel rules with superkernels (that is, kernels which satisfy $\int x^k K(x)dx = 0$ for all $k > 0$).

7 Acknowledgements

This research was sponsored by NSERC grant A0270 and FCAR grant EQ 2904.

References

[1] K. Azuma, "Weighted sums of certain dependent random variables", *Tohoku Math. J.*, vol. 37, pp. 357-367, 1967.

[2] L. Devroye and T.J. Wagner, "On the L1 convergence of the kernel estimators of regression functions with applications in discrimination", *Z. Wahrscheinlichkeitstheorie Verwandte Gebiete*, vol. 51, pp. 18-25, 1980.

[3] L. Devroye and T.J. Wagner, "Distribution-free consistency results in nonparametric discrimination and regression function estimation". *Ann. Statist.* vol. 8, pp. 231-239, 1980.

[4] L. Devroye, "On the almost everywhere convergence of nonparametric regression function estimate", *Ann. Statist.*, vol. 9, pp. 1310-1319, 1981.

[5] —, "Any discrimination rule can have arbitrary bad probability of error for finite sample size", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-4, pp. 154-157, 1982.

[6] —, "Automatic pattern recognition: a study of the probability of error", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, pp. 530-543, 1988.

[7] L. Devroye and L. Györfi, *Nonparametric Density Estimation. The L1 View*, New York: Wiley, 1985.

[8] L. Devroye and A. Krzyżak, "An equivalence theorem for L1 convergence of the kernel regression estimate", *J. Statist. Planning Inference*, vol. 23, pp. 71-82, 1989.

[9] N. Glick, "Sample-based classification procedures using density estimates", *J. Amer. Statist. Assoc.*, vol. 67, pp. 116-122, 1972.

[10] —, "Consistency conditions for probability estimators and integrals for density estimators", *Utilitas Mathematica*, vol. 6, pp. 61-74, 1974.

[11] W. Greblicki, "Asymptotically optimal pattern recognition procedures with density estimates", *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 250-251, 1978.

[12] W. Greblicki, A. Krzyżak and M. Pawlak, "Distribution- free pointwise consistency of kernel regression estimate", *Ann. Statist.*, vol. 12, pp. 1570-1575, 1984.

[13] W. Greblicki, A. Krzyżak and M. Pawlak, "Necessary and sufficient conditions for Bayes risk consistency of the kernel classification rule", unpublished manuscript, 1985.

[14] W. Greblicki and M. Pawlak, "Necessary and sufficient conditions for Bayes risk consistency of a recursive kernel classification rule", *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 408-412, 1987.

[15] L. Devroye, *A course in density estimation.* Boston: Birkhäuser, 1987.

[16] L. Györfi, "On the rate of convergence of nearest neighbor rules", *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 509-512, 1978.

[17] —, "Recent results on nonparametric regression estimate and multiple classification", *Prob. Contr. Inform. Theory*, vol. 10, pp. 509-512, 1978.

[18] —, "The rate of convergence of $k_n - NN$ regression estimates and classification rules", *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 362-364, 1981.

[19] A. Krzyżak and M. Pawlak, "Universal consistency results for Wolverton-Wagner regression function estimates with application in discrimination", *Prob. Contr. Inform. Theory*, vol. 12, pp. 33-42, 1983.

[20] —, "Distribution-free consistency of a nonparametric kernel regression estimate and classification". *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 78-81, 1984.

[21] —, "Almost everywhere convergence of a recursive regression function estimate and classification".*IEEE Trans. Inform. Theory*, vol. IT-30, pp. 91-93, 1984.

[22] A. Krzyżak, "The rates of convergence of kernel regression estimates and classification rules". *IEEE Trans. Information Theory*, vol. IT-32, pp. 668-679, 1986.

[23] W.F. Stout, *Almost Sure Convergence*, New York: Academic, 1974.

[24] C. Stone, "Consistent nonparametric regression", *Ann. Statist.*, vol. 8, pp. 1348-1360, 1977.

[25] C. Stone, "Optimal rates of convergence for nonparametric estimators". *Ann. Statist.*, vol. 8, pp. 1348-1360, 1980.

[26] C. Stone, "Optimal global rates of convergence for nonparametric regression", *Ann. Statist.*, vol. 10, pp. 1040-1053, 1982.

[27] J. Van Ryzin, "Bayes risk consistency of classification procedures using density estimators", *Sankhya*, Series A, vol. 28, pp. 161-170, 1966.

[28] R.L. Wheeden and A. Zygmund, *Measure and integral*, New York: Marcel Dekker, 1977.

[29] C.T. Wolverton and T.J. Wagner, "Asymptotically optimal discriminant functions for pattern classification", *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 258-265, 1969.

NONPARAMETRIC REGRESSION ANALYSIS OF SOME ECONOMIC DATA

Y.P. MACK
Division of Statistics
University of California
Davis California 95616 USA

and

N.J. WULWICK
The Jerome Levy Economics Institute
Bard College
Blithewood, Annandale-On-Hudson
New York 12504 USA

ABSTRACT. We study the relationship between unemployment rate and the rate of change of wage rates based on economic data from 1861 to 1913 in Britain. Various parametric models have been proposed in the past. Here we present a kernel regression approach within the framework of Nadaraya (1964) and Watson (1964). The choices of the bandwidth based on cross-validation and a new local criterion are compared. The validity of the currently accepted Phillips Curve is addressed.

1. Introduction

The study of the relationship between unemployment rate and the rate of change of wage rates (or inflation rate) has always been the subject of great interest to economists and public policy makers. For instance, one would like to know the unemployment level to maintain stable wages, or the unemployment rate for stable product prices, or the corresponding drop in inflation rate with respect to an increase in unemployment rate. A pioneer in quantifying this relationship was A.W. Phillips, who, in 1958, embarked on a systematic study of the unemployment/inflation figures in Britain from 1861 to 1913. (Phillips in fact studied the figures for the period 1861-1957, although he did separate the data into three sub-periods: 1861-1913, 1913-1948, and 1948-1957. We choose to consider the first sub-period here since the procedure behind Phillips' approach in arriving at his model for 1861-1913 is well-known. The other two sub-periods provide too small a data-base for the large sample character of the kernel regression estimation considered in this study to be meaningful.) In hypothesizing his models, Phillips noted:

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 361–374.
© 1991 *Kluwer Academic Publishers.*

(I) "When the demand for labour is high and there are very few unemployed we should expect employers to bid wage rates up quite rapidly ... . On the other hand it appears that workers are reluctant to offer their services at less than prevailing rates when the demand for labour is low and unemployment is high so that wage rates will fall only very slowly. The relation between unemployment and the rate of change of wage rates is therefore likely to be highly nonlinear."

(II) "Thus in a year of rising business activity, with the demand of labour increasing and the percentage unemployment decresasing, employers will be bidding more rigorously for the services of labour than they would be in a year during which the average percentage unemployment was the same but the demand for labour was not increasing. Conversely, in a year of falling business activity, ... the worker will be in a weaker position to press for (wage increases), than they would in a year during which the average percentage unemployment was the same but the demand for labour was not decreasing."

Denoting unemployment rate by U_j and inflation rate by $I_j = \dot{W}_j/W_j$, where $\dot{W}_j = (W_{j+1} - W_{j-1})/2$, $j = 1,\ldots,n = 53$; $j = 1$ corresponds to the year 1861, Phillips proposed the model

$$I_j + a = b\, U_j^{-c} + k\, \dot{U}_j U_j^{-m} \tag{1}$$

with $\dot{U}_j$ defined analogously as $\dot{W}_j$, and a, b, c, m are positive constants. When $\dot{U}_j = 0$ approximately, we have the second model

$$I_j + a = b\, U_j^{-c}. \tag{2}$$

The resulting function estimate which carries his name (the Phillips Curve) was based on (2), and was derived by a most unorthodox (though perhaps not without some intuitive insight) method. He partitioned the range of unemployment (on the x-axis) into six unequally spaced intervals: 0-2, 2-3, 3-4, 4-5, 5-7, 7-11. Then he averaged the unemployment and inflation percentages within each interval, so that six pairs of averages were obtained. Next, he selected trial values for the parameter a. For each such value, the ordinary least squares (OLS) estimates of b and c were calculated based on the six pairs of averages. Finally, the set of estimates with the smallest sum of squares of residuals was chosen. The rationale behind the choice of the grouping of intervals was, ostensibly, to account for the upturns and downturns in each time cycle so that the effect of changing unemployment on inflation tends to be cancelled out by averaging. The ad hoc estimates of the parameters based on (2) given in Phillips (1958) are:

$$\hat{a} = 0.9, \quad \hat{b} = 9.638, \quad \hat{c} = 1.394. \tag{3}$$

The scatter-plot of the raw data with the six average pairs and the Phillips Curve are presented in Figure 1 below.

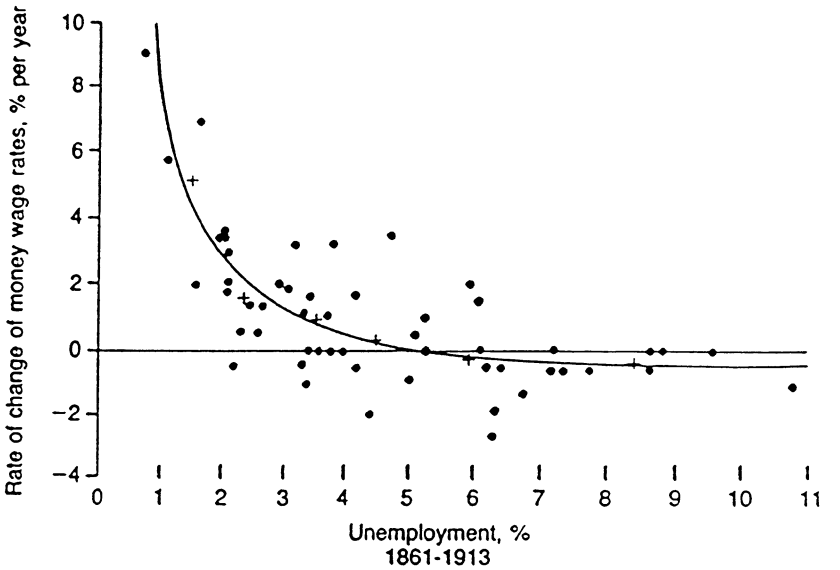


Figure 1. The Original Phillips Curve

2. Controversies

No sooner was Phillips' article published than challenges and counter-rebuttals appeared in regard to the validity of his analysis and conclusions. Controversies, largely from an economic viewpoint though statistical criticisms occasionally appeared, persisted to the present time. (See Desai (1984), Humphrey (1986) and Wulwick (1987, 1989) for some accounts.) The integrity of the data source was questioned by Routh (1959). Lipsey (1960) considered an augmented model of (1) by including an extra U_j^{-2} additive term. Using standard (read "orthodox") statistical methods, he arrived at a curve which virtually coincided with Phillips' Curve. Whether some of these criticisms were justified or not, the Phillips Curve influenced British policy makers in the 1960's. In 1967, the Labour Party Chancellor Callaghan presented a budget aimed at maintaining unemployment at 2.25% to achieve price-level stability -- a notion which politicians of the time associated with the Phillips Curve and related doctrines.

In the U.S., however, from the mid 1960's to the 1970's, there was a prevailing school of thought, championed by the Nobel laureate Milton Friedman that, in the long run, inflation is essentially a monetary

phenomenon, quite unaffected by current unemployment (see Friedman (1968)), thus the Phillips Curve is only a short-run relation. By the end of the 1970's, economists were deeply divided in regard to the effect of increasing unemployment on lowering inflation. There was even a wide-spread claim that the Phillips Curve was, for all practical economic purposes, dead.

Then came the 1980's together with recession and disinflation. Alan Blinder wrote in Business Week in late 1989:

> "The 1980's, however, delivered a resounding rejection of monetarists doctrine. Velocity -- the rate at which money turns over -- behaved erratically. The former close link between money growth and inflation appeared to break. Policymakers now understand that maintaining a constant growth of the money supply does not ensure the economy's health and may even imperil it."
>
> "... high unemployment does indeed arrest an inflation that pessimists had thought unstoppable. And the quantitative dimensions of the trade-off were just what Phillips Curves of the 1970's had predicted. Data from the 1980's verified an old rule of thumb that keeping unemployment one percentage point higher for a year will lower the inflation rate by one-half of a percentage point. Policymakers of the 1990's must now presume that the Phillips Curve trade-off is alive and well in the U.S."

For the purpose of the NATO meeting here, we are more concerned with the validity from a statistical function estimation point of view. Also, a desire to focus our attention on Phillips' original data means that we will have to bypass the criticism of Routh (1959), and hence ignore possible dependency structure requiring a time series approach (Lipsey (1960), Gilbert (1976)). Our nonparametric analysis will therefore be based on the assumption that the 53 bivariate data points are independent and identically distributed.

That aspect settled, the most obvious question is then on the choice of grouping of intervals, seemingly to achieve $\dot{U}_j = 0$ approximately. That Phillips' choice is not unique was pointed out by Wulwick (1989). In that same article, Wulwick also provided convincing arguments rejecting the easy excuse that the crude method of Phillips was employed because of the urgency of time or lack of computing facility. Another question is how contemporary approaches such as nonlinear OLS or maximum likelihood estimation (MLE) compare with Philllips' results. Gilbert (1976) obtained the OLS estimates for model (2):

$$\hat{a} = 0.8826,\ \hat{b} = 8.9387,\ \hat{c} = 1.3837,\ R^2 = 0.6828, \qquad (4)$$

while Oliver (1986) assumed Gaussian additive homoscedastic errors and derived the MLE's of (2), which are in close agreement with the values in (4). In neither of the above two studies were residual analysis made or normality assessed. Additionally, no mention was made concerning the stochastic nataure of the U-variable -- one would assume that any tests of significance were made conditional on the U_j observations.

Finally, there was a philosophical issue raised by Routh (1959) concerning the insistence of fitting a set of noisy data with a very "smooth" curve (smoothness in this context refers to the shape of the curve rather than a differentiability condition). He maintained:

"The double process of simplication (first, the averaging by which six points are derived, then the elimination of the remaining irregularities by the curve-fitting) has produced an effect of smoothness that is largely illusory."

3. Kernel Regression

The discussion in the last section indicates some of the difficulties encountered in fitting a function to a given data set by either crude means or even by some parametric methods. In the sequel, we will attempt a nonparametric estimation based on the Nadaraya (1964) and Watson (1964) version of the kernel method. The regression function is assumed to be of the form $r(u) = E(I|U=u)$. We also assume the unemployments have a density $f(u)$ which is compactly supported and bounded away from 0 from below. To fix notations, define the kernel estimate of $r(u)$ as

$$\hat{r}(u,b) = \frac{\hat{h}(u,b)}{\hat{f}(u,b)}, \tag{5}$$

where

$$\hat{h}(u,b) = (nb)^{-1}[\sum_j w((u-U_j)/b)I_j],$$

and

$$\hat{f}(u,b) = (nb)^{-1}[\sum_j w((u-U_j)/b)]$$

is a kernel estimate of $f(u)$. The function w is the kernel (typically a symmetric function satisfying $\int w = 1$), and $b = b(n)$ is a bandwidth sequence (typically $b \to 0$ and $nb \to \infty$ as $n \to \infty$), and n is the sample size of our data. Some earlier and recent works on the Nadaraya-Watson (NW) estimate include Rosenblatt (1969), Schuster (1972), Schuster and Yakowitz (1979) and Mack and Silverman (1982). More recent works which deal with the important question of bandwidth selection will be mentioned later.

Several variants of (5) exist (for examples, see Mack and Müller

(1989ab).) Our choice in adopting the NW version rather than, for instance, the convolution version is because of our unwillingness to sacrifice variance in exchangae for a simplified bias (see Jennen-Steinmetz and Gasser (1988), Mack and Müller (1989b) and Chu and Marron (1989)). A second decision we made at the outset is to include the stochasticity of the U-variable, so that the so-called fixed-design approach is not appropriate.

As mentioned earlier, the choice of the bandwidth in practice is an important issue, one which has provided the kernel experts with interesting challenges for several decades. It was only in the 1980's when serious advances were made. Broadly speaking, the bandwidth can be chosen according to a global or local perspective. The global viewpoint (i.e., one bandwidth for the entire range of x-coordinates) is represented by the so-called cross-validation method. For the estimation of a density or a regression under the fixed-design assumption, there is by now a better understanding of how the cross-validated bandwidth and estimates employing such a bandwidth behave (see Rice (1984b), Hall and Marron (1987ab), and Härdle, Hall and Marron (1988)). The case of kernel regression when the data are stochastic in both coordinates is a bit more complex. One result from Härdle and Marron (1985) says that the bandwith selected according to a certain version of cross-validation is essentially equivalent to selecting the bandwidth according to an average squared error, whose expectation can be regarded as integrated mean squared error, weighted w.r.t. the density f.

For the local viewpoint, typically the performance is measured by the asymptotic mean squared error at a point. When the first coordinate variable is random as well, as is our economic data, this mean squared error sometimes must be interpreted from the limiting distribution's bias and variance (correctly scaled). In this setting, under second order smoothness of f and r at u, and other appropriate conditions on the kernel and the bandwidth, it can be shown (Schuster (1972)) that the asymptotic bias is given to the first order by

$$\beta_n(b) = 2f(u))^{-1}[r''(u)f(u) + 2r'(u)f'(u)] \int v^2 w(v)dv\ b^2 \qquad (6)$$

and the asymptotic variance is, to the first order, given by

$$\sigma_n^2(b) = (f(u)nb)^{-1}\ s^2(u) \int w^2(v)dv, \qquad (7)$$

where $s^2(u) = \mathrm{Var}(I|U=u)$. Hence the optimal local choice of the bandwidth to minimize the asymptotic mean squared error at u is given by

$$b^*(u) = c(u)n^{-1/5}\ , \qquad (8)$$

where

$$c(u)^5 = [f(u)s^2(u)\int w^2(v)dv][(r''(u)f(u) + 2r'(u)f'(u))\int v^2 w(v)dv]^{-2}.$$

Evidently, the quantitites involving r', r'', f, f' and s^2 are not available in practice. The fact that these quantities can be replaced by initial estimates to form a data-driven version of the theoretically optimal local bandwidth is the content of Mack and Müller (1987). The situation is similar with the fixed-design case, with the bias expression greatly simplified (if the design points are equall-spaced, only $r''(u)$ will appear as a model quantity). Here the method by initially estimating the scale coefficient of the optimal bandwidth, called the "plug-in" method for want of a better term, was shown by Müller and Stadtmüller (1987a) to provide some improvement over cross-validation. Similar results for the fixed-design case were indicated by Müller (1985) and Staniswallis (1989) with a more direct estimate of the local mean square error.

With the NW version of regression estimate, when the data are random in both coordinates, the "plug-in" method in practice becomes a serious undertaking since no less than five model quantities have to be estimated initially. Thus despite the recent endorsement of the "plug-in" method in density estimation (Park and Marron (1990)), some researchers attempted other methods such as the bootstrap (Härdle and Marron (1989)).

Thus for the economic data at hand, we made our third decision to extend the approach of Müller (1985) and Staniswallis (1989) and propose estimates of the bias and variance at a point u as follows:

$$\hat{\beta}(b) = [\sum_j w((u-U_j)/b)]^{-1}[\sum_j w((u-U_j)/b)[\tilde{r}(U_j) - \tilde{r}(u)]] \qquad (9)$$

and

$$\hat{\sigma}^2(b) = [\sum_j w((u-U_j)/b)]^{-2}[\sum_j w^2((u-U_j)/b)\tilde{s}^2(U_j)], \qquad (10)$$

where $\tilde{r}(u)$ and $\tilde{s}^2(u)$ are initial estimates of $r(u)$ and $s^2(u)$, respectively. Our criterion for local bandwidth choice is then based on minimizing the estimated mean squared error

$$\hat{\mu}(b) = \hat{\beta}^2(b) + \hat{\sigma}^2(b). \qquad (11)$$

The estimates $\tilde{r}(u)$ and $\tilde{s}^2(u)$ play a crucial role in local bandwidth choice. Suppose we employ kernel estimates also: Set

$$\tilde{r}(u) = [\sum_j k((u-U_j)/h)]^{-1}[\sum_j k((u-U_j)/h)I_j] \qquad (12)$$

for estimating $r(u)$ with kernel k and bandwidth h, and set

$$\tilde{s}^2(u) = [\sum_j k((u-U_j)/h)]^{-1}[\sum_j k((u-U_j)/h)[I_j-\tilde{r}(u)]^2] \qquad (13)$$

for estimating $s^2(u)$. Under appropriate conditions, Mack (1990) showed that the local bandwidth chosen by minimizing (11) is in an asymptotic sense equivalent to the local bandwidth choice according

to the theoretical expression (8). The initial estimates' bandwidth h should be somewhat slower than b in order to guarantee the proper rate of convergence of the bias term. This "over-smoothing" of the initial estimate was also mentioned in Staniswallis (1989) and Härdle and Marron (1989). However, the bandwidth choice of h based on derivative estimation in the fixed-design setting may not be amenable to the NW kernel estimate when both coordinates of the data are random, and that the conditional variance is generally non-constant.

4. Implementation of Kernel Regression on Phillips' Data

For the NW estimate of the economic data, we employ the optimal mean squared error kernel of Epanechnikov (1969). Some initial thoughts were given to choosing the kernel in a manner similar to Gasser et al (1984), but the attempt was soon abandoned since the marginal density f was not postulated, and we are unwilling to use a constant test value for the conditional variance, as the NW framework assumes neither an additive noise structure nor homoscedasticity.

Therefore our data analysis began with cross-validation, resulting in a bandwidth $b_{CV} = 0.64$. The NW estimate incorporating b_{CV} is displayed in Figure 2. The fit follows the general hyperbolic shape of the Phillips Curve, with slight irregularities between 5% and 7% unemployment rate, as well as an obvious boundary effect on the right end of the support. Modifying the boundaries according to the Richardson extrapolation of Rice (1984a) improved the right side somewhat, with the left side virtually unaffected by the correction. Since we are mostly interested in the region where unemployment is substantially less than 10%, boundary correaction was not carried out for subsequent estimates.

Our implementation of the local adaptation procedure is as follows:

(a) For a given kernel to be used in the initial estimates, obtain the cross-validated bandwidth;

(b) Form initial estimates of r(u) and $s^2(u)$ with the same kernel as in (a), but with several trial values of the bandwidth slightly larger than the cross-validated bandwidth obtained in (a). Decide on an over-smoothing bandwidth, and the corresponding initial estimates.

(c) Fix a kernel for the NW regression. With this kernel, and using initial estimates of (b), form estimates of bias and variance. Then select the minimizing choice of the local bandwidth according to the estimated mean squared error from a possible range $[\underline{b}(u), \overline{b}(u)]$.

(d) With the same kernel as in (c), and with the local bandwidth, form the NW estimate.

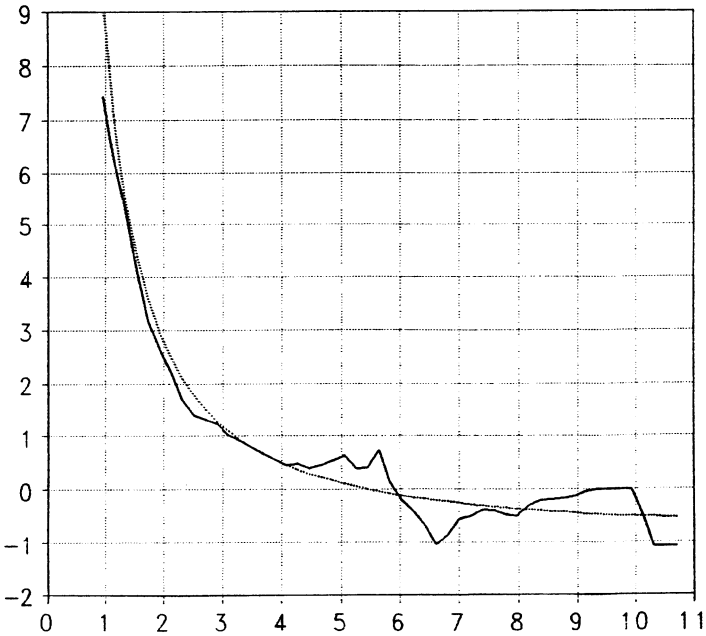


Figure 2. Nadaraya-Watson Fit by Cross-Validation (solid) Phillips Curve (dotted).

For the data at hand, we employed the standard normal as the kernel for the initial estimates. The cross-validated bandwidth for the normal kernel was obtained, and then a slightly larger bandwidths was adopted by experimentation (based on visual inspection of the estimates before obvious artifacts of over-smoothing became evident).

With this slight over-smoothing in the initial estimation, the local bandwidth choice was carried out using the Epanechnikov kernel. Initially, our experience with the Phillips data suggested that the local bandwidth estimate can be quite sensitive to the choice of the interval $[\underline{b}(u), \bar{b}(u)]$. Fortunately the resulting NW fit seemed less so. With $\underline{b}(u) = 0.3$ and $\bar{b}(u) = 4.0$, the local bandwidth estimate is displayed in Figure 3, and the final NW fit is given in Figure 4. Compared with the fit from cross-validation, it can be seen that the irregularities between 5% and 7% unemployment are further reduced. Overall, the local fit indicates an improvement.

Since our nonparametric approach did not result in a major disagreement with the Phillips Curve, we also checked the residuals based on the published model (2) with estimates (3). Based on a kernel density estimate as well as probability plot, the normality assumption does

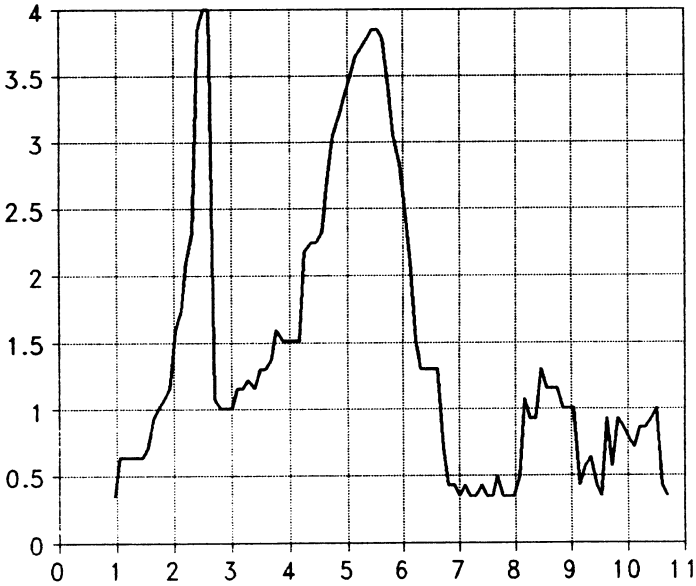


Figure 3: Local Bandwidth Estimate $[\underline{b}(u), \overline{b}(u)] = [0.3, 4.0]$

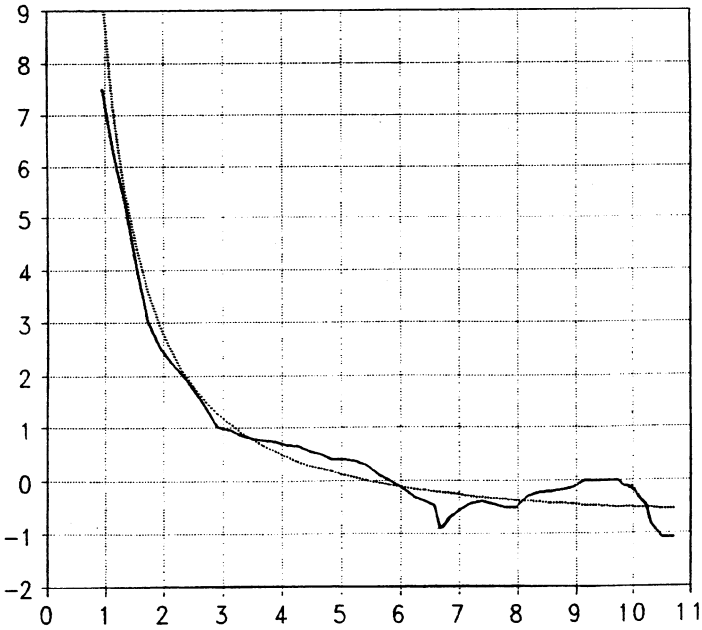


Figure 4: Local Fit (solid) Phillips Curve (dotted).

not appear to have been severely violated, suggesting that MLE may not be unreasonable. However, the graph (Figure 5) of the conditional variance estimate does indicate some heteroscedasticty.

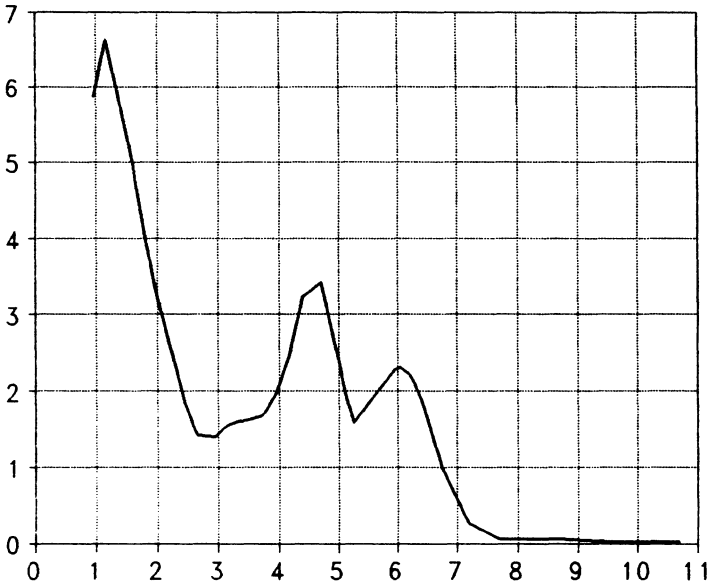


Figure 5. Conditional Variance Estimate.

5. Conclusion

We have presented the procedure employed by Phillips in his 1958 paper which was an ad hoc curve-fitting. We mentioned briefly the controversies surrounding the validity of his analysis. Then we presented a nonparametric approach based on the Nadaraya-Watson paradigm. After a very short discourse on some relevant aspects of this so-called kernel method, we introduced a new data-adaptive technique which is local in spirit, as opposed to the method of cross-validation, which is global. This new local adaptive technology was implemented on Phillips' original data.

The objective of our exercise bringing the analysis of some economic data deep into the territory of what some people might call "esoteric kernel method" is not to reject or validate the Phillips Curve. A more detailed analysis of Phillips' data with emphasis from an economics viewpoint is given in Wulwick and Mack (1990). Here our interest is in constructing a regression estimate with minimum assumptions on the model structure. As it turned out, the model

arrived at by Phillips using a rather "bizarre" method turns out to hold up quite well. Earlier we mentioned the criticism of Lipsey (1960) concerning Phillips' unorthodox method, while Gilbert (1976) supported Phillips from a parametric estimation viewpoint. The motivation of the present study comes from a desire to pay more attention to statistical assumptions, as well as to employ a well-known data set to experiment with a novel data-adaptive function estimation procedure which is free from many parametric regression analysis baggages, and which adopts a perform- ance criterion from a local perspective. Our experimental outcome appears to support the notion that this local method might be a serious competitor to the global method of cross-validation, although one gradually comes to appreciate the elegance and simplicity of the latter, especially from a computational point of view.

Except for the boundary effect on the extreme right, the gross feature of the kernel fit is in general agreement with the Phillips Curve. Where the nonparametric fits differ most are the fluctuations occuring around 5% to 7% unemployment, and the fit by cross-validation a bit more severe compared with the local fit. Some ramifications are that comparing the local fit with the Phillips Curve, the unemployment rate for stable wages occurs around 5.8% rather than 5.5%, and the unemployment rate corresponding to 2% inflation is about 2.25% for both. In any case, these figures will not be significantly different if we include the effect of confidence intervals.

As a final remark concerning our local adaptive procedure, the proper choices of the kernel and the over-smoothing bandwidth for the initial estimation stage in steps (a) and (b) remain to be decided. We hope to pursue further investigation along these lines. Once these issues are settled, the method described here in this article can be a useful data-exploratory tool, no matter if one wants to treat the curve estimates in their own right, or if one seeks a reference point from which parametric models can be built.

Acknowledgement. We thank Jim Dykes and Hong Zhou for computational help.

Postscript. As a result of the NATO ASI, the authors learned of the work of Vieu (1990) on local bandwidth choice for the Nadaraya-Watson kernel regression estimates based on local cross-validation. It would be interesting to investigate if Vieu's criterion is equivalent to the method proposed in the present paper.

Reference to Postscript

Vieu, P. (1990). Nonparametric regression: Optimal local bandwidth choice. To appear in J. Roy. Statist. Soc. B.

REFERENCES

Blinder, A.S. (1989) 'The economic myths that the 1980's exploded', in the column Economic Viewpoint, appeared in the November 27, 1989 issue of Business Week, p. 22.

Chu, C.K. and Marron, J.S. (1989) 'Choosing a kernel regression estimator', manuscript.

Desai, M. (1984) 'Wages, prices and unemployment a quarter century after the Phillips Curve', in Econometrics and Quantitative Economics, eds. Hendry D.F. and Wallis, D.F., Oxford: Blackwell.

Epanechnikov, V.A. (1969) 'Nonparametric estimation of a multivariate probability density', Theory Probab. Appl. 14, 153-158.

Friedman, M. (1968) 'The role of monetrary policy: 1967 Presidential Address of the American Economic Association', Amer. Econ. Rev., March, 1-17.

Gasser, Th., Müller, H.-G., Köhler, W., Molinari, L. and Prader, A. (1984) 'Nonparametric regression analysis of growth curves', Ann. Statist. 12, 210-229.

Gilbert, C.L. (1976) 'The original Phillips Curve estimates', Economica, February, 51-57.

Hall, P. and Marron, J.S. (1987a) 'Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation', Prob. Theory Relat. Fields 74, 567-581.

Hall, P. and Marron, J.S. (1987b) 'On the amount of noise inherent in bandwidth selection for a kernel density estimator', Ann. Statist. 15, 163-181.

Härdle, W. and Marron, J.S. (1985) 'Optimal bandwidth selection in nonparametric regression function estimation', Ann. Statist. 13, 1465-1481.

Härdle, W., Hall, P. and Marron, J. (1988) 'How far are automatically chosen regression smoothing parameters from their optimal?' J. Amer. Statist. Assoc. 83, 86-95.

Härdle, W. and Marron, J.S. (1989) 'Bootstrap simultaneous error bars for nonparametric regression', manuscript.

Humphrey, T.M. (1986) From Trade-offs to Policy Ineffectiveness: A History of the Phillips Curve, Richomond, VA: Federal Reserve Bank of Virginia, October.

Jennen-Steinmetz, C. and Gasser, Th. (1988) 'A unifying approach to nonparametric regression estimation', J. Amer. Statist. Assoc. 83, 1084-1089.

Lipsey, R.G. (1960) 'The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1862-1957', Economica, February, 1-31.

Mack, Y.P. (1990) 'Local bandwidth choice in kernel regression via initial mean squared error estimation: The random X case', manuscript.

Mack, Y.P. and Müller, H.-G. (1987) 'Adaptive nonparametric estimation of a multivariate regression function', J. Multivariate Anal. 23, 169-182.

Mack, Y.P. and Müller, H.-G. (1989a) 'Derivative estimation in nonparametric regression with random predictor variable', Sankhyā Ser. A, 51, 59-72.
Mack, Y.P. and Müller, H.-G. (1989b) 'Convolution type estimators for nonparametric regression'. Statist. Probab. Lett. 7, 229-239.
Mack, Y.P. and Silverman, B.W. (1982) 'Weak and strong uniform consistency of kernel regression estimates', Z. Wahr. verw. Geb. 61, 405-415.
Müller, H.-G. (1985) 'Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators', Statist. & Decisions Supplement 2, 193-206.
Müller, H.-G. and Stadtmüller, U. (1987a) 'Variable bandwidth kernel estimators of regression curves', Ann. Statist. 15, 182-201.
Müller, H.-G. and Stadtmüller, U. (1987b) 'Estimation of heteroscedasticity in regression analysis', Ann. Statist. 15, 610-625.
Nadaraya, E.A. (1964) 'On estimating regression', Theory Probab. Appl. 9, 141-142.
Oliver, F.R. (1986) 'Direct least squares estimation of Phillips' function', J. Applied Statist. 13, 221-225.
Park, B.U. and Marron, J.S. (1990) "Comparison of data-driven bandwidth selectors', J. Amer. Statist. Assoc. 85, 66-72.
Phillips, A.W. (1958) 'The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861-1957', Economica, November, 283-299.
Rice, J. (1984a) 'Boundary modification for kernel regression', Comm. Statist. Th.-Meth. 13, 893-900.
Rice, J. (1984b) 'Bandwidth choice for nonparametric regression estimators', Ann. Statist. 12, 1215-1230.
Rosenblatt, M. (1969) 'Conditional probability density and regression estimates', in Multivariate Analysis II, ed. Krishnaiah, 25-31.
Routh, G. (1959) 'The relation between unemployment and the rate of change of money wages: A comment', Economica, November, 299-315.
Schuster, E. (1972) 'Joint asymptotic distribution of the estimated regression function at a finite number of distinct points', Ann. Math. Statist. 43, 84-88.
Schuster, E. and Yakowitz, S. (1979) 'Contributions to the theory of nonparametric regression, with applications to system identification', Ann. Statist. 7, 139-149.
Staniswallis, J.G. (1989) 'Local bandwidth selection for kernel estimates', J. Amer. Statist. Assoc. 84, 276-283.
Watson, G.S. (1964) 'Smooth regression analysis', Sankhyā Ser. A, 26, 359-372.
Wulwick, N.J. (1987) 'The Phillips Curve: Which? Whose? To do what? How?', South. Econ. J. 53, 834-857.
Wulwick, N.J. (1989) 'Phillips' approximate regression', Oxford Econ. Papers 41, 51-57.
Wulwick, N.J. and Mack, Y.P. (1990). 'A kernel regression of Phillips' data'. Manuscript.

V. Dependent Data

NONPARAMETRIC REGRESSION METHODS FOR REPEATED MEASUREMENTS

Adelchi Azzalini
Dept. of Statistical Sciences
University of Padua
Italy

Adrian Bowman
Dept. of Statistics
University of Glasgow
U.K.

ABSTRACT. The possible use of nonparametric smoothing in the context of repeated measurements analysis is discussed and a test for interaction between time and group is described in detail as a prototype of other possible uses. Autocorrelation of the error terms is considered in the case of an AR(1) dependence structure. The test has essentially exact level when the autocorrelation parameter is known. An example from real data is presented and a set of simulations is used to evaluate the level and the power of the test when the autocorrelation parameter is estimated.

1 Introduction

In the last decade, and particularly in the last few years, there has been an enormous production of research work for repeated measurements analysis, with particular emphasis on continuous data with an assumption of normality. A key feature of repeated measurements data is the presence of time trend and the study of this, and its possible interaction with covariates, is of considerable importance in many cases.

The most common way to handle the time component is to assume that the trend belongs to a certain parametric class of functions. While such an approach has clear advantages, it suffers from the problem of bias when the parametric class is not chosen appropriately. On the other hand, regarding the values of the time trend at the observed time points as unrelated quantities may lead to a large number of parameters and decrease of efficiency, unless the number of time points is small.

These remarks suggest an intermediate approach which retains the smoothness assumption associated with parametric models without prescribing a specific class of functions. Such an approach can be pursued through smooth nonparametric estimation of the time trend. Pooling information from neighbouring time points can be expected to produce more efficient estimates than regarding nearby points as totally unrelated.

Of the many hypothesis testing problems in the framework of repeated measurements, we shall deal with testing the equality of k trend curves, associated with k groups of subjects, which effectively amounts to testing interaction between time and group. However, it is clear that the same technique described here can be adapted to a number of other testing problems, such as the existence of a time effect.

Hart & Wehrly (1986) have already made use of nonparametric regression methods in the context of repeated measurements analysis, but they were concerned with the estimation

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 377–387.
© 1991 *Kluwer Academic Publishers.*

problem from an asymptotic point of view. Raz (1989) considers essentially the same problem of the present paper, and also makes use of nonparametric smoothing, but assesses significance through permutation tests.

2 A Test Based on Smoothing

2.1 The form of the test statistic

For ease of exposition, we consider the balanced case which has k groups of n subjects in each group and each subject is observed at T time points. The response value for subject j in group i at time t will be denoted by y_{ijt} $(i = 1, \ldots, k; j = 1, \ldots, n; t = 1, \ldots, T)$.

In addition to the time trend, denoted by $f_i(t)$ for the i-th group of subjects $(i = 1, \ldots, k)$, we also wish to account for individual subject variability. Practical experience has shown that this component is very often present, although it is generally regarded as a nuisance parameter.

Consider then the following model

$$y_{ijt} = \mu_j + f_i(t) + \varepsilon_{ijt} \tag{1}$$

where μ_j represents the specific effect of the j-th subject of group i, $f_i(t)$ the smooth trend component for the i-th group and the ε_{ijt}'s are 0-mean normal random errors which are assumed to be correlated when they refer to the same subject, and otherwise to be independent. Denote by Ω the covariance matrix of $(\varepsilon_{ij1}, \ldots, \varepsilon_{ijT})$, for each i and j; in particular, we shall consider the case when Ω refers to an AR(1) process with lag-1 autocorrelation ρ.

In constructing a test statistic for the null hypothesis that

$$f_i(t) = f(t) \qquad (i = 1, \ldots, k; t = 1, \ldots, T) \tag{2}$$

where $f(t)$ represents an unknown smooth function, one would like to ensure invariance with respect to the μ_j's in (1). This is easily achieved by working with the residuals

$$u_{ijt} = y_{ijt} - \overline{y}_{ij\cdot} \tag{3}$$

where $\overline{y}_{ij\cdot}$ is the average value of the (i, j)-th subject; also write $\overline{u}_{it} = \sum_j u_{ijt}/n$. Notice that, since we have allowed the possibility that the μ_j's are random effects, then any constant component of Ω is automatically removed in considering the u_{ijt}'s. Furthermore, if (1) contained a group effect, then transformation (3) would remove it; therefore, for the purpose of the present paper, such an effect has not been inserted in (1).

The specific group trend can be estimated by various nonparametric regression estimates. Let us consider for instance the Nadaraya-Watson estimate

$$\hat{f}_i(t) = \frac{\sum_{t'} K\left(\frac{t-t'}{h}\right) \overline{u}_{it}}{\sum_{t'} K\left(\frac{t-t'}{h}\right)} \tag{4}$$

where $K(\cdot)$ is a kernel function and h is a smoothing parameter; however, the method described below is applicable to any estimator which is linear in the data, not only to those of type (4). Analogously, under the hypothesis (2), $f(t)$ can be estimated by

$$\hat{f}_\cdot(t) = \frac{\sum_i \sum_{t'} K\left(\frac{t-t'}{h}\right) \overline{u}_{it}}{\sum_i \sum_{t'} K\left(\frac{t-t'}{h}\right)} = \frac{\sum_{t'} K\left(\frac{t-t'}{h}\right) \left(\sum_i \overline{u}_{it}/n\right)}{\sum_{t'} K\left(\frac{t-t'}{h}\right)}. \tag{5}$$

Then a natural test statistic to consider for hypothesis (2) is

$$F = \frac{\sum_i \sum_t [\hat{f}_i(t) - \hat{f}_.(t)]^2}{\sum_i \sum_j \sum_t (u_{ijt} - \overline{u}_{it})^2} \tag{6}$$

high values being significant. Clearly, this statistic replaces the trend estimate of the standard statistic from linear model theory with a nonparametric estimate. Althought not exactly of the same form, (6) stems from the same logic underlying the pseudo-likelihood ratio test considered by Azzalini, Bowman & Härdle (1989).

As well as being invariant with respect to the μ_j's, (3) is scale invariant and, under the null hypothesis, its distribution does not depend on the true form of $f(t)$. Moreover, under the null hypothesis, the bias of the estimate $\hat{f}_i(t)$ cancels out with the identical bias of $\hat{f}_.(t)$ in the numerator of (6), and so this unpleasant effect of the Nadaraya-Watson estimator, as well as of virtually all nonparametric estimators, is avoided.

2.2 Distribution of the test statistic

In this section, we develop a technique for computing the null distribution of F at a given point w. On writing

$$y = (y_{111}, \ldots, y_{11T}, y_{121}, \ldots, y_{knT})',$$

we have

$$\Pr(F > w) = \Pr(y'A_0 y - w y' A_1 y > 0) = \Pr(y'Ay > 0) \tag{7}$$

where A_0, A_1 are suitable matrices associated to the numerator and denominator quadratic forms of F, and $A = A_0 - wA_1$. Now the problem is reduced to the computation of the distribution of a quadratic form of y, treating y as if it had 0-mean and correlation matrix equal to

$$\mathrm{var}(y) = I_{nk} \otimes \Omega$$

because of the invariance properties of F.

Johnson and Kotz (1972, p.150–3) provide some standard results on the distribution of quadratic forms like $y'Ay$. Specifically, evaluation of (4) can be shown to be equivalent to computation of the distribution of a linear combination of independent χ^2 variates with coefficients given by the eigenvalues of $(I_{nk} \otimes \Omega)A$. Various algorithms and routines exist for computing the distribution of such a linear combination; a commonly used routine which implements an algorithm due to Imhof is given by Koerts and Abrahamse (1969).

However, for most practical purposes, the exact evaluation of (4) is at the same time unnecessary and computationally burdensome. It is widely recognised that one can get a reasonably accurate approximation to a distribution function through replacement of the original distribution by a member of a parametric class, matching the first three or four moments with the original ones; Pearson (1963) and Solomon & Stephens (1978) provide some supporting numerical evidence. In our case, from formula (22) of Johnson & Kotz (1972, p.153), the s-th cumulant of $y'Ay$ is

$$\kappa_s = 2^{s-1}(s-1)!\,\mathrm{tr}((I_{nk} \otimes \Omega)A)^s \tag{8}$$

which does not require explicit computation of the eigenvalues of $(I_{nk} \otimes \Omega)A$. In the numerical work of this paper, the Johnson family of curves has been fitted to the distribution of $y'Ay$, matching the first four moments with the aid of the routines of Hill (1976) and Hill, Hill & Holder (1976). Comparison with the exact computations based on the routine

of Koerts and Abrahams (1969), shows the approximate probabilities based on the Johnson curves to be sufficiently accurate for practical purposes in this type of problem.

In the following, we shall derive an explicit expression for A. Additionally, since the matrices involved in (8) are square of order nkT and these matrices must be raised at successive powers, we shall make simplifications to reduce the actual computations to matrices of order T.

Consider first A_0. Denote by $\hat{f}_i$ the T-vector whose t-th component is given by (4), and by $\hat{f}.$ the T-vector whose t-th component is given by (5); further let

$$\overline{u}_i = \begin{pmatrix} \overline{u}_{i1} \\ \vdots \\ \overline{u}_{iT} \end{pmatrix}, \quad \overline{u} = \begin{pmatrix} \overline{u}_1 \\ \vdots \\ \overline{u}_k \end{pmatrix},$$

and $\overline{u}. = \sum_i \overline{u}_i/n$, whose dimensions are $T \times 1, kT \times 1, T \times 1$, respectively. Since $\hat{f}_i(t)$ is a linear function of the $\overline{u}_{it}$'s, there is a $T \times T$ smoothing matrix W such that

$$\hat{f}_i = W\overline{u}_i$$

so we can write

$$\begin{aligned} \begin{pmatrix} \hat{f}_1 - \hat{f}. \\ \vdots \\ \hat{f}_k - \hat{f}. \end{pmatrix} &= (I_k \otimes W) \begin{pmatrix} \overline{u}_1 - \overline{u}. \\ \vdots \\ \overline{u}_k - \overline{u}. \end{pmatrix} \\ &= (I_k \otimes W)(I_{kT} - k^{-1} J_k \otimes I_T)\overline{u} \\ &= (I_k \otimes W)(M_k \otimes I_T)\overline{u} \\ &= (M_k \otimes W)\overline{u} \end{aligned}$$

where J_k is the $k \times k$ matrix with all 1's, and $M_k = I_k - k^{-1}J_k$. Since

$$\overline{u} = n^{-1}(I_k \otimes 1'_n \otimes I_n)(I_{nk} \otimes M_T)y$$

where 1_n denotes the $n \times 1$ vector of all 1's, then the numerator of F can be written as

$$\begin{aligned} \|(M_k \otimes W)\overline{u}\|^2 &= \|n^{-1}(M_k \otimes W)(I_k \otimes 1'_n \otimes I_T)(I_{nk} \otimes M_T)y\|^2 \\ &= \|n^{-1}(M_k \otimes W)(I_k \otimes 1'_n \otimes M_T)y\|^2 \end{aligned}$$

Finally we have

$$\begin{aligned} A_0 &= n^{-2}(I_k \otimes 1_n \otimes M_T)[M_k \otimes (W'W(1'_n \otimes M_T))] \\ &= n^{-2}[M_k \otimes J_n \otimes (M_T W' W M_T)]. \end{aligned}$$

The denominator of F can be written as

$$\begin{aligned} &\sum_i \sum_j \sum_t (y_{ijt} - \overline{y}_{ij.} - \overline{y}_{i.t} + \overline{y}_{i..})^2 = \\ &\quad = \|y - T^{-1}(I_{nk} \otimes J_T)y - n^{-1}(I_k \otimes J_n \otimes I_n)y + (nT)^{-1}(I_k \otimes J_{nT})y\|^2 \\ &\quad = \|I_k \otimes (I_n \otimes I_T - T^{-1} I_n \otimes J_T - n^{-1} J_n \otimes J_T + (nT)^{-1} J_n \otimes J_T)y\|^2 \end{aligned}$$

where the "dot" notation means averaging with respect to the corresponding index. After some simplification, one obtains

$$A_1 = I_k \otimes M_n \otimes M_T.$$

We are now able to obtain the relevant terms of (8), namely

$$\begin{aligned} \mathrm{var}(y)A &= n^{-2}(I_k \otimes I_n \otimes \Omega)[M_k \otimes J_n \otimes (M_T W' W M_T) - wI_k \otimes M_n \otimes M_T)] \\ &= n^{-2} M_k \otimes J_n \otimes \Omega M_T W' W M_T - wI_k \otimes M_n \otimes \Omega M_T \end{aligned}$$

leading to

$$\mathrm{tr}\{\mathrm{var}(y)A\} = n^{-1}(k-1)\,\mathrm{tr}(\Omega M_T W' W M_T) - wk(n-1)\mathrm{tr}(\Omega M_T)$$

using the fact that $\mathrm{tr}(B \otimes C) = \mathrm{tr}(B)\mathrm{tr}(C)$. In taking the powers of $\mathrm{var}(y)A$, much simplification occurs because the A_0 and A_1 are orthogonal, a property shared with the quadratic forms of standard linear models theory. After some algebra, one gets

$$\mathrm{tr}[\{\mathrm{var}(y)A\}^s] = n^{-s}(k-1)\mathrm{tr}\{(\Omega M_T W' W M_T)^s\} + (-w)^s k(n-1)\mathrm{tr}\{(\Omega M_T)^s\} \quad (9)$$

for s=2,3,4.

It is not difficult to relax the assumption of equal group size; we only sketch its handling. Denote by $n_1, \ldots, n_k$ the sample sizes of the k groups and let

$$V = \mathrm{diag}(1'_{n_1}/n_1, \ldots, 1'_{n_k}/n_k),$$

a block diagonal matrix of order $N = \sum_i n_i$. The previous definition of $\overline{u}$ must then be replaced by

$$\overline{u} = (V \otimes I_T)(I_N \otimes M_T)y.$$

The denominator of F in matrix notation is

$$\|y - (I_N \otimes T^{-1}J_T)y - (L \otimes I_T)y + (L \otimes T^{-1}J_T)y\|^2$$

where

$$L = \mathrm{diag}(J_{n_1}/n_1, \ldots, J_{n_k}/n_k).$$

3 Practical implementation

Expression (9) coupled with (8) allow us to compute (7) quite accurately when Ω is known and h is specified. Since in practice Ω is not known, an estimate is required. This issue and the choise of h will now be discussed.

3.1 Estimation of ρ

As already mentioned, we shall restrict ourselves to the case when Ω corresponds to the correlation matrix of an AR(1) process with parameter ρ. This assumption is clearly restrictive and may be especially questioned in a nonparametric framework when one would rather like to relax assumptions. Apart from cases where there are specific reasons to believe an AR(1) structure is adequate, such a choice can be viewed as a practical means of accounting

for the bulk of the correlation structure through a first order approximation. In addition, the components of Ω are essentially nuisance parameters, and so accurate estimation is a subsidiary issue.

In order to estimate ρ, and hence Ω, independently of the μ_j's and of the $\hat{f}_i(t)$'s, consider the quantities

$$d_{ijt} = y_{ijt} - \overline{y}_{ij\cdot} - \overline{y}_{i\cdot t} + \overline{y}_{i\cdot\cdot}$$

which is obtained by subtracting the mean group value for each time point and then subtracting the subject mean from the new values (or *vice versa*). The d_{ijt}'s can be used as basic quantities to obtain an estimate of ρ, for example through their sample lag-1 autocorrelation.

In this framework, however, one must bear in mind that T is usually not large and severe bias will be present. This remark leads to consideration of how Ω relates to the covariance matrix of $(d_{ij1}, \ldots, d_{ijT})'$. Subtraction of the group mean at each time point does not alter the correlation structure of the data, since Ω is transformed into $\frac{n-1}{n}\Omega$, but subtraction of the subject mean does lead to severe bias. Simple corrective action is provided by a suitable modification of Daniels (1956) estimate, namely

$$\tilde{\rho} = \frac{2\sum_i \sum_j \sum_{t=2}^{T} d_{ijt} d_{ij(t-1)}}{\sum_i \sum_j \sum_{t=2}^{T} (d_{ijt}^2 + d_{ij(t-1)}^2)}$$

which suffers from less bias than the sample autocorrelation. Further improvement is provided by two modifications obtained by Azzalini & Frigo (1990), the simpler of which is

$$\tilde{\rho}_c = \min\left(\frac{\tilde{\rho} + c}{1 - c}, 1\right)$$

where $c = (T-1)^{-1}$; the 'min' is taken to avoid values above 1. Although $\tilde{\rho}_c$ was originally obtained for the case of a single group of subjects, it is the length T which is the more influential factor in the bias of $\tilde{\rho}$. The above modification can therefore be expected to work well even in the present case.

3.2 Choice of the bandwidth

Another issue connected with the present methodology is the choice of the smoothing parameter h. This choice does not matter when the null hypothesis is true, since the level of the test is essentially exact for values of h.

The choice of h becomes relevant when dealing with the alternative hypothesis. It is intuitive that in order to make use of the smoothness assumption, and hence to improve power, h must not be very small. On the other hand, a large h would remove all nonlinear features of the data. Instead of selecting a single value of h, we have found it preferable to consider a plot of the observed significance level versus the smoothing parameter, which we call a 'significance trace', removing the extreme values of h. Visual inspection of the significance trace is then recomended.

If a single value of h is required, a rough guide is to select a value which averages some 8 or 12 data points, with most of the weight on the central 6 or 8 points. Using a normal kernel this recomendation would give h about $2/T$ or $3/T$. In the rest of the paper, we shall express h as a multiple of $1/T$, to ease comparison between cases with different values of T.

We illustrate the use of the significance trace with the aid of data of Andersen, Jensen & Schou (1981) who refer to 10 patients observed at $T = 14$ equally spaced time points. These data have been analysed by Olkin & Vaeth (1981) among others, dividing the 10 patients in $k = 2$ groups of $n = 5$ subjects each; the group mean values for the two groups are displayed in Figure 1a. Figure 1b shows the significance trace for these data, providing strong evidence of interaction between time and groups. This conclusion is different from the one of Olkin & Vaeth (1981), but they have analysed the data only at 3 time points. Notice that, for very low values of h, the observed significance decreases appreciably, in relative terms, confirming empirically the previous remark on the inappropriateness of very small values of h.

While nonparametric regression curves $\hat{f}_i(t)$ corresponding to values of h from 0.05 to 0.5 are different in curvature and amplitude, it is remarkable that the observed significances produced by the pseudo-likelihood ratio test are very similar. This underlines the difference between estimation and inference. Another appealing feature of this procedure is that the numerator of (6) is free of bias, as already remarked.

4 Simulation results

Some simulation work has been conducted with several goals:

- to evaluate the effect on the level of the test of replacing the exact value of ρ by its estimate $\tilde{\rho}_c$;
- to explore the effect on the power of different choices of h;
- to examine the behaviour of the test when Ω is incorrectly assumed to of AR(1) structure.

The simulations have been conducted using a normal kernel. In all cases, $k = 2$ and $T = 18$ were set, while n was 6 or 12 and the number of replicates was 1000. Some samples were generated under the null hypothesis; hence $f(t) = 0$ was used, because of the invariance property of F in this case. These cases have a 0 in the column headed by H. Other samples, with $H = 1$, were generated under the alternative hypothesis; in this case the interaction term was

$$\pm(0.4\log(t) - 0.8088) \qquad (t = 1, \ldots, 18)$$

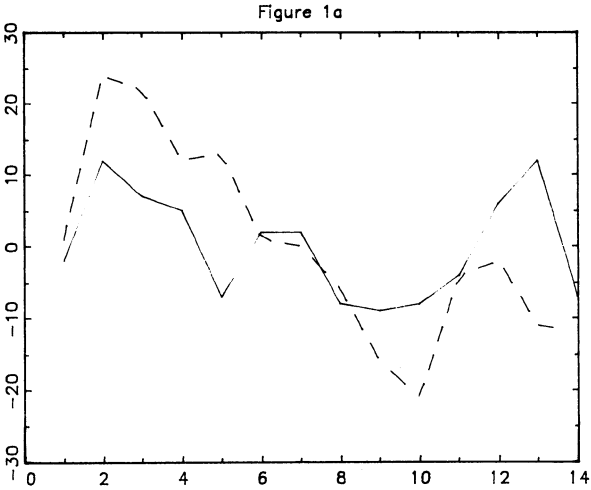
where the + applies to one group and the − sign to the other, and the constant 0.8088 centers the logaritmic curves around 0; this form of the interaction term is the same as that of Raz (1989, p. 860). For each generated set of data, the observed significance level was evaluated and the null hypothesis rejected if this was less than 5%.

Beside generating error terms from the supposed AR(1) structure, some simulations generated data errors from an AR(2) model of the form

$$\varepsilon_t = \alpha_1\varepsilon_{t-1} + \alpha_2\varepsilon_{t-2} + \eta_t$$

where the η_t's are white noise N(0,1). Many simulations refer to $\alpha_1 = 0.99, \alpha_2 = -0.4$, the same case, quite distinct from AR(1), considered by Raz (1989). Oher simulations explore the behaviour of the test for other combinations of (α_1, α_2).

The results of the simulations are summarised in Table 1 which gives the estimated mean value of $\tilde{\rho}_c$, the power of the test in percentage when ρ is supposed known and the corresponding power when ρ when is estimated from the data.



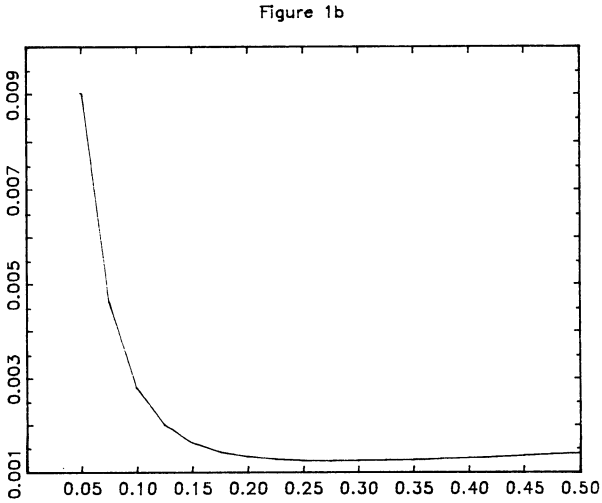


Figure 1: For the data of Andersen, Jensen & Schou (1981), Figure 1a shows the mean value at each given time point, separately for the two groups, Figure 1b plots the significance trace.

Table 1: Estimated power (γ) of F at 5% level

H	α_1	α_2	n	h	$E(\tilde{\rho}_c)$	$\gamma(\rho)\%$	$\gamma(\tilde{\rho}_c)\%$
0	0.00	0.0	6	0.10	0.000	5.2	5.1
0	0.50	0.0	6	0.10	0.489	5.0	6.1
0	0.50	0.0	12	0.10	0.492	5.4	5.9
0	-0.50	0.0	6	0.10	-0.492	4.7	5.0
0	0.99	-0.4	6	0.05	0.758	3.3	2.3
0	0.99	-0.4	6	0.10	0.758	2.3	1.3
0	0.99	-0.4	12	0.10	0.760	1.7	1.2
0	0.00	0.8	6	0.10	-0.357	15.8	40.5
0	0.00	-0.8	6	0.10	-0.054	0.0	0.0
1	0.00	0.0	6	0.01	0.001	79.0	78.9
1	0.00	0.0	6	0.05	0.001	96.9	96.7
1	0.00	0.0	6	0.10	0.000	97.9	97.2
1	0.00	0.0	6	0.20	0.001	98.0	98.1
1	0.50	0.0	6	0.05	0.489	57.3	56.9
1	0.50	0.0	6	0.10	0.488	57.7	58.1
1	0.50	0.0	6	0.20	0.489	59.2	60.1
1	-0.50	0.0	6	0.10	-0.492	100.0	100.0
1	0.00	0.0	12	0.10	0.002	100.0	100.0
1	0.50	0.0	12	0.10	0.491	89.0	89.1
1	0.71	0.0	12	0.10	0.688	58.1	56.6
1	0.99	-0.4	12	0.10	0.760	44.0	36.7

Inspection of the first 4 rows of Table 1 shows that, if the correlation structure of the data is indeed of AR(1) type ($\alpha_2 = 0$) the actual level of the proposed methods is not appreciably affected by the fact that ρ is replaced by $\tilde{\rho}_c$.

Another pleasant feature is that changing h over a fairly wide range of values does not seem to affect the power, as can be seen from the first 7 rows which have $H = 1$. An ecception occurs when h is 0.01, therefore showing the improvement provided by smoothing.

When the assumption of AR(1) correlation structure is actually incorrect ($\alpha_2 \neq 0$), the test can have an actual level very different from the nominal one. An assessment on the resulting loss of power can be obtained by comparing the last two rows of Table 1 which have nearly equal lag-1 autocorrelation. We recall that, for an AR(2) process, the lag-1 autocorrelation is $\alpha_1/(1-\alpha_2)$. Further work is necessary to make the test more robust to incorrect specification of the autocorrelation structure, and possibly from the assumption of normality.

Acknowledgements

This work has been carried out over a series of visits between the author's institutions. The financial support of the Science and Engeneering Research Council and of the Italian Ministry of Education is gratefully acknowledged.

References

Andersen, A.H., Jensen, E.B. & Schou, G. (1981). Two way analysis of variance with autocorrelated errors. *Int. Statist. Rev.* **49**, 153–7.

Azzalini, A., Bowman, A. W. & Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76**, 1–11.

Azzalini, A. & Frigo, A. C. (1990). An explicit nearly unbiased estimate of the AR(1) parameter for repeated measurements. *J. Time Series Analysis*, to appear.

Daniels, H. E. (1956). The approximate distribution of serial correlation coefficients. *Biometrika* **43**, 169–85.

Hart, J.D. & Wehrly, T.E. (1986). Kernel regression estimation using repeated measurements data. *J. Amer. Statist. Assoc.* **81**, 1080–8.

Hill, I. D. (1976). Algorithm AS100: Normal-Johnson and Johnson-Normal transformations. *Appl. Statist.* **25**, 190–2.

Hill, I. D., Hill, R. & Holder, R. L. (1976). Algorithm AS99: Fitting Johnson curves by moments. *Appl. Statist.* **25**, 180–9.

Koerts, J. & Abrahamse, A. P. J. (1969). *On the theory and application of the general linear model.* Rotterdam: Rotterdam University Press.

Johnson, N. L. & Kotz, S. (1972). *Distributions in statistics: continuous multivariate distributions.* New York: Wiley.

Olkin, I. & Vaeth, M. (1981). Maximum likelihood estimation in a two-way analysis of variance with correlated errors in one classification. *Biometrika* **68**, 653–60.

Pearson, E. S. (1963). Some problems arising in approximating to probability distributions using moments. *Biometrika* **50**, 94–111.

Raz, J. (1989). Analysis of repeated measurements using nonparametric smoothers and randomization tests. *Biometrics* **45**, 851–71.

Solomon, H. & Stephens, M. A. (1978). Approximation to density functions using Pearson curves. *Technometrics* **73**, 153–160.

NONPARAMETRIC PREDICTION FOR UNBOUNDED ALMOST STATIONARY PROCESSES

Denis BOSQ
Université Pierre et Marie Curie
Laboratoire de Statistique
Théorique et Appliquée
4, place Jussieu
75252Paris cedex 05
France

ABSTRACT. Recent results on nonparametric prediction imply that the underlying process is bounded. This condition is rather cumbersome for applications.

Here, we study a class of predictors for unbounded mixing processes. This class contains the classical convolution kernel and orthogonal series predictors.

On the other hand, numerical results show that it is not necessary to remove seasonal fluctuations before using the non parametric prediction method. We make a theoretical study of this property by considering almost stationary processes.

Finally, we deal with the practical use of the method and compare it with the BOX-JENKINS method.

1. Introduction

In the present paper, we study a class of nonparametric predictors for a mixing process, based on regression estimates.

Such an approach has been investigated by BOSQ (1983a, 1983b), COLLOMB (1984, 1985), BOSQ-DELECROIX (1985), SARDA-VIEU (1985,), COLLOMB-HÄRDLE (1986), COLLOMB-HÄRDLE-HASSANI (1987), TRUONG-STONE (1988), BOSQ (1989), GANNOUN (1989), CARBON-DELECROIX (1989).
There is a very large number of related papers devoted to regression estimation. An extensive bibliography can be found in GYORFI-HÄRDLE-SARDA-VIEU (1989) and HÄRDLE (1990).

All the recent results suppose that the data belong to a fixed compact set. This is rather cumbersome for the applications. On the contrary, our results are valid for unbounded processes.

AMS Classification and Key Words : 62G - Density estimation, regression estimation, prediction, nonstationary processes.

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 389–403.
© 1991 *Kluwer Academic Publishers.*

On the other hand, numerical results show that it is not necessary to remove seasonal fluctuations before using the nonparametric prediction method. We make a theoretical study of this property by considering almost stationary processes.

We say that an $\mathbb{R}^d$-valued process $(X_t,\ t \in \mathbb{Z})$ is <u>almost stationary</u> if there exists a deterministic bounded sequence $(s(t),\ t \in \mathbb{Z})$ and a strictly stationary process $(Y_t,\ t \in \mathbb{Z})$ such that

$$X_t = s(t) + Y_t \quad , \quad t \in \mathbb{Z} \tag{1}$$

Let $X_t = \left(X_t^{(1)}\ ,\ X_t^{(2)}\right)$, $t \in \mathbb{Z}$ be such a process. We define the <u>pseudo-regression</u> $\bar{m}$ of $\left(X_t^{(2)}\right)$ with respect to $\left(X_t^{(1)}\right)$ and show that, for suitable $\delta > 0$ and (x_n), we have

$$n^{\delta} \sup_{\|x\| \leq x_n} \left| m_n(x) - \bar{m}(x) \right| \xrightarrow{a.s.} 0 \ , \tag{2}$$

where m_n is a nonparametric regression estimate.

From this, we infer asymptotic results for predictors in the p-Markovian case. Finally, we consider practical situations and exhibit the superiority of the nonparametric method on the so-called BOX-JENKINS's method.

It should be noticed that we do not try to estimate s . For this problem least squares or spectral methods could be used (see ROSENBLATT (1985) and GOURIEROUX-MONFORT (1990)). Here we emphasize robustness of nonparametric predictors when small deviations from stationarity occur.

2. Pseudo-density estimation

Let $(K_r\ ,\ r \in I)$ be a family of measurable $\mathbb{R}^{2d}$-$\mathbb{R}$ mappings indexed by I , an unbounded subset of $\mathbb{R}_+$. We make the following assumptions :

$[\mathbf{A}_1.$ $\quad \sup_{(x,u) \in \mathbb{R}^{2d}} |K_r(x,u)| \leq a_r < +\infty \quad , \quad r \in I$

$[\mathbf{A}_2.$ $\quad \sup_{u \in \mathbb{R}^d} |K_r(x',u) - K_r(x,u)| \leq b_r \|x'-x\|^{\ell} \ , \ r \in I$

Here a_r, b_r, ℓ are positive constants and $\|.\|$ denotes the Euclidean norm in $\mathbb{R}^d$.

We consider an almost stationary process $(X_t,\ t \in \mathbb{Z})$ satisfying (1) and suppose that

A_3. Y_t admits a density f and there exists a <u>pseudo-density</u> $\bar{f}$ defined by $\sup_{x\in\mathbb{R}^d} |\bar{f}_n(x) - \bar{f}(x)| \to 0$ as $n \to \infty$, where

$$\bar{f}_n(x) = \frac{1}{n}\sum_{t=1}^{n} f(x-s(t)) .$$

Note that if f is bounded and s periodic, A_3 is fulfilled and $\bar{f}(.) = \frac{1}{T}\sum_{t=1}^{T} f(. - s(t))$ where T denotes a period for s. If f satisfies a Lipschitz condition and if $s(n) \to s$ (respectively, $n^{-1}\sum_{t=1}^{n} |s(t)| \to 0$), then $\bar{f}(.) = f(.-s)$ (respectively, $\bar{f} = f$).

A_4. For each $\delta_0 \in \mathbb{R}_+$, we have for $r \to \infty$, $r \in I$

$$\sup_{|\delta|\leq\delta_0} \sup_{x\in\mathbb{R}^d} \left| \int_{\mathbb{R}^d} K_r(x+\delta, u+\delta) f(u) - f(x) \right| \to 0 .$$

Note that if $K_r(x,u)$ is a function of $x-u$, then "$\sup_{|\delta|\leq\delta_0}$" may be omitted.

A_5. (Y_t) is geometrically strong mixing.

Now we define a mean-density estimate by

$$f_n(x) = \frac{1}{n}\sum_{t=1}^{n} K_{r_n}(x, X_t) , \quad x \in \mathbb{R}^d , \tag{3}$$

then we have

Proposition 2.1. Under $A_1 \to A_5$, if $r_n \to +\infty$ and if for each $\eta > 0$,

$$\sum_n b_{r_n}^{d/\ell} x_n^d \operatorname{Log} n \exp\left(-\eta \frac{n^\lambda}{a\, r_n}\right) < +\infty \tag{4}$$

where $0 < \lambda < \frac{1}{2}$, then

$$\sup_{\|x\|\leq x_n} |f_n(x) - \bar{f}(x)| \xrightarrow{a.s.} 0 \tag{5}$$

Corollary 2.2. If $r_n = c_1 n^a$, $a_r = c_2 r^b$, $b_r = c_3 r^c$ with a,b,c,c_1,c_2,c_3 strictly positive and $ab < \frac{1}{2}$, then if

$$n^\delta \sup_{\|x\|\leq n^\gamma} |E f_n(x) - \bar{f}(x)| \to 0 , \tag{6}$$

where $0 \leq \delta < \frac{1}{2} - ab$ and $\gamma > 0$, we obtain

$$n^{\delta} \sup_{\|x\| \leq n^{\gamma}} |f_n(x) - \bar{f}(x)| \xrightarrow{a.s.} 0 . \tag{7}$$

Condition (6) has to be checked in each special case. A rather general criterion based on a "three functions" theorem can be found in BOSQ (1988).

Corollary 2.2. applies to the well-known PARZEN-ROSENBLATT <u>kernel estimate</u>

$$f_n^K(x) = \frac{1}{n} \sum_{t=1}^{n} r_n^d \, K(r_n(x-X_t)) , \quad x \in \mathbb{R}^d . \tag{8}$$

Here K is an integrable bounded function with $\int K(u)du = 1$ which satisfies a Lipschitz condition.

Corollary 2.3.

a) If f is uniformly continuous and if $\frac{r_n^d}{n^{\beta}} \operatorname{Log} n \to 0$ for some $\beta \in \left]0, \frac{1}{2}\right[$, then

$$\sup_{\|x\| \leq n^{\gamma}} \left| f_n^K(x) - \bar{f}(x) \right| \xrightarrow{a.s.} 0 , \quad \gamma > 0 . \tag{9}$$

b) Suppose that $\|v\| K(v)$ is integrable, f satisfies a Lipschitz condition and for some $\eta > 0$,

$$n^{\eta} \sup_{x \in \mathbb{R}^d} \left| \bar{f}_n(x) - \bar{f}(x) \right| \to 0 . \tag{10}$$

Then, if $r_n = cn^a$, $0 < a < (2d)^{-1}$ and $0 < \delta < \min\left(\eta, a, \frac{1}{2} - ad\right)$, we have, for each $\gamma > 0$,

$$n^{\delta} \sup_{\|x\| \leq n^{\gamma}} \left| f_n^K(x) - \bar{f}(x) \right| \xrightarrow{a.s.} 0 . \tag{11}$$

c) If in addition $\|x\|^q f(x)$ is bounded for some $q > d$ and $|K|$ is ultimately decreasing, then

$$\sup_{x \in \mathbb{R}^d} \left| f_n^K(x) - \bar{f}(x) \right| \xrightarrow{a.s.} 0 . \tag{12}$$

Note that PELIGRAD (1987), ROUSSAS (1988), BOSQ (1987, 1988) have obtained similar results in the stationary case.

Corollary 2.4. If $r_n = c_1 n^a$, $0 < a < \frac{1}{2}$, if (10) is satisfied and if for some $\delta_0 > 0$

$$\sup_{|\delta| \leq \delta_0} \sum_{j>r} \left| \int e_j(x) \; f(x-\delta)\,dx \right| \leq c_2 \; r^{-\alpha} , \tag{13}$$

then

$$n^{\delta} \sup_{x \in E} \left| f_n^p(x) - \bar{f}(x) \right| \xrightarrow{a.s.} 0 , \tag{14}$$

where $0 < \delta < \min\left(\eta, \alpha a, \frac{1}{2} - a\right)$.

Note that, as is usual in the projection method, d does not appear explicitely.

3. Pseudo-regression estimation

Consider an $\mathbb{R}^d \times \mathbb{R}$-valued almost stationary process $\left(X_t^{(1)}, X_t^{(2)}\right)$, $t \in \mathbb{Z}$ satisfying

$$X_t^{(i)} = s^{(i)}(t) + Y_t^{(i)} \quad , \quad t \in \mathbb{Z} \; , \; i=1,2, \tag{15}$$

where $(s^{(1)}, s^{(2)})$ is bounded and $\left(Y_t^{(1)}, Y_t^{(2)}\right)$, $t \in \mathbb{Z}$ is integrable and strictly stationary.

Assume the density $g(.,.)$ of $\left(Y_t^{(1)}, Y_t^{(2)}\right)$ exists and define

$$\varphi(u) = \int_{\mathbb{R}} v \; g(u,v) \; dv \quad , \; x \in \mathbb{R}^d , \tag{16}$$

and

$$\bar{\varphi}_n(u) = \frac{1}{n} \sum_{t=1}^{n} \int v \; g(u-s^{(1)}(t), \; v-s^{(2)}(t))\,dv, \; u \in \mathbb{R}^d . \tag{17}$$

We consider the following assumptions

A'3. There exists a function $\bar{\varphi}$ such that
$$\lim_{n\to\infty} \sup_{x \in \mathbb{R}^d} \left| \bar{\varphi}_n(x) - \bar{\varphi}(x) \right| = 0 .$$

A'4. For each δ_0 positive, we have, for $r \to +\infty$, $r \in I$,
$$\sup_{|\delta| \leq \delta_0} \sup_{x \in \mathbb{R}} \left| \int v \; K_r(x+\delta, u+\delta) g(u,v)\,du \; dv - \varphi(x) \right| \to 0 .$$

A'5. $\left(Y_t^{(1)}, Y_t^{(2)}\right)$, $t \in \mathbb{Z}$ is geometrically strong mixing.

A6. Let f_2 denotes the density of $Y_t^{(2)}$. There exists $\alpha > 2$ such that $\sup_{y \in \mathbb{R}} |y|^{\alpha} f_2(y) < +\infty$.

Note that

$$\int v\, g(u-s^{(1)}(t),\ v-s^{(2)}(t))\,dv = \varphi(u-s^{(1)}(t)) + s_t^{(2)} f_2(u-s^{(2)}(t)) \tag{18}$$

Using (18), it is easy to prove the following properties.

a) If φ and f_2 are bounded and if $s^{(1)}$ and $s^{(2)}$ are periodic with period T, then

$$\bar{\varphi}(u) = \frac{1}{T}\sum_{t=1}^{T}\int v\, g(u-s^{(1)}(t),\ v-s^{(2)}(t))\,dv\ . \tag{19}$$

b) If φ satisfies a Lipschitz condition and if f_2 is bounded, then $\frac{1}{n}\sum_{t=1}^{n}(|s^{(1)}(t)|+|s^{(2)}(t)|) \to 0$ implies $\bar{\varphi} = \varphi$.

Now we define an estimate for $\bar{\varphi}$ by

$$\varphi_n(x) = \frac{1}{n}\sum_{t=1}^{n} X_t^{(2)}\, K_{r_n}\left(x,\ X_t^{(1)}\right)\ ,\ x \in \mathbb{R}^d\ . \tag{20}$$

The consistency of φ_n is given by the following

Proposition 3.1.

a) Let $A_1, A_2, A_3', A_4', A_5'$ hold and let $Y_t^{(2)}$ be bounded. Then condition (4) implies

$$\sup_{\|x\|\leq x_n} \left|\varphi_n(x) - \bar{\varphi}(x)\right| \xrightarrow{a.s.} 0\ . \tag{21}$$

b) In addition, if $r_n = c_1\, n^a$, $a_r = c_2\, r^b$, $b_3 = c_3\, r^c$ with all the constants strictly positive and a $b < \frac{1}{2}$ and if there exists $\gamma > 0$ and $0 < \delta < \frac{1}{2} - ab$ such that

$$n^{\delta} \sup_{\|x\|\leq n^{\gamma}} \left|E\,\varphi_n(x) - \bar{\varphi}(x)\right| \to 0\ . \tag{22}$$

then

$$n^{\delta} \sup_{\|x\|\leq n^{\gamma}} \left|\varphi_n(x) - \bar{\varphi}(x)\right| \xrightarrow{a.s.} 0\ . \tag{23}$$

Proposition 3.2.

Let $A_1, A_2, A_3', A_5', A_6$ hold and let :

- $r_n \to +\infty$,
- $n^{\delta} \sup_{\|x\|\leq x_n} \left|E\,\varphi_n(x) - \bar{\varphi}(x)\right| \to 0$ for some δ in $\left[0, \frac{1}{2}\right[$, (24)
- $\alpha > 5$,

- $r_n = c_1 n^a$, $a_r = c_2 r^b$, $b_r = c_3 r^c$ where c_1, c_2, c_3, a, b, c are strictly positive and $a\,b + \delta < \frac{\alpha - 5}{2(\alpha-1)}$.

  Then, for each $\gamma > 0$,

$$n^{\delta} \sup_{\|x\| \leq n^{\gamma}} \left| \varphi_n(x) - \bar{\varphi}(x) \right| \xrightarrow{a.s.} 0 . \tag{25}$$

Finally suppose that $\left(Y_t^{(1)}\right)$, $(s^{(1)}(t))$ fulfil A_3, denote by $f_{(1)}$ the density of $Y_t^{(1)}$ and set $\bar{f}_{n,1}(x) = \frac{1}{n} \sum_{t=1}^{n} f_{(1)}(x - s^{(1)}(t))$, and $\bar{f}_1$ for the limit of $\bar{f}_{n,1}$ and suppose that $f_{(1)}$ fulfills A_4.

Then we define the <u>pseudo-regression</u> $\bar{m}$ by

$$\bar{m}(x) = \frac{\bar{\varphi}(x)}{\bar{f}_{(1)}(x)} 1_{\bar{f}_1(x) \neq 0} , \quad x \in \mathbb{R}^d , \tag{26}$$

and its estimate by

$$m_n(x) = \frac{\varphi_n(x)}{f_n^{(1)}(x)} 1_{f_n^{(1)}(x) \neq 0} , \quad x \in \mathbb{R}^d , \tag{27}$$

where

$$f_n^{(1)}(x) = \frac{1}{n} \sum_{t=1}^{n} K_{r_n}\left(x, X_t^{(1)}\right) , \quad x \in \mathbb{R}^d . \tag{28}$$

In order to obtain consistency in the unbounded case, we consider the following assumption.

A_7. There exist $\lambda > 0$, $c > 0$, $w > 0$ such that

$$\inf_{\|x\| \leq x_n} \bar{f}_1(x) \geq c\, n^{-w} ,$$

where x_n tends to infinity and satisfies $x_n \leq n^{\lambda}$ for all n.

Now we state

Proposition 3.3.

Let $A_1, A_2, A_3, A_4, A_3', A_4', A_5', A_6, A_7$ hold with $0 < w < \frac{\alpha - 9}{2(\alpha-1)}$ and let K_r be strictly positive.

Suppose that for some $\delta > w + \frac{2}{\alpha-1}$,

$$n^{\delta} \left[\sup_{\|x\| \leq x_n} \left| E\, f_n(x) - \bar{f}_1(x) \right| + \sup_{\|x\| \leq x_n} \left| E\, \varphi_n x) - (\bar{\varphi}(x) \right| \right] \to 0 . \tag{29}$$

Then, if $r_n = c_1 n^a$, $a_r = c_2 r^b$, $b_r = c_3 r^c$ with $0 < ab < \frac{\alpha - 9}{2(\alpha-1)} - w$, we have

$$n^{\delta'} \sup_{\|x\| \leq x_n} \left| m_n(x) - \bar{m}(x) \right| \xrightarrow{a.s.} 0 \tag{30}$$

for some $\delta' > 0$.

If K_r is not positive (39) must be replaced by (40) (see the proofs of Propositions 3.3 and 3.5) and it is necessary to make the additional assumption.

A8. $\sup_{\|x\| \leq x_n} |\bar{m}(x)| = o(n^{\zeta})$ for each $\zeta > 0$.

Note that (30) remains valid if, for some $k > 0$, $E\left(e^{k|Y_t^{(2)}|}\right) < \infty$ and $\delta > w$, $0 < ab < \frac{1}{2} - w$.

For the classical PARZEN-ROZENBLATT estimate,

$$m_n^K(x) = \sum_{t=1}^{n} X_t^{(2)} \, r_n^d \, K\left(r_n\left(x - X_t^{(1)}\right)\right) \, / \, f_n^K(x) \tag{31}$$

we make the assumption

A9. The conditional density $f^{\left(2\atop 1\right)}(u|v)$ of $Y_t^{(1)}$ with respect to $Y_t^{(2)}$ satisfies the Lipschitz condition

$$\left| f^{\left(2\atop 1\right)}(u'|v) - f^{\left(2\atop 1\right)}(u|v) \right| \leq c \, \|u' - u\|$$

where c does not depend on v ,

and we take K strictly positive and satisfying a Lipschitz condition. Then we obtain

Proposition 3.4. Let A_3, A_3', A_5, A_6, A_9 hold and suppose that, for some $\eta > 0$,

$$n^{\eta}\left[\sup_{\|x\| \leq x_n} \left| \bar{f}_n^{(1)}(x) - \bar{f}^{(1)}(x) \right| + \sup_{\|x\| \leq x_n} \left| \bar{\varphi}_n(x) - \bar{\varphi}(x) \right| \right] \to 0 \ . \tag{32}$$

Then, if $r_n = c_1 \, n^a$ with $0 < a < \frac{1}{d}\left[\frac{\alpha - 9}{2(\alpha - 1)} - w\right]$

and $w < \min(a, \eta) - \frac{2}{\alpha - 1}$,

we have, for some $\delta' > 0$,

$$n^{\delta'} \sup_{\|x\| \leq x_n} \left| m_n(x) - \bar{m}(x) \right| \xrightarrow{a.s.} 0 \ . \tag{33}$$

Note that if $E\left(\exp\left(k \, Y_t^{(2)}\right)\right) < +\infty$, $(k > 0)$, we can choose a such that $0 < a < \frac{1}{d}\left(\frac{1}{2} - w\right)$ and $w < \min(a, \eta)$.

On the other hand, if $f_{(1)}$ and φ are bounded and the s_i's are periodic, then (32) holds for any $\eta < 1$.

In the following, $\bar{m}_n^p$ is the projection regression estimate associated with the kernel $K_r = \sum_{j=1}^{r} e_j \times e_j$. Here the e_j's are supposed to be continuous and $E = \mathbb{R}^d$.

Proposition 3.5. If $A_1 \to A_4$, A_5' , $A_6 \to A_8$ hold, if in addition the Fourier series of $\bar{\varphi}_n$ and $\bar{\varphi}$ satisfy

$$n^{\beta} \sup_{\|x\| \leq x_n} \left| \sum_{j=1}^{c_1 n^a} \int e_j(y)(\bar{\varphi}_n(y) - \bar{\varphi}(y))dy \cdot e_j(x) \right| \to 0 ,$$

and

$$n^{\beta} \sup_{\|x\| \leq x_n} \left| \sum_{j=c_1 n^a+1}^{\infty} \int e_j(y)\bar{\varphi}(y)dy \cdot e_j(x) \right| \to 0$$

for some $\beta > 0$; and if, finally, conditions of Corollary 2.4. are satisfied, we have

$$n^{\delta'} \sup_{\|x\| \leq x_n} \left| \bar{m}_n(x) - \bar{m}(x) \right| \xrightarrow{a.s.} 0$$

for some $\delta' > 0$.

Conditions ensuring rates of uniform convergence for Hermite series can be found in many papers (see e.g., LIERO (1989)). A_7 and A_8 are valid for Gaussian processes (see below).

4. Prediction

Let $(Z_t, t \in \mathbb{Z})$ be a real valued almost stationary process :

$$Z_t = s(t) + Y_t \ , \ t \in \mathbb{Z}, \tag{34}$$

where $|s(t)| \leq s$ and (Y_t) is strictly stationary. Given observations $Z_1, \ldots, Z_N$, we wish to construct a predictor of Z_{N+h} $(h \geq 1)$. For this purpose, we choose a positive integer p and try to approximate

$$Z_{N+h}^{*} = E(Z_{N+h} | Z_N, \ldots, Z_{N-p+1}) \ .$$

Set

$$X_t^{(1)} = (Z_t, \ldots, Z_{t+p-1}) \ , \ X_t^{(2)} = Z_{t+p-1+h}, t=1, \ldots, N+p-1+h=n \ .$$

We obtain the previous scheme and define the predictor as

$$\hat{Z}_{N+h} = m_n(X_n) \ . \tag{35}$$

Theorem 4.1.

a) Let assumptions of Corollary 2.2. and Proposition 3.1.b. hold. Then, if in addition (Z_t) is stationary, if $|Z_t| \leq d$ andif its density is strictly positive on $[-d,+d]$, we have for some $\delta > 0$,

$$N^{\delta} \left| \hat{Z}_{N+h} - Z_{N+h}^{*} \right| \xrightarrow{a.s.} 0 \ . \tag{36}$$

b) Let assumptions of Proposition 3.3. hold. Then, for some $\delta > 0$,

$$N^{\delta} \left|\hat{Z}_{N+h} - Z^{*}_{N+h}\right| 1_{\|(Z_{N+p-1},\dots,Z_N)\| \leq x_n} \xrightarrow{a.s.} 0 ,$$

and, if

$$N^{\delta'} P(\|(Z_{N+p-1},\dots,Z_N)\| > x_n) \to 0 , \tag{37}$$

then

$$N^{\min(\delta,\delta')} \left|\hat{Z}_{N+h} Z^{*}_{N+h}\right| \xrightarrow{P} 0 . \tag{38}$$

Example : If (Y_t) is a Gaussian Markov real process and if $E\, Y_t = 0$, $E\, Y_t^2 = 1$, then

$$\inf_{|x| \leq (w\mathrm{Log}n)^{1/2}} \hat{f}_1(x) > \frac{e^{-s^2}}{\sqrt{2\pi}} n^{-w} ,$$

$$P(|Z_n| > (w\mathrm{Log}n)^{1/2}) < \frac{1}{\sqrt{2\pi}} \frac{e^{-s^2}}{s+(w\mathrm{Log}n)^{1/2}} n^{-w} ,$$

and (37) is valid.

Furthermore, A_8 hold as $\sup_{\|x\| \leq x_n} |\bar{m}(x)| = O((\mathrm{Log}n)^{1/2})$.

5. Applications

The underlying (implicit) model for (Y_t) in (34) is a nonlinear p-autoregressive process. Then, after a possible differenciation (or fractional differenciation), we must estimate p . For this purpose, choose integers t_0 and p_0 and define

$$\Delta_p = \sum_{t=t_0}^{N-h} \left(\hat{Z}^{(p)}_{t+h} - Z_{t+h}\right)^2 \quad ; \; p=1,2,\dots,p_0 ,$$

where $\hat{Z}^{(p)}_{t+1}$ is the predictor (35) associated with the data $Z_1,\dots,Z_t$ and the order p . The selected p (say $\hat{p}$) satisfies

$$\Delta_{\hat{p}} = \min_{1 \leq p \leq p_0} \Delta_p .$$

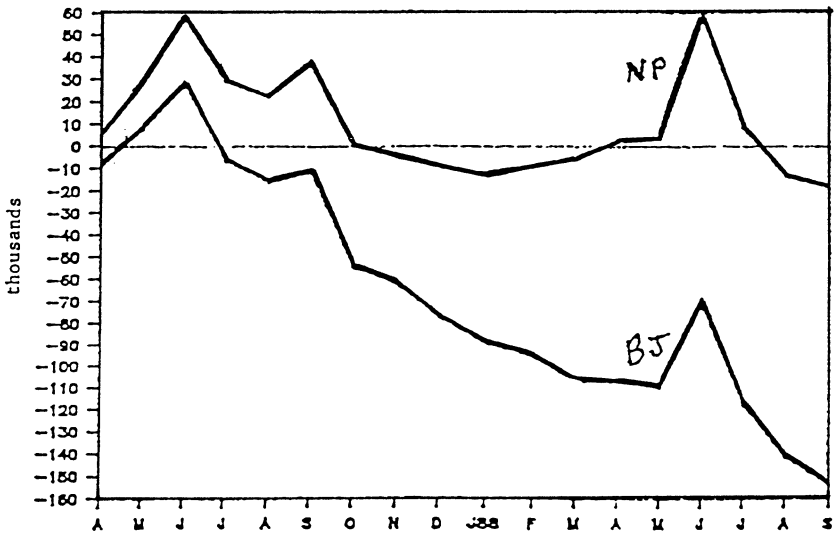
In practice, results are somewhat astonishing. The methods leads to large p's which is in contradiction with the "parcimony principle".

Our method works more efficiently than the classical BOX-JENKINS's one because determination of a seasonal component is not required. Moreover, choosing a large p allows to improve large future prediction as can be seen below. For an extensive practical study of the nonparametric prediction method, see CARBON-DELECROIX (1989).

ıparison of B.J. and non parametric methods for prediction of number of french cars registrations in 1987-1988

horizon	data	BOX-JENKINS ((4,1,0),4,1,0)) prediction	NON PARAMETRIC r = 36 prediction
1	192100	183108	197142
2	156700	173807	179143
3	151200	170549	180682
4	195900	161850	167878
5	146100	136339	138748
6	129600	134416	144142
7	232300	189057	195925
8	197800	190014	192538
9	208900	193875	204616
10	160600	148572	156283
11	160000	153919	164498
12	218000	206162	221286
13	189000	188580	197147
14	184000	181256	185602
15	141600	179818	195450
16	210000	163564	160263
17	157100	133880	135351
18	146700	134378	141147

Absolute errors



Proofs (sketches).

Proposition 2.1. We have

$$E\, f_n(x) = \frac{1}{n}\sum_{t=1}^{n} \int K_{r_n}(x,u)\, f(u-s(t))\,du ,$$

so that

$$\Delta_n = \sup_{x\in\mathbb{R}^d} \left| E\, f_n(x) - \frac{1}{n}\sum_{t=1}^{n} f(x-s(t)) \right|$$

$$\leq \sup_{\delta\in s(\mathbb{Z})} \sup_{x\in\mathbb{R}^d} \left| \int K_{r_n}(x+\delta, u+\delta)\, f(u)\,du - f(x) \right| .$$

Since $s(\mathbb{Z})$ is a compact set, A_4 implies $\Delta_n \to 0$. Thenby A_3, we get $\sup_{x\in\mathbb{R}^d} \left| E\, f_n(x) - \bar{f}(x) \right| \to 0$.

It remains to show that

$$\Delta_n' = \sup_{\|x\|\leq x_n} |f_n(x) - E\, f_n(x)| \to 0 .$$

It is easy to see that we can take $\|.\|$ as the sup norm of $\mathbb{R}^d$, and (x_n) in such a manner that $\nu_n = b_{r_n}^{1/\ell}\, x_n (\mathrm{Log}\, n)^{1/\ell}$ is an integer. Now cover $\{x : \|x\| \leq x_n\}$ by ν_n^d closed hypercubes : $\left\{x : \|x - x_{jn}\| \leq d_n\, \nu_n^{-1}\|\right\}$, $1 \leq j \leq \nu_n^d$.

By A_2, we have

$$|f_n(x) - f_n(x_{jn})| \leq \frac{1}{\mathrm{Log}\, n} , \quad 1 \leq j \leq \nu_n^d ,$$

and

$$|E\, f_n(x) - E\, f_n(x_{jn})| \leq \frac{1}{\mathrm{Log}\, n} , \quad 1 \leq j \leq \nu_n^d .$$

Therefore

$$\Delta_n' \leq \sup_{1\leq j\leq \nu_n^d} |f_n(x_{jn}) - E\, f_n(x_{jn})| + \frac{2}{\mathrm{Log}\, n} .$$

Now, using an exponential inequality (CARBON (1983)) for each $|f_n(x_{jn}) - E\, f_n(x_{jn})|$, we get, for $\varepsilon > 0$ and n large,

$$P\left(\Delta_n' \geq \varepsilon\right) \leq 3\, \nu_n^d \exp\left(-\frac{\varepsilon}{56\, a_{r_n}}\, n^{\gamma}\right) ,$$

hence the result. ∎

Corollary 2.2. Clear. ∎

Corollary 2.3.

a) Here $a_r = \sup|K|\, r^d$, $b_r = O(r^{d+1})$ and by BOCHNER lemma (see, e.g., BOSQ-LECOUTRE (1987), p. 61), A_4 is satisfied. Thus the proof is clear.

b) We get

$$\left|Ef_n(x)-\bar{f}_n(x)\right| = \frac{1}{n}\left|\sum_{t=1}^{n}\int r_n^d\, K(r_n(x-u))[f(u-s(t)-f(x-s(t))]\right|$$

$$\leq M_f\, r_n^{-1}\int \|v\|K(v)\,dv$$

where M_f is the Lipchitz coefficient for f. Therefore, if $\delta < a$,

$$n^{\delta}\sup_{x\in\mathbb{R}^d}\left|E\, f_n(x) - \bar{f}_n(x)\right| \to 0 ,$$

and the result follows easily.

c) It is easy to see that $\bar{f}$ is uniformly continuous ; then it suffices to show that $\sup_{\|x\| > n^{\gamma}}\left|f_N^K(x)\right| \xrightarrow{a.s.} 0$. This can be done, adapting the proof of BOSQ (1988), p. 10-12. We omit details. ∎

Corollary 2.4. The proof is straightforward because (13) implies

$$|E\, {}_n(x) - \bar{f}_n(x)| \leq \sup_{j,x}|e_j(x)| \cdot \sum_{j>r_n}\left|\int e_j\, \bar{f}_n\right| = O(n^{-\alpha a}) . \quad ∎$$

Proposition 3.1. Similar to the proof of Proposition 2.1. ∎

Proposition 3.2. As in Proposition 2.1, we make a covering of $\{x:\|x\| \leq n^{\gamma}\}$ by ν_n^d closed hypercubes. Here $\nu_n = b r_n^{1/\ell}\, n^{\gamma}\, n^{\zeta/\ell}$ for some $\zeta > 1+\delta$.

Then we obtain the bound

$$D_n = n^{\delta}\sup_{\|x\|\leq n^{\gamma}}|\varphi_n(x) - E\,\varphi_n(x)|$$

$$\leq n^{\delta}\sup_{1\leq j\leq \nu_n^d}|\varphi_n(x_{jn})-E\varphi_n(x_{jn})| + n^{\delta-3}\left(\frac{1}{n}\sum_{t=1}^{n}\left(\left|X_t^{(2)}\right|+E\left|X_t^{(2)}\right|\right)\right)$$

$$\leq A_n + B_n .$$

Using again CARBON (1983) exponential inequality, we get $A_n \xrightarrow{a.s.} 0$. As for B_n, we use A_7 and conclude the Proof as in BOSQ (1988), p. 16-18. ∎

Proposition 3.3. Write

$$m_n(x) - \bar{m}(x) = \frac{\bar{f}_1(x)-f_n(x)}{\bar{f}_1(x)}\, m_n(x) + \frac{\varphi_n(x)-\bar{\varphi}(x)}{\bar{f}_1(x)} . \quad (39)$$

Therefore, using A_7,

$$\sup_{\|x\|\leq x_n} \left| m_n(x) - \bar{m}(x) \right| \leq c^{-1} n^{w} \left[\sup_{\|x\|\leq x_n} \left| \bar{f}_1(x) - f_n(x) \right| \sup_{\|x\|\leq x_n} m|_n(x)| + \sup_{\|x\|\leq x_n} \left| \varphi_n(x) - \bar{\varphi}(x) \right| \right] \tag{40}$$

A bound for $\sup_{\|x\|\leq x_n} |m_n(x)|$ can be found as in BOSQ (1988) p. 24 by using the positivity of K_r . Now we can apply Corollary 2.2. and Proposition 3.2. to arrive at the conclusion holds.■

Proposition 3.4. It suffices to verify that assumptions of Proposition 3.3. hold. ■

Proposition 3.5. Instead of (39), we use

$$\bar{m}_n(x) - \bar{m}(x) = \frac{\varphi_n(x) - \bar{\varphi}(x)}{f_n(x)} \bar{f}_1(x) + \bar{m}(x) \frac{\bar{f}_1(x) - f_n(x)}{f_n(x)} \tag{41}$$

and the result follows by standard armuments. ■

Theorem 4.1. Clear. ■

Acknowledgement : I would like to thank Professor G. Roussas for helpful comments and D. Tilly.

References

BOSQ D. (1983a) - Nonparametric prediction in stationary processes - in Lectures notes in Stat. 16, 69-84 - Springer-Verlag.

BOSQ D. (1983b) - Sur la prédiction non paramétrique de variables aléatoires et de mesures aléatoires - Z. Wahrscheinlichkeitstheirie und verw. Gebiete., 64, 541-553.

BOSQ D. (1987) - La statistique non paramétrique des processus. Rend. sem. Mat. Univers. Politecn. Torino, 45, 2, 1-24.

BOSQ D. (1988) - Prévision non paramétrique d'un processus stationnaire non borné. Publ. L.S.T.A. 81, 28 p. Univ. Pierre et Marie Curie - Paris.

BOSQ D. (1989) - Estimation et prévision non paramétrique d'un processus stationnaire. C.R. Acad. Sci. Paris I, 308, 453-456.

BOSQ D. - DELECROIX M. (1985) - Prediction of a Hilbert valued random variable. Stochastic Process. Appl., 19, 271-280.

BOSQ D. - LECOUTRE J.P. (1987) - Théorie de l'estimation fonctionnelle - 342 p. - Economica - Paris.

BOX G.E.P. - JENKINS G.M. (1976) - Time series Analysis : Forecasting and control - Holden Day.

CARBON M. (1983) - Inégalité de Bernstein pour les processus fortement mélangeants non nécessairement stationnaires. C.R. Acad. Sci. Paris I, 297, 303-306.

CARBON M. - DELECROIX M. (1989) - Non parametric forecasting in time series : a computational point of view. To appear.

COLLOMB G. (1984) - Propriétés de convergence presque complète du prédicteur à noyau. Z. Wahrscheinlichkeits-theorie und verw. Gebiete, 66, 441-460.

COLLOMB G. (1985) - Non-parametric time series analysis and prediction : uniform almost sure convergence. Statistics 2, 297-307.

COLLOMB G. HÄRDLE W. (1986) - Strong uniform convergence rates in robust non parametric time series analysis. Stochastic Process Appl. 23, 77-89.

COLLOMB G. - HÄRDLE W. - HASSANI S. (1987) - A note on prediction via estimation of the conditional mode function. J. Statist. Plann. Inference 15, 227-236.

GANNOUN A. (1989) - Prévision par médiane conditionnelle. Ph. D. Thesis - Paris.

GOURIEROUX A. - MONFORT A. (1990) - Séries temporelles et modèles dynamiques. Economica - Paris.

GYÖRFI L. - HÄRDLE W. - SARDA P. - VIEU P. (1989) - Nonparametric curve estimation from time series. Lecture Notes in Stat. 60 - Springer Verlag.

HÄRDLE W. (1990) - Applied nonparametric regression. Cambridge Univ. Press - To appear.

LIERO H. (1989) - Strong uniform consistency of non parametric regression function estimates. Prob. theory and related fields, 82, 587-614.

PELIGRAD M. (1987) - Properties of uniform consistency of the kernel estimator of density and of regression - Preprint.

ROSENBLATT M. (1985) - Stationary sequences and random fields - BIRKHAUSER.

ROUSSAS G.G. (1988) - Nonparametric estimation in mixing sequences of random variables. J. of Statist. Plann. and Inference, 18, 135-149.

SARDA P. - VIEU P. (1985) - Nonparametric regression estimation. Application to prediction. Proceeding of the 4th EYSM. Pliska - Studia Mathematica Bulgarica.

TRUONG Y.K. - STONE C.J. (1988) - Nonparametric time series prediction - Preprint.

MONTE CARLO AND TURBULENCE

K. HELLAND
4647 T Highway 280 East Suite 150
Birmingham, Alabama 35242 USA

K. S. LII
Department of Statistics
University of California, Riverside
Riverside, California 92521 USA

M. ROSENBLATT
Department of Mathematics
University of California, San Diego
La Jolla, California 92093 USA

ABSTRACT. In a prior paper an estimate of three-dimensional energy transfer was attempted in grid-generated turbulence by using estimates of second and third order spectra and employing Monte Carlo estimates of integrals. The methods used in such a Monte Carlo simulation are considered in greater detail. To see how well such a technique using Monte Carlo works, a smaller problem involving second order spectra with the results known analytically is employed. Some general remarks are also made on Monte Carlo estimates of Fourier transforms.

1. INTRODUCTION

1.1. Energy Transfer

In the model of homogeneous turbulence the velocity field $\vec{v}(\vec{x},t) = (v_1, v_2, v_3)^*$ (A^* is the conjugate transpose of the matrix A) is considered a random solution of the Navier-Stokes equation

$$\partial\vec{v}(\vec{x},t)/\partial t + (\vec{v}\cdot\nabla)\vec{v} = -\tfrac{1}{\rho}\nabla p + \nu\nabla^2\vec{v}$$

and the continuity equation

$$\vec{\nabla}\cdot\vec{v} = 0$$

that is a stationary stochastic process relative to spatial displacements (in $\vec{x} = (x_1, x_2, x_3)$). In the equations written above ∇ is the gradient operator, ρ the density, p the thermodynamic pressure, ν the kinematic viscocity and ∇^2 the Laplacian. The expression

$$\vec{v}\cdot\nabla = \sum_{i=1}^{3} v_i \frac{\partial}{\partial x_i}$$

and it is understood to be applied to each component of $\vec{v}$ in $(\vec{v}\cdot\nabla)\vec{v}$. If one has turbulence (typically with very low viscocity) apparently the same conditions do not lead to the same fluid flow and so a stochastic model has been proposed (see Batchelor (1953) for a discussion of homogeneous turbulence). Foias and Temam (1980) have established the existence of homogeneous statistical solutions of the Navier-Stokes equation under appropriate conditions. On taking the divergence of the nonlinear equation one finds that

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 405–418.
© 1991 *Kluwer Academic Publishers.*

$$\frac{1}{\rho}\nabla^2 p = -\sum_{i,j=1}^{3} \frac{\partial^2 v_i v_j}{\partial x_i \partial x_j}\,.$$

This allows the pressure term p in the Navier-Stokes equation to be eliminated.

Assume that $E\vec{v}(\vec{x},t) = 0$. Finiteness of second order moments implies that $\vec{v}(\vec{x},t)$ has the Fourier representation

$$\vec{v}(\vec{x},t) = \int e^{i\vec{k}\cdot\vec{x}}\, d\vec{z}(\vec{k},t) \tag{1}$$

with $\vec{k} = (k_1, k_2, k_3)$ the spatial wave number vector. The second order spectral matrix is $f(\vec{k},t) = (f_{\alpha,\beta}(\vec{k},t);\ \alpha,\beta = 1,2,3)$ with

$$\begin{aligned} f(\vec{k},t)\delta_{\vec{k},\vec{k}'}\, dk &= E\{d\vec{z}(\vec{k},t)^*\, d\vec{z}(\vec{k}',t)\} \\ &= (E\, dz_\alpha(\vec{k},t)^*\, dz_\beta(\vec{k}',t);\quad \alpha,\beta = 1,2,3) \end{aligned} \tag{2}$$

$\delta_{\vec{k},\vec{k}'}$ the Kronecker delta symbol and $\vec{z}(\vec{k},t) = (z_1(\vec{k},t),\ z_2(\vec{k},t),\ z_3(\vec{k},t))$ a process of orthogonal increments. The energy associated with the wave number vector $\vec{k}$ is $\sum_{i=1}^{3} f_{ii}(\vec{k},t)$. From this point on time dependence is implicit and there will be no explicit mention of t. Existence of appropriate moments whenever required will be assumed. Let

$$\begin{aligned} Q_{\alpha,\beta}(\vec{k},\vec{k}')\, dk\, dk' &= k_\alpha \operatorname{Im} E\{dz_\alpha(\vec{k}-\vec{k}')\, z_\beta(-\vec{k})\, dz_\beta(\vec{k}')\} \\ &= k_\alpha B_{\alpha\beta\beta}(\vec{k},\vec{k}')\, dk\, dk',\ \alpha,\beta = 1,2,3 \end{aligned}$$

with the $B_{\alpha\beta\beta}(\vec{k},\vec{k}')$ bispectra. Set

$$Q(\vec{k},\vec{k}') = \sum_{\alpha,\beta} Q_{\alpha,\beta}(\vec{k},\vec{k}')\,. \tag{3}$$

One can show that

$$\frac{\partial}{\partial t}\left(\frac{1}{2}\sum_{i=1}^{3} f_{ii}(\vec{k})\right) = \int Q(\vec{k},\vec{k}')\, dk' - \nu|\vec{k}|^2 \sum_{i=1}^{3} f_{ii}(\vec{k}) \tag{4}$$

(see Batchelor (1953)) and for this reason $Q(\vec{k},\vec{k}')$ has been interpreted as the net mean rate of energy transfer from $d\vec{k}'$ to $d\vec{k}$. The term in the last equation involving $Q(\vec{k},\vec{k}')$ couples the wave number vectors $\vec{k}$ and $\vec{k}'$ and is due to the nonlinearity of the Navier-Stokes equation. Set

$$E(k) = \int_{|\vec{k}|=k} \frac{1}{2}\sum_{i=1}^{3} f_{ii}(\vec{k})\, dk$$

and

$$P(k,k') = P(|\vec{k}|,|\vec{k}'|) = \int_{|\vec{k}|=k}\int_{|\vec{k}'|=k'} Q(\vec{k},\vec{k}')\, dk\, dk'\,.$$

Then

$$\frac{\partial E(k)}{\partial t} = \int P(k,k')\, dk' - 2\nu k^2 E(k)$$

follows from equation (4).

A homogeneous velocity field may also be isotropic. Let $x^{(1)},\ldots,x^{(s)}$ be s locations with components of velocity in directions $\alpha^{(1)},\ldots,\alpha^{(s)}$ at these locations. Consider a rotation or reflection that puts the locations and directions in a new configuration. The homogeneous random field is isotropic if the joint probability distribution of the velocity components in the new configuration is the same as those in the old configuration, for any such rotation or reflection.

1.2. Homogeneity and Isotropy

In the experiment an x-wire was set up at the origin. One of the x-wire components measured in the downstream direction and the other vertically upwards (perpendicular to the downstream direction). Two single wire probes were set up in the downstream direction at optional locations in the vertical plane of observation. The locations of the single wire probes were specified by a Monte Carlo procedure. The object was to obtain an estimate of $P(k,k')$ assuming isotropy by the Monte Carlo procedure and Fourier integral estimates based on it (see K. S. Lii et al(1982)).

Before giving a more extensive discussion of the estimation procedure, we note some of the relations implied by homogeneity and isotropy. Downstream and transverse correlations f and g are

$$\begin{aligned} u^2 f(r) &= E\{v_1(0,0,0)\, v_1(r,0,0)\} \\ u^2 g(r) &= E\{v_2(0,0,0)\, v_2(r,0,0)\} \end{aligned} \tag{5}$$

with

$$u^2 = \mathrm{var}\,(v_1(0,0,0)) = \mathrm{var}\,(v_2(0,0,0))$$

(the subscript 1 corresponds to the downstream direction). Isotropy implies that the covariances

$$\begin{aligned} R_{i,j}(\vec{r}) &= E\{v_i(\vec{x})v_j(\vec{x}+\vec{r})\} \\ &= u^2\left(\frac{f(r)-g(r)}{r^2} - r_i r_j + g(r)\delta_{i,j}\right) \end{aligned}$$

with $\vec{r} = (r_1, r_2, r_3)$, $r = |r|$ as well as

$$g = f + \tfrac{1}{2} r f'$$

where $f' = \partial f/\partial r$ (see Batchelor (1953)).

Isotropy (together with homogeneity) also imply that

$$\begin{aligned} B_{222}(\vec{k},\vec{k}') &= B_{111}(k_2,k_2';k_1,k_1';k_3,k_3') \\ B_{333}(\vec{k},\vec{k}') &= B_{111}(k_3,k_3';k_2,k_2';k_1,k_1') \\ B_{122}(\vec{k},\vec{k}') &= B_{211}(k_2,k_2';k_1,k_1',k_3,k_3') \\ B_{311}(\vec{k},\vec{k}') &= B_{211}(k_1,k_1';k_3,k_3',k_2,k_2') \\ B_{133}(\vec{k},\vec{k}') &= B_{211}(k_2,k_2';k_3,k_3',k_1,k_1') \\ B_{233}(\vec{k},\vec{k}') &= B_{211}(k_3,k_3';k_2,k_2',k_1,k_1') \\ B_{322}(\vec{k},\vec{k}') &= B_{211}(k_3,k_3';k_1,k_1',k_2,k_2') \end{aligned} \tag{6}$$

where it is understood that

$$B_{\alpha\beta\beta}(\vec{k},\vec{k}') = B_{\alpha\beta\beta}(k_1,k_1';k_2,k_2';k_3,k_3') \,.$$

The relations (6) imply that $Q(\vec{k},\vec{k}')$ can be specified in terms of B_{111} and B_{211}.

1.3. The Experiment

In the experiment turbulence was set up downstream in the wind tunnel by an air stream with mean downstream velocity U. The velocities $v_1(0,0,0;t)$, $v_2(0,0,0;t)$, $v_1(0,x_2,x_3;t)$ and $v_1(0,x_2',x_3';t)$ (mean velocities are subtracted) are measured as functions of the time t at the locations $(0,0,0)$, $(0,x_2,x_3)$, and $(0,x_2',x_3')$. These fluctuation velocities as functions of time are interpreted as functions of the downstream coordinate x_1 using Taylor's hypothesis.

$$\vec{v}(x_1,x_2,x_3;t) = \vec{v}(x_1+U\tau,x_2,x_3;t+\tau) \,.$$

This hypothesis is a "frozen-flow" approximation assuming the turbulent velocity field is swept past at a uniform speed U.

Let

$$b_{\alpha\beta\beta}(k_1,k_1';x_2,x_2';x_3,x_3') = \frac{1}{(2\pi)^2}\iint e^{i(k_1x_1-k_1'x_1')}E\{v_\alpha(0,0,0)v_\beta(x_1,x_2,x_3)v_\beta(x_1',x_2',x_3')\}\,dx_1\,dx_1'$$

Then

$$B_{\alpha\beta\beta}(\vec{k},\vec{k}') = \frac{1}{(2\pi)^4}\iiiint \exp\{i(k_2r_2+k_3r_3-k_2'r_2'-k_3'r_3')\} \, b_{\alpha\beta\beta}(k_1,k_1';r_2,r_2';r_3,r_3')\,dr_2\,dr_3\,dr_2'\,dr_3' \,.$$

Statistical estimates of $b_{111}(k_1,k_1';x_2,x_2';x_3,x_3')$ and $b_{211}(k_1,k_1';x_2,x_2';x_3,x_3')$ are obtained for the points (x_2,x_3), (x_2',x_3') chosen by Monte Carlo by using bispectral density estimates (see Brillinger and Rosenblatt (1967) and K. S. Lii et al (1982) for the details). Making use of these, Monte Carlo Fourier transform estimates of B_{111} and B_{211} are then used. The relations (6) and (3) are then employed to obtain an estimate of $Q(\vec{k},\vec{k}')$. A further Monte Carlo integration (by uniformly distributing points in wave number shells) is used to get an estimate of $P(k,k')$ utilizing (3).

2. MONTE CARLO

2.1. Integrals

Suppose we wish to estimate the integral

$$I = \int h(x)\,dx \,.$$

One can get an unbiased estimate by considering

$$I_n = \frac{1}{n}\sum_{i=1}^{n}\frac{h(X_i)}{g(X_i)}$$

where the X_i's are independent and identically distributed with probability density $g(x)$. Then

$$EI_n = \int h(x)\,dx = I$$

and

$$\operatorname{var}(I_n) = \tfrac{1}{n}\left[\int \tfrac{h^2(x)}{g(x)}\,dx - I^2\right] . \tag{7}$$

This limitation to unbiased estimates is an idealization since in our problem the estimates are biased since the bispectral estimates initially made use of in our program are intrinsically biased. Nonetheless, we continue with a discussion of unbiased estimates since it can give us some limited insight.

Minimizing (7) amounts to minimizing

$$n \operatorname{var}(I_n) + \lambda \int g(x)\,dx$$

with respect to g. The first variation is

$$-\int \tfrac{h^2(x)}{g^2(x)}\,\delta g(x)\,dx + \lambda \int \delta g(x)\,dx$$

leading one to

$$g(x) = |h(x)| / \left\{\int |h(x)|\,dx\right\}^{1/2} .$$

The minimum is then

$$\operatorname{var}(I_n) = n^{-1}\left[\left(\int |h(x)|\,dx\right)^2 - \left(\int h(x)\,dx\right)^2\right] .$$

Of course, this is somewhat of an illusion since in the case $h(x)$ is positive it assumes one already has $\int h(x)\,dx$ available. In realistic situations one chooses a density $g(x)$ on the basis of the limited information available.

The problem of real interest is that of estimating a Fourier transform

$$I(\vec{\lambda}) = \int e^{i\vec{x}\cdot\vec{\lambda}} h(\vec{x})\,dx .$$

Again assuming an unbiased sampling procedure with density $g(\vec{x})$ one is led to an estimate

$$I_n(\vec{\lambda}) = \tfrac{1}{n}\sum_{i=1}^{n} h(X_i) e^{iX_i\cdot\vec{\lambda}} / g(X_i)$$

where the X_i's are independent identically distributed random k-vectors (in the k-dimensional case) with density $g(\vec{x})$. The estimate is unbiased with variance

$$\operatorname{var}(I_n(\vec{\lambda})) = n^{-1}\left[\int \tfrac{|h(\vec{x})|^2}{g(\vec{x})}\,dx - \left|\int e^{\vec{x}\cdot\vec{\lambda}} h(\vec{x})\,dx\right|^2\right] .$$

The minimizing sampling density $g(\vec{x})$ is again given by

$$g(\vec{x}) = |h(\vec{x})| / \left\{\int |h(\vec{z})|\,dz\right\}^{1/2}$$

leading to the minimal variance

$$\operatorname{var}(I_n(\vec{\lambda})) = n^{-1}\left[\left(\int |h(\vec{x})|\,dx\right)^2 - \left|\int e^{i\vec{x}\cdot\vec{\lambda}} h(\vec{x})\,dx\right|^2\right] \tag{8}$$

at $\vec{\lambda}$. However, if we wish to have an unbiased estimate making use of an a priori guess for the density $g(\vec{x})$ we are led to variance (7).

2.2. Deviation

The deviation in the 1-dimensional case between I_n and I (with factor $\sqrt{n}$) is

$$\sqrt{n}(I_n - I) = \left[n^{-1} \sum_{i=1}^{n} \frac{h(X_i)}{g(X_i)} - \int h(x)\,dx \right]$$
$$= \sqrt{n} \int \frac{h(x)}{g(x)}\, d[G_n(x) - G(x)]$$

where

$$G(x) = \int_{-\infty}^{x} g(u)\,du$$

and $G_n(x)$ is the corresponding sample distribution function. We assume sufficient smoothness so that the indicated operations can be carried out. One can integrate by parts if $h(x)/g(x) \to 0$ as $|x| \to \infty$ and $(h(x)/g(x))'$ is integrable to obtain

$$\sqrt{n}(I_n - I) = -\int \sqrt{n}[G_n(x) - G(x)] \left(\tfrac{h(x)}{g(x)}\right)' dx\,.$$

By the result of Komlos, Major and Tusnady (1975) one can replace the last integral by

$$-\int B_G(x) \left(\tfrac{h(x)}{g(x)}\right)' dx = \int \tfrac{h(x)}{g(x)}\, dB_G(x)$$

with $B_G(x)$ a Brownian bridge process corresponding to distribution G with at most error $O(\log n/\sqrt{n})$. In the case of the Fourier transform $h(x)$ is replaced by $e^{ix\lambda}h(x)$ and $(h(x)/g(x))'$ by $(e^{ix\lambda}h(x)/g(x))'$.

Let us look at the two dimensional case of the Fourier integral. Then integrating by parts we have

$$\begin{aligned} \sqrt{n}\,(I_n(\vec{\lambda}) - I(\vec{\lambda})) &= \sqrt{n} \int e^{i\vec{x}\cdot\vec{\lambda}} \frac{h(\vec{x})}{g(\vec{x})}\, d[G_n(\vec{x}) - G(\vec{x})] \\ &= -\sqrt{n} \int \left(e^{i\vec{x}\cdot\vec{\lambda}} \frac{h(\vec{x})}{g(\vec{x})} \right)_{x_1} d_{x_2}[G_n(\vec{x}) - G(\vec{x})]\,dx_1 \end{aligned} \tag{9}$$

if $h(\vec{x})/g(\vec{x}) \to 0$ uniformly as $|x| \to \infty$ and the last integral can be rewritten with a further integration by parts as

$$\sqrt{n} \int \left(e^{i\vec{x}\cdot\vec{\lambda}} \tfrac{h(\vec{x})}{g(\vec{x})} \right)_{x_1,x_2} \sqrt{n}\,[G_n(\vec{x}) - G(\vec{x})]\,dx_1\,dx_2 \tag{10}$$

if $(e^{i\vec{x}\cdot\vec{\lambda}}h(\vec{x})/g(\vec{x}))_{x_1} \to 0$ uniformly as $|x| \to \infty$ and $(e^{i\vec{x}\cdot\vec{\lambda}}h(\vec{x})/g(\vec{x}))_{x_1,x_2}$ is integrable. The integral (10) can be approximated by

$$\int B_G(\vec{x})(e^{i\vec{x}\cdot\vec{\lambda}}h(\vec{x})/g(\vec{x}))_{x_1,x_1}\,dx_1\,dx_2 \tag{11}$$

with an error at most $O((\log n)^2/n^{1/2})$ (using the remark of Tusnady (1977)) where $B_G(\vec{x})$ is a 2-dimensional Brownian bridge process corresponding to distribution G.

It should be understood that in the 1-dimensional case the Wiener process $W_G(\cdot)$ corresponding to distribution G is $W_G(x) = W(G(x))$ and so the Brownian bridge corresponding to G is

$$B_G(x) = W_G(x) - G(x)W_G(\infty) .$$

The 2-dimensional Wiener process $W(x,y)$, $0 \le x, y$, has covariance

$$\min(x,x') \min(y,y') .$$

The 2-dimensional Wiener process $W_G(\cdot)$ corresponding to G is

$$W_G(x,y) = W(G(x,y))$$

and this implies that the 2-dimensional Brownian bridge corresponding to G is

$$B_G(x,y) = W_G(x,y) - G(x,y)W_G(\infty,\infty) .$$

In the 2-dimensional case the integration by parts can be reversed so as to express (11) as

$$\begin{aligned} &\int e^{i\vec{x}\cdot\vec{\lambda}} \frac{h(\vec{x})}{g(\vec{x})}\, dB_G(\vec{x}) \\ &\quad = \int e^{i\vec{x}\cdot\vec{\lambda}} \frac{h(\vec{x})}{g(\vec{x})}\, dW_G(\vec{x}) - W_G(\infty,\infty) \int e^{i\vec{x}\cdot\vec{\lambda}} \frac{h(\vec{x})}{g(\vec{x})}\, dG(\vec{x}) . \end{aligned} \tag{12}$$

Of course, the corresponding representation is valid in the k-dimensional case. The error in making such an approximation with the Brownian bridge in the k-dimensional case can be shown to be no greater than $O(\{\log n\}^{1/2}\, n^{-1/2k})$ (see P. Massart (1989)).

2.3. Measures of Deviation

The indication is that the normalized deviation (9) can be approximated by (12). If this approximation is good enough a global measure of the normalized deviation could be obtained by integrating the square of (12) over $[0,T]^k$ with $T = T(n) \to \infty$ sufficiently slowly as $n \to \infty$. The principal contribution will come from the first term of (12) rather than the second. The mean of the first term is

$$T^k \int \tfrac{h(\vec{x})^2}{g(\vec{x})}\, d\vec{x}$$

while the variance is to the first order

$$T^k \int r(\vec{u})^2\, du$$

(if $r(\vec{u}) \in L^2$) where

$$r(\vec{u}) = \int e^{i\vec{u}\cdot\vec{x}} \tfrac{h(\vec{x})^2}{g(\vec{x})}\, dx .$$

The variance is thus to the first order proportional to

$$T^k \int \tfrac{h(\vec{x})^4}{g(\vec{x})^2}\, dx .$$

A standard argument indicates that the integral of the square of (12) over $[0,T]^k$ centered and appropriately normalized is asymptotically normal as $T = T(n) \to \infty$ at a sufficiently

slow rate. The asymptotic distribution of the maximal deviation of (12) over $[0,T]^k$ can be determined using techniques similar to those employed in Bickel and Rosenblatt (1973).

3. A SECOND ORDER MONTE CARLO

3.1 Second Order Spectra

In order to get some feeling for the potential of Monte Carlo methods in an experimental context like that in the paper of Lii et al (1982), it was thought to be sensible to consider a much simpler problem in which simple theoretical computations could be carried out and compared with corresponding Monte Carlo integral estimates. If we let

$$\phi_{ij}(\vec{k}) = (8\pi^3)^{-1} f_{ij}(\vec{k})$$

then

$$E(k) = 2\pi\, k^2 \sum_{i=1}^{3} \phi_{ii}(\vec{k})$$

and

$$\phi_{ij}(\vec{k}) = A(k)(k_i k_j - k^2 \delta_{ij}) = \phi_{ji}(\vec{k})\,.$$

Notice that

$$\phi_{11}(k_2, k_1, k_3) = \phi_{22}(k_1, k_2, k_3)$$

while

$$\phi_{33}(k_1, k_2, k_3) = \phi_{11}(k_3, k_1, k_2)\,.$$

Thus ϕ_{22}, ϕ_{33} can be expressed in terms of ϕ_{11}. A form that has been proposed for the transverse correlation as valid for low Reynolds number turbulence is

$$f(r) = 1/(1 + r^2/\alpha^2)\,.$$

Then the corresponding transverse correlation is

$$g(r) = f(r)\left[1 - \tfrac{r^2}{\alpha^2} f(r)\right]$$

and the analytic computation of

$$\begin{aligned}\varphi_{11}(k_1, r_2, r_3) &= \frac{1}{2\pi}\int u^2 \left(\frac{f(r) - g(r)}{r^2}\, r_1^2 + g(r)\right) e^{-ik_1 r_1}\, dr_1 \\ &= \frac{u^2}{2\pi}\int f(r)\left[1 - f(r)\left(\frac{r_2^2 + r_3^2}{\alpha^2}\right)\right] e^{-ik_1 r_1}\, dr_1\end{aligned}$$

can be carried out leading to

$$\text{Re } \varphi_{11}(k_1, r_2, r_3) = \frac{u^2\alpha}{2\xi}\left[1 - \frac{r_2^2 + r_3^2}{2\alpha^2\xi^2}(1 + \alpha\xi|k_1|)\right] \cdot \exp(-\alpha\xi|k_1|)\,, \tag{13}$$

$$\text{Im } \varphi_{11}(k_1, r_2, r_3) = 0$$

where

$$\xi^2 = 1 + \tfrac{r_2^2 + r_3^2}{\alpha^2}\,.$$

Also using the relation

$$E(k) = \frac{1}{\pi}\int_0^\infty u^2 f(r) k^2 r^2 \left(\tfrac{\sin kr}{kr} - \cos kr\right) dr \tag{14}$$

(see Batchelor (1953)) one can see that

$$E(k) = \tfrac{u^2\alpha^2}{2}(k + \alpha k^2)e^{-\alpha k}\,. \tag{15}$$

The object is to use relations (13) and (14) together with a Monte Carlo estimate of $E(k)$ based on the Fourier representation and compare the result to the precise answer (15). There are several variables involved in the Monte Carlo estimate. One of these is the length $2H$ of the side of the square in the (x_2, x_3) space that the Monte Carlo sampling with uniform density is carried out on. A second is the number of n_r of sample points in the Fourier integral estimate. A third variable is the number of points n_k in the shell integration. The Monte Carlo estimate of $\varphi_{11}(\vec{k})$ is

$$\tilde{\phi}_{11}(\vec{k}) = \frac{(2H)^2}{4\pi^2 n_r}\sum_{v_r} \text{Re } \varphi_{11}(k_1, r_2, r_3) \cdot \cos(k_2 r_2 + k_3 r_3)$$

where $v_r = (r_2, r_3)$ is one of the n_r spatial vectors chosen by Monte Carlo.

Four sets of graphs are given to illustrate this dependence on these variables. In the first set $H = 4.0$, $n_k = 100$, $n_{rec} = 100$ (number of repetitions) while n_r varies through 200, 300, 400, 500, 1,000 and 2,000. It is seen that with this choice of length H at $n_r = 2,000$ one has very good agreement with the precise computation. In the second and third sets $H = 6.0$ and $H = 8.0$ respectively and it appears that the density of points n_r in the Monte Carlo estimation is not sufficient to give good agreement. In the last set of graphs n_k is varied with $H = 4.0$ and $n_r = 200$. Here it seems that the number of points n_k is not too relevant with respect to getting good agreement.

4. ACKNOWLEDGEMENT

This research was partially supported by Office of Naval Research through contract N00014-85-k-0468 and grant N00014-90-J1371.

5. REFERENCES

Batchelor, G. K. (1953) The Theory of Homogeneous Turbulence, Cambridge University Press, Cambridge.

Foias, C. and Temam, R. (1980) 'Homogeneous statistical solutions of Navier-Stokes equations,' Indiana J. Math. **29**, 913–957.

Lii, K. S., Helland, K. N. and Rosenblatt, M. (1982) 'Estimating three-dimensional energy transfer in isotropic turbulence,' J. Time Series Anal. **3**, 1–28.

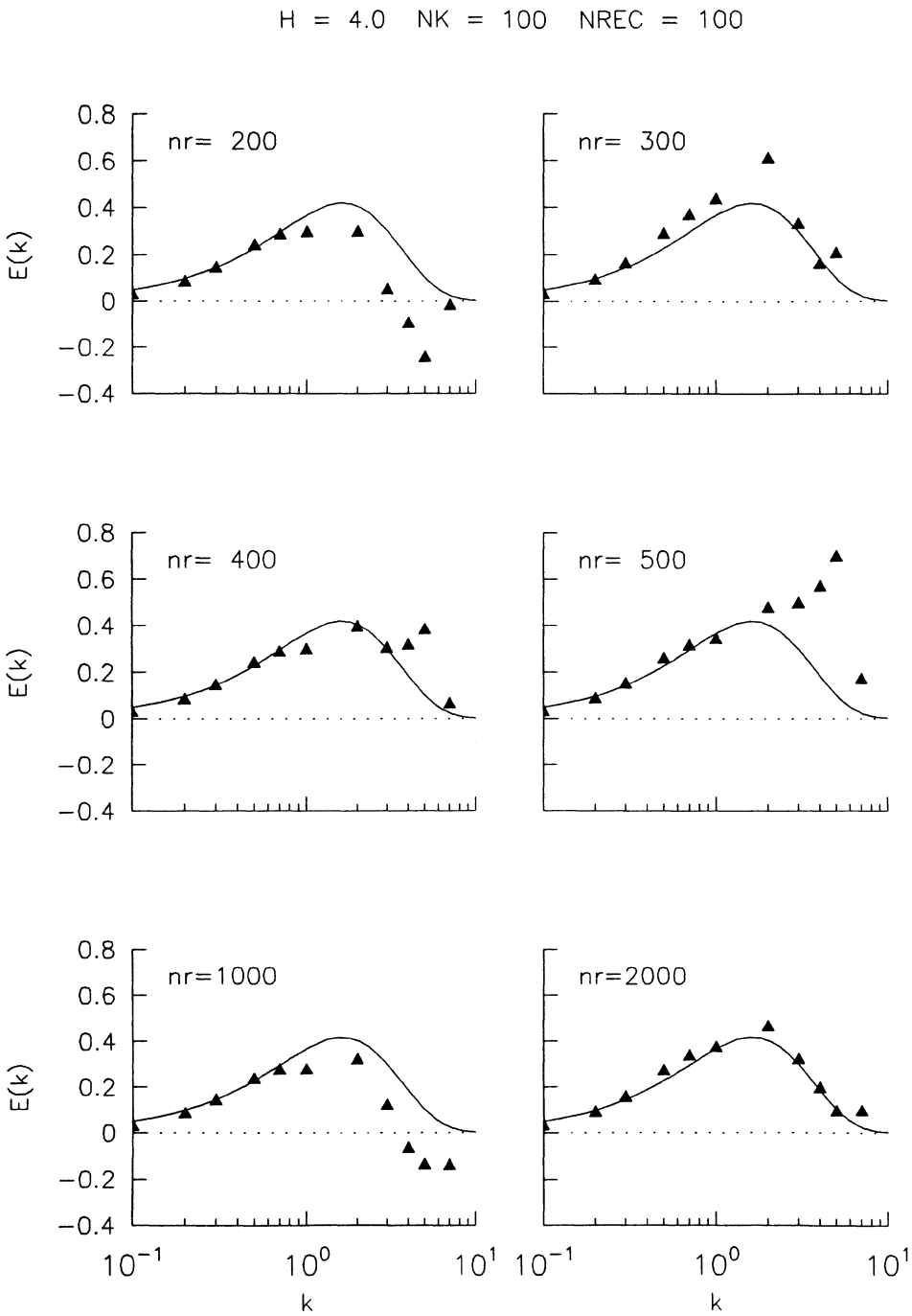
Brillinger, D. and Rosenblatt, M. (1967) 'Computation and interpretation of kth order spectra,' in B. Harris, (ed.) Spectral Analysis of Times Series, J. Wiley, New York, 189–232.

Komlos, J. Major, P., and Tusnady, G. (1975) 'An approximation of partial sums of independent RV's and the sample D.F.I.,' Z. Wahrscheinlichkeitstheorie und verw. Geb. **32**, 111–131.

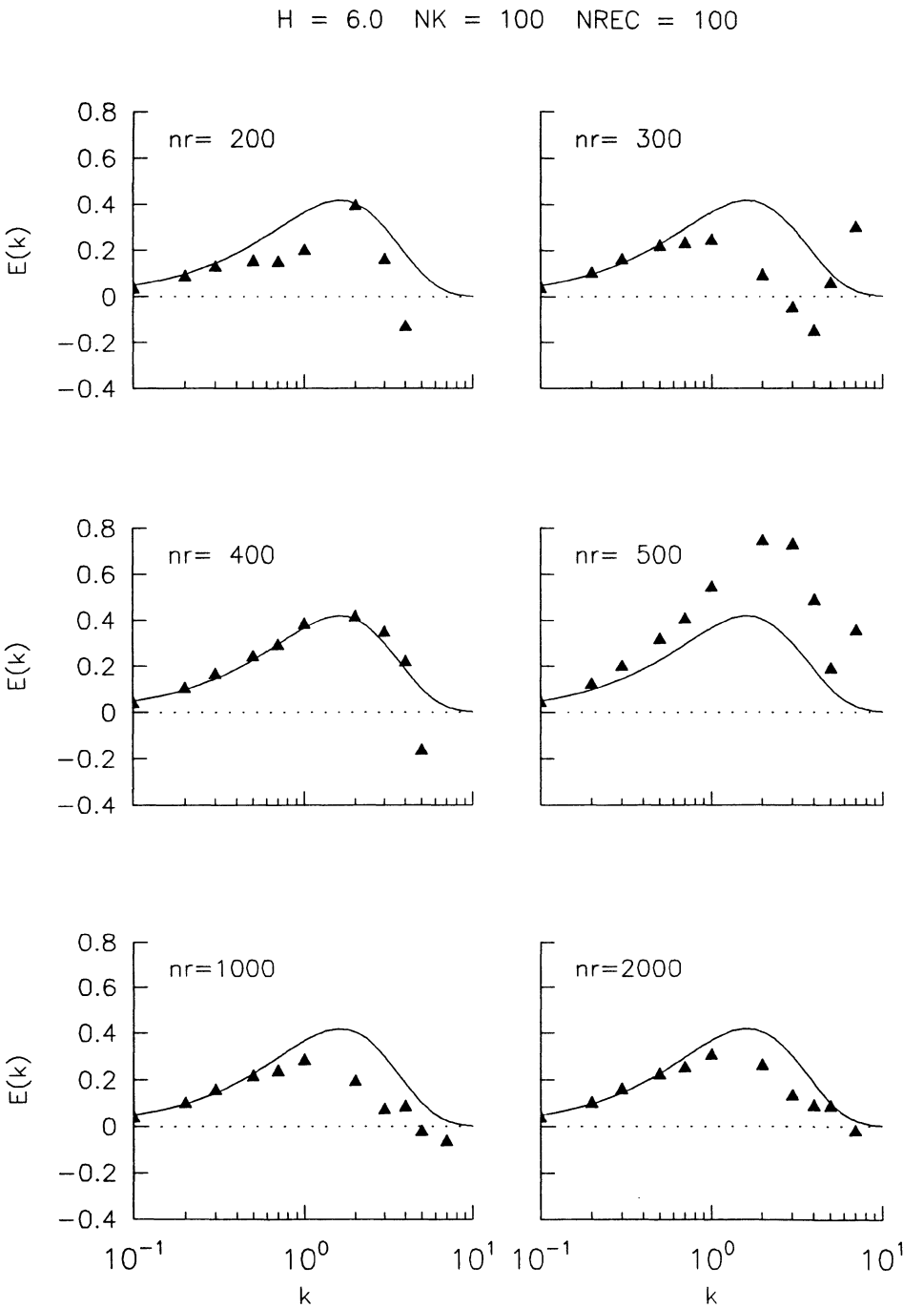
Tusnady, G. (1977) 'A remark on the approximation of the sample DF in the multidimensional case,' Per. Math. Hung. **8**, 53–55.

Massart, P. (1989) 'Strong approximation for multivariate empirical and related processes, via KMT constructions,' Ann. Prob., **17**, 266–291.

Bickel, P. and Rosenblatt, M. (1973) 'Two-dimensional random fields,' in P. Krishnaiah (ed.) Multivariate Analysis III, Academic Press, 3–15.

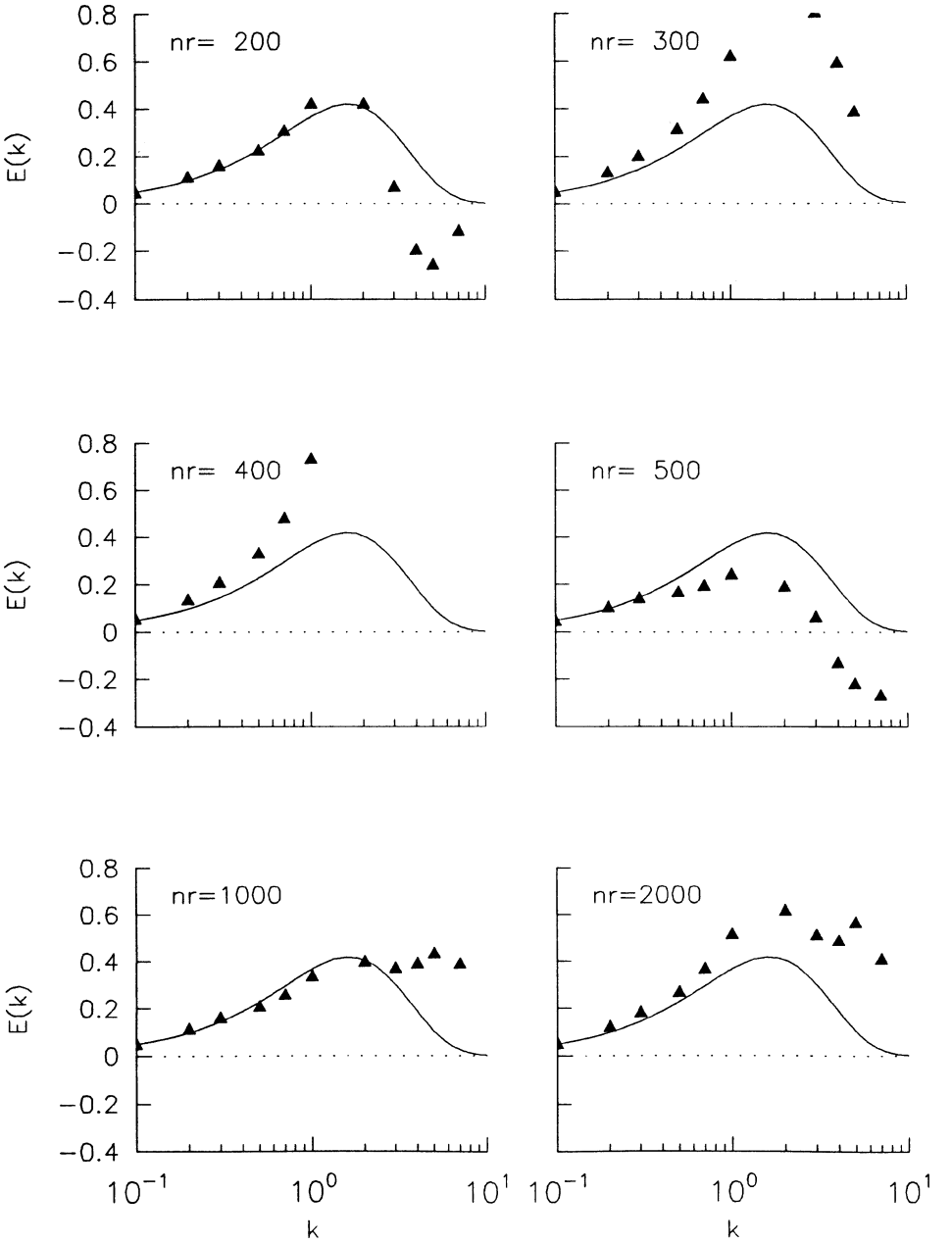

H = 4.0 NK = 100 NREC = 100
nr= 200
nr= 300
nr= 400
nr= 500
nr=1000
nr=2000
E(k)
0.8
0.6
0.4
0.2
0
−0.2
−0.4
10^{-1}
10^{0}
10^{1}
k

Set No. 1.

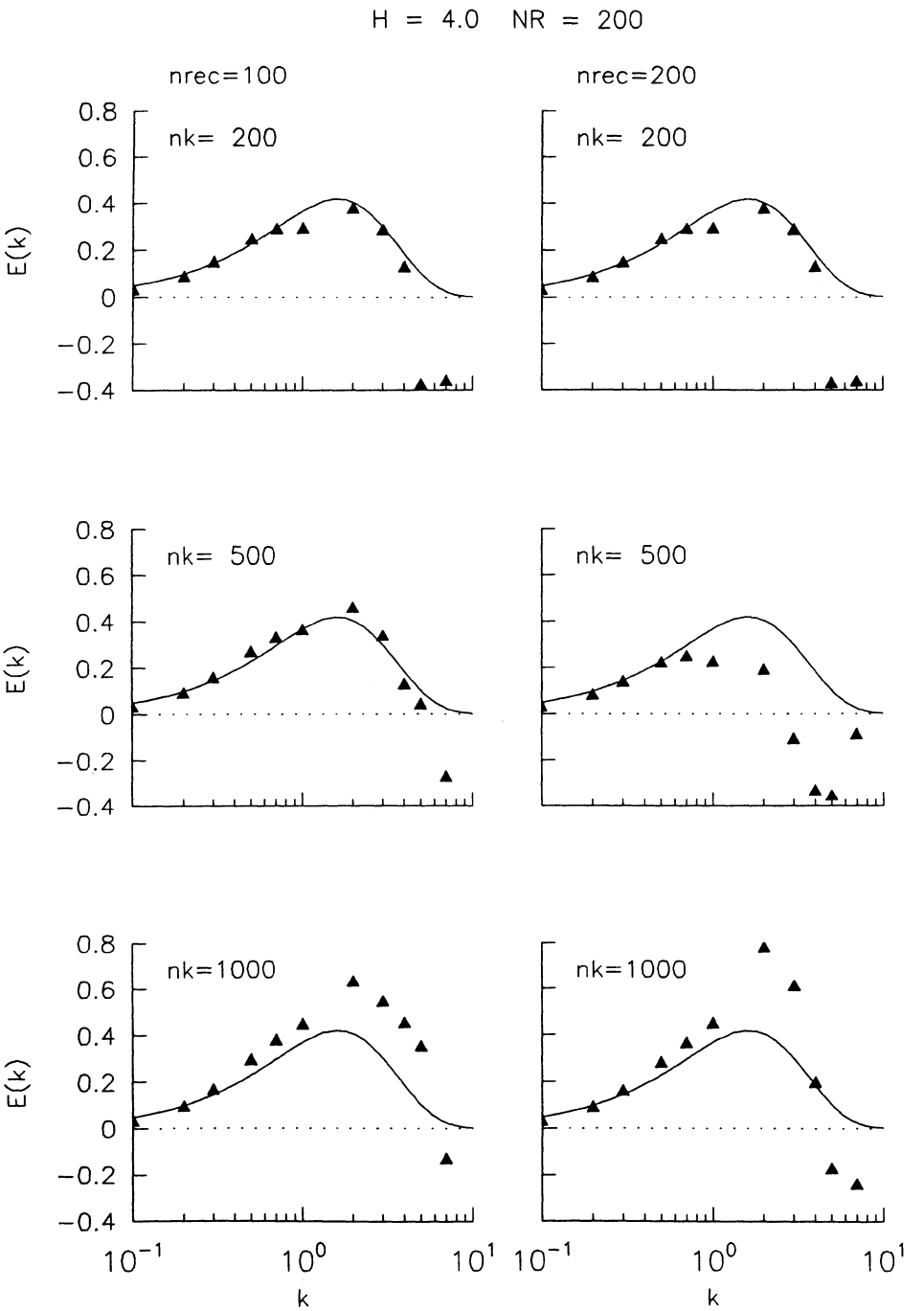

H = 6.0 NK = 100 NREC = 100
nr= 200
nr= 300
nr= 400
nr= 500
nr=1000
nr=2000
E(k)
0.8
0.6
0.4
0.2
0
−0.2
−0.4
10^{-1}
10^{0}
10^{1}
k

Set No. 2.

H = 8.0 NK = 100 NREC = 100



Set No. 3.


H = 4.0 NR = 200
nrec=100
nrec=200
nk= 200
nk= 200
nk= 500
nk= 500
nk=1000
nk=1000
E(k)
E(k)
E(k)
0.8
0.6
0.4
0.2
0
−0.2
−0.4
10^{-1}
10^{0}
10^{1}
k
k

Set No. 4.

KERNEL DENSITY ESTIMATION UNDER A LOCALLY MIXING CONDITION

Tuan Dinh PHAM
Laboratory of Modelling
and Computation, CNRS
University of Grenoble
FRANCE

Lanh Tat TRAN
Mathematics Department
Indiana University
Bloomington, IN 47405
U.S.A.

ABSTRACT. A dependence condition weaker than the strong mixing condition is formulated. Kernel density estimates of stationary sequences satisfying this condition are investigated. The uniform convergence of the estimates is obtained under weak assumptions. The trade off between the rates at which the bandwidths and dependence coefficients tend to zero is explicitly determined. A method to estimate the rate of decay of the dependence coefficients is also proposed.

1. Introduction.

The nonparametric estimation of a probability density $f(x)$ is an interesting problem in statistical inference and has an important role in communication theory and pattern recognition. An account of this information can be found in Papantoni-Kazakos and Kazakos (1977). The literature dealing with density estimation when the observations are independent is extensive. The reader is referred to Wegman (1972) and Devroye and Györfi (1985) for background material. The purpose of this paper is to investigate kernel type density estimators when the observations are dependent. For a bibliography and relevant papers, see Roussas (1969, 1988), Györfi (1981), Robinson (1983, 1987), Yoshihara (1984), Yakowitz (1985, 1987), Masry (1986), Hart and Vieu (1988), Györfi, Härdle, Sarda and Vieu (1989).

Let $X_t, t = \ldots, -1, 0, 1, \ldots$ be a strictly stationary sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and taking values in R^d. Let $\mathcal{F}_t$ be the σ–field generated by X_t. Throughout the paper, we assume that X_t satisfies a general dependence condition to be defined below:

Definition 1.1. Let Z be the set of integers. Let $S \subset Z$ and let $\mathcal{F}_S$ denote the σ–field generated by $X_t, t \in S$. The process X_t is said to be locally mixing (LM) if for all $S \subset Z$ and $t \in Z - S$,

$$\alpha(S,t) \equiv \sup\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_S, B \in \mathcal{F}_t\} \leq C\|S\|^\theta \phi(\hat{d}(s,t)).$$

for some $C > 0, \theta \geq 0$ and some function $\phi : [1, \infty) \to [0, \infty)$ with $\phi(\hat{d}) \downarrow 0$ as $\hat{d} \to \infty$. Here $\|S\|$ denotes the cardinality of the set S and $\hat{d}(S,t)$ denotes the Euclidean distance between the set S and the point t.

The definition of the LM property is reminiscent of the strong mixing property defined by Rosenblatt (1956a) as follows:

Let $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_n^\infty$ denote respectively the σ–fields generated by $X_t, t \leq 0$ and by $X_t, t \geq n$. Then X_t is strong mixing if

$$\alpha(n) = \sup\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_n^\infty, B \in \mathcal{F}_{-\infty}^0\} \to 0.$$

The strong mixing condition is well known to be weaker than many other dependence conditions, for example, the absolutely regular condition or the ϕ–mixing condition. The class of strong mixing processes is also larger than the class of asymptotically uncorrelated processes. Many

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 419–430.
© 1991 Kluwer Academic Publishers.

stochastic processes are known to be strong mixing. Gorodettski (1977) has obtained various conditions for linear processes to be strong mixing. Under weak assumptions autoregressive moving average processes are strong mixing with $\alpha(n)$ decaying to zero at an exponential rate (see Pham and Tran (1985)). For more information on strong mixing processes, see Rosenblatt (1956a), Ibragimov (1962) or Roussas (1988). The LM property is weaker than the strong mixing property. Indeed, if X_t is strong mixing then X_t satisfies the LM property with $\theta = 0$.

Recently, the study of strong mixing processes has generated a considerable amount of interest among applied researchers. The reason is at least partly due to the fact that many stochastic processes, among them various useful time series models, satisfy the strong mixing property. The results on strong mixing processes normally assume certain rates of decay of the mixing coefficients. To apply them to real life data, some information about these rates would be useful. One possibility in time series analysis is to fit the data to some known time series models, for example, autoregressive moving average time series models or bilinear models, and then use the mixing rates that these models are known to have. This technique is, however, limited in scope and fails when the data does not fit any known time series models. For an interesting example where the data does not fit any existing models, see Yakowitz (1987). The LM condition appears easier to employ than the strong mixing condition in this situation. This will be discussed in Section 3.

Let $f(x)$ be the density of X_1 which is assumed to be continuous. As an estimator of $f(x)$ we shall consider the kernel estimate defined by

$$f_n(x) = (nh_n^d)^{-1} \sum_{j=1}^{n} K((x - X_j)/h_n),$$

Here K is a kernel function and $\{h_n\}$ is a sequence of bandwidths tending to zero as n tends to infinity.

Our results show that $f_n(x)$ converges to $f(x)$ uniformly on compact sets at rates only marginally slower than those in the independent case when $\phi(n)$ decays to zero at an exponential rate, which is the case for a wide class of autoregressive moving average time series models and bilinear models. The appropriate bandwidths to achieve these rates of convergence are explicitly given.

Parameter estimation in time series analysis is usually carried out under the Gaussian assumption. It is often useful to check whether the density of a time series is Gaussian or close to it and also to determine the nature of its non Gaussianity. In this case the interest is often on the overall behavior of the density. Kernel density methods have an important role in density estimation. The case of independent observations has been studied extensively since the works of Rosenblatt (1956b) and Parzen (1962). Under dependent situations, kernel type density estimators are considered in Roussas (1969) for Markov processes, Robinson (1984), Roussas (1988) and Tran (1989a, b, 1990a, b) for various mixing processes. For some applications of density estimation under dependence, the reader is referred to Roussas (1989, 1990) and Izenman and Tran (1990).

The method of proving the results in this paper is based on approximations of dependent r.v.'s by independent ones. The blocking method employed here is rather nonstandard and is distinct from the usual method of blocking r.v.'s into alternate long and short blocks (see Tran (1989b, 1990a, b)). The letter C will be used to denote constants whose values are unimportant and may vary. All limits are taken as $n \to \infty$ unless indicated otherwise.

2. The LSM property and convergence of $f_n(x)$.

The kernel function K is assumed to satisfy the following conditions:

Assumption 1. Assume $\int K(x)dx = 1$ and K has an integrable radial majorant, i.e., $Q(x) \equiv \sup\{|K(y)| : \|y\| > \|x\|\}$ is integrable, where $\|x\| = \max(|x_1|, |x_2|, \ldots, |x|_d)$.

Assumption 2. K satisfies the Lipschitz condition

$$|K(x) - K(y)| < C\|x - y\|.$$

Note that Assumptions 1 and 2 imply that K is bounded by a constant $\tilde{K}$. Let

$$K_n(x) = (1/h_n^d)K(x/h_n)$$

and

$$\eta_j(x) = K_n(x - X_j)\,,\ \Delta_j(x) = \eta_j(x) - E\eta_j(x). \tag{2.1}$$

Then

$$f_n(x) = (1/n)\sum_{j=1}^{n} K_n(x - X_j) = (1/n)\sum_{j=1}^{n} \eta_j(x).$$

Lemma 2.1. Suppose h_n tends to zero and Assumption 1 holds. Then

$$h_n^d \sup_{x\in D} E\Delta_1^2(x) \le C,$$

where C is a constant independent of n and D is a compact subset of R^d.

Proof. Clearly

$$h_n^d E\Delta_1^2(x) \le (1/h_n^d)\int K^2((x-u)/h_n)f(u)du.$$

Under Assumption 1, by the Lebesgue density theorem (see Stein (1970)),

$$\lim_{n\to\infty}\int (1/h_n^d)K^2((x-u)/h_n)f(u)du = f(x)\int K^2(u)du.$$

The proof of the lemma follows since $\int K^2(u)du < \infty$ and since f is bounded on compacts. We will need the following result from Bradley (1983):

Lemma 2.2. Suppose X and Y are random variables taking their values on $\mathcal{L}$ and R respectively where $\mathcal{L}$ is a Borel space; suppose U is a uniform–$[0,1]$ r.v. independent of (X,Y); and suppose ξ and ν are positive numbers such that $\xi \le \|Y\|_\nu < \infty$. Then there exists a real–valued r.v. $Y^* = f(X,Y,U)$, where f is a measurable function from $\mathcal{L}\times R\times[0,1]$ into R, such that

(i) Y^* is independent of X,

(ii) the probability distributions of Y and Y^* are identical, and

(iii) $P(|Y^* - Y| \ge \xi)$

$$\le 18(\|Y\|_\nu/\xi)^{\nu/(2\nu+1)}[\sup|P(A\cap B) - P(A)P(B)| : A\in\mathcal{B}(X), B\in\mathcal{B}(Y)]^{2\nu/(2\nu+1)},$$

where $\mathcal{B}(X), \mathcal{B}(Y)$ are the σ–fields generated by X and Y respectively; and where $\|Y\|_\nu = (E|Y|^\nu)^{1/\nu}$.

Theorem 2.1. (i) Suppose Assumptions 1 and 2 hold and X_t satisfies the LM property. Let $m = m(n)$ be a sequence of nondecreasing positive integers with $m/n \to 0$. Assume

$$m(nh_n^d)^{-1} \log n \to 0. \tag{2.2}$$

Let θ and ϕ be as defined in Definition 1.1. Let

$$\Psi(n) = (m \log n)^{1/2}(nh_n^d)^{-1/2}, \tag{2.3}$$

and suppose h_n tends to zero in a manner that

$$h_n^{-d(d+2+\tau)/2} m^{-(d+\tau-4\theta\tau)/2} (\log n)^{-(d+\tau)/2} n^{(d+\tau+2)/2} \phi^{2\tau}(m) g(n) \to 0 \tag{2.4}$$

for some $1/2 > \tau > 0$ and some function $g(n) > 0$ with $\sum_{n=1}^{\infty}(1/g(n)) < \infty$. Then

$$\sup_{x \in D} |f_n(x) - Ef_n(x)| = O(\Psi(n)) \text{ a.s.}$$

for any compact set D.
(ii) If in addition $|f(x) - f(y)| \le C\|x - y\|, \int \|x\||K(x)|dx < \infty$ and

$$h_n^{1+(d/2)}(mn^{-1} \log n)^{-1/2} = O(1); \tag{2.5}$$

then

$$\sup_{x \in D} |f_n(x) - f(x)| = O(\Psi(n)) \text{ a.s.} \tag{2.6}$$

Proof. Note that

$$f_n(x) - Ef_n(x) = (1/n) \sum_{j=1}^{n} \Delta_j(x).$$

Let $a = h_n^{(d+1)}\Psi(n)$. Since D is compact, it can be covered with, say v cubes I_k's having sides of length a and center at x_k. Using the value of $\Psi(n)$ in (2.3),

$$\begin{aligned} v \le C(h_n^{(d+1)}\Psi(n))^{-d} &= C(h_n^{(d/2)+1}(m \log n)^{1/2} n^{-1/2})^{-d} \\ &= Ch_n^{-(d^2+2d)/2}(m \log n)^{-d/2} n^{d/2}. \end{aligned} \tag{2.7}$$

Now

$$\begin{aligned} \sup_{x \in D} |f_n(x) - Ef_n(x)| &\le \max_{1 \le k \le v} \sup_{x \in I_k} |f_n(x) - f_n(x_k)| \\ &+ \max_{1 \le k \le v} |f_n(x_k) - Ef_n(x_k)| + \max_{1 \le k \le v} \sup_{x \in I_k} |Ef_n(x_k) - Ef_n(x)|. \end{aligned} \tag{2.8}$$

For $x \in I_k$, by Assumption 2,

$$|f_n(x) - f_n(x_k)| \le Ch_n^{-(d+1)}\|x - x_k\| \le Ch_n^{-(d+1)}a = O(\Psi(n)) \text{ a.s.} \tag{2.9}$$

Therefore,

$$\max_{1\le k\le v}\sup_{x\in I_k}|f_n(x)-f_n(x_k)| = O(\Psi(n)) \quad \text{a.s.} \tag{2.10}$$

$$\max_{1\le k\le v}\sup_{x\in I_k}|Ef_n(x_k)-Ef_n(x)| = O(\Psi(n)) \quad \text{a.s.}$$

By (2.8) and (2.10), it remains to show that

$$\max_{1\le k\le v}|f_n(x_k)-Ef_n(x_k)| = O(\Psi(n)) \quad \text{a.s.} \tag{2.11}$$

Denote n/m by ℓ for brevity. Define $\bar{\Delta}_i(x) = \Delta_i(x)/n$, and let

$$Z_i(x) = \bar{\Delta}_i(x) + \bar{\Delta}_{i+m}(x) + \ldots + \bar{\Delta}_{i+([\ell]-1)m}(x)$$

for $i = 1, 2, \ldots, m$ where $[\ell]$ indicates the integer part of ℓ . Choose

$$\epsilon \equiv \epsilon(n) = 4\beta\Psi(n), \tag{2.12}$$

where β is a large positive constant to be prescribed later on in (2.32).

If n is a multiple of m, then $\sum_{i=1}^m Z_i(x) = \sum_{i=1}^n \bar{\Delta}_i(x)$ and

$$\begin{aligned} &P\left(\max_{1\le k\le v}|f_n(x_k)-Ef_n(x_k)|>\epsilon\right) \\ &= P\left(\max_{1\le k\le v}|\sum_{j=1}^n \bar{\Delta}_j(x_k)|>\epsilon\right) \le \sum_{k=1}^v P\left(|\sum_{j=1}^n \bar{\Delta}_j(x_k)|>\epsilon\right) \\ &\le \sum_{k=1}^v P[|Z_i(x_k)|>\epsilon/m \text{ for some } 1\le i\le m] \\ &\le \sum_{k=1}^v\sum_{i=1}^m P[|Z_i(x_k)|>\epsilon/m] \le vm\max_{1\le k\le v}P[|Z_1(x_k)|>\epsilon/m] \end{aligned} \tag{2.13}$$

since the $Z_i(x_k)$'s are identically distributed.

If n is not a multiple of m, then

$$\begin{aligned} &P\left[\max_{1\le k\le v}|f_n(x_k)-Ef_n(x_k)|>\epsilon\right] \\ &\le \sum_{k=1}^v P[|Z_i(x_k)|>\epsilon/(2m) \text{ for some } 1\le i\le m] \\ &+ P[|\bar{\Delta}_{[\ell]m+1}+\ldots+\bar{\Delta}_n|>\epsilon/2]. \end{aligned}$$

However,

$$|\bar{\Delta}_{[\ell]m+1}+\ldots+\bar{\Delta}_n| \le 2\tilde{K}m(nh_n^d)^{-1}.$$

But

$$\begin{aligned} &2\tilde{K}m(nh_n^d)^{-1}\epsilon^{-1} \le Cm(nh_n^d)^{-1}(m\log n)^{-1/2}(nh_n^d)^{1/2} \\ &= Cm^{1/2}(nh_n^d)^{-1/2}(\log n)^{-1/2} = C(m(nh_n^d)^{-1}(\log n)^{-1})^{1/2} \to 0, \end{aligned}$$

by (2.2). Therefore, $P[|\bar{\Delta}_{[\ell]m+1} + \ldots + \bar{\Delta}_n| > \epsilon/2]$ is zero for sufficiently large n. Thus we may assume without loss of generality that n is a multiple of m.

We will write $Z_1(x_k)$ and $\bar{\Delta}_i(x_k)$ simply as Z_1 and $\bar{\Delta}_i$ for brevity. Lemma 2.2 will now be employed to approximate the r.v.'s $\bar{\Delta}_1, \bar{\Delta}_{1+m}, \ldots, \bar{\Delta}_{1+(\ell-1)m}$ by independent identically distributed r.v.'s $W_1, \ldots, W_\ell$ as follows: by enlarging the probability space if necessary, introduce a sequence $(U_1, U_2, \ldots)$ of independent uniform $[0,1]$ random variables, which are independent of our given sequence $\bar{\Delta}_1, \bar{\Delta}_{1+m}, \ldots, \bar{\Delta}_{1+(\ell-1)m}$. Let $\nu = \tau/(1-2\tau)$ so that $\nu/(2\nu+1) = \tau$. Define $W_1 = \bar{\Delta}_1$. By Lemma 2.2, for $\xi \le \|\bar{\Delta}_1\|_\nu$ and for each $\ell \ge j \ge 2$, there exists a r.v. W_j which is a measurable function of $\bar{\Delta}_1, \bar{\Delta}_{1+m}, \ldots, \bar{\Delta}_{1+(j-1)m}, U_{j-1}$ such that W_j is independent of $\bar{\Delta}_1, \bar{\Delta}_{1+m}, \ldots, \bar{\Delta}_{1+(j-2)m}$, has the same distribution as $\bar{\Delta}_{1+(j-1)m}$ and satisfies

$$P[|\bar{\Delta}_{1+(j-1)m} - W_j| \ge \xi] \le 18(\|W_j\|_\nu/\xi)^\tau \{m^\theta \phi(m)\}^{2\tau}. \tag{2.14}$$

Let

$$\eta = \epsilon/(2m). \tag{2.15}$$

Then

$$\begin{aligned} P[|Z_1| > \epsilon/m] &= P[|\sum_{j=0}^{\ell-1} \bar{\Delta}_{1+jm}| > \eta] \\ &\le P[|\sum_{j=1}^{\ell} W_j| > \eta/2] + P[|\sum_{j=1}^{\ell} \bar{\Delta}_{1+(j-1)m} - W_j| > \eta/2]. \end{aligned} \tag{2.16}$$

Define

$$\lambda \equiv \lambda(n) = ((n^2/\ell)h_n^d \log n)^{1/2}. \tag{2.17}$$

By Markov inequality and the independence of the W_j's,

$$\begin{aligned} P[|\sum_{j=1}^{\ell} W_j| > \eta] &= P[\sum_{j=1}^{\ell} W_j > \eta] + P[\sum_{j=1}^{\ell} -W_j > \eta] \\ &\le e^{-\lambda\eta} \left(\Pi_{j=1}^{\ell} E[\exp \lambda W_j] + \Pi_{j=1}^{\ell} E[\exp(-\lambda W_j)]\right). \end{aligned} \tag{2.18}$$

Now $|W_j| \le \tilde{K}/(nh_n^d)$ a.s. since $\Delta_{1+(j-1)m} \le \tilde{K}/h_n^d$ a.s. and W_j has the same distribution as $\bar{\Delta}_{1+(j-1)m} = n^{-1}\Delta_{1+(j-1)m}$. Thus

$$\lambda|W_j| \le \tilde{K}((n^2/\ell)h_n^d \log n)^{1/2}(nh_n^d)^{-1} = \tilde{K}(m(nh_n^d)^{-1}\log n)^{1/2}. \tag{2.19}$$

By assumption $m(nh_n^d)^{-1}\log n \to 0$. Thus $\lambda|W_j| \le 1/2$ for large n. Consequently, there exists an N such that

$$\exp(\lambda W_j) \le 1 + \lambda W_j + W_j^2\lambda^2 \quad \text{and } \exp(-\lambda W_j) \le 1 - \lambda W_j + W_j^2\lambda^2$$

for $n > N$. Hence, for large n,

$$\begin{aligned} &P[|\sum_{j=1}^{\ell} W_j| > \eta] \\ &\le e^{-\lambda\eta} \left(\Pi_{j=1}^{\ell}[1 + \lambda^2 EW_j^2] + \Pi_{j=1}^{\ell} E[1 + \lambda^2 EW_j^2]\right) \\ &\le e^{-\lambda\eta} \left(\Pi_{j=1}^{\ell} \exp[\lambda^2 EW_j^2] + \Pi_{j=1}^{\ell} \exp[\lambda^2 EW_j^2]\right) \\ &= 2\exp(-\lambda\eta)\exp\Big(\sum_{j=1}^{\ell} \lambda^2 EW_j^2\Big) = 2\exp(-\lambda\eta + \lambda^2 \sum_{j=1}^{\ell} EW_j^2). \end{aligned} \tag{2.20}$$

By (2.20), for large n,

$$P[|\sum_{j=1}^{\ell} W_j| > \eta/2] \leq 2\exp((-\lambda\eta/2) + \lambda^2 \sum_{j=1}^{\ell} EW_j^2). \tag{2.21}$$

By Lemma 2.1,

$$\sum_{j=1}^{\ell} EW_j^2 \leq C\ell h_n^{-d} n^{-2}. \tag{2.22}$$

From (2.21) and (2.22),

$$P[|\sum_{j=1}^{\ell} W_j| > \eta/2] \leq 2\exp((-\lambda\eta/2) + C\lambda^2 \ell h_n^{-d} n^{-2}). \tag{2.23}$$

A simple computation shows that

$$C\lambda^2 \ell h_n^{-d} n^{-2} = C \log n \tag{2.24}$$

by the choice of λ in (2.17) above. By (2.12), (2.15) and (2.3),

$$\begin{aligned} \eta &= 2\beta\Psi(n)/m = 2\beta(m\log n)^{1/2}(nh_n^d)^{-1/2}/m \\ &= 2\beta(\log n)^{1/2}(nh_n^d)^{-1/2}(n/\ell)^{-1/2} = 2\beta\log n((n^2/\ell)h_n^d \log n)^{-1/2}. \end{aligned} \tag{2.25}$$

Thus

$$\lambda\eta/2 = \beta \log n. \tag{2.26}$$

From (2.23), (2.24) and (2.26),

$$P[|\sum_{j=1}^{\ell} W_j| > \eta/2] \leq 2\exp((-\beta + C)\log n) = 2n^{-(\beta+C)}. \tag{2.27}$$

Turning now to the last term of (2.16) and using (2.14), we have for $\eta/(2\ell) \leq \|\bar{\Delta}_1\|_\nu$

$$\begin{aligned} P[|\sum_{j=1}^{\ell}(\bar{\Delta}_{1+m(j-1)} - W_j)| > \eta/2] &\leq \ell \max_{1\leq j\leq \ell} P[|\bar{\Delta}_{1+m(j-1)} - W_j| > \eta/(2\ell)] \\ &\leq C\ell \max_{1\leq j\leq \ell} (\ell/\eta)^\tau \|W_j\|_\nu^\tau \{m^\theta \phi(m)\}^{2\tau} \\ &\leq C\ell(\ell/\eta)^\tau n^{-\tau} h_n^{-d\tau} \{m^\theta \phi(m)\}^{2\tau} \end{aligned} \tag{2.28}$$

since $|W_j| \leq \tilde{K}/(nh_n^d)$ a.s.

Note that

$$(\ell/\eta)^\tau n^{-\tau} h_n^{-d\tau} \geq C\epsilon^{-\tau} h_n^{-d\tau}.$$

Now $\epsilon^{-\tau} h_n^{-d\tau} \to \infty$ since $\epsilon = 4\beta\Psi(n) \to O$ by (2.2) and $h_n \to 0$. Thus

$$(\ell/\eta)^\tau n^{-\tau} h_n^{-d\tau} \to \infty.$$

For $\eta/(2\ell) > \|\bar{\Delta}_1\|_\nu$, we then have by (2.14)

$$\begin{aligned}\max_{1\le j\le \ell} P[|\bar{\Delta}_{1+m(j-1)} - W_j)| > \eta/(2\ell)] &\le \max_{1\le j\le \ell} P[|\bar{\Delta}_{1+m(j-1)} - W_j| > \|\bar{\Delta}_1\|_\nu] \\ &\le C\{m^\theta \phi(m)\}^{2\tau} \\ &\le C(\ell/\eta)^\tau n^{-\tau} h_n^{-d\tau}\{m^\theta\phi(m)^{2\tau}\}.\end{aligned}$$

Thus again (2.28) holds.

From (2.13), (2.16), (2.27) and (2.28),

$$\begin{aligned}&P\left[\max_{1\le k\le v} |f_n(x_k) - Ef_n(x_k)| > \epsilon\right] \\ &\le Cvmn^{-(\beta+C)} + Cvm\ell(\ell/\eta)^\tau n^{-\tau} h_n^{-d\tau}\{m^\theta\phi(m)\}^{2\tau} \\ &\le Cvmn^{-(\beta+C)} + Cvm^{1+2\theta\tau}\ell(\ell/\eta)^\tau n^{-\tau} h_n^{-d\tau}\phi^{2\tau}(m) \equiv Cq_{1n} + Cq_2(n), \text{ say.}\end{aligned} \tag{2.29}$$

A simple computation using (2.7) yields

$$\begin{aligned}q_{1n} &\le Ch_n^{-(d^2+2d)/2}(m\log n)^{-d/2} n^{d/2} mn^{-(\beta+C)} \\ &\le Ch_n^{-(d^2+2d)/2} n^{d/2} mn^{-(\beta+C)} \le Ch_n^{-(d^2+2d)/2} n^{d/2} n^{1-(\beta+C)} \\ &= Ch_n^{-(d^2+2d)/2} n^{1-(\beta+C)+(d/2)}.\end{aligned} \tag{2.30}$$

By (2.2), we have $h_n > Cn^{-1/d}$ since $nh_n^d \to \infty$. Thus, by (2.30)

$$q_{1n} \le Cn^{(d^2+2d)/(2d)} n^{1-(\beta+C)+(d/2)}. \tag{2.31}$$

From (2.31), it is easily seen that

$$\sum_{n=1}^{\infty} q_{1n} < \infty, \tag{2.32}$$

if β is chosen sufficiently large.

Turning now to the last term of (2.29), we have

$$\begin{aligned}q_{2n} &= vm^{1+2\theta\tau}\ell^{1+\tau}\eta^{-\tau}n^{-\tau}h_n^{-d\tau}\phi^{2\tau}(m) \\ &= vm^{1+2\theta\tau}(n/m)^{1+\tau}\eta^{-\tau}n^{-\tau}h_n^{-d\tau}\phi^{2\tau}(m) \\ &= vm^{(2\theta-1)\tau}n(\epsilon/(2m))^{-\tau}h_n^{-d\tau}\phi^{2\tau}(m) \\ &\le Cvm^{2\theta\tau}n\epsilon^{-\tau}h_n^{-d\tau}\phi^{2\tau}(m) \\ &\le C(h_n^{(d+1)}\Psi(n))^{-d}m^{2\theta\tau}n\epsilon^{-\tau}h_n^{-d\tau}\phi^{2\tau}(m) \\ &\le Ch_n^{-d(d+1+\tau)}\{(m\log n)^{1/2}(nh_n^d)^{-1/2}\}^{-d-\tau}m^{2\theta\tau}n\phi^{2\tau}(m) \\ &= Ch_n^{-d(d+1+\tau)}m^{-(d+\tau-4\theta\tau)/2}(\log n)^{-(d+\tau)/2}n^{(d+\tau+2)/2}h_n^{d(d+\tau)/2}\phi^{2\tau}(m) \\ &= Ch_n^{-d(d+2+\tau)/2}m^{-(d+\tau-4\theta\tau)/2}(\log n)^{-(d+\tau)/2}n^{(d+\tau+2)/2}\phi^{2\tau}(m).\end{aligned} \tag{2.33}$$

By (2.4) and (2.33),

$$\sum_{n=1}^{\infty} q_{2n} < \infty \tag{2.34}$$

since $\sum_{n=1}^{\infty}(1/g(n)) < \infty$. The proof of (i) follows from the Borel Cantelli Lemma and (2.29), (2.32) and (2.34).

(ii) By the Lipschitz condition of f and since $\int K(x)dx = 1$ and $\int \|x\|K(x)dx < \infty$, following Roussas (1988, p.141), we have

$$\sup |Ef_n(x) - f(x)| \leq Ch_n.$$

The proof of (ii) follows since $h_n = O(\Psi(n))$ by (2.5).

3. Discussion and examples.

In the independent case, it is not hard to see from the proof of Theorem 2.1 that m can actually be chosen equal to 1. In this case, Condition (2.2) reduces to

$$nh_n^d(\log n)^{-1} \to \infty. \tag{3.1}$$

If Assumptions 1,2 are met and (3.1) hold, then

$$\sup_{x\in D} |f_n(x) - Ef_n(x)| = O\big((\log n)^{1/2}(nh_n^d)^{-1/2}\big).$$

Choose h_n so that h_n tends to zero in such a manner that $h_n(\log n)^{-1/(d+2)}n^{1/(d+2)}$ is bounded away from zero and infinity; then the convergence rate becomes

$$(\log n)^{1/2}n^{-1/2}(\log n)^{-d/(2(d+2))}n^{d/(2(d+2))} = (n^{-1}\log n)^{1/(2+d)},$$

which is the optimal rate of uniform convergence for density estimators (see Stone (1983)). The rate of convergence obtained in the independent case is thus sharp.

Example 3.1. Consider the important case that $\phi(\hat{d}) = e^{-s\hat{d}}$ for some $s > 0$. Choose $m = \beta \log n$ and

$$g(n) = n\log n(\log\log n)^{1+\epsilon}. \tag{3.2}$$

for some $\varepsilon > 0$. Then $\sum_{n=1}^{\infty}(1/g(n)) < \infty$. Since (2.2) implies that $h_n > Cn^{-1/d}$, it is easily seen that (2.4) is satisfied if β is sufficiently large. Theorem 2.1 then gives the following:

Proposition 3.1. Suppose Assumptions 1 and 2 hold and X_t satisfies the LM property with $\phi(\hat{d}) = e^{-s\hat{d}}$ and (ii) of Theorem 2.1 holds. If in addition $nh_n^d(\log n)^{-2} \to \infty$, then

$$\sup_{x\in D} |f_n(x) - f(x)| = O(\log n(nh_n^d)^{-1/2}) \qquad \text{a.s.}$$

Example 3.2. Suppose $\phi(\hat{d}) = \hat{d}^{-\gamma}$ for some $\gamma > 0$. Choose $h_n = n^{-\rho}$ for some $1/d > \rho > 0$. Let $g(n)$ be as in (3.2). Then (2.4) is satisfied if

$$n^{\rho d(d+2+\tau)/2}m^{-(d+\tau-4\theta\tau)/2}(\log n)^{-(d+\tau)/2}n^{(d+\tau+2)/2}m^{-2\gamma\tau}n\log n(\log\log n)^{1+\epsilon} \to 0,$$

or equivalently,

$$n^{\{(\rho d+1)(d+2+\tau)+2\}/2}m^{-(d+\tau-4\theta\tau+4\gamma\tau)/2}(\log n)^{-(d+\tau)/2}\log n(\log\log n)^{1+\epsilon} \to 0,$$

which is in turn satisfied if $m = n^{\mu}$ with

$$\mu > \frac{(\rho d + 1)(d + 2 + \tau) + 2}{d + \tau - 4\theta\tau + 4\gamma\tau} \tag{3.3}$$

for some $0 < \tau < 1/2$. The right hand side of (3.3) is continuous as a function of τ at $\tau = 1/2$. Thus (3.3) is satisfied for some $0 < \tau < 1/2$ if it is satisfied for $\tau = 1/2$, that is,

$$\mu > \frac{(\rho d + 1)(2d + 5) + 4}{2d + 1 - 4\theta + 4\gamma}. \tag{3.4}$$

For (3.4) to have a solution, we need the right hand side to be smaller than 1, that is,

$$(\rho d + 1)(2d + 5) + 4 < 2d + 1 - 4\theta + 4\gamma,$$

or

$$2\rho d^2 + 5\rho d + 8 + 4\theta < 4\gamma,$$

which reduces to

$$\gamma > (2\rho d^2 + 5\rho d + 8 + 4\theta)/4. \tag{3.5}$$

Noting that

$$n^{\mu}(nh_n^d)^{-1} \log n = n^{\mu - 1 + \rho d} \log n,$$

and

$$\Psi(n) = (n^{\mu - 1} \log n)^{1/2} h_n^{-d/2} = (n^{\mu - 1 + \rho d} \log n)^{1/2},$$

we have by Theorem 2.1:

Proposition 3.2. Suppose Assumptions 1 and 2 hold with $h_n = n^{-\rho}$ for some $1/d > \rho > 0$ and (ii) of Theorem 2.1 holds. Suppose X_t satisfies the LM property with $\phi(\hat{d}) = \hat{d}^{-\gamma}$ for some γ satisfying (3.5). If in addition $\mu < 1 - \rho d$ for some μ satisfying (3.4), then

$$\sup_{x \in D} |f_n(x) - f(x)| = O((n^{\mu - 1 + \rho d} \log n)^{1/2}).$$

In Proposition 3.2, we assume X_t satisfies the LM property with $\phi(\hat{d}) = \hat{d}^{-\gamma}$ for some γ satisfying (3.5). It would be of interest to find a method to check whether (3.5) is likely to be satisfied. Let us assume $d = 1$ for simplicity. Then (3.5) reduces to

$$\gamma > (7\rho + 8 + 4\theta)/4 \tag{3.6}$$

Assume for the moment that X_t is a Gaussian process. Let $S \subset Z$ and let $H(S)$ be the closure in mean square of the vector space generated by linear combinations of r.v.'s X_t with $t \in S$. Define

$$r(H(S), H(t)) = \sup\{|\mathrm{corr}(g, h)| : g \in H(S), h \in H(t)\},$$

with $\mathrm{corr}(g, h)$ the correlation of the r.v.'s g, h. By Lemma 2 of Rosenblatt (1985, p. 75),

$$\alpha(S, t) \leq r(H(S), H(t)) \leq 2\pi\alpha(S, t).$$

Thus if the LM condition holds with $\phi(\hat{d}) = \hat{d}^{-\gamma}$ for some $\gamma > 0$, then using Definition 1.1,

$$r(H(S), H(t)) \leq C\|S\|^{\theta}\hat{d}^{-\gamma}. \tag{3.7}$$

In the Gaussian case, $r(H(S), H(t))$ is simply the multiple correlation between the r.v.'s $X_s, s \in S$ and X_t. Estimates of $r(H(S), H(t))$ for various S and t can be obtained from relevant sample multiple correlations. A curve of the form $C\|S\|^{\theta}\hat{d}^{-\gamma}$ can be fitted to these point estimates to obtain estimates, say $\hat{\theta}$ and $\hat{\gamma}$, of θ and γ. If the Gaussian assumption is not badly violated, we still can use $\hat{\theta}$ and $\hat{\gamma}$ as rough estimates of θ and γ. Condition (3.6) is probably satisfied if $\hat{\gamma}$ is much greater than $(7\rho + 8 + 4\hat{\theta})/4$.

Recently, density estimation under a weak dependence condition referred to as the strong mixing condition in the locally transitive sence (SMLT) was investigated by Tran (1990a, b). The LM condition here is weaker than the SMLT condition. Estimating the rate of decay of the SMLT coefficients requires estimates of canonical correlations while here only multiple correlation coefficients are involved. However, sharper convergence rates of f_n to f can be obtained under the SMLT condition than under the weaker LM condition.

References

Bradley, R.C. (1983). Approximation theorems for strongly mixing random variables. Michigan Math. J. 30, 69-81.

Devroye, L. and Gyorfi (1985). Nonparametric density estimation: the L_1 view. New York: Wiley.

Gorodetskii, V.V. (1977). On the strong mixing properties for linear sequences. Theory probab. Appl. 22 411-413.

Györfi, L. (1981). Strong consistent density estimate from ergodic sample. J. Multivar. Anal. 11, 81-84.

Györfi, L. Härdle, W., Sarda, P. and Vieu, P. (1989). Nonparametric Curve Estimation from Time Series. New York: Springer-Verlag.

Hart, J.D. and Vieu P. (1988). Data driven bandwidth choice for density estimation based on dependent data. Technical report, Department of Statistics, Texas A&M University, College Station, Texas.

Ibragimov, I.A. (1962) Some limit theorems for stationary sequences. Theor. Probab. Appl. 4, 347–382.

Izenman, A.J. and Tran, L.T. (1990). Kernel estimation of the survival function and hazard rate under weak dependence. J. Statist. plann. Inf. 24, 233-247.

Masry, E. (1986). Recursive probability density estimation for weakly dependent stationary processes. IEEE Trans. Inform. Theory, Vol IT 18, 254-267.

Papantoni-Kazakos, P. and Kazakos, D. (1977). Nonparametric methods in communications. New York: Marcel Dekker.

Parzen, E. (1962). On estimation of a probability density function and mode. Ann Math. Statist 33, 1065-1076.

Pham, D.T. (1987). The mixing property of bilinear and generalized random coefficient autoregressive models. Stoch. Proc. Appl. 23, 291-300.

Pham, D.T. and Tran, L.T. (1985). Some strong mixing properties of time series models. Stoch. Proc. Appl. 19, 297-303.

Robinson, P.M. (1983). Nonparametric estimators for time series. J. Time Series Anal. 4, 185-197.

Robinson, P.M. (1987). Time series residuals with application to probability density estimation. J. Time Series Anal. 3, 329-344.

Rosenblatt, M. (1956a). A central limit theorem and a strong mixing condition. Proc. Nat. Acad. Sci. U.S.A. 42, 43-47.

Rosenblatt, M. (1956b). Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. 27, 832-837.

Rosenblatt, M. (1985). Stationary Sequences and Random Fields. Birkhauser, Boston.

Roussas, G.G. (1969). Non-parametric estimation of the transition distribution of a Markov process. Ann. Inst. Statist. Math., 21, 73-87.

Roussas, G.G. (1988). Nonparametric estimation in mixing sequences of random variables. J. Statist. Plann. Inf. 18, 135-149.

Roussas, G.G. (1989). Hazard rate estimation under dependence conditions. J. Statist. Plann. Inf. 22, 81-93.

Roussas, G.G. (1990). Asymptotic normality of the kernel estimate under dependence conditions: Applications to hazard rate. J. Statist. Plann. Inf. 25, 81-104.

Stein, E.M. (1970). Singular Integrals and Differentiability Properties of Functions. Princeton University Press, Princeton, New Jersey.

Stone, C.J. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function and its derivatives. Recent Advances in Statistics: Papers in honor of H. Chernoff, ed. by M.H. Rezvi, J.S. Rustagi, and D. Siegmund, 393-406.

Tran, L.T. (1989a). The L_1 convergence of kernel density estimates under dependence. Canadian Journ. Statist. 17, 197-208.

Tran, L.T. (1989b). Recursive kernel density estimation under dependence. IEEE Trans. Inform. Theory, Vol 35, 1103-1108.

Tran, L.T. (1990a). Density estimation under dependence. Statist. Prob. Lett. 10, 193-201.

Tran, L.T. (1990b). On multivariate variable kernel density estimates for time series. Manuscript.

Wegman, E.J. (1972). Nonparametric probability density estimation: I. A summary of available methods. Technometrics, Vol. 14, 533-546.

Yakowitz, S. (1985). Nonparametric density estimation, prediction and regression for Markov sequences. J. Amer. Statist. Assoc. 80, 215-221.

Yakowitz, S. (1987). Nearest–neighbor methods for time series analysis. J. Time Series Anal. 8, 235-247.

Yoshihara, K. (1984). Density estimation for samples satisfying a certain absolute regularity condition. J. Statist. plann. Inf. 9, 19-32.

NONPARAMETRIC ESTIMATION OF SURVIVAL FUNCTIONS UNDER DEPENDENT COMPETING RISKS

M. BHASKARA RAO
Department of Statistics
North Dakota State University
FARGO, ND 58105, USA

G. JOGESH BABU
and
C. RADHAKRISHNA RAO
Department of Statistics
Pennsylvania State University
UNIVERSITY PARK, PA 16802, USA

ABSTRACT. In a certain target population, the individuals will die due to either Cause 1 or Cause 2 with probabilities π and $(1-\pi)$, respectively. Let F_1 and F_2 be the life time distributions of individuals who die off due to Causes 1 and 2, respectively. In any random sample of individuals from the population, subjects can leave the study at random times. In this paper, we derive nonparametric estimates of π, F_1 and F_2 using such censored data and study some of their properties. The model that suggests itself encapsulating the essentials of the problem is more general than the usual competing risks model.

1. INTRODUCTION

The work that is carried out in this paper is the culmination of our effort to analyze certain data set whose outline was presented to us. The data set consisted of three types of observations. Every member of a certain target population either dies of cancer or other causes. A proportion of π of the population will die of cancer and the proportion $(1-\pi)$ of other causes. The proportion π is unknown. Let F_1 be the distribution function of the life span of an individual

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 431–441.
© 1991 Kluwer Academic Publishers.

selected at random from the sub-population who will die of cancer and F_2 that of the individual from the sub-population who will die of other causes. In order to estimate π, F_1 and F_2, we select n individuals at random from the population and monitor their life times. In the course of this project, some individuals might leave the study at random times, i.e., censoring takes place. At the end of the study, we will have three types of observations.

(a) Individuals who died of cancer and their life times.
(b) Individuals who died of other casuses and their life times.
(c) Individuals who left the study and the times at which they left the study.

There is another dimension to this problem. The research project is financed for a fixed length of time, M, say. Some individuals could be alive at the end of the project period. For simplicity, we ignore this aspect of the problem and assume that every one in the sample can be monitored until one's death. (The period M can easily be incorporated in our model.)

We formulate the above problem mathematically as follows. A certain population is under scrutiny. Every individual in the population dies due to one of two known causes of death : Cause 1 and Cause 2. An individual is selected at random from the population. Let

$$\begin{aligned} \bar{\Delta} &= 1 \quad \text{if he dies of Cause 1,} \\ &= 0 \quad \text{if he dies of Cause 2.} \end{aligned}$$

Let $P(\bar{\Delta} = 1) = \pi$ and $P(\bar{\Delta} = 0) = 1-\pi$. Let X be the life span of the individual. Let F_1 be the conditional distribution of $X \mid \bar{\Delta} = 1$ and F_2 that of $X \mid \bar{\Delta} = 0$. Let F be the distribution function of X. Then

$$F(x) = \pi F_1(x) + (1-\pi)F_2(x), \quad 0 \le x < \infty.$$

Let C be the time (Censoring variable) at which the individual leaves the study and G its distribution function. Assume that C and $(X,\bar{\Delta})$ are independent.

Associated with each individual we thus have a triplet $(X,\bar{\Delta},C)$ which is not observable in its entirety. In practice, we observe the pair (Y,Δ) of random variables for each individual in the population enshrining the types of observations mentioned above, where

$$
\begin{aligned}
Y &= X \quad \text{if} \quad X \le C \quad \text{and} \quad \bar{\Delta} = 1, \\
&= X \quad \text{if} \quad X \le C \quad \text{and} \quad \bar{\Delta} = 0, \\
&= C \quad \text{if} \quad X > C,
\end{aligned}
$$

and

$$
\begin{aligned}
\Delta &= 1 \quad \text{if} \quad X \le C \quad \text{and} \quad \bar{\Delta} = 1, \\
&= 0 \quad \text{if} \quad X \le C \quad \text{and} \quad \bar{\Delta} = 0, \\
&= -1 \quad \text{if} \quad X > C.
\end{aligned}
$$

The variable Δ indicates the nature of the observation, i.e., death due to Cause 1 $(\Delta = 1)$ or death due to Cause 2 $(\Delta = 0)$ or left the study $(\Delta = -1)$. The variable records either the death time or the time at which the individual left the study.

Let f_1, f_2 and g be the probability density functions of the probability measures induced by F_1, F_2 and G, respectively, with respect to a common dominating measure μ_1. Let μ_2 be the counting measure on $\{-1,0,1\}$. Then the joint probability density function h of Y and Δ with respect to the product measure $\mu_1 x \mu_2$ is given by

$$
\begin{aligned}
h(y,1) &= \pi\, f_1(y)\, P(C \ge y), \\
h(y,0) &= (1-\pi)\, f_2(y)\, P(C \ge y), \qquad (1.1) \\
h(y,-1) &= g(y)\, P(X > y), \quad 0 \le y < \infty.
\end{aligned}
$$

This joint density function is useful in estimation discussed in Section 2.

Let (y_1,δ_1), (y_2,δ_2), $\cdots$, (y_n,δ_n) be n independent realizations of (Y,Δ). The main objective of the paper is to estimate π, F_1 and F_2 using the given data. Let $\mathfrak{F}$ be the collection of all distribution functions on $[0,\infty)$. We want to maximize the likelihood of the data over the parameter space $[0,1] \times \mathfrak{F} \times \mathfrak{F} \times \mathfrak{F}$ of

(π, F_1, F_2, G). The family $\mathcal{F}$ of distributions is undominated. The estimation problem falls into the realm of nonparametrics. We will use ideas of Kiefer and Wolfowitz (1956) to derive the generalized maximum likelihood estimates of π, F_1 and F_2 in Section 2. Section 3 discusses estimability problem of the parameters. Section 4 presents a numerical illustration.

Now we will explain how the above model is more general than what is called the "competing risks" model. See Kalbfleisch and Prentice (1980) for ideas and analysis concerning the competing risks model. One may identify the individual as a machine, Cause 1 and Cause 2 as components of the machine and then promote the idea that the causes are competing for the life of the individual. The machine will breakdown if any one of the components fails. Let X_1 be the life span of Component 1 and X_2 that of Component 2 with distribution functions F_1 and F_2, respectively. Let X be the life span of the machine and F its distribution function. One may identify X_1 with the conditioned random variable $X \mid \bar{\Delta} = 1$ and X_2 with $X \mid \bar{\Delta} = 0$. The analogy ends here. In the competing risks model, X_1 and X_2 are assumed to be independent. In this case, π is deterministic and is given by

$$\pi = P(X_1 \leq X_2)$$

and also $$F(x) = \pi F_1(x) + (1-\pi)F_2(x), \quad 0 \leq x < \infty.$$

In the competing risks model, π has a definitive structure and fixed, where as in our model π has no structure. The lack of structure on π makes X_1 and X_2 dependent which is realistic and natural in our problem.

The above framework is natural to use under certain medical premises when it is believed that every individual will eventually either die of cancer or heart diseae and every other form of death such as murder, suicide, accidental death etc. is treated as censoring the true cause of death : cancer or heart disease. The proportion π of individuals who will die of cancer would vary from country to country,

from state to state, from community to community etc. Under the same premises, it is believed in the dictum : YOU ARE WHAT YOU EAT. Every food, vegetarian or not, contains carcinogens and fat. In a laboratory experiment involving mice, medical researchers study how a particular composition of foods consisting of carcinogens and fats at predetermined levels fed to the mice affects the proportion π of mice dying of cancer. The results of this paper are helpful to them in estimating π.

2. GENERALIZED MAXIMUM LIKELIHOOD ESTIMATES

Before we derive the desired estimates, we present some salient features of generalized maximum likelihood method of estimation. See Johansen (1978) for a good account of these ideas.

Suppose X is a random variable or vector whose probability law is unknown but known to belong to a well defined class $\mathcal{P}$ of probability measures. Let x be a realization of X. We distinguish two cases.

Case 1. The family is dominated. This means that there is a σ-finite measure μ such that $P \ll \mu$ for every P in $\mathcal{P}$. Let $dP/d\mu$ be the Radon-Nikodym derivative of P with respect to μ (called the probability density function of X under P with respect to μ). A model $\hat{P}$ in $\mathcal{P}$ is said to be a maximum likelihood estimate if

$$[d\hat{P}/d\mu](x) \geq [dP/d\mu](x)$$

for all P in $\mathcal{P}$. Case 2. The family $\mathcal{P}$ is undominated. This means that there is no measure μ such that $P \ll \mu$ for all P in $\mathcal{P}$. Consequently, we cannot have a probability density function of X under P with respect to one common reference measure to compute the likelihood of the data x under each model P. Kiefer and Wolfowitz (1956) introduced the following definition. A model $\hat{P} \in \mathcal{P}$ is said to be a generalized maximum likelihood estimate if

$$[d\hat{P}/d(P+\hat{P})](x) \geq [dP/d(P+\hat{P})](x) \quad \text{for all } P \text{ in } \mathcal{P}. \qquad (2.1)$$

Two comments are in order on this definition.

1. If $\mathcal{P}$ is dominated, both definitions of likelihood estimates

agree.

2. If $\hat{P}(\{x\}) > 0$, i.e., under $\hat{P}$ the data $\{x\}$ has positive probability of occurence, then $\hat{P}$ is a generalized maximum likelihood estimate if

$$\hat{P}(\{x\}) \geq P(\{x\}) \quad \text{for all } P \text{ in } \mathcal{P}. \tag{2.2}$$

The second comment made above has a practical value. In order to find a generalized maximum likelihood estimate, we look at models in $\mathcal{P}$ under which the given data $\{x\}$ have positive probability of occurence. Then we need to maximize this probability over such a class of models. Johansen (1978) followed this line of attack to demonstrate that the Kaplan-Meier estimate (1958) of the survival function under random censoring is indeed a generalized maximum likelihood estimate.

Let (y_1,δ_1), (y_2,δ_2), $\cdots$, (y_n,δ_n) be n independent realizations of (Y,Δ). As has been pointed out earlier, the parameter space of the problem is $[0,1] \times \mathcal{F} \times \mathcal{F} \times \mathcal{F}$ whose typical element is characterized by (π, Dist. of $X \mid \bar{\Delta} = 1$, Dist. of $X \mid \bar{\Delta} = 0$, Dist. of C). Let P, Q, and R be any distributions in $\mathcal{F}$ of $X \mid \bar{\Delta} = 1$, $X \mid \bar{\Delta} = 0$, and C, respectively, under which the above data have a positive probability of occurence. The following summarization of data is helpful to write down the likelihood of the data.

<u>SUMMARY STATISTICS</u>

Let $u_1 < u_2 < \cdots < u_k$ be distinct values among $y_1, y_2, \cdots, y_n$. Let $u_0 = 0$ if $u_1 > 0$. Let

d_i = number died due to Cause 1 at time u_i,

s_i = number died due to Cause 2 at time u_i,

c_i = number who left the study at time u_i, and

$n_i = \sum_{j \geq i} (d_j + s_j + c_j)$

= number of subjects in the study active throughout the time interval $[0,u_i)$,

$i = 1,2, \cdots ,k$. Note that $n_1 = n$.

Let for each $i = 1,2,\cdots,k$,

$$p_i = P_P(X = u_i \mid \bar{\Delta} = 1); \quad p_i' = P_P(u_{i-1} < X < u_i \mid \bar{\Delta} = 1);$$

$$q_i = P_Q(X = u_i \mid \bar{\Delta} = 0); \quad q_i' = P_Q(u_{i-1} < X < u_i \mid \bar{\Delta} = 0); \text{ and}$$

$$r_i = P_R(C = u_i).$$

In the definition of p_1' and q_1', the left end point u_0 is included in the interval under the probabilities. If $u_1 = 0$, take $p_1' = 0 = q_1'$.

Let $$p_{k+1} = P_P(X > u_k \mid \bar{\Delta} = 1) \quad \text{and}$$

$$q_{k+1} = P_Q(X > u_k \mid \bar{\Delta} = 0).$$

Observe that $\sum_{i=1}^k (p_i + p_i') + p_{k+1} = 1 = \sum_{i=1}^k (q_i + q_i') + q_{k+1}$. The likelihood of the data under (π,P,Q,R), following (1.1), is given by

$L(\pi,P,Q,R)$

$$= \Pi_{i=1}^k [\pi\, p_i\, P_R(C \geq u_i)]^{i} [(1-\pi)\, q_i\, P_R(C \geq u_i)]^{i} [r_i\, P_{P,Q}(X > u_i)]^{i}$$

$$= \{\Pi_{i=1}^k r_i^{c_i} [P_R(C \geq u_i)]^{d_i+s_i}\} \times$$

$$\{\Pi_{i=1}^k [\pi p_i]^{d_i} [(1-\pi)q_i]^{s_i} [\pi P_P(X > u_i|\bar{\Delta} =1) + (1-\pi)P_Q(X > u_i|\bar{\Delta}=0)]^{c_i}.$$

We note that the likelihood factorises into a product of two partial likelihoods given by

$$L_1(R) = \Pi_{i=1}^k r_i^{c_i} [P_R(C \geq u_i)]^{d_i+s_i} \quad \text{and}$$

$L_2(\pi,P,Q) =$

$$\Pi_{i=1}^k [\pi p_i]^{d_i} [(1-\pi)q_i]^{s_i} [\pi(\textstyle\sum_{j\geq i+1}(p_j+p_j')+p_{k+1}) + (1-\pi)(\sum_{j\geq i+1}(q_j+q_j')+q_{k+1})]^{c_i}.$$

From now on, we will concentrate on deriving the generalized maximum likelihood estimates of π, and the conditional distributions of

$X \mid \bar{\Delta} = 1$ and $X \mid \bar{\Delta} = 0$. Thus we need to maximize $L_2(\pi,P,Q)$ with respect to $0 \le \pi \le 1$, p_i, $p_i' \ge 0$ and q_i, $q_i' \ge 0$ subject to the conditions

$$\sum_{i=1}^{k} (p_i + p_i') + p_{k+1} = 1 = \sum_{i=1}^{k} (q_i + q_i') + q_{k+1}.$$

we will simplify this problem of maximization as follows. Let P^* and Q^* be two distributions of $X \mid \bar{\Delta} = 1$ and $X \mid \bar{\Delta} = 0$, respectively, given by

$$P_{P^*}(X = u_i \mid \bar{\Delta} = 1) = p_i + p_i',$$

$$P_{P^*}(X > u_k \mid \bar{\Delta} = 1) = p_{k+1},$$

$$P_{Q^*}(X = u_i \mid \bar{\Delta} = 0) = q_i + q_i',$$

and $$P_{Q^*}(X > u_i \mid \bar{\Delta} = 0) = q_{k+1}.$$

It can be verified that $L_2(\pi,P^*,Q^*) \ge L_2(\pi,P,Q)$. Consequently, we can assume, without loss of generality, that each $p_i' = 0$ and each $q_i' = 0$. Thus we need to maximize

$L_2(\pi,P,Q) =$

$$\prod_{i=1}^{k} [\pi p_i]^{d_i} [(1-\pi)q_i]^{s_i} [\pi\sum_{j\ge i+1} p_j + (1-\pi)\sum_{j\ge i+1} q_j]^{c_i}$$

with respect to $0 \le \pi \le 1$, p_i's ≥ 0, q_i's ≥ 0 and subject to the conditions

$$\sum_{i=1}^{k+1} p_i = 1 = \sum_{i=1}^{k+1} q_i.$$

To solve the above maximization problem, we employ the following transformation . For $i = 1,2,\bullet\bullet\bullet,k+1$, let

$$a_i = \pi p_i/[\pi\sum_{j\ge i} p_j + (1-\pi)\sum_{j\ge i} q_j]$$

and $$b_i = (1-\pi)q_i/[\pi\sum_{j\ge i} p_j + (1-\pi)\sum_{j\ge i} q_j],$$

It can be verified that, for $i = 1,2,\cdots,k+1$,

$$\pi p_i = a_i \Pi_{j=1}^{i-1} (1 - a_j - b_j), \quad (1-\pi)q_i = b_i \Pi_{j=1}^{i-1} (1 - a_j - b_j)$$

and $$\pi \Sigma_{j\geq i}\, p_j + (1-\pi)\Sigma_{j\geq i}\, q_j = \Pi_{j=1}^{i-1} (1 - a_j - b_j) \tag{2.3}$$

with the convention that empty product = 1. Under the invertible transformation, the likelihood L_2 works out to be

$$L_2(\pi,P,Q) = \Pi_{i=1}^{k} a_i^{d_i} b_i^{s_i} (1 - a_i - b_i)^{n_i-d_i-s_i}.$$

We maximize L_2 over $0 \leq a_i \leq 1$, $0 \leq b_i \leq 1$ and $a_i + b_i \leq 1$ for each $i = 1,2,\cdots,k$. Thus L_2 is maximized at

$$a_i = d_i/n_i \quad \text{and} \quad b_i = s_i/n_i \tag{2.4}$$

for $i = 1,2,\cdots,k$. The desired estimates $\hat{\pi}$, $\hat{F}_1$ and $\hat{F}_2$ of π, the conditional distributions of $X \mid \bar{\Delta} = 1$ and $X \mid \bar{\Delta} = 0$, respectively, have to satisfy, from (2.3) and (2.4), the following likelihood equations. For $i = 1,2,\cdots,k$,

$$\begin{aligned} \hat{\pi}\, \hat{p}_i &= (d_i/n_i)\, \Pi_{j=1}^{i-1} \{1 - [(d_j + s_j)/n_j]\}, \\ (1-\hat{\pi})\, \hat{q}_i &= (s_i/n_i)\, \Pi_{j=1}^{i-1} \{1 - [(d_j + s_j)/n_j]\}, \\ \hat{p}_{k+1} &= 1 - \Sigma_{j=1}^{k} \hat{p}_j \quad \text{and} \quad \hat{q}_{k+1} = 1 - \Sigma_{j=1}^{k} \hat{q}_j. \end{aligned} \tag{2.5}$$

In the next section, we discuss the solvability of these likelihood equations.

3. ESTIMABILITY OF π, F_1 and F_2

The estimability of π, F_1 and F_2 hinges on the value of c_k. If $c_k = 0$, i.e., no one left the study at the last observed time u_k of activity, one can check that $\hat{p}_{k+1} = 0 = \hat{q}_{k+1}$. This follows from (2.3) that

$$\hat{\pi}\, \hat{p}_{k+1} + (1-\hat{\pi})\, \hat{q}_{k+1} = \Pi_{j=1}^{k} \{1 - [(d_j + s_j)/n_j]\} = 0$$

and $n_k = d_k + s_k$. In this case, the likelihood equations have a unique solution given by

$$\hat{\pi} = \sum_{i=1}^{k} \hat{\pi}\, \hat{p}_i = \sum_{i=1}^{k} (d_i/n_i) \prod_{j=1}^{i-1} \{1 - [(d_j + s_j)/n_j]\},$$

$$\hat{p}_i = (1/\hat{\pi})\, (d_i/n_i) \prod_{j=1}^{i-1} \{1 - [(d_j + s_j)/n_j]\}, \qquad (2.6)$$

$$\hat{q}_i = (1/\hat{\pi})\, (s_i/n_i) \prod_{j=1}^{i-1} \{1 - [(d_j + s_j)/n_j]\} \quad \text{for } i = 1,2,\cdots,k.$$

Consequently, $\hat{F}_1(x) = \sum_{\{1 \le i \le k;\ u_i \le x\}} \hat{p}_i$ and

$$\hat{F}_2(x) = \sum_{\{1 \le i \le k;\ u_i \le x\}} \hat{q}_i \quad \text{for } 0 \le x < \infty.$$

On the other hand, if $c_k > 0$, i.e., if there is atleast one in the sample who left the study at the last observed time u_k, the likelihood equations have more than one solution. Note that we can estimate $\pi\, p_{k+1} + (1-\pi)\, q_{k+1}$ but not the individual probabilities p_{k+1} and q_{k+1}.

However, we can estimate the following probability. Let $M > 0$ be fixed. Let $\pi_M = P$(a randomly selected person dies due to Cause 1 during $[0,M]$) $= P(X \le M, \bar{\Delta} = 1) = \pi\, F_1(M)$, which is estimable.

The asymptotic properties of $\hat{\pi}_M$ and other estimators of some other allied parameters are deferred to another research report by Babu, C.R. Rao and M.B. Rao (1989). Bootstrap is also discussed in this paper.

Before closing this section, we would like to emphasize that the generalized maximum likelihood estimates of F and G are the usual Kaplan-Meier estimates and the above calculations bear testimony to this effect.

4. A NUMERICAL EXAMPLE

Suppose a sample of size $n = 7$ yields the following summary statistics.

u_i :	1.7	2.3	4.4	4.5	4.9	6.0	6.1
d_i :	1	0	0	1	0	0	0
s_i :	0	1	0	0	0	1	1
c_i :	0	0	1	0	1	0	0
n_i :	7	6	5	4	3	2	1

Observe that only 2 out of 7 died of Cause 1. A crude estimate of π could be 2/7 which needs to be adjusted taking into account the fact that two subjects in the sample left the study. Since $c_k = 0$, we have a maximum likelihood estimate of π available. Indeed, from (2.6),

$$\hat{\pi} = 9/28 > 2/7.$$

REFERENCES

BABU, G.J., RADHAKRISHNA RAO, C.R. and BHASKARA RAO, M. (1989). Nonparametric estimation of Specific Occurence/Exposure Rate in Survival Analysis, a preprint.

JOHANSEN, S. (1978). The product limit estimator as a maximum likelihood estimator, Scand.J.Stat., 5, 195-199.

KALBFLEISCH, J.D. and PRENTICE, R.L. (1980). The statistical analysis of failure data, John Wiley, New York.

KAPLAN, E.L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations, J.Amer.Stat.Assoc., 53, 457-481.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, Ann.Math.Stat., 27, 887-906.

ESTIMATION OF TRANSITION DISTRIBUTION FUNCTION AND ITS QUANTILES IN MARKOV PROCESSES: STRONG CONSISTENCY AND ASYMPTOTIC NORMALITY

GEORGE G. ROUSSAS
Division of Statistics
University of California
Davis, California 95616 USA

ABSTRACT. Let $X_1,\ldots,X_{n+1}$ be the first n+1 random variables from a strictly stationary Markov process which satisfies certain additional regularity conditions. On the basis of these random variables, a nonparametric estimate of the one-step transition distribution function is shown to be uniformly (in the main argument) strongly consistent and asymptotically normal. Furthermore, for any $p\in(0,1)$, a natural nonparametric estimate of the p-th quantile of the distribution function just mentioned is proven to be strongly consistent. The class of Markov processes studied includes the Markov processes usually considered in the literature; namely, processes which either satisfy Doeblin's hypothesis, or, more generally, are geometrically ergodic.

1. Introduction

Let the real-valued random variables (r.v.'s) $X_1,X_2,\ldots$ be defined on a probability space $(\Omega,\mathcal{A},P)$ and suppose they constitute a strictly stationary Markov process satisfying some additional requirements. Suppose the distributions of X_1 and (X_1,X_2) have probability density functions (p.d.f.'s), $p(\cdot)$ and $q(\cdot,\cdot)$, respectively, with respect to the appropriate Lebesgue measures. Then the one-step transition p.d.f. is given by $t(y|x) = q(x,y)/p(x)$, where throughout the paper we consider x's for which $p(x)>0$. The quantities $p(x)$ and $q(x,y)$ may be estimated nonparametrically by the (kernel) estimates $p_n(x)$ and $q_n(x,y)$, respectively, given below.

$$p_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\Big(\frac{x-X_i}{h_n}\Big), \tag{1.1}$$

$$q_n(x,y) = \frac{1}{nh_n^2}\sum_{i=1}^{n} K\Big(\frac{x-X_i}{h_n}\Big)K\Big(\frac{y-X_{i+1}}{h_n}\Big), \tag{1.2}$$

where $0<h_n \to 0$, as $n\to\infty$. Then the natural (nonparametric) estimate for $t(y|x)$ is

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 443–462.
© 1991 *Kluwer Academic Publishers.*

$$t_n(y|x) = \frac{q_n(x,y)}{p_n(x)} . \tag{1.3}$$

The transition p.d.f. $t(\cdot|\cdot)$ and its estimate $t_n(\cdot|\cdot)$ are of paramount interest in a Markov processes context, and so is the (one-step) transition distribution function (d.f.) $G(\cdot|\cdot)$. A reasonable (nonparametric) estimate of $G(y|x)$ is given by

$$G_n(y|x) = \int_{-\infty}^{y} t_n(x'|x)dx', \qquad y\in \mathbb{R}. \tag{1.4}$$

The estimates defined by (1.1)-(1.4) were actually introduced in Roussas (1969a, 1969b) and various asymptotic properties were also established. Yakowitz (1979) proposed an alternative estimate of $G(y|x)$. In connection with the estimate in (1.4), it was shown, in particular, that

$$\sup\{|G_n(y|x)-G(y|x)|;\ y\in \mathbb{R}\} \to 0 \tag{1.5}$$

in probability, and that, for $0<p<1$,

$$\xi_n(p,x) \to \xi(p,x) \tag{1.6}$$

in probability, as $n\to\infty$, where $\xi(p,x)$ is the (unique, by assumption) p-th quantile of $G(\cdot|x)$ and $\xi_n(p,x)$ is a suitable (nonparametric) estimate of $\xi(p,x)$.

In the present paper, it is proposed to strengthen the convergences in (1.5) and (1.6) to almost sure convergence, and also establish asymptotic normality for a suitably normalized version of $G_n(y|x)$. Asymptotic normality of $\xi_n(p,x)$ has already been established in Roussas (1969b). Results similar to those in Theorems 2.1 and 2.2 have also been obtained recently by Samanta (1989) in the framework of independent identically distributed pairs of r.v.'s. Recursive nonparametric estimates of conditional densities and expectations of stationary processes have been obtained by Masry (1989), under both ρ-mixing and α-mixing. The discussion does not included uniform strong consistency of the conditional d.f., and the topics of quantiles and asymptotic normality are not being dealt with.

An assumption usually associated with a (strictly) stationary Markov process is that of the existence of a single ergodic set which has no cyclically moving subsets. This basic assumption will be incorporated in the conditions to be stated below. In addition, it is often assumed that the process satisfies Doeblin's hypothesis (D) (see, for example, Doob (1953), page 192). Actually, hypothesis (D), along with the existence of a single ergodic set without cyclically moving subsets, is usually denoted by (D_o) and is an often made assumption (see page 221 in the reference just cited). The results obtained in Roussas (1969a, 1969b) were established under condition (D_o). Hypothesis (D) is a rather strong one and not easily verifiable. This perhaps was the motivation for replacing it by a weaker assumption, the so-called geometric ergodicity. For its definition and basic properties, the reader is referred to Rosenblatt (1970).

It is known that, if a stationary Markov process satisfies condition (D_0), then it is ϕ-mixing with mixing coefficient $\gamma\rho^n$ for some $\gamma>0$ and some $\rho\in(0,1)$. If the process is geometrically ergodic; that is, if it satisfies condition G_p for some $p\in[1,\infty]$, then the process is ρ-mixing with mixing coefficient not larger than $M\rho^n$ for some $M>0$ and some $\rho\in(0,1)$. For the definition of ϕ-mixing, ρ-mixing, and a discussion of the above stated results, the interested reader is referred, for example, to Roussas and Ioannides (1987), Definitions 1.2, 1.3, and Examples 3.3, 3.4. The results just mentioned and the fact that ϕ-mixing implies ρ-mixing motivates the assumption made below that the underlying Markov process is ρ-mixing. Actually, the results in the paper will be derived under the assumption that the maximal correlation $\rho(n)$ is only of the order of $O(n^{-\nu})$ $(\nu>1)$ rather than being exponential. This weaker assumption imposes some restrictions on the choice of the bandwidth b_n (see relations (6.14)), which may be relaxed, if $\rho(n)$ is allowed to be of exponential order.
The main results of the paper, Theorems 2.1-2.3, are stated in Section 2. The proofs of Theorems 2.1 and 2.2 are given in Section 4, and that of Theorem 2.3 in Section 6. Throughout the paper, limits are taken as $n\to\infty$, unless otherwise stated. Integrals will be taken over the entire appropriate spaces, unless otherwise indicated. Finally, the letter C will be used to represent constants which are unimportant and may vary from place to place.

2. Assumptions and Statement of Main Results

The assumptions under which the results in this paper are derived are gathered together below for easy reference. It will be usually indicated which of these assumptions are used in the various steps of the derivations.

Assumptions

(A1) (i) The real-valued r.v.'s X_n, $n\geq1$, are defined on a probability space $(\Omega,\mathcal{A},P)$ and constitute a strictly stationary Markov process with a single ergodic set and no cyclically moving subsets.

(ii) The distributions of X_1 and (X_1,X_2) have continuous p.d.f.'s $p(\cdot)$ and $q(\cdot,\cdot)$, respectively, with respect to appropriate Lebesgue measures.

(iii) The process $\{X_n\}$ is ρ-mixing with maximal correlation $\rho(n) = O(n^{-\nu})$, $\nu>1$.

(A2) K is a p.d.f. defined on $\mathbb{R}$ such that:

(i) K is bounded.

(ii) $|x|K(x) \to 0$ as $|x| \to \infty$.

(iii) $\int xK(x)dx = 0$ and $\int x^2K(x)dx<\infty$.

(A3) $\{h_n\}$ is a sequence of real numbers such that:

(i) $0<h_n \to 0$.

(ii) For an arbitrarily large $\ell\geq 2$, $\sum_{n=1}^{\infty} \left(n^{1/2}h_n\right)^{-\ell} < \infty$.

(iii) $nh_n^5 \to 0$.

(A4) For $p\in(0,1)$ and $x\in\mathbb{R}$, the one-step transition d.f. of the process, $G(\cdot|x)$, has a unique p-th quantile $\xi(p,x)$.

(A5)(i) $p(\cdot)$ has a continuous and bounded second order derivative.

(ii) The p.d.f. $q(\cdot,\cdot)$ has continuous second order partial derivatives, denoted by $q''_{ij}(\cdot,\cdot)$, $i,j=1,2$, such that

$$\int |q''_{ij}(x,y)|dy \leq C \text{ (independent of } x\in\mathbb{R}).$$

(iii) $p(x) \leq C$, $x\in\mathbb{R}$, and $|q_{1,i}(x,y) - p(x)p(y)|\leq C$, $i\geq 2$, $x,y \in \mathbb{R}$.

Comments on the assumptions

1. The Markovian assumption of the underlying process is explicitly used only once, in the proof of Lemma 5.5. That proof can certainly be carried out under ρ-mixing alone by properly strengthening other assumptions. We chose, however, to retain the Markovian assumption because it is under such an assumption that conditional quantities are fully meaningful and of paramount importance.
2. The assumption of continuity of the p.d.f.'s $p(\cdot)$ and $q(\cdot,\cdot)$ is not critical. Results may be stated for suitable sets of continuity points, but making statements to this effect each time, becomes a cumbersome undertaking.
3. The assumption of ρ-mixing may easily be relaxed to α-mixing at the expense of strengthening some other assumptions. We chose, however, to work in such a framework, for, as already explained, usual assumptions made on Markov processes lead to ρ-mixing.
4. Assumption (A3)(ii) may be replaced by an exponential bound, but the present bound is enough for the purposes of this paper.
5. The second part of assumption (A5)(iii) can be replaced by (uniform in $i\geq 2$) boundedness of the joint p.d.f.'s $q_{1,i}(x,y)$. The assumption made, however, seems more natural.
In all that follows, when appropriate, it will be assumed that the arguments are carried out for $x\in\mathbb{R}$ with $p(x)>0$ without explicitly mentioning it. With this in mind, we have then.

Theorem 2.1. Let assumptions (A1); (A2)(i),(ii); and (A3)(i),(ii) be satisfied, and let $G_n(y|x)$ be defined by (1.4). Then

$$\sup\{|G_n(y|x)-G(y|x)|;\ y\in\mathbb{R}\} \to 0 \quad \text{a.s.},\ x\in\mathbb{R}.$$

Also,

Theorem 2.2. Let assumptions (A1); (A2)(i),(ii); (A3)(i),(ii); and (A4) be satisfied, let $G_n(y|x)$ be defined by (1.4), and (in order to

fix ideas) let $\xi_n(p,x)$ be the smallest p-th quantile of $G_n(y|x)$, $0<p<1$. Then

$$\xi_n(p,x) \to \xi(p,x), \quad \text{a.s.}, \ x\in \mathbb{R}.$$

Regarding the asymptotic normality of $G_n(y|x)$, we have

<u>Theorem 2.3</u>. Let assumptions (A1)-(A5) be satisfied. Then, for $x,y\in \mathbb{R}$,

$$(nh_n)^{1/2}[G_n(y|x)-G(y|x)] \xrightarrow{d} N(0,\sigma^2(x,y)), \quad x,y \in \mathbb{R},$$

where

$$\sigma^2(x,y) = \frac{B(x,y)[p(x)-B(x,y)]}{p^3(x)} \int K^2(u)du \tag{2.1}$$

$$= \frac{\begin{vmatrix} p(x) & B(x,y) \\ B(x,y) & B(x,y) \end{vmatrix}}{p^3(x)} \int K^2(u)du \ (>0).$$

Results discussed in this paper may be modified and/or generalized in various ways. To mention two or three of them, one may consider the recursive versions, or estimate conditional quantities of k steps in the future, or to examine rates of convergence when appropriate. Details will be presented elsewhere.

3. Some Preliminary Results

Define $B(x,y)$ and $B_n(x,y)$ by

$$B(x,y) = \int_{-\infty}^{y} q(x,x')dx', \ B_n(x,y) = \int_{-\infty}^{y} q_n(x,x')dx', \ x,y \in \mathbb{R}. \tag{3.1}$$

Then

$$\sup_{y\in \mathbb{R}} |G_n(y|x)-G(y|x)| \le \frac{1}{p_n(x)} \sup_{y\in \mathbb{R}} |B_n(x,y)-B(x,y)|$$

$$+ \frac{1}{p_n(x)} |p_n(x)-p(x)|. \tag{3.2}$$

Since, under the assumptions of Theorem 2.1, $p_n(x) \xrightarrow{a.s.} p(x)$, in order to establish the theorem, it suffices to show that

$$\sup_{y\in \mathbb{R}} |B_n(x,y)-B(x,y)| \xrightarrow{a.s.} 0. \tag{3.3}$$

Actually, as will be argued below, it suffices to show that

$$B_n(x,y) \xrightarrow{a.s.} B(x,y) \tag{3.4}$$

from which (3.3) will be concluded. To this end, it suffices to show that

$$\mathcal{E}B_n(x,y) \to B(x,y), \text{ and } B_n(x,y)-\mathcal{E}B_n(x,y) \xrightarrow{a.s.} 0. \tag{3.5}$$

For simplicity, write h rather than h_n and use (1.2) and (3.1) to obtain

$$B_n(x,y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\bar{K}\left(\frac{y-X_{i+1}}{h}\right), \text{ where } \bar{K}(z) = \int_{-\infty}^{z} K(u)du. \tag{3.6}$$

Hence,

$$\mathcal{E}B_n(x,y) = \int \bar{K}\left(\frac{y-s}{h}\right)a_n(x,s)ds, \text{ where } a_n(x,s) = \int K(u)q(x-hu,s)du. \tag{3.7}$$

Clearly,

$$(0\le)\bar{K}\left(\frac{y-s}{h}\right)a_n(x,s) \le a_n(x,s) \tag{3.8}$$

with

$$a_n(x,s) \to q(x,s) \text{ and } \int a_n(x,s)ds = \int K(u)p(x-hu)ds \to p(x). \tag{3.9}$$

Furthermore,

$$\bar{K}\left(\frac{y-s}{h}\right)a_n(x,s)I_{(-\infty,y)}(s) \to q(x,s), \ \bar{K}\left(\frac{y-s}{h}\right)a_n(x,s)I_{(y,\infty)}(s) \to 0. \tag{3.10}$$

From relations (3.8)-(3.10), it follows, by means of the generalized Dominated Convergence Theorem (see Loève (1963), page 162) that

$$\mathcal{E}B_n(x,y) = \int \bar{K}\left(\frac{y-s}{h}\right)a_n(x,s)ds \to \int_{-\infty}^{y} q(x,s)ds = B(x,y).$$

That is, the following lemma has been established.

<u>Lemma 3.1</u>. Under assumptions (A1)(i),(ii); (A2)(i),(ii); and (A3)(i),

$$\mathcal{E}B_n(x,y) \to B(x,y), \quad x,y \in \mathbb{R}.$$

The convergence in (3.4) is justified in the following

<u>Lemma 3.2</u>. Under assumptions (A1); (A2)(i); and (A3)(i),(ii),

$$B_n(x,y) - \mathcal{E}B_n(x,y) \xrightarrow{a.s.} 0, \quad x,y \in \mathbb{R}.$$

<u>Proof</u>. For i=1,...,n, set

$$Y_{ni} = Y_{ni}(x,y) = \frac{1}{h}\left\{\left[K\left(\frac{x-X_i}{h}\right)\bar{K}\left(\frac{y-X_{i+1}}{h}\right)\right] - \mathcal{E}\left[K\left(\frac{x-X_i}{h}\right)\bar{K}\left(\frac{y-X_{i+1}}{h}\right)\right]\right\},$$

so that $\mathcal{E}Y_{ni} = 0$ and $|Y_{ni}| \le Ch^{-1}$ for all i, n, and by (3.6),

$$B_n(x,y) - \mathcal{E}B_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} Y_{ni}.$$

Under the assumption of the lemma, Theorem 4.1 in Roussas (1988) applies and gives

$$P\left[\left|B_n(x,y)-B(x,y)\right|>\varepsilon\right] = P\left(\left|\sum_{i=1}^{n} Y_{ni}\right| > \varepsilon n\right) \le \left(C\varepsilon^{-\ell}\right)/\left(n^{1/2}h\right)^{\ell}$$

for any $\ell\ge 2$. The Borel-Cantelli lemma completes the proof. ∎

Combining the previous two lemmas, we have (3.4); that is,

<u>Corollary 3.1</u>. Under assumptions (A1); (A2)(i),(ii); and (A3)(i),(ii),

$$B_n(x,y) \xrightarrow{a.s.} B(x,y), \quad x,y \in \mathbb{R}.$$

For later reference, let us also record the following

<u>Lemma 3.1</u>. Under assumptions (A1); (A2)(i),(ii); and (A3)(i),(ii),

$$p_n(x) \xrightarrow{a.s.} p(x), \quad x \in \mathbb{R}.$$

<u>Proof</u>. First, by (A1)(i),(ii),

$$\mathcal{E}p_n(x) = \int K(u)p(x-hu)du \to p(x),$$

and then $p_n(x) - \mathcal{E}p_n(x) \xrightarrow{a.s.} 0$, since, by (A1)(iii) and (A3)(ii),

$$\sum_{n=1}^{\infty} P\left[\left|p_n(x)-\mathcal{E}p_n(x)\right|>\varepsilon\right] = \sum_{n=1}^{\infty} P\left(\left|\sum_{i=1}^{n} Z_{ni}\right|>\varepsilon n\right) \le \left(C\varepsilon^{-\ell}\right)\sum_{n=1}^{\infty}\left(n^{1/2}h\right)^{-\ell} < \infty,$$

where $Z_{ni} = Z_{ni}(x) = \frac{1}{h}\left[K\left(\frac{x-X_i}{h}\right)-\mathcal{E}K\left(\frac{x-X_i}{h}\right)\right]$. ∎

4. Proof of Theorems 2.1 and 2.2

<u>Proof of Theorem 2.1</u>. By suppressing x, since it remains fixed thoughout the arguments, define $H(y)$ and $H_n(y)$ by

$$H(y) = \frac{B(x,y)}{p(x)}, \qquad H_n(y) = \frac{B_n(x,y)}{p(x)}. \tag{4.1}$$

Then $H(\cdot)$ is a continuous (proper) d.f. (whereas $H_n(\cdot)$ need not be), and by Corollary 3.1,

$$H_n(x,y) \xrightarrow{a.s.} H(y), \quad y \in \mathbb{R}. \tag{4.2}$$

The proof that (4.2) implies

$$\sup_{y \in \mathbb{R}} \left| H_n(x,y) - H(y) \right| \xrightarrow{a.s.} 0 \tag{4.3}$$

is reminiscent of the familiar proof of the a.s. uniform convergence of the empirical d.f. to the population d.f. For the sake of brevity, the detailed arguments are omitted, but they can be found in Roussas (1990b). ∎

We now proceed with the

Proof of Theorem 2.2. Write ξ and ξ_n instead of $\xi(p,x)$ and $\xi_n(p,x)$, respectively, and, for $k=1,2,\ldots$, set $\varepsilon_k = 1/k$. Define $\delta_k = \delta(\varepsilon_k)$ by

$$\delta_k = \min\{G(\xi+\varepsilon_k|x)-G(\xi|x),\ G(\xi|x)-G(\xi-\varepsilon_k|x)\}.$$

Then δ_k is positive because of assumption (A4), and, clearly, $\delta_k \downarrow 0$ as $\varepsilon_k \downarrow 0$. Furthermore,

$$|G(\xi_n|x)-G(\xi|x)| \le \sup_{y \in \mathbb{R}} |G_n(y|x)-G(y|x)| \overset{df.}{=} g_n. \tag{4.4}$$

From the definition of δ_k, it is also clear that,

$$\bigcap_{k\ge1} \bigcup_{n\ge1} \bigcap_{\nu\ge0} (g_{n+\nu} \le \delta_k) \subseteq \bigcap_{k\ge1} \bigcup_{n\ge1} \bigcap_{\nu\ge0} (|\xi_{n+\nu}-\xi| \le \varepsilon_k). \tag{4.5}$$

Since $\varepsilon_k \downarrow 0$ implies $\delta_k \downarrow 0$, the left-hand side of (4.5) is the set of convergence to 0 of g_n. By Theorem 2.1, the probability of this set is equal to 1. Then so is the probability of the set on the right-hand side of (4.5), which is the set of convergence of ξ_n to ξ. ∎

5. Some Further Auxiliary Results

In this section, some additional auxiliary results are derived, necessary to establish the asymptotic normality stated in Theorem 2.3. By (3.1),

$$(nh)^{1/2}[G_n(y|x)-G(y|x)] = (nh)^{1/2}\left[\frac{B_n(x,y)}{p_n(x)} - \frac{B(x,y)}{p(x)}\right]. \tag{5.1}$$

Define the function $H: \mathbb{R} \to \mathbb{R}$ by $H(y_1,y_2) = \dfrac{y_2}{y_1}$, $y_1 \ne 0$.

Then (5.1) is rewritten as follows

$$(nh)^{1/2}[G_n(y|x)-G(y|x)] = (nh)^{1/2}[H(T_n)-H(\theta)], \tag{5.2}$$

where

$$T_n = T_n(x,y) = (p_n(x),\ B_n(x,y)),\ \theta = \theta(x,y) = (p(x),\ B(x,y)). \tag{5.3}$$

Hence, by the standard Cramér-Wold device (see, for example, Theorem (ii), page 321, in Rao (1965)), all we have to show is that, for all $c_1, c_2 \in \mathbb{R}$ with $c_1^2 + c_2^2 \neq 0$,

$$\xi_n(x,y) \overset{\text{df.}}{=} (nh)^{1/2}\{c_1[p_n(x)-p(x)] + c_2[B_n(x,y)-B(x,y)]\} \xrightarrow{d} N\left(0,\ \tau_c^2\right), \tag{5.4}$$

where

$$\tau_c^2 = \tau_c^2(x,y) = (c_1,\ c_2)C(x,y)(c_1,\ c_2)^t = [c_1^2 p(x)+2c_1c_2B(x,y)+c_2^2B(x,y)]\int K^2(u)du \tag{5.5}$$

and "t" denotes transpose.

Let

$$\psi_n(x,y) \overset{\text{df.}}{=} (nh)^{1/2}\{c_1[p_n(x)-\mathcal{E}p_n(x)]+c_2[B_n(x,y)-\mathcal{E}B_n(x,y)]\}. \tag{5.6}$$

Then (5.4) is equivalent to relations (5.7) and (5.8) below; namely,

$$\xi_n(x,y)-\psi_n(x,y) = (nh)^{1/2}\{c_1[\mathcal{E}p_n(x)-p(x)] + c_2[\mathcal{E}B_n(x,y)-B(x,y)]\} \to 0, \tag{5.7}$$

and

$$\psi_n(x,y) \xrightarrow{d} N\left(0,\ \tau_c^2\right). \tag{5.8}$$

The convergence in (5.7) is established in the following

<u>Proposition 5.1</u>. Let assumptions (A1)(i),(ii); (A2)(i),(ii); (A3)(i),(iii); and (A5)(i),(ii) be satisfied. Then, for $x,y \in \mathbb{R}$,

(i) $(nh)^{1/2}[\mathcal{E}p_n(x)-p(x)] \to 0$, and (ii) $(nh)^{1/2}[\mathcal{E}B_n(x,y)-B(x,y)] \to 0$,

so that the convergence in (5.7) holds.

<u>Proof</u>. A brief justification is as follows:

(i) $$(nh)^{1/2}[\mathcal{E}p_n(x)-p(x)] = (nh)^{1/2}\int K(u)[p(x-hu)-p(x)]du = (nh^5)^{1/2}\iint_0^1(1-\lambda)u^2K(u)p''(x-\lambda hu)d\lambda du,$$

and

$$\iint_0^1(1-\lambda)u^2K(u)p''(x-\lambda hu)d\lambda du \to \frac{1}{2}\ p''(x)\int u^2K(u)du$$

by a repeated application of the Dominated Convergence Theorem. Since $nh^5 \to 0$, the proof is complete.

(ii) The proof of this part is more complicated. From (3.1), it is easily seen that,

$$\mathcal{E}B_n(x,y) = \int_{-\infty}^{y} \iint K(u)K(v)q(x-hu,x'-hv)\,du\,dv\,dx'.$$

Then, denoting by $q_i'(\cdot,\cdot)$ and $q_{ij}''(\cdot,\cdot)$ the first and second order derivatives of $q(\cdot,\cdot)$, $i,j = 1,2$, we have

$$(nh)^{1/2}[\mathcal{E}B_n(x,y)-B(x,y)] = (nh^5)^{1/2}[I_1(x,y) + I_2(x,y) + I_3(x,y)],$$

where

$$I_1(x,y) = \int_{-\infty}^{y} \iiint_0^1 (1-\lambda)u^2K(u)K(v)q_{11}''(x-\lambda hu,x'-\lambda hv)\,d\lambda\,du\,dv\,dx',$$

$$I_2(x,y) = 2\int_{-\infty}^{y} \iiint_0^1 (1-\lambda)uvK(u)K(v)q_{12}''(x-\lambda hu,x'-\lambda hv)\,d\lambda\,du\,dv\,dx',$$

$$I_3(x,y) = \int_{-\infty}^{y} \iiint_0^1 (1-\lambda)v^2K(u)K(v)q_{22}''(x-\lambda hu,x'-\lambda hv)\,d\lambda\,du\,dv\,dx'.$$

But

$$|I_1(x,y)| = \left|\int_{-\infty}^{y-\lambda hv} \iiint_0^1 (1-\lambda)u^2K(u)K(v)q_{11}''(x-\lambda hu,t)\,d\lambda\,du\,dv\,dt\right|$$

$$\le \int_0^1 (1-\lambda)\int u^2K(u)\int |q_{11}''(x-\lambda hu,t)|\,dt\,du\,d\lambda$$

$$\le C\cdot\frac{1}{2}\int u^2K(u)\,du \text{ finite (by assumptions (A2)(iii) and (A5)(ii)),}$$

and similarly for $I_2(x,y)$ and $I_3(x,y)$. Then assumption (A3)(iii) completes the proof.∎

As a first step in proving (5.8), we have to obtain the asymptotic variance of $\psi_n(x,y)$. More precisely, we will show

<u>Proposition 5.2</u>. Let assumptions (A1); (A2)(i),(ii);(A3)(i); and (A5)(iii) be satisfied. Then, for $x,y \in \mathbb{R}$,

$$\operatorname{var}[\psi_n(x,y)] \to \tau_c^2(x,y), \text{ where } \tau_c^2(x,y) \text{ is given by (5.5).}$$

The proof of this proposition is the culmination of a series of lemmas to be stated and established below. To start with, use relations (5.6), (3.6) and (1.1) to obtain,

$$\operatorname{var}[\psi_n(x,y)] = \frac{c_1^2}{h}\operatorname{var}\left[K\left(\frac{x-X_1}{h}\right)\right] + \frac{c_2^2}{h}\operatorname{var}\left[K\left(\frac{x-X_1}{h}\right)\bar{K}\left(\frac{y-X_2}{h}\right)\right] +$$

$$\frac{2c_1c_2}{h}\operatorname{cov}\left[K\left(\frac{x-X_1}{h}\right),K\left(\frac{x-X_1}{h}\right)\bar{K}\left(\frac{y-X_2}{h}\right)\right] +$$

$$\frac{2c_1^2}{nh}\sum_{1\le i<j\le n}\operatorname{cov}\Big[K\Big(\frac{x-X_i}{h}\Big),\ K\Big(\frac{x-X_j}{h}\Big)\Big]\ +$$

$$\frac{2c_1c_2}{nh}\sum_{1\le i<j\le n}\operatorname{cov}\Big[K\Big(\frac{x-X_i}{h}\Big),\ K\Big(\frac{x-X_j}{h}\Big)\bar{K}\Big(\frac{y-X_{j+1}}{h}\Big)\Big]\ +$$

$$\frac{2c_1c_2}{nh}\sum_{1\le i<j\le n}\operatorname{cov}\Big[K\Big(\frac{x-X_i}{h}\Big)\bar{K}\Big(\frac{y-X_{i+1}}{h}\Big),K\Big(\frac{x-X_j}{h}\Big)\Big]\ +$$

$$\frac{2c_2^2}{nh}\sum_{1\le i<j\le n}\operatorname{cov}\Big[K\Big(\frac{x-X_i}{h}\Big)\bar{K}\Big(\frac{y-X_{i+1}}{h}\Big),K\Big(\frac{x-X_j}{h}\Big)\bar{K}\Big(\frac{y-X_{j+1}}{h}\Big)\Big]. \quad (5.9)$$

The lemmas below take care of the asymptotic behavior of the expressions on the right-hand side of relation (5.9).

<u>Lemma 5.1</u>. Under the assumptions of Proposition 5.2 and for $x\in\mathbb{R}$,

$$\frac{1}{h}\operatorname{var}\Big[K\Big(\frac{x-X_1}{h}\Big)\Big] \to p(x)\int K^2(u)\,du.$$

<u>Proof</u>. Indeed,

$$\frac{1}{h}\operatorname{var}\Big[K\Big(\frac{x-X_1}{h}\Big)\Big] = \int K^2(u)p(x-hu)\,du-h\Big[\int K(u)p(x-hu)\,du\Big]^2$$

$$\to p(x)\int K^2(u)\,du. \quad \blacksquare$$

<u>Lemma 5.2</u>. Under the assumptions of Proposition 5.2 and for $x,y\in\mathbb{R}$,

$$\frac{1}{h}\operatorname{var}\Big[K\Big(\frac{x-X_1}{h}\Big)\bar{K}\Big(\frac{y-X_2}{h}\Big)\Big] \to B(x,y)\int K^2(u)\,du.$$

<u>Proof</u>. First,

$$\frac{1}{h}\Big\{\mathcal{E}\Big[K\Big(\frac{x-X_1}{h}\Big)\bar{K}\Big(\frac{y-X_2}{h}\Big)\Big]\Big\}^2 \le \frac{1}{h}\Big[\mathcal{E}K\Big(\frac{x-X_1}{h}\Big)\Big]^2$$

$$\le h\Big[\int K(u)p(x-hu)\,du\Big]^2 \to 0. \quad (5.10)$$

Next,

$$\frac{1}{h}\,\mathcal{E}\Big[K\Big(\frac{x-X_1}{h}\Big)\bar{K}\Big(\frac{y-X_2}{h}\Big)\Big]^2 = \int K^2(u)\Big[\int \bar{K}^2\Big(\frac{y-t}{h}\Big)q(x-hu,t)\,dt\Big]du, \quad (5.11)$$

and, by a suitable application of the generalized Dominated Convergence Theorem (see corresponding lemma in Roussas (1990b) for details),

$$\int \bar{K}^2\Big(\frac{y-t}{h}\Big)q(x-hu,t)\,dt \to B(x,y).$$

Finally,

$$(0\le)K^2(u)\int \bar{K}^2\left(\frac{y-t}{h}\right)q(x-hu,t) \le K^2(u)q(x-hu,t)dt = K^2(u)p(x-hu)$$

with

$$K^2(u)p(x-hu) \to K^2(u)p(x) \text{ and } \int K^2(u)p(x-hu)du \to p(x)\int K^2(u)du.$$

Therefore,

$$\int K^2(u)\left[\int \bar{K}^2\left(\frac{y-t}{h}\right)q(x-hu,t)dt\right]du \to B(x,y)\int K^2(u)du. \tag{5.12}$$

Relations (5.10)-(5.12) complete the proof. ∎

Also,

<u>Lemma 5.3</u>. Under the assumptions of Proposition 5.2 and for $x,y\in\mathbb{R}$,

$$\frac{1}{h}\operatorname{cov}\left[K\left(\frac{x-X_1}{h}\right),K\left(\frac{x-X_1}{h}\right)\bar{K}\left(\frac{y-X_2}{h}\right)\right] \to B(x,y)\int K^2(u)du.$$

<u>Proof</u>. In the first place,

$$\frac{1}{h}\left[\mathcal{E}K\left(\frac{x-X_1}{h}\right)\right]\mathcal{E}\left[K\left(\frac{x-X_1}{h}\right)\bar{K}\left(\frac{y-X_2}{h}\right)\right] \le \frac{1}{h}\left[\mathcal{E}K\left(\frac{x-X_1}{h}\right)\right]^2 = h\left[\int K(u)p(x-hu)du\right]^2 \to 0.$$

Next,

$$\frac{1}{h}\,\mathcal{E}\left[K^2\left(\frac{x-X_1}{h}\right)\bar{K}\left(\frac{y-X_2}{h}\right)\right] = \int K^2(u)\left[\int \bar{K}\left(\frac{y-t}{h}\right)q(x-hu,t)dt\right]du \to B(x,y)\int K^2(u)du,$$

arguing as in (5.11)-(5.12) with $\bar{K}^2\left(\frac{y-t}{h}\right)$ there replaced by $\bar{K}\left(\frac{y-t}{h}\right)$. The proof then follows. ∎

In dealing with the remaining terms on the right-hand side of (5.9), the following familiar technique (see, Masry (1986)) will be needed. Namely, define S_1, S_2 by:

$$S_1 = \{(i,j)\,|\,i,j\in\{1,\dots,n\},\ 1\le j-i\le c_n\},$$

$$S_2 = \{(i,j)\,|\,i,j\in\{1,\dots,n\},\ c_n+1\le j-i\le n-1\},$$

where $\{c_n\}$ is any sequence of positive integers such that $c_n\to\infty$ and $c_nh_n\to 0$. To fix ideas, let us take

$$c_n = [1/h_n^{1-\lambda}],\ 0<\lambda<1, \tag{5.13}$$

where $[x]$ indicates the integral part of x.

<u>Lemma 5.4</u>. Under the assumptions of Proposition 5.2 and for $x,y\in\mathbb{R}$,

$$\frac{1}{nh}\sum_{1\le i<j\le n}\operatorname{cov}\left[K\left(\frac{x-X_i}{h}\right),\ K\left(\frac{x-X_j}{h}\right)\right]\to 0.$$

Proof. Indeed,

$$\sum_{1\le i<j\le n}\operatorname{cov}\left[K\left(\frac{x-X_i}{h}\right),\ K\left(\frac{x-X_j}{h}\right)\right]=\Sigma_{S_1}\operatorname{cov}\left[K\left(\frac{x-X_i}{h}\right),\ K\left(\frac{x-X_j}{h}\right)\right]+$$

$$\Sigma_{S_2}\operatorname{cov}\left[K\left(\frac{x-X_i}{h}\right),\ K\left(\frac{x-X_j}{h}\right)\right]\overset{df.}{=}J_{1n}+J_{2n}. \tag{5.14}$$

Recall that $q_{i,j}(\cdot,\cdot)$ denotes the joint p.d.f. of X_i,X_j. Then

$$|J_{1n}|\le h^2\Sigma_{S_1}\iint K(u)K(v)\,|q_{i,j}(x-hu,x-hv)-p(x-hu)p(x-hv)|\,du\,dv$$

$$\le Ch^2\Sigma_{S_1}1\le Cnc_nh^2. \tag{5.15}$$

Therefore, by (5.13),

$$\frac{1}{nh}\,|J_{1n}|\le Cc_nh\to 0. \tag{5.16}$$

Next, using known inequalilties for the covariance under ρ-mixing (see, for example, relation (1.8) in Roussas and Ioannides (1987)), we obtain

$$\frac{1}{nh}\,|J_{2n}|\le C\int K^2(u)p(x-hu)\,du\cdot\sum_{\ell=c_n}^{\infty}\ell^{-\nu}\to 0. \tag{5.17}$$

Relations (5.14), (5.16), and (5.17) complete the proof. ∎

Also,

Lemma 5.5. Under the assumptions of Proposition 5.2 and for $x,y\in\mathbb{R}$,

$$\frac{1}{nh}\sum_{1\le i<j\le n}\operatorname{cov}\left[K\left(\frac{x-X_i}{h}\right),K\left(\frac{x-X_j}{h}\right)\bar{K}\left(\frac{y-X_{j+1}}{h}\right)\right]\to 0. \tag{5.18}$$

Proof. Splitting the summation on the left-hand side of (5.18) in two parts over S_1 and S_2, denoting the corresponding quantities by $\bar{J}_{1n}$ and $\bar{J}_{2n}$, and using the Markovian property, we have, as in the previous lemma,

$$\frac{1}{nh}\,|\bar{J}_{1n}|\le Ch\Sigma_{S_1}\int K(y_i)K(y_j)\left[\int t(x_{i+1}|x-hy_j)\,dx_{j+1}\right]dy_j\,dy_i$$

$$=Ch\Sigma_{S_1}1\to 0\quad\text{as in (5.15).} \tag{5.19}$$

The proof that $(nh)^{-1}|\bar{J}_{2n}|\to 0$ is identical to that in (5.17) since $\bar{K}(\cdot)$ is bounded (by 1). The proof of the lemma is concluded.∎

Finally,

Lemma 5.6. Under the assumptions of Propositions 5.2 and for $x,y\in\mathbb{R}$,

$$\frac{1}{nh}\sum_{1\le i<j\le n}\operatorname{cov}\Big[K\Big(\frac{x-X_i}{h}\Big)\bar{K}\Big(\frac{y-X_{i+1}}{h}\Big),\ K\Big(\frac{x-X_j}{h}\Big)\bar{K}\Big(\frac{y-X_{j+1}}{h}\Big)\Big]\to 0. \tag{5.20}$$

Proof. Split the summation on the left-hand side of (5.20) in two parts over S_1 and S_2, denote the corresponding quantities by J^*_{1n} and J^*_{2n}, and observe that

$$\mathcal{E}\Big[K\Big(\frac{x-X_i}{h}\Big)\bar{K}\Big(\frac{y-X_{i+1}}{h}\Big)K\Big(\frac{x-X_j}{h}\Big)\bar{K}\Big(\frac{y-X_{j+1}}{h}\Big)\Big]$$

$$\le h^2\iiiint K(y_i)K(y_j)q_{i,i+1,j,j+1}(x-hy_i,x_{i+1},x-hy_j,x_{j+1})dy_i dx_{i+1}dy_j dx_{j+1}$$

$$= h^2\iint K(y_i)K(y_j)q_{i,j}(x-hy_i,x-hy_j)dy_i dy_j$$

$$\le Ch^2\iint K(y_i)K(y_j)dy_i dy_j = Ch^2. \tag{5.21}$$

Likewise,

$$\mathcal{E}\Big[K\Big(\frac{x-X_i}{h}\Big)\bar{K}\Big(\frac{y-X_{i+1}}{h}\Big)\Big]\cdot\mathcal{E}\Big[K\Big(\frac{x-X_j}{h}\Big)\bar{K}\Big(\frac{y-X_{j+1}}{h}\Big)\Big]$$

$$= h^2\int K(y_i)p(x-hy_i)dy_i\ \cdot\ \int K(y_j)p(x-hy_j)dy_j \le Ch^2. \tag{5.22}$$

From (5.21) and (5.22), it follows that $(nh)^{-1}|J^*_{1n}|\to 0$. That $(nh)^{-1}|J^*_{2n}|\to 0$ follows as in the previous two lemmas. The proof of the lemma is complete. ∎

We may now combine these lemmas in order to justify the

Proof of Proposition 5.2. It is an immediate consequence of Lemmas 5.1-5.6 by way of relation (5.9). ∎

6. Proof of Theorem 2.3

As was explained at the beginning of the previous section, it suffices to show asymptotic normality for $\xi_n(x,y)$, given by (5.4); or just asymptotic normality for $\psi_n(x,y)$, given by (5.6), in view of Proposition 5.1. This is done by following the familiar steps of rewriting $\psi_n(x,y)$, by means of (1.1) and (3.6), and splitting the

resulting sum into large (α-) blocks and small (β-) blocks. In what follows, it will be notationally convenient to write K_i and $\bar{K}_{i+1}$ instead of $K\left(\frac{x-X_i}{h}\right)$ and $\bar{K}\left(\frac{y-X_{i+1}}{h}\right)$, respectively. Then

$$\psi_n = \psi_n(x,y) = \frac{1}{(nh)^{1/2}} \sum_{i=1}^{n} Z_{ni}, \tag{6.1}$$

where

$$Z_{ni} = Z_{ni}(x,y) = c_1(K_i - \mathcal{E}K_i) + c_2[(K_i\bar{K}_{i+1}) - \mathcal{E}(K_i\bar{K}_{i+1})]. \tag{6.2}$$

Also, set

$$S_n = S_n(x,y) = \sum_{i=1}^{n} Z_{ni} = (nh)^{1/2}\psi_n. \tag{6.3}$$

It will be shown that

$$\frac{1}{(nh)^{1/2}} S_n \xrightarrow{d} N\left(0,\ \tau_c^2\right), \text{ where } \tau_c^2 \text{ is given by (5.6).} \tag{6.4}$$

As usually, $\alpha = \alpha(n)$, $\beta = \beta(n)$ are positive integers tending to infinity and $\alpha+\beta \le n$, and $\mu = \mu(n)$ is the largest integer for which $\mu(\alpha+\beta) \le n$. Split $\Delta_n = \{1,\ldots,n\}$ into the following partitions

$$\left.\begin{aligned} \Delta'_{nm} &= \{(m-1)(\alpha+\beta)+1,\ldots,(m-1)(\alpha+\beta)+\alpha, \quad m=1,\ldots,\mu\}, \\ \Delta''_{nm} &= \{(m-1)(\alpha+\beta)\alpha+1,\ldots,\mu(\alpha+\beta), \quad m=1,\ldots,\mu\}, \\ \Delta'''_{n\mu} &= \{\mu(\alpha+\beta)+1,\ldots,n\}. \end{aligned}\right\} \tag{6.5}$$

Set

$$\left.\begin{aligned} y_{nm} &= \sum_{i\in\Delta'_{nm}} Z_{ni} = \sum_{i=k_m}^{k_m+\alpha-1} Z_{ni}, \quad k_m = (m-1)(\alpha+\beta)+1, \\ y'_{nm} &= \sum_{j\in\Delta''_{nm}} Z_{nj} = \sum_{j=\ell_m}^{\ell_m+\beta-1} Z_{nj}, \quad \ell_m = (m-1)(\alpha+\beta)+\alpha+1, \\ y'_{\mu+1} &= \sum_{k\in\Delta'''_{nm}} Z_{nk} = \sum_{k=\mu(\alpha+\beta)+1}^{n} Z_{nk}, \end{aligned}\right\} \tag{6.6}$$

so that

$$S_n = S'_n + S''_n + S'''_n, \tag{6.7}$$

where

$$S'_n = \sum_{m=1}^{\mu} y_{nm}, \quad S''_n = \sum_{m=1}^{\mu} y'_{nm}, \quad S'''_n = y'_{\mu+1}. \tag{6.8}$$

It is shown below that

$$\frac{1}{nh}\left[\mathcal{E}(S_n'')^2 + \mathcal{E}(S_n''')^2\right] \to 0 \quad \text{and} \quad \frac{1}{(nh)^{1/2}} S_n' \xrightarrow{d} N\left(0, \tau_c^2\right) .$$

Then the desired result will follow by means of (6.3), (6.4) and (6.7). Before going any further, let us make a digression regarding a possible choice of the quantities α and β. To this effect, let us take

$$\alpha = [n^{\delta}], \ \beta = [n^{\delta'}], \ 0<\delta'<\delta<1. \tag{6.9}$$

Then

$$\mu \approx n^{1-\delta}, \ \frac{\beta\mu}{n} \approx n^{\delta'-\delta} \to 0, \ \frac{\alpha\mu}{n} \to 1. \tag{6.10}$$

For later reference, let us set Z_i for Z_{ni} and observe that

$$\begin{aligned} \operatorname{var}(Z_1) &= c_1^2 \operatorname{var}(K_1) + c_2^2 \operatorname{var}(K_1\bar{K}_2) + 2c_1c_2 \operatorname{cov}\left(K_1, K_1\bar{K}_2\right) \\ &\le \left(c_1^2 + c_2^2 + 4|c_1c_2|\right)\mathcal{E}K_1^2 \\ &= h\left(c_1^2 + c_2^2 + 4|c_1c_2|\right)\int K^2(u)p(x-hu)du \le Ch \end{aligned} \tag{6.11}$$

(C, depending on x).

Also, let us record the following result, following immediately from Lemmas 5.4-5.6. Namely,

<u>Corollary 6.1</u>. Under the assumptions of Proposition 5.2 and for $x, y \in \mathbb{R}$,

$$\frac{1}{nh} \sum_{1\le i<j\le n} \left| \operatorname{cov}\left(Z_i, Z_j\right)\right| \to 0. \quad \blacksquare$$

We now proceed with the formulation of the following

<u>Lemma 6.1</u>. Under the assumptions of Proposition 5.2 and for $x, y \in \mathbb{R}$,

$$\text{(i)} \ \frac{1}{nh} \sum_{m=1}^{\mu} \operatorname{var}\left(y_{nm}'\right) \to 0, \quad \text{(ii)} \ \frac{1}{nh} \sum_{1\le i<j\le\mu} \left|\operatorname{cov}\left(y_{ni}', y_{nj}'\right)\right| \to 0.$$

<u>Proof</u>. By means of (6.11),

$$\operatorname{var}\left(y_{nm}'\right) = \sum_{i=\ell_m}^{\ell_m+\beta-1} \operatorname{var}\left(Z_i\right) + 2 \sum_{\ell_m\le k<\ell\le\ell_m+\beta-1} \operatorname{cov}\left(Z_k, Z_\ell\right)$$

$$\leq C\beta h + 2 \sum_{\ell_m \leq k < \ell \leq \ell_m + \beta - 1} |\mathrm{cov}(Z_k, Z_\ell)| .$$

Therefore

$$\frac{1}{nh} \sum_{m=1}^{\mu} \mathrm{var}(y'_{nm}) \leq C\left(\frac{\beta\mu}{n}\right) + \frac{2}{nh} \sum_{m=1}^{\mu} \sum_{\ell_m \leq k < \ell \leq \ell_m + \beta - 1} |\mathrm{cov}(Z_k, Z_\ell)|$$

$$\leq C\left(\frac{\beta\mu}{n}\right) + \frac{2}{nh} \sum_{1 \leq i < j \leq n} |\mathrm{cov}(Z_i, Z_j)| ,$$

and the expression on the right-hand side above tends to 0, by (6.10) and Corollary 6.1.

(ii) Clearly,

$$\frac{1}{nh} \sum_{1 \leq i < j \leq \mu} \left|\mathrm{cov}(y'_{ni}, y'_{nj})\right| \leq \frac{C}{nh} \sum_{1 \leq i < j \leq \mu} \sum_{k=\ell_i}^{\ell_i + \beta - 1} \sum_{\ell=\ell_j}^{\ell_j + \beta - 1} |\mathrm{cov}(Z_k, Z_\ell)|$$

$$\leq \frac{C}{nh} \sum_{1 \leq i < j \leq n} |\mathrm{cov}(Z_i, Z_j)| \to 0, \text{ as above.}$$

The proof is complete. ∎

The following result may now be established.

<u>Proposition 6.1</u>. Under the assumptions of Proposition 5.2 and for $x, y \in \mathbb{R}$,

$$\frac{1}{nh} \mathcal{E}(S''_n)^2 \to 0 \quad \text{and} \quad \frac{1}{nh} \mathcal{E}(S'''_n)^2 \to 0.$$

<u>Proof</u>. Indeed,

$$\mathcal{E}(S''_n)^2 = \sum_{m=1}^{\mu} \mathrm{var}(y'_{nm}) + 2 \sum_{1 \leq i < j \leq \mu} \mathrm{cov}(y'_{ni}, y'_{nj}),$$

and the proof of the first assertion follows by Lemma 6.1. Next, by (6.11),

$$\frac{1}{nh} \mathcal{E}(S'''_n)^2 \leq \frac{1}{nh} \sum_{i=\mu(\alpha+\beta)+1}^{n} \mathrm{var}(Z_i) + \frac{2}{nh} \sum_{\mu(\alpha+\beta)+1 \leq i < j \leq n} |\mathrm{cov}(Z_i, Z_j)|$$

$$\leq C\,\frac{n - \mu(\alpha+\beta)}{n} + \frac{2}{nh} \sum_{1 \leq i < j \leq n} |\mathrm{cov}(Z_i, Z_j)|$$

$$= C\left(1 - \frac{\alpha\mu}{n} - \frac{\beta\mu}{n}\right) + \frac{2}{nh} \sum_{1 \leq i < j \leq n} |\mathrm{cov}(Z_i, Z_j)| \to 0$$

on account of (6.10) and Corollary 6.1. ∎

The following result will also be essential in the sequel.

Proposition 6.2. Under the assumptions of Proposition 5.2 and for $x, y \in \mathbb{R}$,

$$\frac{1}{nh} \sum_{m=1}^{\mu} \operatorname{var}\left(y_{nm}\right) \to \tau_c^2 ,$$

where τ_c^2 is given in (5.5).

Proof. Clearly,

$$\frac{1}{(nh)^{1/2}} S_n' = \frac{1}{(nh)^{1/2}} S_n - \frac{1}{(nh)^{1/2}}\left(S_n'' + S_n'''\right) = \psi_n - \frac{1}{(nh)^{1/2}}\left(S_n'' + S_n'''\right),$$

so that

$$\frac{1}{nh} \mathcal{E}(S_n')^2 = \mathcal{E}\psi_n^2 + \frac{1}{nh} \mathcal{E}\left(S_n'' + S_n'''\right)^2 - \frac{2}{nh} \mathcal{E}\left[S_n\left(S_n'' + S_n'''\right)\right] . \tag{6.12}$$

By Proposition 6.1, relation (6.3), Proposition 5.2, and (6.12),

$$\frac{1}{nh} \mathcal{E}(S_n')^2 \to \tau_c^2 . \tag{6.13}$$

On the other hand,

$$\mathcal{E}(S_n')^2 = \operatorname{var}(S_n') = \sum_{m=1}^{\mu} \operatorname{var}\left(y_{nm}\right) + 2\sum_{1\le i<j\le\mu} \operatorname{cov}\left(y_{ni}, y_{nj}\right),$$

and

$$\frac{1}{nh} \sum_{1\le i<j\le\mu} \left|\operatorname{cov}\left(y_{ni}, y_{nj}\right)\right| \le \frac{C}{nh} \sum_{1\le i<j\le n} \left|\operatorname{cov}\left(Z_i, Z_j\right)\right| \to 0.$$

Therefore the identity

$$\frac{1}{nh} \mathcal{E}(S_n')^2 = \frac{1}{nh} \sum_{m=1}^{\mu} \operatorname{var}\left(y_{nm}\right) + \frac{2}{nh} \sum_{1\le i<j\le\mu} \operatorname{cov}\left(y_{ni}', y_{nj}'\right),$$

and relation (6.13) provide the desirable result. ∎

From this point on, the completion of the proof of the theorem follows the familiar pattern. An approximation to $(nh)^{-1/2}S_n'$ by independent r.v.'s (see, for example, relation (3.1) in Roussas (1990a)) requires the condition $\mu\rho(\beta) \to 0$. For the choice of α, β suggested in (6.9) (see also (6.10)), this condition is satisfied, provided $\nu > (1-\delta)/\delta'$. The second step in the proof is that of applying the Normal Convergence Criterion. Its validity is secured by the condition $\left(n^{1-\delta}h^2\right)^{-1} \to \infty$ (see arguments just before

Proposition 3.1 in Roussas (1990a)). Since also $nh^5 \to 0$, by assumption (A3)(iii), we may consider the following two cases: $\left(n^{1-\delta}h^2\right)^{-1} \le nh^5$ and $\left(n^{1-\delta}h^2\right)^{-1} > nb^5$. These inequalities are equivalent to $h \ge n^{(\delta-2)/7}$ and $h < n^{(\delta-2)/7}$, respectively, with $\delta < \frac{3}{5}$. To summarize then: Let $\rho(n)$ and $h=h_n$ be as below:

$$\left.\begin{array}{l} \rho(n) = O(n^{-\nu}), \quad \nu > (1-\delta)/\delta', \quad 0<\delta'<\delta<1, \text{ and either} \\ \quad h_n \ge n^{(\delta-2)/7} \ \left(\delta < \frac{3}{5}\right) \text{ with } nh_n^5 \to 0, \text{ or} \\ \qquad h_n < n^{(\delta-2)/7} \text{ with } n^{1-\delta}h_n^2 \to \infty \ \left(\delta < \frac{3}{5}\right). \end{array}\right\} \tag{6.14}$$

Then $(nh)^{-1/2}S_n' \xrightarrow{d} N\left(0, \tau_c^2\right)$. By Proposition 6.1 and relation (6.7), the same is true of $(nh)^{-1/2}S_n$. Finally, by means of (6.3), $\psi_n \xrightarrow{d} N\left(0, \tau_c^2\right)$. On the basis of the arguments presented in (5.1)-(5.8), this last convergence establishes Theorem 2.3. ∎

Remark 6.1. (i) For the popular choice of the bandwidth $h_n = Cn^{-\theta}$ $\left(C>0, \theta>0\right)$, the last two conditions in (6.14) become as follows, respectively:

$$\frac{1}{5} < \theta \le \frac{2-\delta}{7} \quad \left(\delta < \frac{3}{5}\right), \text{ and } \frac{2-\delta}{7} < \theta < \frac{1-\delta}{2} \quad \left(\delta < \frac{3}{5}\right).$$

(ii) It is noteworthy that for suitable choices of δ and δ', subject to the restrictions in (6.14), ν may get arbitrarily close to 1. For example, for $\delta = \frac{599}{1,000}\left(< \frac{600}{1,000} = \frac{3}{5}\right)$, and $\delta' = \frac{400}{1,000}$, we have $\frac{1-\delta}{\delta'} = 1.0025$. Thus, $\nu \geqq 1.0026$.

For this value of δ, the inequalities for θ in part (i) are close to the following: $0.2 < \theta \le 0.20014$ and $0.20015 < \theta < 0.2005$, respectively. That is, θ may get arbitrarily close to 1/5.

Acknowledgement

This research was partially supported by a grant from the Office of Research and Technology of Greece.

References

Doob, J.L. (1953) 'Stochastic Processes'. John Wiley and Sons, Inc. New York.

Loève, Michel (1963) 'Probability Theory', 3rd Ed. Van Nostrand Co., Inc. Princeton, New Jersy.

Masry, Elias (1986) 'Recursive probability density estimation for weakly dependent stochastic processes'. IEEE Trans. Inform. Theory 18, 254-267.

Masry, Elias (1989) 'Nonparametric estimation of conditional probability densities and expectations of stationary processes: Strong consistency and rates'. Stochastic Process. Appl. 32, 109-127.

Rao, C.R. (1965) 'Linear Statistical Inference and Its Applications', 1st Ed. John Wiley and Sons, Inc., New York.

Rosenblatt, M. (1970) 'Density estimates and Markov sequences'. In Nonparametric Techniques in Statistical Inference, M. Puri, Ed. Cambridge Unversity Press, Oxford.

Roussas, G.G. (1969a) 'Nonparametric estimation in Markov processes'. Ann. Inst. Statist. Math. 21, 73-78.

Roussas, G.G. (1969b) 'Nonparametric estimation of the transition distribution function of a Markov process'. Ann. Math. Statist. 40, 1386-1400.

Roussas, G.G. (1988) 'A moment inequality of S_{n,k_n} for triangular arrays of random variables under mixing conditions, with applications'. In Statistical Theory and Data Analysis II, pages 273-292, K. Matusita, Ed., North-Holland.

Roussas, G.G. (1990a) 'Asymptotic normality of the kernel estimate under dependence conditions'. J. Statist. Plann. Inference 25, 81-104.

Roussas, G.G. (1990b) 'Estimation of transition distribution function and its quantiles in Markov processes: Strong consistency and asymptotic normality'. Technical Report No. 197, Division of Statistics, University of California, Davis.

Roussas, G.G. and Ioannides, D. (1987) 'Moment inequalities for mixing sequences of random variables'. Stochastic Anal. Appl. 5(1), 61-120.

Samanta, M. (1989) 'Non-parametric estimation of conditional quantiles'. Statist. Probab. Lett. 7(5), 407-412.

Yakowitz, S. (1979) 'Nonparametric estimation of Markov transition functions'. Ann. Statist. 7, 671-679.

L_1 STRONG CONSISTENCY FOR DENSITY ESTIMATES IN DEPENDENT SAMPLES

G. B. SILVEIRA
London School of Economics
Dept of Statistics
Houghton st
London WC2A 2AE UK

ABSTRACT: We present necessary conditions for L_1 strong consistency of Kernel Density Estimates on uniform mixing samples. Those conditions turn out to be the same as in the i.i.d. case. We also study L_1 consistency for conditional density estimates.

1. Introduction

The statistical problem of density estimation plays an important role in a number of very active research fields such as pattern recognition and communication theory. There is a vast literature on density estimation based on dependent or independent samples. For a good survey of general results for independent samples we refer the reader to the book of Prakasa Rao [10]. For dependent samples, unfortunately, the results are much more scattered and, as far as we know, there is hardly any single volume providing a comprehensive study of the subject. See, however the recent book by Györfi et al. [9].

Asymptotic nontrivial results on density estimation are, quite often, demonstrated under a number of technical conditions which may lead a non–specialist to think of these results as being artificial. Devroye & Györfi in their book **Nonparametric Density Estimation, The L_1 View** , present a number of rather neat results, which hold valid under weak and simple conditions. The L_1 view for density estimation generates a very elegant theory, at least from a mathematical point of view. The results in Devroye and Györfi's book are based on i.i.d. samples only. Some of them have been extended to weakly dependent samples. Examples are Györfi [8] and Tran [11]. Here as in the independent case the regularity conditions are very weak.

In both of the last two works mentioned above, the aim was to present sufficient conditions for L_1 complete convergence of kernel density estimates (Györfi also considered histogram density estimates). Our purpose in this paper is to study necessary conditions for the L_1 strong convergence of kernel density estimates on weakly dependent samples. In addition, we show a kind of L_1 complete convergence for kernel estimates of conditional densities.

The remaining parts of this paper are organized as follows: in Section 2 we present the estimates and dependence conditions we are going to deal with. In Section 3 we present some technical lemmas. Finally in Section 4 we demonstrate

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 463–474.
© 1991 *Kluwer Academic Publishers.*

the main results.

2. Dependence Conditions and Results

2.1 WEAK DEPENDENCE CONCEPTS

Let $\{ X_n \}$, $n \geq 1$ be a strictly stationary stochastic process. We say that $\{ X_n \}$ is a strong mixing (φ–mixing resp.) process with mixig coefficients $\{ \alpha_n \}$ ($\{ \varphi_n \}$, respectively) , $n \geq 1$, if :

$$| P(A \cap B) - P(A)P(B) | \leq \alpha_k \text{ , (} | P(B|A) - P(B) | \leq \varphi_k \text{ , respectively.)}$$

for all $A \in \sigma\{ X_m , 1 \leq m \leq n \}$, $B \in \sigma\{ X_j, j \geq n+k+1 \}$. When $\alpha_k \leq a\exp(-ck)$, where $a>0$ and $c>0$, we say that the process is exponentially strong mixing.

The class of strong mixing processes contains properly the uniform or φ–mixing class. For a discussion on the hierarchy of mixing conditions we refer the reader to Bradley et al [2]. An important absence in the intersection of the above classes is the ARMA process. As a matter of fact only very special autoregressive processes of order 1 are uniform mixing (in which case the autoregressive function is necessarily bounded). See Doukhan et al. [5] for a discussion. On the other hand, many ARMA process are exponentially strong mixing, which represents quite a gain in generality over the uniform mixing processes. See Gorodetski [7] and Withers [12] for a discussion on linear processes and strong mixing conditions.

2.2 CONDITIONAL AND UNCONDITIONAL KERNEL ESTIMATES

Kernel estimates are probably the best known among all nonparametric density estimates. In particular, when dealing with dependent data, they are, in general quite more tractable than their rivals.

Let $\{ Z_n \}$, $n \geq 1$, be a stationary process taking values in $\mathbb{R}^d$. Let f be the common density and let

$$W: \mathbb{R}^d \to \mathbb{R}$$

be a real–valued function which integrates to 1. We define the Kernel Density Estimate (KDE) of f based on $\{ Z_m , 1 \leq m \leq n \}$ by:

$$f_n(z) = (nb^d)^{-1} \sum_{k=1}^{n} W((z - Z_k)/b) , z \in \mathbb{R}^d,$$

where $b = b(n)$ is a positive real number (usually referred to as bandwidth or smoothing parameter). When $d = c_1 + c_2$, $Z_n = (X_n , Y_n)$, where we are considering row vectors, and Y_n admits of a conditional density given X_n we estimate the conditional density by the ratio of the respective unconditional estimates. Namely, let

$$z = (x , y) \in \mathbb{R}^{c_1 + c_2}.$$

Also, let $f(z) = g(x)h(y|x)$, where h is the conditional density of Y_n given

X_n and g is the density of X_n. We estimate $h(y|x)$ by $f_n(z)/g_n(x)$, where :

$$f_n(z)= (nb^{c_1+c_2})^{-1} \sum_{k=1}^{n} W((x-X_k)/b) V((y-Y_k)/b) \text{ , and}$$

$$g_n(x)= (nb^{c_1})^{-1} \sum_{k=1}^{n} W((x-X_k)/b) ,$$

where W and V are real–valued functions integrating to 1. We can now state our first result:

Theorem 1: *Let $\{ Z_n \}$, $n\geq 1$ be a strictly stationary process taking values in $\mathbb{R}^{c_1+c_2}$, with $c_1, c_2 > 0$. Assume the existence of the conditional density of Y_1 given X_1 and the existence of a density for X_1, and consider the above defined estimate for the conditional density. Then ,if either*

(H1) $\{ Z_n \}$ is a uniform mixing process and $b(n) + (nb(n)^{c_1+c_2})^{-1} = o(1)$,

or

(H2) $\{ Z_n \}$ is a exponential strong mixing process and $\log^3(n)\ (nb(n)^{c_1+c_2})^{-1} = o(1)$,

we have

$$\int\int | h_n(y|x) - h(y|x) | \ dy\ \mu(dx) = o(1), \text{ a.s. as } n\to\infty .$$

Where μ is the distribution of X_1.

For unconditional density estimates we have:

Theorem 2: *Let $\{X_n\}$ be a φ–mixing process such that $\Sigma\varphi_n^q < \infty$, for some $q\geq 1$, and let $\{X_{-n}\}$ be also φ–mixing. If $J_n\to 0$ in probability then $b+1/(nb^d)\to 0$, as $n\to\infty$, where*

$$J_n = \int | f_n - f |$$

3. Auxiliary Results

Lemma 1: *Let $\{ A_n \}$ be dependent events on some probability space $(\Omega, \mathfrak{F}, P)$, let M_i be the σ–field generated by A_1 , ... , A_i , $M_0= \{ \phi, \Omega \}$ and let $p_i = P[A_i| M_{i-1}]$. For $i\leq n$ let N be the number of A_i's that occur, i.e. ,*

$$N = \sum_{i=1}^{n} I[\, A_i \,] .$$

Let $a \leq b$ be non-negative real numbers, and for $0<\varepsilon<1$, suppose that

$$a \leq p_1 + \dots + p_n \leq b \quad \text{and} \quad p_1^2 + \dots + p_n^2 \leq \varepsilon .$$

Let N^* be Poisson distributed with mean a. Then

$$d(\, N , N^* \,) \leq \beta\varepsilon + b - a ,$$

where $d(\, X, Y \,)$ is the variation distance between the random variables X and Y and β is a function of ε that remains bounded as $\varepsilon \to 0$.

Proof : See Freedman[6] .

Now we are going to introduce some notation for the proofs of the next Lemma and Theorem 2. As will be seen in Section 4 the proof of Theorem 2 uses the classical partition of the sample into small and large blocks. Namely, we are going to split the set of the first n integers into alternate blocks of length l_n (large blocks) and s_n (small blocks). Let

$$\Lambda_{ni} = \{\, (i-1)(l_n+s_n)+1, \dots , (i-1)(l_n+s_n)+l_n \,\} , \; 1 \leq i \leq r_n ,$$

where r_n is the largest integer j so that $(j-1)(l_n+s_n)+l_n < n$. Also let

$$\Psi_{ni} = \{\, (i-1)(l_n+s_n)+l_n+1, \dots , i(l_n+s_n) \,\} , \; 1 \leq i \leq r_n , \text{ and}$$

$$\Psi_{nr_n} = \{\, (r_n-1)(l_n+s_n)+l_n+1, \dots , n \,\} .$$

For $1 \leq k \leq n$ and belonging to some Λ_{ni} let $\gamma(k)$ be the order of the large block to which k belongs (that is $\gamma(k)=i$ iff $k \in \Lambda_{ni}$). For $1 \leq k \leq n$ belonging to some Ψ_{ni}, let $\delta(k)=i$ iff $k \in \Psi_{ni}$. We will also need the following convention: for each large block Λ_{ni} , let Γ_{ni} be the innermost portion of Λ_{ni} far apart from the boundaries by at least s_n "numbers". Namely let

$$\Gamma_{ni} = \Lambda_{ni} - [\, \{\, (i-1)(l_n+s_n)+1 , \dots , (i-1)(l_n+s_n)+s_n \,\} \cup$$

$$\cup \{\, (i-1)(l_n+s_n)+l_n-s_n+1 , \dots , (i-1)(l_n+s_n)+l_n \,\} \,] .$$

Finally, let

$$\Lambda = \bigcup_{i=1}^{r_n} \Lambda_{ni} \quad , \quad \Psi = \bigcup_{i=1}^{r_n} \Psi_{ni} \quad \text{and} \quad \Gamma = \bigcup_{i=1}^{r_n} \Gamma_{ni},$$

where we have dropped the subscript n from Λ, Ψ and Γ for notational convenience.

Lemma 2: *Suppose $\{X_n\}$, $n \geq 1$, is a φ–mixing process such that $\{\varphi_n\} \in l^q$. Consider the following array of events:*

$$\begin{array}{l} \mathbf{A}_{1b(1)} \\ \mathbf{A}_{1b(2)}\ \mathbf{A}_{2b(2)} \\ \cdots \\ \mathbf{A}_{1b(n)}\ \mathbf{A}_{2b(n)} \cdots \mathbf{A}_{nb(n)} \\ \cdots \end{array}$$

and suppose that:

(i) for each n and k ($1 \leq k \leq n$) we have $\mathbf{A}_{kb(n)} \in \sigma(X_k)$.

(ii) for each n and k we have $P[\ \mathbf{A}_{1b(n)}\ ] = P[\ \mathbf{A}_{kb(n)}\ ]$.

(iii) $\lim_n nP[\ \mathbf{A}_{1b(n)}\ ] = \beta < \infty$.

For $1 \leq i \leq r_n$, set

$$S_i = \bigcup_{j \in \Psi_{ni}} \mathbf{A}_{jb(n)}\ ,\ L_i = \bigcup_{j \in \Lambda_{ni}} \mathbf{A}_{jb(n)}\ ,\ M_i = \bigcup_{j \in \Gamma_{ni}} \mathbf{A}_{jb(n)}\ .$$

Then:

(A) $\liminf_n P[\ \Sigma' I[S_i] = 0\ ] = 1$,

(B) $\liminf_n P[\ \Sigma' I[L_i] = 0\ ] \geq \exp(-\beta)$,

(C) $\liminf_n P[\ \Sigma' I[M_i] = 0\ ] \geq \exp(-\beta)$,

Where the summations above are over i in $\{\ 1, \ldots, r_n\ \}$.

Proof: As the technique of proof is virtually the same for cases (A), (B) and (C), we will only discuss (B). We have:

$$P[\ L_j\ ] \leq l_n P[\ \mathbf{A}_{1b(n)}], \ 1 \leq j \leq r_n, \tag{3.1}$$

$$|\ P[L_j | L_1, \ldots, L_{j-1}] - P[L_j]\ | \leq \varphi_{s_n}\ . \tag{3.2}$$

From (3.1) and (3.2), we have

$$r_n P[L_1] - r_n \varphi_{s_n} \le \sum_{j=1}^{r_n} P[\, L_j | \, L_1, \dots, L_{j-1} \,] \le$$

$$r_n P[\, L_1 \,] + r_n \varphi_{s_n} \tag{3.3}$$

and

$$\sum_{j=1}^{r_n} P^2[\, L_j | L_1, \dots, L_{j-1} \,] \le 2r_n \left(P^2[L_1] + \varphi_{s_n}^2 \right) \le$$

$$\le 2r_n l_n^2 P^2[\, A_{1b} \,] + 2r_n \varphi_{s_n}^2 . \tag{3.4}$$

Now choose α so that $(1-\alpha)q<\alpha<1$, and let

$$l_n = O(n^{\alpha}),\ s_n = O(n^{(1-\alpha)q}),\ \text{so that } r_n = O(n^{(1-\alpha)}).$$

Since $\{\varphi_n\} \in l^q$ and $\lim_n nP[A_{1b}] = \beta$, we have that the right hand side of (3.4) goes to 0 as n goes to infinity. Assumptions of Lemma 1 are in force and we conclude that :

$$|\, P[\, I_{L_1} + \dots + I_{L_r} = 0 \,] - P[\, Z = 0 \,] \,| \le$$

$$\le \beta(2r_n l_n^2 P^2[A_{1b}] + r_n \varphi_{s_n}^2) + 2r_n \varphi_{s_n} = o(1),$$

where $Z \sim$ Poisson($r_n P[L_1] - r_n \varphi_{sn}$). Thus

$$P[\, I_{L_1} + \dots + I_{L_{rn}} = 0 \,] = \exp(-r\, P[L\,] + o(1)) + o(1).$$

Therefore

$$\liminf P[\, I_{L_1} + \dots + I_{L_{r_n}} = 0 \,] = \exp(-\limsup r_n P[L_1]) \ge$$

$$\ge \exp(-\limsup r_n l_n P[A_{1b}] = \exp(-\beta),$$

by (3.1) and assumption (iii). □

4. Proofs of Theorems

Proof of Theorem 1: By Theorem 1 in Györfi [8] and Corolary 2.1 in Tran [11] we have:

$$\int | g_n(x) - g(x) | \, dx \to 0, \text{ a.s. } \quad \text{as } n \to \infty . \tag{4.1}$$

Therefore

$$\int\int f(x,y)/g(x) \, | g(x) - g_n(x) | \, dy \, dx \to 0, \text{ a.s.} \tag{4.2}$$

or

$$\int\int f(x,y) \, | 1 - g_n(x)/g(x) | \, dx \, dy \to 0, \text{ a.s.} \tag{4.3}$$

But

$$\int\int | f_n(x,y) - f(x,y) | \, dx \, dy \to 0, \text{ a.s. .} \tag{4.4}$$

By (4.3) and (4.4) we have:

$$\int\int g_n(x) \, | f_n(x,y)/g_n(x) - f(x,y)/g(x) | \, dy \, dx$$

$$= \int\int | f_n(x,y) - f(x,y).g_n(x)/g(x) | \, dy \, dx \to 0. \tag{4.5}$$

Now notice that for fixed x both f_n/g_n and f/g are densities on $\mathbb{R}^d$, so that

$$\int\int g(x) \, | f_n(x,y)/g_n(x) - f(x,y)/g(x) | \, dy \, dx$$

$$= \int\int_{x:\, g(x) \le g_n(x)} + \int\int_{x:\, g(x) \ge g_n(x)}$$

$$\le \int\int g_n(x) \, | f_n(x,y)/g_n(x) - f(x,y)/g(x) | \, dy \, dx$$

$$+ \; 2\int_A g(x) \int_B (f(x,y)/g(x) - f_n(x,y)/g(x)) \, dy \, dx \tag{4.6}$$

Where $A = \{ x: g(x) > g_n(x) \}$ and $B = \{ y: f(x,y)/g(x) > f_n(x,y)/g_n(x) \}$. In the above we have used Scheffe's Theorem according to which $\int | u - v | = 2 \int \max\{ (u - v), 0 \}$, if u and v are densities. Now the second term on the right hand side of (4.6) is less than

$$2\int_A\int_B f(x,y)-f_n(x,y)\,dy\,dx \le 2\iint |\,f_n-f\,|.$$

The last inequality, (4.6) and (4.5) imply the result.

Proof of Theorem 2: The proof consists of two parts. For the first part we must show that $J_n\to 0$, in probability, for some density f implies $b(n)\to 0$. The proof of this result can be found in Devroye et al (1985), and therefore is omitted. In the second part, we must show that $J_n\to 0$, in probability, for some density f implies that as $n\to\infty$ we have $n[b(n)]^d\to\infty$.

The proof of the latter part will be carried out by showing that no subsequence of $\{n[b(n)]^d\}$ can converge to a finite limit.Suppose we have such a subsequence of $\{nb^d\}$ having finite, limit, say, β. Since $J_n\le 2$ and $J_n\to 0$, in probability, we have

$$E(J_n) = E[\int |\,f_n(x)-g(x)\,|\,dx\,] = o(1).$$

Now let $\{\,X_n^*\,\}$, $n\ge 1$, be an independent copy of the process $\{\,X_n\,\}$, $n\ge 1$. The superscript "*" on f_n below is to be understood as the estimated density by the former process. We have

$$\int |\,f_n(x)-g(x)\,|\,dx + \int |\,|\,f_n^*(x)-g(x)\,|\,dx \ge \int |\,f_n(x)-f_n^*(x)\,|\,dx$$

Therefore $J_n\to 0$, in probability and J_n being uniformly bounded, imply

$$\int |\,f_n^*(x)-f_n(x)\,|\,dx = o(1), \text{ in probability, and thus}$$

$$\int E|\,f_n(x)-f_n^*(x)\,|dx = E\int |\,f_n(x)-f_n^*(x)\,|dx = o(1).$$

Let R be any positive number such that:

$$\int_{B_R(0)} W(z)dz > 0, \tag{4.7}$$

where $B_r(a)$ stands for the Open Ball in $\mathbb{R}^d$ centered at a and with radius r. Let $A=A(x)$ be the event that

$$X_k-x \notin B_{bR}(0),\ k\epsilon\Lambda \text{ and } X_k^*-x \notin B_{bR}(0),\ k\epsilon\Psi.$$

We have :

$$E\int |\overset{*}{f}_n - f_n| = \int E|\overset{*}{f}_n - f_n| \geq \int E[(\overset{*}{f}_n - f_n)I[A]]$$

$$= n^{-1}\Sigma' \int E[(W_b(x-\overset{*}{X}_k) - W_b(x-X_k))I[A]\,dx$$

$$+ n^{-1}\Sigma'' \int E\,W_b(x-\overset{*}{X}_k)I[A]\,dx$$

$$- n^{-1}\Sigma'' \int E\,W_b(x-X_k)I[A]\,dx$$

$$= (I) + (II) + (III), \tag{4.8}$$

where Σ' (Σ'', respectively) stands for the summation over all k in Γ ($\{1, 2, \ldots, n\} - \Gamma$, respectively) and $W_b(z)=(1/b^d)W(z/b)$. We will need some further notation. Let

$$D(x) := [X_j \notin B_{bR}(x),\ j \in \Lambda], \tag{4.9}$$

$$D_{-k}(x) := [X_j \notin B_{bR}(x),\ j \in \Lambda - \Lambda_{n\gamma(k)}], \quad k \in \Lambda, \tag{4.10}$$

$$D^*(x) := [\overset{*}{X}_j \notin B_{bR}(x),\ j \in \Psi], \tag{4.11}$$

$$D^*_{-k}(x) := [\overset{*}{X}_j \notin B_{bR}(x),\ j \in \Psi - \Psi_{n\delta(k)}], \quad k \in \Psi, \tag{4.12}$$

For $k \in \Gamma$ we have, by the repeated use of a well known inequality for covariances of φ–mixing processes (see, e.g., Billingsley [1]):

$$E[W_b(x-\overset{*}{X}_k)I[A]] = E[W_b(x-\overset{*}{X}_k)I[D^*]]\,P[D]$$

$$\geq E[W_b(x-\overset{*}{X}_k)]P[D^*]\,P[D] - 6\varphi_{s_n} E[W_b(x-\overset{*}{X}_k)]\,P[D]. \tag{4.13}$$

Also,

$$E[W_b(x-X_k)I[A]] = E[W_b(x-X_k)I[D]].P[D^*]$$

$$\leq E[W_b(x-X_k)\,I[X_k \notin B_{bR}(x)]\,I[D_{-k}]]\,P[D^*]$$

$$\leq E[W_b(x-X_k)\,I[X_k \notin B_{bR}(x)]]\,P[D_{-k}]\,P[D^*]$$

$$+ 6E[\, W_b(x-X_k)\, I[\, X_k \notin B_{bR}(x)\,]\,]\, P[D^*]\, \varphi_{s_n}$$

$$= E[\, W_b(x-X_k)\, I[\, X_k \notin B_{bR}(x)\,]\,]\, P[D]\, P[D^*]$$

$$+ 6E[\, W_b(x-X_k)\, I[\, X_k \notin B_{bR}(x)\,]\,]\, P[D^*]\, \varphi_{s_n}$$

$$+ E[\, W_b(x-X_k) I[X_k \notin B_{bR}(x)]\,]\, P[D^*]\, [P[D_{-k}]-P[D]]. \tag{4.14}$$

From (4.13) and (4.14) we see that:

$$(I) \geq a_n \int E[\, W_b(x-X_k^*)-W_b(x-X_k)\, I[\, X_k \notin B_{bR}(x)\,]\,]\, P[D^*]\, P[D]\, dx$$

$$+ a_n T_n\,, \tag{4.15}$$

where T_n is the sum on the right hand side in (4.13) and (4.14) minus

$$\int E[\, W_b(x-X_k^*)-W_b(x-X_k).I[\, X_k \notin B_{bR}(x)\,]\,]\, P[D^*]\, P[D]\, dx$$

and $a_n = n^{-1}(\, l_n - 2s_n\,)r_n$. Now the integral on the right hand side of (4.15) can be written as

$$a_n \int f(y) \int_{B_R(0)} W(z).P[D^{*+}].P[D^+]\, dz\, dy, \tag{4.16}$$

Where D^{*+} (D^+, respectively) stands for $[\, X_j^* \notin B_{bR}(y-zb),\ j\epsilon\Psi\,]$ ($[\, X_j \notin B_{bR}(y-zb),\ j\epsilon\Lambda\,]$). By Lebesgue's density theorem we have

$$\lim\ nP[\, X_j \epsilon B_{bR}(y-zb)\,] = \beta R^d V f(y)\,, \tag{4.17}$$

Where V is the volume of the unit ball in R^d. Now by means of Lemma 4, Fatou's Lemma and (4.17), we get:

$$\underline{\lim}\ a_n \int E[\, W_b(x-X_k^*)-W_b(x-X_k)\, I[\, X_k \notin B_{bR}(x)]\,]\, P[D^*]\, P[D]\, dx$$

$$\geq \int f(y) \int_{B_R(0)} W(z) \exp(-\beta R^d V f(y))\, dz\, dy > 0. \tag{4.18}$$

Also, by a further application of Lebesgue's density theorem, it is easily seen that:

$$\int E[\, W_b(x-X_k)\, I[\, X_k \notin B_{bR}(x)\,]\,]\, P[D^*]\, [\, P[D_{-k}]-P[D]\,]dx = o(1).$$

Indeed, the above integrands are all majorized by the integrable function $u(x) = E[W_b(x-X_k)]$. On the other hand,

$$P[\, D_{-k}\,] - P[\, D\,] \leq l_n P[\, X_1 \in B_{bR}(x)\,] = o(1)$$

by Lebesgue's density theorem, our assumption that nb^d is $O(1)$ and the choice of $l_n = O(n^{\alpha})$. The other components of T_n are also readily seen to be $o(1)$ as they are dominated by

$$\varphi_{s_n} \int E[\, W_b(x-X_1)\,]dx \leq \varphi_{s_n} = o(1)$$

where the inequality follows by an application of Young's inequality, according to which:

$$\int f * g \leq \int f . \int g$$

Thus

$$\liminf (I) \geq \int f(y) \exp(-\beta R\, Vf(y))\, dy \int_{B_R(0)} W(z)\, dz > 0\, .$$

By a further application of Young's inequality, we have

$$(III) \leq 3s_n n^{-1} \int E[\, W_b(x-X_k)\,]\, dx \leq 3s_n n^{-1} = o(1).$$

Now from (4.18), $T_n = o(1)$, $(III) = o(1)$ and $(II) \geq 0$, so that we obtain

$$\liminf E \int |\, f_n - f_n^*\, | > 0,$$

which cannot hold. Therefore no subsequence of $\{nb^d\}$ can have a finite limit and thus $\lim nb^d = \infty$. □

5 Acknowledgements

This paper is part of the author's doctoral dissertation. The supervision of Professor Peter Robinson proved extremely valuable and is sincerely acknowledged. This

article is based on research funded by the Economic and Social Research Council (ESRC), reference number: R000231441.

6. References

1. Billingsley, P. (1968): Convergence of Probability Measures. John Wiley: New York.
2. Bradley, R. C., Janson, S., Wlodzimiers, B. (1987): " On Dominations Between Measures of Dependence ", Journal of Multivariate Analysis, 23, 303–311.
3. Devroye, L. (1983): " The equivalence of weak, strong and complete convergence in L_1 for kernel density estimates ". Annals of Statistics, 11, 896–904.
4. Devroye, L. and Györfi, L. (1985): " Nonparametric Density Estimation". John Wiley: New York.
5. Doukhan, P. , Ghindes, M. (1980): " Etude du Processus X(n)=f(X(n–1)) + e(n) ", Thesis at Univ. Paris Sud, Orsay, France.
6. Freedman, D. (1974): " The Poisson approximation for dependent events ", Annals of Probability , 2, 256–269.
7. Gorodetski, V. V. (1977): " On the Strong Mixing Properties for Linear Sequences ". Theory of Probability and its Applications, 22, 411–413.
8. Györfi, L. (1987): " Density Estimation from Dependent Sample ", In Statistical Data Analysis Based on the L_1 Norm and Related Methods, Y Dodge Ed. North–Holland, pp. 393–402.
9. Györfi, L., Härdle, W., Sarda, P. & Vieu, P. (1989): " Nonparametric Curve Estimation from Time Series ", Springer–Verlag, New York.
10. Prakasa Rao, B. L. S. (1987): " Asymptotic Theory of Statistical Inference " John Wiley: New York.
11. Tran, L. T. (1989): " The L_1 Convergence of Kernel Density Estimates under Dependence ", Canadian Journal of Statistics, 17, 197–208.
12. Withers, C. S. (1981): " Central Limit Theorems for Dependent Variables I " Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete, 57, pp 509–534.

VI. Time Series Analysis, Signal Detection

NONPARAMETRIC STATISTICAL SIGNAL DETECTION PROBLEMS

R. Ahmad
Department of Statistics and Modelling Science
University of Strathclyde
26 Richmond Street, Livingstone Tower
Glasgow G1 1XH, Scotland, UK

ABSTRACT. Distribution-free signal detection techniques are investigated for three most commonly occurring families of distributions. These are Pareto and uniform renewal processes, and stochastic processes with stationary independent symmetric increments.

1. Introduction and Preliminaries

The general signal detection problems, in essence, can be treated as testing statistical hypotheses, see Bell (1964a), Bell-Ahmad (1983), Bell-Ahmad-Park-Lui (1987) and Choi-Bell-Ahmad-Park (1989). The methodology also includes parametric, semiparametric and nonparametric estimation of parameters, relevant functional estimation and hypotheses testing. In signal detection problems the data is usually composed of i.i.d. random variables with a common probability distribution belonging to a class of distributions. The observed variables are interarrival times of some event or signal of interest. Unlike general estimation and testing problems in statistics and probability, for the data in signal detection problems it is undesirable to transform data prior to the actual statistical analysis. This is because the data in such situations consists of real times and is very crucial to the actual inference.

A major part of statistical methodology in signal detection is directed towards models in which the data received are Gaussian, Poissonian or Gaussian mixture. The first two families have relatively "thin tails" which restricts their use in signal detection problems. Under these models the classical statistical techniques and their extensions can, in essence, be adapted to signal detection problems.

In view of the many applications and the desirability of having probability distributions with "thick tails", we consider three classes of distributions. These are: (i) Pareto renewal processes family, (ii) uniform renewal processes class, and (iii) stochastic processes with stationary independent symmetric increments. In practice these processes appear in many signal detection problems such as sonar, radar, electronic detection; various medical diagonosis e.g. abnormal heart beats, strokes, seizures, etc.; radiation and other hazards; natural hazards e.g. floods, high winds, tornadoes etc. Other examples include disasters and catastrophes in industrial and commercial spheres e.g. collapse of big firms, exceptional rise and fall in the stock exchange markets, among others.

The treatment of various statistical problems here is essentially that of an expository overview incorporating some recent new developments. The organisation of the paper is as follows. In section two the basic statistical concepts are treated. The next three sections, respectively, briefly develop some detection procedures for the above three families of stochastic processes.

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 477–492.
© 1991 *Kluwer Academic Publishers.*

2. Basic Statistical Concepts

In the sequel we shall consider the following three classes.

(a) Pareto Renewal Processes: $\Omega_o = \{F(x) : P(X \leq x) = 1 - (\frac{\theta}{x})^{\psi},\ x > \theta > 0, \psi > 0\}$.

Remark 2.1. The Pareto distribution was introduced by Vilfredo Pareto (1848-1923) during 1897. Other researchers who used this distribution include Pigou (1932), Mandelbrot (1960,1963), Fisk (1961) and Johnson and Kotz (1970).

(b) Uniform Renewal Processes: $\Omega_1 = \{F(x) : F = U(0, \theta), \theta > 0\}$.

(c) Stochastic Processes with Stationary Independent Symmetric Increments: $\Omega_2 = \{F(x) : F(x) + F(-x) = 1 \quad \text{for all} \quad x\}$.

Three types of detection problems will be treated. These are outlined below in terms of pure noise (PN) or H_o and noise-plus-signal $(N + S)$ or H_1, see Bell (1964a).

(i) *Goodness-of-fit Detectors* : Here we detect the problem of $H_o : \mathcal{L} = \mathcal{L}_o \in \Omega$ against $H_1 : \mathcal{L} \neq \mathcal{L}_o$, where $\mathcal{L}_o$ is completely specified probability law of the underlying process.

(ii) *Class-fit Detectors* : This involves testing $H_o : \mathcal{L} \in \Omega$ versus $H_1 : \mathcal{L} \notin \Omega$.

(iii) *Two-sample Detectors* : This problem is to detect $H_o : \mathcal{L}_1 = \mathcal{L}_2$ against $H_1 : \mathcal{L}_1 \neq \mathcal{L}_2$, where $\mathcal{L}_i \in \Omega$, $i = 1, 2$.

We have chosen the above setting because of, firstly, its wide applicability in many signal detection situations, and secondly, because of limited available space. However, several of these ideas can be generalised by considering different underlying stochastic processes, introducing multi-dimensional processes, and by choosing different detection hypotheses such as dependence, randomness k-sample problems etc. The generic data is denoted by $\mathbf{Z} = (X_1, \ldots, X_n)$ or $\mathbf{Z} = (X_1, \ldots, X_n; Y_1, \ldots, Y_m)$.

For many practical situations the decision as to whether a signal is present or not is optimally based on the data solely through the minimal sufficient statistics (MSS), $S(\mathbf{Z})$. However, in many other cases one needs a statistic complementary to the MSS; this is called the maximal statistical noise (MSN) and is denoted by $N(\mathbf{Z})$ depending upon the model in the problem.

2.1 BASIC DATA TRANSFORMATION AND MAXIMAL STATISTICAL NOISE (BDT and MSN)

Definition 2.1. Let $N(\mathbf{Z})$ be a (vector- valued) statistic independent of $S(\mathbf{Z})$ and such that $\delta(\mathbf{Z}) = [\mathbf{S}(\mathbf{Z}), \mathbf{N}(\mathbf{Z})]$ is $1 - 1$ almost everywhere w.r.t. Ω. Then, (i) $\delta(\mathbf{Z})$ is called the BDT; and (ii) $N(\mathbf{Z})$ is called the MSN.

Clearly $S(\mathbf{Z})$ contains all the relevant information about the parameter; and it will be seen that $N(\mathbf{Z})$ contains all the relevant information about the structure of the process. From $\delta(\mathbf{Z})$ one should almost always be able to reconstruct the original data, $\mathbf{Z}$. For various examples see Bell-Ahmad (1983), Bell-Ahmad-Park- Lui (1987) and Choi-Bell-Ahmad-Park (1989).

2.2 TYPES OF DISTRIBUTION-FREE-NESS

There are three distinct types of distribution-free statistics which arise for many types of statistical hypotheses testing and signal detection problems.

Definition 2.2. (i) A statistic $T(\mathbf{Z})$ is said to be nonparametric distribution-free (NPDF) w.r.t. a family Ω^* of stochastic processes if there exists a single distribution function $G(\cdot)$ such that for all probability laws $\mathcal{L}$ in Ω^*

$$P(T(\mathbf{Z}) \leq t) = G(t).$$

(ii) A family of statistics, $\{T_1(\mathbf{Z};\mathcal{L})\}$, which is indexed by the elements of Ω^* is called parametric distribution-free (PDF) w.r.t. Ω^* if there exists a distribution function, $G(\cdot;T_1) \equiv G_1(\cdot)$, such that

$$P(T_1(Z;\mathcal{L}) \leq t_1|\mathcal{L}) = G_1(t_1).$$

(Note that each NPDF is PDF).

(iii) A parametric distribution-free statistic $T(\cdot;\cdot)$ is called strongly distribution-free (SDF) w.r.t. Ω^* (and in the PN-situation: $\mathcal{L} = \mathcal{L}_o$) if the probability of false dismissal (PFD) i.e.; $\beta(\cdot)$ depends on $\mathcal{L}_1$ only through $\mathcal{L}_1 o \mathcal{L}_o^{-1}$, where the symbol "$o$" denotes the composition of the involved functions.

Example 2.1. Let $\Omega^* = \Omega_o$, and $\mathbf{Z} = (\mathbf{Z_1}, \ldots, \mathbf{Z_N})$ be the first N interarrival times with $Z_i = X_i$ for $1 \leq i \leq n$, and $Z_{n+r} = Y_r$ for $1 \leq r \leq m$, where $N = n + m$. Let $\hat{\psi}_1$ and $\hat{\psi}_2$ be the MLE's of ψ_1 and ψ_2, that is,

$$\hat{\psi}_1 = n[\sum_{j=1}^{n} logX_j - nlog\theta_o]^{-1}, \quad \hat{\psi}_2 = m[\sum_{j=1}^{m} logY_j - mlog\theta_o]^{-1}.$$

Set $T_1 = n\hat{\psi}_2/m\hat{\psi}_1$ and $T_2 = 2n\psi/\hat{\psi}_1$. Then T_1 is NPDF w.r.t. Ω^* with $G = F_{2n,2m}$; and T_2 is PDF w.r.t. Ω^* with $G_1 = \chi^2_{2m}$. For SDF statistics see Bell (1964b) and Choi-Bell-Ahmad-Park (1989) section 4. Notice that T_1 is a function of the data $\mathbf{Z}$ only through the MSN, $N(\mathbf{Z})$; and T_2 is a function of the MSS, $S(\mathbf{X})$.

The following theorem gives the relationship between the MSN, MSS and DF statistics.

Theorem 2.1. Let Ω^* be a family of c.d.f.'s admitting a MSS, $S(\mathbf{Z})$ for data $\mathbf{Z} = (\mathbf{Z_1}, \ldots, \mathbf{Z_n})$. Then, (i) $T(h, \mathbf{Z}) = \mathbf{h}(\mathbf{N}(\mathbf{Z}))$ is NPDF w.r.t. Ω^* for each measurable function $h(\cdot)$; and (ii) $T^*(h^*, F, S(\mathbf{Z})) = \mathbf{h}^*(\mathbf{F}, \mathbf{S}(\mathbf{Z}))$, where $\mathbf{Z}$ is governed by G, is PDF w.r.t. Ω^* for each measurable function $h^*(\cdot)$.

3. Detection Procedures for the Family Ω_o

Several types of applications entail point processes with "dead" times after each event. One such class of stochastic processes is Ω_o. An additional interesting property of the interarrival time distrirutions is that they are "thick-tailed" relative to the corresponding distributions for Gaussian and Poisson processes. Further, a variety of tail thicknesses, one for each ψ-value, is possible. Also, the family Ω_o is a good approximating class for a subclass of Levy (1925) stable processes on R^1_+. Notice that the exact expressions for various densities in this subclass are not easily available. On the basis of these facts Ω_o is an important class of signal detection models and problems. For example, recently Davidson and Smith (1990) use such models for exceedances over high thresholds environmental extremes, in nuclear siting policy, hydrology and oceanography.

3.1 CLASS-FIT PROBLEM FOR $\Omega_o(\theta, \psi)$

In 1933 Kolmogorov introduced the $K-S$ statistic, $D_n(F_o) = \sup_x |F_n(x) - F_o(x)|$, for known continuous cdfs $F_o(\cdot)$ and their respective empirical cdfs $F_n(\cdot)$. This statistic and its various versions have been used widely for goodness-of- fit problems. However, in many signal detection problems $F_o(\cdot)$ is known only up to a nuisance parameter, or equivalently is known only to be a member of a specific family. Lilliefors (1967,1969), Srinivasan (1970), Durbin (1973) and Choi (1981) introduced modified versions of the $K-S$ statistic for such situations (for example, exponential, normal and uniform cases).

To test $H_o : \mathcal{L}(\cdot; \theta, \psi) \in \Omega_o$ against $H_1 : \mathcal{L}(\cdot; \theta, \psi) \notin \Omega_o$, we shall define two statistics:

$$\hat{D}_n = \sup_x |F_n(x) - \hat{F}(x)| = \sup_x |F_n(x) - F(x; \hat{\theta}, \hat{\psi})|$$

where $(\hat{\theta}, \hat{\psi})$ are MLEs of (θ, ψ), and

$$\tilde{D}_n = \sup_x |F_n(x) - \tilde{F}(x)|,$$

where $\tilde{F}(x) = P(X_1 \leq x | S(\mathbf{Z}))$, that is $\tilde{F}$ is the Rao-Blackwell estimate of F.

Naturally these two modifications are applicable where the underlying family Ω admits estimators $\hat{F}$ and $\tilde{F}$. The statistics $\hat{D}, \tilde{D}$ for one-sample, 2-sample, and several other problems are asymptotically equivalent to the $K-S$ statistic D. However, for small samples $\tilde{D}$ turns out to be more sensitive to any deviations from H_o. The statistic $\hat{D}$ and its various equivalent versions can be found in Bell et al. (1987,1989). Here, we re-emphasize the statistic $\tilde{D}$ through the following useful result.

Theorem 3.1. Let $X_1, \ldots, X_n$ be a random sample with an unknown distribution from the family $\Omega_o(\theta, \psi)$. Then

$$(i) \qquad \begin{aligned} \tilde{F}_n(x) &= 0, \quad x < X_{(1)} \\ &= 1/n, \quad x = X_{(1)} \\ &= 1 - (\frac{n-1}{n})(1 - \frac{\hat{\psi}}{n} \log \frac{x}{X_{(1)}})^{n-1}, \quad X_{(1)} < x < X_{(1)} e^{n/\hat{\psi}} \\ &= 1, \quad x \geq X_{(1)} e^{n/\hat{\psi}} \end{aligned}$$

$$(ii) \qquad \tilde{D}_n = \sup_x |F_n(x) - \tilde{F}(x)|$$

$$= (\frac{n-1}{n}) \sup_{0<u<\sum_{j=1}^{n-1} U_{(j)}} \left| \frac{1}{n-1} \sum_{j=1}^{n-1} \in (u - U_{(j)}) - \{1- \right.$$

$$\left. - [1 - \frac{u}{(n-1)\bar{u}}]^{n-1}\} \right|,$$

where $\in (y) = 1$ if $y \geq 0$ and zero otherwise, and $U_{(j)}$, denotes the jth order statistic from $U(0,1)$.

The critical values of this statistic cannot be obtained from the known existing tables. However, these can easily be obtained by the Monte Carlo simulation method for various sample sizes and significant levels (probability of false alarm). For two-sample inference for various cases, see Bell et al. (1987).

4. Detection Techniques for the Family Ω_1

Before we develop some detection procedures for renewal processes with uniform interarrival times with nuisance parameter θ, it seems worthwhile to point out some other detection procedures in the case of $U(o,\theta), \theta > 0$ distribution. For detecting $H_o : \theta \leq \theta_o$ against $H_1 : \theta > \theta_o$ any test, $\phi(\mathbf{Z})$, is uniformly most powerful (UMP) at the significance level α for which $E_{\theta_o}\phi(\mathbf{Z}) = \alpha$, $E_\theta\phi(\mathbf{Z}) \leq \alpha$ for $\theta \leq \theta_o$; and $\phi(\mathbf{Z}) = \mathbf{1}$ when $X_{(n)}$ is greater than θ_o and zero otherwise.

For detecting $H_o : \theta = \theta_o$ against $H_1 : \theta \neq \theta_o$ a unique UMP unbiased test exists and is given by $\phi(\mathbf{Z}) = \mathbf{1}$ when $X_{(n)} > \theta_o$ or $X_{(n)} \leq \theta_o n\sqrt{\alpha}$, and $\phi(\mathbf{Z}) = \mathbf{0}$ otherwise. Similarly, a UMP detection procedure for the case $H_o : \theta = \theta_o$ against $H_1 : \theta < \theta_o$ can be developed and combined with the one sided case discussed above. For further details see Choi et al. (1989).

4.1 THE KOLMOGOROV-SMIRNOV-TYPE DETECTION PROCEDURES FOR RENEWAL PROCESSES WITH UNIFORM INTERARRIVALS (NUISANCE PARAMETERS CASE)

The object here is to develop the Kolmogorov-Smirnov-type detection procedures for "H_o : $F = U(0,\theta)$ for some $\theta > 0$" that is the case $PN_1 : \mathcal{L}_F \in \Omega_1$, where $X_1, X_2, \ldots, X_n, \ldots$ are the interarrival times of a renewal process with distribution law $\mathcal{L}_F \equiv \mathcal{L}$.

If the value of θ were specified as θ_o then, in addition to statistics employed above, a natural Kolmogorov-Smirnov statistic is

$$D_n = \sup_x |F(x, \theta_o) - F_n(x)|,$$

where $F_n(x)$ is the sample cpf. However, in the situation under discussion the value of θ is unknown, and, therefore, θ is a nuisance parameter. In the case of nuisance parameter structure, we define two variants of Kolmogorov-Smirnov-type statistics. These are called, respectively, Lilliefors-type statistics and Srinivasan-type statistics; see Lilliefors (1967, 1969), Srinivasan (1970).

The above two types of statistics are examples of NPDF statistics w.r.t. an appropriate family of distributions Ω^*, such as $\{\text{normals}\}$, $\{\text{exponentials}\}$, and $\Omega(URP)$.

Let a probability law $\mathcal{L}_o(X;\theta_1;\theta_2,\ldots,\theta_k)$, say, with distribution function $F_o(x;\theta_1,\ldots,\theta_k)$ be such that $T_1,T_2,\ldots,T_k$ are the joint sufficient statistics for the vector parameter $\theta = (\theta_1,\ldots,\theta_k)$. For a fixed real x_1 define the random variable $S(X_1) = I_{\{X_1 \le x_1\}}$, where I_A is the indicator function of the set A. Clearly $S(X_1)$ is an unbiased estimator of $F_o(x;\theta_1,\theta_2,\ldots,\theta_k)$ under $H_o : F = F_o$. Consequently, from the general theory of sufficient statistics it follows that the statistic

$$\tilde{F}_\theta(x) = E\{S(X_1)|T_1,\ldots,T_k\}$$

is an unbiased estimator of $F_o(x;\theta)$ with a smaller variance than $S(X_1)$. If in addition $(T_1,\ldots,T_k)$ are complete, then $\tilde{F}_\theta(x)$ is the unique MVUE of $F_o(x;\theta)$. If the distribution of statistics like $\hat{D}_n$ or $\tilde{D}_n$ does not depend on θ, then a statistic of this type would be an apropriate statistic for detecting a composite hypothesis F.

Now we specialize these statistics to the case of renewal processes with uniform interarrivals, where the nuisance parameter is $\theta > 0$.

In the treatment of signal detection problems the objective here is to detect

$$PN : \mathcal{L}(\mathbf{Z}) \in \mathbf{\Omega_1} \quad \textbf{against} \quad (\mathbf{N}+\mathbf{S}) : \mathcal{L}(\mathbf{Z}) \notin \mathbf{\Omega_1}$$

where $\mathbf{Z} = (\mathbf{X_1},\ldots,\mathbf{X_n})$ are the interarrival times.

Recall that the general rule in constructing detection procedures for problems like PN and other variants is:

(i) to use the MSS when the PN formulation entails a completely specified stochastic process law (as in the case with MODEL : PN), and

(ii) employ the MSN when the PN formulation entails membership in a family of stochastic processes laws.

4.2 LILLIEFORS-TYPE DETECTION STATISTIC

This statistic was earlier defined to be of the form:

$$\hat{D}_n = \sup_x |\hat{F}_\theta(x) - F_n(x)|,$$

where $\hat{F}_\theta(x)$ is the MLE of $F_\theta(x)$, and $F_n(x)$ is the cumulative sample distribution, that is, $F_n(x) = (\text{number of } X_j's \le x)/n = n^{-1}\sum_{j=1}^n \in (x - X_j)$, where $\epsilon(u) = 1$ if $u \ge 0$ and zero otherwise.

The MLE of θ is $X_{(n)} = \max(X_1,\ldots,X_n)$, which also happens to be MSS. Note that $T \equiv X_{(n)}$ is an example of a PDF statistic and $T(Z;\mathcal{L}_{\theta_o}) = X_{(n)}/\theta$ is an example of a NPDF statistic. Clearly, we have

$$\hat{F}_\theta(X) = \begin{cases} 0 & \text{if } x < 0, \\ \dfrac{x}{X_{(n)}}, & 0 \le x \le X_{(n)} \\ 1 & \text{if } x > X_{(n)} \end{cases}$$

Therefore, the Lilliefors type detection statistic can be written as

$$\begin{aligned}\hat{D}_n &= \sup_x |F_n(x) - \hat{F}_\theta(x)| \\ &= \sup_{0\le x\le X_{(n)}} |\frac{1}{n}\sum_{j=1}^{n} \in [x - X_{(j)}] - \frac{x}{X_{(n)}}| \\ &\overset{\mathcal{L}}{=} \sup_{0\le x\le X_{(n)}} |\frac{1}{n}\sum_{j=1}^{n-1} \in [\frac{x}{X_{(n)}} - \frac{X_{(j)}}{X_{(n)}}] - \frac{x}{X_{(n)}}| \\ &\overset{\mathcal{L}}{=} \sup_{0\le u\le 1} |\frac{1}{n}\sum_{j=1}^{n-1} \in [u - U_{(j)}] - u|\end{aligned}$$

when $u = x/X_{(n)}$ and $U_{(j)} = X_{(j)}/X_{(n)}$ and $\overset{\mathcal{L}}{=}$ means "it has the same probability $\mathcal{L}$ or distribution".

Before we give the actual decision rule, we sum up the basic distribution structure as below.

Theorem 4.1. Under the Class-fit hypothesis, the following are valid:

(i) The statistic $T(Z, \mathcal{L}_{\theta_o}) = \dfrac{X_{(n)}}{\theta}$ has the distribution given by $P\{X_{(n)}/\theta \le u\} = u^n$, $0 < u < 1$.

(ii) If $R(X_j)$ denotes the rank of X_j then the vector variable $R = [R(X_1), \ldots, R(X_n)]$ is distributed uniformly over the permutations of $\{1, 2, \ldots, n\}$. This we write as $R \sim D - U\{S_n(1, 2, \ldots, n)\}$.

(iii) The vector

$$V^* = [\frac{X_{(1)}}{X_{(n)}}, \ldots, \frac{X_{(n-1)}}{X_{(n)}}] = (V_1^*, \ldots, V_{n-1}^*)$$

has the same law as $[(U_{(1)}, ..., U_{(n-1)}]$, where $U_1, \ldots, U_{n-1}$ are i.i.d. random variables from $U(0, 1)$. That is, $V^* \overset{\mathcal{L}}{=} [U_{(1)}, \ldots, U_{(n-1)}]$, where $U_{(j)}$ is the jth order statistic.

Decision Rule. For detecting $PN_1 : \mathcal{L} \in \Omega_1$ versus $N + S : \mathcal{L} \notin \Omega_1$ decide $N + S$, if and only if, $\hat{D}_n > d^*$, where $P\{\hat{D}_n \le d^*\} = 1 - \alpha$, and $\alpha = P\{\text{False Alarm}\}$.

Lilliefors (1967, 1969) have given some critical values in other cases but not for a uniform renewal process. Choi (1981) has computed some tables for this case.

4.3 SRINIVASAN-TYPE DETECTION STATISTICS

Consider the same detection problem as in subsection (4.2) above. This type of statistic (other than Ω_1) was originally suggested by Lieberman and Resnikoff (1955) and later on Srinivasan (1970) investigated some cases in detail. Here, the statistic is

$$\tilde{D}_n = \sup_x |F_n(x) - \tilde{F}_\theta(x)|$$

where $\tilde{F}_\theta(x)$ is the MVUE (minimum variance unbiased estimate) of $F_\theta(x)$. If $\mathcal{L} \in \Omega_1$, then by the theory of sufficiency, one gets

$$\begin{aligned}
\tilde{F}_\theta(x) &= E\{I_{[X_1 \le x]} | X_{(n)}\} \\
&= P\{X_1 \le x | X_{(n)}\} \\
&= n^{-1} \sum_{j=1}^{n} P(X_j \le x | X_{(n)}) \\
&= n^{-1} \sum_{j=1}^{n} P(X_{(j)} \le x | X_{(n)}) \\
&= n^{-1} \{\sum_{j=1}^{n-1} P(X_{(j)} \le x | X_{(n)}) + P(X_{(n)} \le x | X_{(n)})\} \\
&= n^{-1} \sum_{j=1}^{n-1} P(X_{(j)} \le x | X_{(n)})
\end{aligned}$$

Finally,

$$\tilde{F}_\theta(x) = \begin{cases} \frac{n-1}{n} \frac{x}{X_{(n)}}, & 0 < x < X_{(n)} \\ 1, & x \ge X_{(n)} \\ 0, & \text{otherwise.} \end{cases}$$

Consequently, the detection statistic of this type becomes

$$\begin{aligned}
\tilde{D}_n &= \sup_x |F_n(x) - \tilde{F}_\theta(x)| \\
&= \sup_{0 \le x \le X_{(n)}} |n^{-1} \sum_{j=1}^{n-1} \in [\frac{x}{X_{(n)}} - \frac{X_{(j)}}{X_{(n)}}] - (\frac{n-1}{n}) \frac{x}{X_{(n)}}| \\
&= \sup_{0 \le u < 1} (\frac{n-1}{n}) |(n-1)^{-1} \sum_{j=1}^{n-1} \in [(u - u_{(j)}) - u].
\end{aligned}$$

Thus, we note that the detection statistic $\tilde{D}_n \overset{\mathcal{L}}{=} (\frac{n-1}{n}) D_{n-1}$, the Kolmogorov-Smirnov statistic with size $n-1$, for all $n \ge 2$ provided $H_o : \mathcal{L} \in \Omega_1$ holds.

Decision Rule. For detecting $PN : \mathcal{L} \in \Omega_1$ against $N + S : \mathcal{L} \notin \Omega_1$; decide $N + S$, if and only if $\tilde{D}_n > \tilde{d}$, where $P\{\tilde{D}_n \le \tilde{d}\} = 1 - \alpha$, and α is PFA.

Srinivasan (1970) computed critical values for the normal and exponential cases, and Choi (1981) for the uniform case. For critical values in connection with detection statistic $\tilde{D}_n$, one can use the standard Kolmogorov-Smirnov table through the relation $\tilde{D}_n = (\frac{n-1}{n}) D_{n-1}$.

4.4 A DETECTION STATISTIC BASED ON MAXIMAL STATISTICAL NOISE

We have already defined a MSN statistic, $N(\mathbf{Z})$, earlier. For the Ω_1 case, that is for $PN : \mathcal{L} \in \Omega_1$ the maximal statistical noise statistic is

$$\begin{aligned}(\mathbf{V}^*, \mathbf{R}) &= (V_1^*, V_2^*, \ldots, V_{n-1}^*; R_1, \ldots, R_n) \\ &= (\frac{X_{(1)}}{X_{(n)}}, \ldots, \frac{X_{(n-1)}}{X_{(n)}}; R(X_1), \ldots, R(X_n)),\end{aligned}$$

where $R(X_j)$ denotes the rank of X_j. Note that in PN situation here we have a NPDF structure for the underlying statistic which leads us to the natural detection statistic candidate $(\mathbf{V}^*, \mathbf{R})$. Furthermore, we shall see that if a detection procedure does not involve the parameter in the model under consideration, then it is solely based on MSN statistic. This motivates one to define a third detection statistic D_n^* as follows:

$$D_n^* = \sup_{0 \le u \le 1} |(n-1)^{-1} \sum_{j=1}^{n-1} \in (u - u_{(j)}) - u|$$

where $u = x/X_{(n)}$ and $U_{(j)} = X_{(j)}/X_{(n)}$. Obviously, D_n^* is the statistic which is based on MSN, $(\mathbf{V}^*, \mathbf{R})$. The statistic D_n^* has the Kolmogorov-Smirnov distribution for a sample of size $n-1$, that is, $D_n^* = D_{n-1}$ under $PN : \mathcal{L} \in \Omega_1$. For critical values one can use the standard K-S table with the sample size $n-1$.

Decision Rule. For detection $PN_2 : \mathcal{L} \in \Omega_1$ versus $N + S : \mathcal{L} \notin \Omega_1$; decide $N + S$, if and only if, $D_n^* = D_{n-1} > d$ where $P\{D_{n-1} \le d\} = 1 - \alpha$ and $\alpha = P$ (False Alarm).

In essence, the detection procedure structure for this Model : PN can be summarized by the following result.

Theorem 4.2. Under $H_o : \mathcal{L} \in \Omega_1$;

(*i*) $\quad D_n^* = D_{n-1}, \quad$ for $\quad n \ge 2,$

(*ii*) $\quad \tilde{D}_n = (\frac{n-1}{n}) D_{n-1}, \quad$ for all $\quad n \ge 2,$

(*iii*) $\quad \hat{D}_n = \sup_{0 \le u < 1} |(\frac{n-1}{n}) F_{n-1}(u) - u|, \quad$ where $\quad F_{n-1}(.)$

is the sample cpf based on $u_1, \ldots, u_{n-1}$ which are i.i.d. $U(0,1)$.
Since $\frac{n-1}{n} \to 1$ as $n \to \infty$, it follows that:

Theorem 4.3. Asymptotically (as $n \to \infty$) the statistics $D^*, \tilde{D}$ and $\hat{D}$ have the same distribution as the usual Kolmogorov-Smirnov statistic D_∞.

4.5 DETECTION PROCEDURE FOR TWO SAMPLE PROBLEMS

The received basic data (BD) set is

$$Z = (X_1, \ldots, X_m; Y_{m+1}, \ldots, Y_{m+n}) = (Z_1, \ldots, Z_N).$$

The detection problem for this situation is

$$H_o(PN); \mathcal{L}_1 = \mathcal{L}_2(\mathcal{L}_i \in \Omega_1, \quad i = 1, 2),$$

$$H_o(N+S) : \mathcal{L}_1 \neq \mathcal{L}_2(\mathcal{L}_i \in \Omega_1, \quad i = 1, 2).$$

For this case the natural detection statistic will be a NPDF statistic and should depend on the MSN $N(\mathbf{Z}) = (\mathbf{V}^*, \mathbf{R})$, where

$$\begin{aligned}(\mathbf{V}^*, \mathbf{R}) &= (V_1^*, V_2^*, \ldots, V_{N-1}^*; R_1, R_2, \ldots, R_N) \\ &= \{\frac{Z_{(1)}}{Z_{(N)}}, \frac{Z_{(2)}}{Z_{(N)}}, \ldots, \frac{Z_{(N-1)}}{Z_{(N)}}; R(Z_1), \ldots, R(Z_N)\}\end{aligned}$$

Under H_o, $\mathbf{Z}_{(\cdot)} = (\mathbf{Z}_{(1)}, \ldots, \mathbf{Z}_{(N)})$ and $\mathbf{R}$ are independent since one can write their joint p.d.f. as equal to the product of the marginal p.d.fs., that is,

$$f_{\mathbf{Z}(\cdot)}(\mathbf{z}_{(\cdot)}, \mathbf{r}) = f_{\mathbf{Z}(\cdot)}(\mathbf{z}_{(\cdot)}).\mathbf{f}_{\mathbf{R}}(\mathbf{r})$$

Note, that under H_o:

$$f_{Z(\cdot)}(\mathbf{z}_{(\cdot)}) = N! \sum_{j-1}^{N} f_{Z_1}(z_{(j)}) I_A$$

where $A\{(z_{(1)}, \ldots, z_{(N)}) : 0 \leq z_{(1)} < z_{(2)} < \ldots < z_{(N)} \leq \theta = \theta_1 = \theta_2\}$; and $\mathbf{R}$ is distributed uniformly over the permutation set $\{S_N(1, 2, \ldots, N)\}$. Furthermore, $\mathbf{Z}_{(\mathbf{N})}$ and $\mathbf{V}^*$ are independent; and so are $\mathbf{R}$ $(\mathbf{Z})$ and $\mathbf{V}^*$.

In this case one can develop the likelihood detection procedure, and show that the resulting detection statistic is based on the maximal statistical noise (MSN), $N(\mathbf{Z}) = (\mathbf{V}^*, \mathbf{R})$. The details are somewhat intricate and can be found in Choi-Bell-Ahmad- Park (1989, pp.17-28).

5. Statistical Signal Detection for Stochastic Processes with Stationary Independent Symmetric Increments

We shall denote this family by Ω(SISI) or Ω_2 as defined earlier. This class includes stochastic processes with probability laws such as Wiener, Compound Poisson, random walk, and symmetric stable laws, among others. For a general treatment of this family see Bell and Ahmad (1983). In particular, there point estimation and confidence bounds for $F_{\mathcal{L}}$ with $\mathcal{L}$ in Ω(SISI) are investiaged. In essence, it turns out that when the underlying distribution function is symmetric about zero (and this is no real restriction for many applications), then one can achieve a confidence region having 50% less interval length as compared to the usual Kolmogorov-Smirnov type regions.

5.1 BASIC TERMINOLOGY, SUFFICIENT STATISTICS AND NONPARAMETRIC PROPERTY

Let $\mathbf{Z} = (\mathbf{X_1}, \ldots, \mathbf{X_n})$ be the vector of increments that is interarrival times. Now, let $S_1(\mathbf{Z}) = (|\mathbf{X}|(\mathbf{1}), \ldots, |\mathbf{X}|(\mathbf{n}))$, be a vector of ordered absolute values; let $N_1(\mathbf{Z}) = (\epsilon, \mathbf{R}^*)$, where

$$\epsilon = [\epsilon(X_1), \ldots, \epsilon(X_n)], \quad \text{and} \quad \mathbf{R}^* = [R(|X_1|), \ldots, R(|X_n|)],$$

where $\epsilon(u) = 1$ or 0 whether $u \geq 0$ or $u < 0$, and $R(|X_j|)$ denotes the rank of $|X_j|$. Finally, let $\delta_1(\mathbf{Z}) = [\mathbf{S_1(Z), N_1(Z)}]$.

One can now easily establish the following result.

Theorem 5.1 (i) $S_1(\mathbf{Z})$ is the MSS for Ω(SISI), and it is complete. (ii) $S_1(\mathbf{Z})$, ϵ and R^* are mutually independent. (iii) $\delta_1(\cdot)$ is 1-1 a.e. and is the BDT for Ω(SISI), and (iv) $N_1(\cdot)$ is MSN for Ω.

The general rule for employing these statistics in signal detection is as follows: (A) for situations involving a specific law (e.g. PN : $\mathcal{L} = \mathcal{L}_o$) use a PDF statistic based on the MSS; (B) for cases involving the underlying structure of the family (e.g., PN : $\mathcal{L} \in \Omega$ or PN : $\mathcal{L}_1 = \mathcal{L}_2$) employ an NPDF statistic based on the MSN. This overall principle will be followed in the sequel. However, in order to make efficient use of this principle, one should consider several alternate versions of the MSS and the MSN.

5.2 DISTRIBUTIONS, PERMUTATIONS AND ALTERNATE STATISTICS

The only versions of MSS and MSN, known to the authors, are those related to maximal invariants or permutation statistics. The statistics $S_1(\cdot)$ and $N_1(\cdot)$ as defined above are maximal invariants. The set of permutations of interest here is the Sign-Time group, S_n^*.

Definition 5.1. (a) $S_n^* = \{$all permutations of coordinates of $\mathbf{Z}$ and changes of signs for coordinates.$\}$. (b) $S_n^*(\mathbf{Z}) = \{\pi(\mathbf{Z}) : \pi \in \mathbf{S_n^*}\}$ is called the orbit of $\mathbf{Z}$.

Theorem 5.2. (i) S_n^* is a wreath product group and its order of cardinality is $(n!)(2^n)$. (ii) The orbit $S_n^*(\mathbf{Z})$ is a MSS and it contains $(n!)(2^n)$ points for almost every $\mathbf{Z}$.

Definition 5.2. (a) A measurable real-valued function $h(\cdot)$ is called a Bell-Pitman (BP) function wrt a set, S^*, of permutations, and a family Ω^*, of stochastic processes laws if $P\{h(\mathbf{Z}) = \mathbf{h}(\pi\mathbf{Z})|\mathcal{L}\} = \mathbf{0}$ unless $\mathbf{Z} = \pi\mathbf{Z}$, for all $\pi \in S^*$ and all $\mathcal{L} \in \Omega^*$. (ii) A function defined by $R(h(\mathbf{Z})) = \sum_{\pi \in \mathbf{S^*}} \in \{\mathbf{h(Z)} - \mathbf{h}(\pi\mathbf{Z})\}$ is called a BP statistic wrt S^* and Ω^*.

Example 5.1. Let $\mathbf{Z} = (\mathbf{X_1}, \ldots, \mathbf{X_n})$, and set $h(\mathbf{Z}) = \sum_{\mathbf{j=1}}^{\mathbf{n}} \mathbf{jX_j}$. Then $h(\mathbf{Z})$ and $R(h(\mathbf{Z}))$ are, respectively, a BP function and a BP statistic wrt S_n^* and Ω(SISI).

Now one can establish the following useful result.

Theorem 5.3. (MSN Theorem). (i) If $h(\cdot)$ is BP function wrt S_n^* and Ω(SISI), then $R(h(\cdot))$ is (a version of) MSN. (ii) T^* is NPDF wrt Ω(SISI), if and only if, there exists a measurable function $W(\cdot)$ and BP function h^* such that $T^*(\mathbf{Z}) = \mathbf{W(R(h^*(Z)))}$. (iii) There are an infinite number of (mutually equivalent) versions of MSN. In particular, $N_1(\mathbf{Z})$, of section 3, can be expressed in terms of at least one BP statistic. Furthermore, if $\mathcal{L} \in \Omega$(SISI), $P\{R(h(\mathbf{Z})) = \mathbf{r}\} = [n!(2^n)]^{-1}$ for $1 \leq r \leq (n!)2^n$.

If F satisfies $F(x) + F(-x) = 1$ for all x, let $G_F(z)$ be defined as zero for $z < 0$ and $2F(z) - 1$ for $z > 0$. Set $G_n^*(z) = n^{-1} \sum_{j=1}^{n} \in (z - |X|(j))$. Now, we have three versions of MSS: (i) $S_1(\mathbf{Z})$, (ii) $S_2^*(\mathbf{Z})$, and (iii) $S_3(\mathbf{Z}) = \mathbf{G_n^*}(\cdot)$. Similarly, for MSN we have (i) $N_1(\mathbf{Z})$ and (ii) $N_h(\mathbf{Z}) = \mathbf{R(h(Z))}$, where $h(\cdot)$ is a BP function wrt S_n^* and Ω(SISI). The following theorem, which follows from Birnbaum and Rubin (1954) and Bell (1964a,b), will be used in constructing all procedures for one of the signal detection models in the sequel.

Theorem 5.4. (PDF Theorem). Each statistic of the form: $\phi[G_F(|X|(1)), \ldots, G_F(|X|(n))]$ or $\psi^*(G_n^*(\cdot))$ is PDF wrt Ω(SISI).

One final statistical tool to be used later is the technique which is a slight generalization of the method proposed by Durbin (1961) and Bell and Doksum (1965); and is employed to avoid certain distributional problems. Let $\delta(\mathbf{Z}) = [\mathbf{S(Z), N(Z)}]$, where $\delta(\cdot)$, $S(\mathbf{Z})$ and $N(\mathbf{Z})$ are, respectively, the BDT, MSS and MSN of a family Ω^* of stochastic process laws. Let Z be data governed by a law $\mathcal{L}$ (usually unknown) in Ω^*, and let $\mathbf{Y}$ be generated by a law $\mathcal{L}_o$ (known) in Ω^* and independent of $\mathbf{Z}$.

Definition 5.3. (a) $\mathbf{Y}$ is called extraneous statistical noise (ESN). (b) $\mathbf{Y}^* = \delta^{-1}[S(\mathbf{Y}), N(\mathbf{Z})]$ is called randomized statistical noise (RSN).

Example 5.2. Let $\mathbf{Z} = (\mathbf{Z_1}, \ldots, \mathbf{Z_n})$ be the increments of data governed by a law $\mathcal{L}_F$ in Ω(SISI). Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be the increments of a Wiener process with $\sigma = .25$. Then $Z_1, \ldots, Z_n$ are i.i.d. with unknown $F(\cdot)$; but $Y_1, \ldots, Y_n$ are i.i.d. with Φ. From $S_1(\mathbf{Z})$, $N_1(\mathbf{Z})$ and $\delta(\mathbf{Z})$ one can form: $\delta^{-1}[|Y|(1), \ldots, (|Y|(n);\ \epsilon(\mathbf{Z});\ \mathbf{R}^*(\mathbf{Z}))] = [Y_1^*, \ldots, Y_n^*]$, where $Y_j^* = [2 \in (Z_j) - 1]/Y_j|(R(|Z_j|))$. This means each Z is replaced by that Y, which has the same absolute value rank and same sign. Then it can be proved that $Y_1^*, \ldots, Y_n^*$ are i.i.d. N(0,1). These ideas are formalized in the result below.

Theorem 5.5. (Randomized Noise Theorem). Let $\mathbf{Z} = (\mathbf{X_1}, \ldots, \mathbf{X_n})$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be independent, and be generated by laws $\mathcal{L}$ and $\mathcal{L}^*$ in Ω^*, respectively.

Let $\delta(.)$, $S(\mathbf{Z})$, $N(\mathbf{Z})$ be respectively the BDT, MSS and MSN for Ω^*. If $\mathbf{Y}^* = (Y_1^*, \ldots, Y_n^*) = \delta^{-1}[S(\mathbf{Y}), N(\mathbf{Z})]$, then Y^* has the same law as Y, that is $Y^* \overset{\mathcal{L}}{=} Y \sim \mathcal{L}^* \in \Omega^*$.

Thus, the distribution of the ESN, $\mathbf{Y}$, has been imposed on the data while the MSN of the data is preserved. Now, we are able to treat the pertinent inference problems.

5.3 GOODNESS-OF-FIT DETECTORS FOR Ω_2

The situation to be treated here is: PN : $\mathcal{L} = \mathcal{L}_o \in \Omega$(SISI) against N+S : $\mathcal{L} \neq \mathcal{L}_o$. One notes that the law $\mathcal{L}_o$ and the sampling spacing Δ, uniquely determine the common distribution F_o, of the independent symmetric increments $X_j = Z(j\Delta) - Z((j-1)\Delta)$, for $j = 1, 2, \ldots, n$ where $Z(0) = 0$.

Example 5.3. From Bell and Ahmad (1983) in example 2.1 $\mathcal{L}_o$ is such that $F_o = N(0, \sigma^2\Delta)$. In example 2.2, F_o is the distribution function with characteristic function $\phi(t) = \exp\{-\lambda\delta [1 - \phi^*(t)]\}$ with $\phi^*(t) = (\theta t)^{-1} \sin t\theta$. In example, 2.3, if $\Delta = 3$, then F_o is that distribution whose characteristic function is $\phi(t) = (1 + t^2/\lambda)^{-3}$.

For a given constant sampling spacing λ, one can rephrase the detection problem as: $PN^* : F = F_o \in \Omega^*$ versus $N + S : F \neq F_o$ where $F_o(x) + F_o(-x) = 1$ for all x. In this case one should base the decision rule on the MSS $S_1(\mathbf{Z})$ and a PDF statistic using the distribution G_{F_o}, where $G_F(x) = 2F(x) - 1, x > 0$. In view of the Birnbaum and Rubin (1954) and Bell (1964b) ideas, one should consider decisions based on statistics of the form $\psi[G_{F_o}(|X|(1)), \ldots, G_{F_o}(|X|(n))]$. Some of these and other decision rules are as follows.

Decision Rule 5.1. Decide $N + S$ iff $\sup_z |G_n^*(z) - G_{F_o}(z)| > d^*$, where $G_n^*(z) = n^{-1} \sum_{j=1}^n \in (z - |X|(j))$ and $d^* = d^*(\alpha, n)$ is chosen so as to achieve PFA,α.

Decision Rule 5.2. Decide $N + S$ iff $\int |G_n^*(z) - G_{F_o}(z)|^2 dG_{F_o}(z) > w^* = w^*(\alpha, n)$.

Of course, several other goodness-of-fit statistics could have been used, see Bell (1964b; Model I). If one knows that the signal has the effect of yielding a specified distribution, H, for $\mathbf{X}^* = (X_1^*, \ldots, X_k^*)$, where $X_j^* = |X|(j)$, then one would prefer a decision rule of the form below.

Decision Rule 5.3. Decide $N + S$ iff $h(x_1^*, \ldots, x_k^*) > c^* \prod_{j=1}^k g_{F_o}(x_j^*)$, where $c^* = c(\alpha, k)$ and $g_{F_o}(\cdot)$ and $h(\cdot)$ are the appropriate densities.

5.4 CLASS-FIT DETECTORS FOR Ω_2

Here one has the detection problem $PN : \mathcal{L} \in \Omega(\text{SISI})$ versus $N + S : \mathcal{L} \notin \Omega(\text{SISI})$. Since this problem does not concern a specific value of the parameter (or law), one is led to NPDF statistics and MSN as defined in sections 3 and 4.

Decision Rule 5.4. (Sign Detector). Decide $N + S$, if and only if, $T_n^*(Z) = \sum_{j=1}^n \in (X_j) \geq c_1$ or $\leq c_2$, where $c_j = c_j(\alpha, n)$.

This rule makes no use of the rank vector in $N_1(\mathbf{Z}) = (\epsilon, \mathbf{R}^*)$. In order to make use of all of the components of $S_1(\mathbf{Z})$, one introduces the Wilcoxon one-sample statistic.

Decision Rule 5.5. (Wilcoxon Detector). Decide $N + S$, if and only if, $W_n^*(Z) = \sum_{j=1}^n \in ((X_j)R(|X_j|)) \leq k_1$ or $\geq k_2$, where $k_j = k_j(\alpha, n)$.

The statistics employed in the two decision rules above involve statistics which are well-tabulated. However, these statistics both have discrete distribution, and it is necessary to turn to randomized procedures if one wishes to achieve certain particular PFA's. One such procedure is based on ESN, $\mathbf{Y} = (Y_1, \ldots, Y_n)$, independent of the data and generated by a known law $\mathcal{L}$ in $\Omega(\text{SISI})$. Let $Y_1, \ldots, Y_n$ be i.i.d. $N(0, 1)$. This corresponds to data $\mathbf{Z} = \{\mathbf{Z}(\boldsymbol{\Delta}), \mathbf{Z}(\mathbf{2\Delta}), \ldots, \mathbf{Z}(\mathbf{n\Delta})\}$ from Wiener- Levy process satisfying $\mu(t) \equiv 0$ and $\sigma^2\Delta = 1$. When the process $\{Z(t)\}$ is independent of the data so are the $Y's$.

From previous development, one knows that the BDT, $\delta(\cdot)$, is such that $\delta(\mathbf{Z}) = [\mathbf{S_1(Z), N_1(Z)}]$, where $S_1(\mathbf{Z}) = [|\mathbf{X}|(\mathbf{1}), \ldots, |\mathbf{X}|(\mathbf{n})]$ and $N_1(\mathbf{Z}) = [\epsilon(\mathbf{X_1}), \ldots, \epsilon(\mathbf{X_n}); \mathbf{R}(|\mathbf{X_1}|), \ldots, \mathbf{R}(|\mathbf{X_n}|)]$. In order to apply the randomized noise theorem (Theorem 5.5), one forms $\mathbf{Y}^* = (Y_1^*, \ldots, Y_n^*) = \delta^{-1}[S_1(\mathbf{Y}^*), N_1(\mathbf{Z})]$. In this case $Y_j^* = [2 \in (X_j) - 1]|Y_j|R(|X_j|))$. This means $N_1(\mathbf{Z}) = \mathbf{N_1}(\mathbf{Y}^*)$. By Theorem 5.5, $Y_1^*, \ldots, Y_n^*$ are i.i.d. $N(0, 1)$ under $PN : \mathcal{L} \in \Omega(\text{SISI})$. One useful decision rule based on the Kolmogorov-Smirnov statistics is then as below.

Decision Rule 5.6. Decide $N+S$, if and only if, $\sup_z |\frac{1}{n}\sum_{j=1}^{n} \in (z - Y_j^*) - \Phi(z)| > d^*$ where $d^* = d^*(\alpha, n)$.

Again one notes that in the above decision rule one could have used any goodness-of-fit statistic. A different type of detection procedure is developed if one employs a different version of the MSN.

Let S_n^* (see Definition 5.1) be the set of sign-time permutations of coordinates of $\mathbf{Z} = (\mathbf{X_1}, \ldots, \mathbf{X_n})$, and let $h(\mathbf{Z}) = \sum_{\mathbf{j=1}}^{\mathbf{n}} \mathbf{jX_j}$. Then one has the following result.

Theorem 5.6. (a) The function $h(\cdot)$ is a BP function wrt Ω(SISI) and S_n^*. (b) $R(h(\mathbf{Z}))$, as in Definition 4.2, is MSN. (c) Under PN, $R(h(\mathbf{Z})) \sim \mathbf{DU}\{\mathbf{1}, \mathbf{2}, \ldots, \mathbf{k}^*\}$ i.e., has a discrete uniform distribution over the integers $\{1, 2, \ldots, k^*\}$, where $k^* = n!2^n$.

Decision Rule 5.7. Decide $N+S$, if and only if, $R(h(\mathbf{Z})) \leq b_1$ or $\geq b_2$, where $b_1 = n!2^n(\frac{\alpha}{2})$ and $b_2 = 1 + (1 - \frac{\alpha}{2})n!2^n$. Of course, one could construct an analogous decision for each BP function.

5.5 TWO-SAMPLE DETECTORS FOR Ω_2

The detection problem here becomes $PN : \mathcal{L}_1 = \mathcal{L}_2$ versus $N+S : \mathcal{L}_1 \neq \mathcal{L}_2$ where $\mathcal{L}_i \in \Omega$(SISI), $i = 1, 2$. Consider the situation where $\mathbf{X} = (X_1, \ldots, X_m)$ is generated by $\mathcal{L}_1$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is generated by $\mathcal{L}_2$ and is independent of $\mathbf{X}$. Under PN, $\mathbf{Z} = (\mathbf{X}, \mathbf{Y}) = (\mathbf{Z_1}, \ldots, \mathbf{z_{m+n}})$ is distributed as a random sample from an unknown $F(\cdot)$, symmetric w.r.t. zero. Since no specific value of a paramter or law is involved, one needs a NPDF statistic. Such a statistic must be based on some version of the MSN, e.g., $S_1(\mathbf{Z}) = (\epsilon, \mathbf{R}^*)$.

Decision Rule 5.8. Decide $N+S$, if and only if, $W = \sum_{j=1}^{n} R(|Z_j|) \geq b_1$ or $\leq b_2$ where $b_j = b_j(\alpha, m, n)$.

Decision Rule 5.9. Decide $N+S$, if and only if,

$$\sup_z \left| \frac{1}{m}\sum_{j=1}^{m} \in (z - |X|(j)) - \frac{1}{n}\sum_{j=1}^{n} \in (z - |Y|(j)) \right| > d^* = d(\alpha, m, n).$$

Decision Rule 5.10. Decide $N+S$, if and only if,

$$\frac{1}{m}\sum_{j=1}^{m} V(R|X_j|) - \frac{1}{n}\sum_{j=1}^{n} V(R|Y_j|) \geq a_1 \text{ or } \leq a_2,$$

where $V_1, \ldots, V_{m+n}$ are i.i.d. Φ and independent of data.

The first and second rules above correspond to the Wilcoxon-Rank-Sum statistic, and the two-sample Kolmogorov-Smirnov statistic, respectively. The last decision rule is based on Theorem 5.5. In essence the techniques here, although concerned with Ω(SISI), are closely related to many of the usual two-sample detectors see Bell (1964a, Model II). Two more detectors for this case are discussed in Bell and Ahmad (1982). One is a modified two-sample detector, which turns out to be a most powerful invariant detector against specific $N+S$ alternatives. The other is a chi-squared type detector, and employs the symmetry

of the underlying process. For various decision rules developed in the last three sections, illustrative numerical examples are given in Bell and Ahmad (1982).

Acknowledgement. The author gratefully acknowledges partial financial support form NATO Advanced Study Institute under the Directorship of Professor G. Roussas.

REFERENCES

Bell, C.B. (1964a). ***Automatic Distribution- Free Statistical Signal Detection.*** Research Report US Navy, Electronics Laboratory, San Diego.

Bell, C.B. (1964b). Some Basic theorems for distribution-free statistics. ***Ann. Math. Statist.***, 35, 150-156.

Bell, C.B., Doksum, K. (1965). Some new distribution-free statistics. ***Ann. Math. Statist***, 36, 203-214.

Bell, C.B., Ahmad, R. (1982). Signal detection for stochastic processes with stationary independent symmetric increments. ***Report, Series in Statistics and Biostatistics***, pp.44, Department of Mathematical Sciences, San Diego State University, California, USA.

Bell, C.B., Ahmad R. (1983). Statistical signal detection for stochastic processes with stationary independent symmetric increments. In: ***Developments in Statistics and Its Applications*** (Eds. A.M. Abouammoh, E.A. El-Neweihi, E-E.A. Aly, M.A. Al-Osh), 459-477, King Saud University Press, Riyadh, Saudi Arabia.

Bell, C.B., Ahmad, R., Park, C.J., Lui, R. (1987). Signal detection for Pareto renewal processes. ***Pakistan J. Statist.***, A, 3(2), 1-53.

Birnbaum, Z.W., Rubin, H. (1954). On distribution- free statistics. ***Ann. Math. Statist.***, 25, 593-598.

Choi, Y.J. (1981). ***Kolmogorov-Smirnov Test with Nuisance Parameters in the Uniform Case.*** MS Thesis, University of Washington.

Choi, Y.J., Bell, C.B., Ahmad, R., Park, C.J. (1989). Signal detection for uniform renewal processes. ***Pakistan J. Statist.***, A, 5(1), 1-49.

Davidson, A.C., Smith, R.L. (1990). Models for exceedances over high thresholds. ***J. Roy. Statist.***, B, 52(3), 393-442.

Durbin, J. (1961). Some methods of constructing exact tests. ***Biometrika***, 48, 41-55.

Durbin, J. (1973). ***Distribution Theory of Tests Based on the Sample Distribution Function.*** SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia.

Fisk, P.R. (1961). The graduation of income distributions. ***Econometrica,*** 29, 171-185.

Johnson, N.L., Kotz, S. (1970). ***Continuous Univariate Distributions - I.*** Wiley.

Levy, P. (1925). ***Cacul des Probabilities.*** Gauthier-Villars, Paris.

Lieberman, G.J., Resnikoff, G.J. (1955). Sampling plans for inspection by variables. ***J. Amer. Statist. Assoc.***, 55, 144-147.

Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. ***J. Amer. Statist. Assoc.***, 62, 399-402.

Lilliefors, H.W. (1969). On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. ***J. Amer. Statist. Assoc.***, 64, 387-389.

Mandelbrot, B. (1960). The Pareto-Levy law and the distribution of income. ***Inter. Econ. Rev.***, 1, 79- 106.

Mandelbrot, B. (1963). The variation of some other speculative prices. *J. Business,* University of Chicago, USA, 36, 394-419.

Pareto, V. (1897). *Cours d'Economie Politique*, Rouge and Cie, Lausanne and Paris.

Pigou, A.C. (1932). *The Economics of Welfare.* McMillan, London.

Renyi, A. (1953). On the theory of order statistics. *Acta Math. Acad. Sci. Hung;* 4, 191-231.

Srinivasan, R. (1970). An approach to testing the goodness-of-fit of incompletely specified distributions. *Biometrika*, 57, 605-611.

FUNCTIONAL IDENTIFICATION IN NONLINEAR TIME SERIES

Bjørn Auestad and Dag Tjøstheim

Department of Mathematics
University of Bergen
5007 Bergen
NORWAY

ABSTRACT. We study functional identification of nonlinear time series models of type $X_t = g(X_{t-1}, \ldots, X_{t-p}) + e_t$. Nonparametric estimation is used to estimate the function g and to single out the most significant lags in a stepwise analysis. We also define projections at a fixed lag of the model and show how this concept can be used to identify the model among familiar nonlinear time series classes. The case of additive models is given special emphasis. The results are illustrated by simulations and by a real data example. Some asymptotics are derived, partly by heuristic means, and the bootstrap is introduced in one of the examples. We indicate the possibility of using this approach as a first step in a general empirical modelling of time series.

1 Introduction

In the last decade a multitude of nonlinear time series models have been suggested. Among them are threshold and exponential autoregressive models, bilinear models, ARCH models, RCA and doubly stochastic models. A very comprehensive survey is given in the recent book by Tong (1990).

Much of the research has been concentrated on studying specific properties such as stationarity, consistency and distributional properties of estimates. Moreover, these models have been fitted to selected data sets. Much less work has gone into the problem of identifying the model, i.e. trying to decide which class of models a given observed series belongs to. Directly related to this is the problem of constructing good indicators — in particular graphical ones — that would make it possible to distinguish between linear and various nonlinear models.

In classical linear modelling the autocorrelation function and partial autocorrelation function are used to identify the type of linear model at hand; possibly in conjunction with an Akaike-type order determination criterion. Similar identification attempts for nonlinear series have been fragmentary. Plots of the conditional mean and the conditional variance are perhaps the most natural tools to use in such a context, and such plots have been emphasized in the books by Tong (1983,1990). A somewhat different approach, using graphics based on state space models, has been taken by Haggan et al (1984).

The main purpose of this paper is to try to make a more systematic approach to the identification problem. Work along these lines was started in Auestad and Tjøstheim (1990),

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 493–507.
© 1991 *Kluwer Academic Publishers.*

but it was to a large extent limited to first order models. In this paper we look at the fairly general model

$$X_t = g(X_{t-1}, \ldots, X_{t-p}) + e_t, \tag{1.1}$$

where $\{X_t\}$ is the observed process and $\{e_t\}$ a residual process. This is a nonlinear and nonparametric generalization of the familiar linear autoregressive model.

The task we look at is to identify p, determine significant lags and the functional form of g. The analysis may stop with an estimate of g, or, if g can be recognized as one of the known parametric models (linear or nonlinear), it may be reestimated in parametric form.

The problem we pose is quite formidable and several of our suggestions are of a tentative nature. Throughout we use nonparametric methods, but at this stage there are a number of unsolved problems connected to choice of optimal smoothing and asymptotic properties. However, as evaluated by simulations, our methods seem to work fairly well granted that the sample size is not too small.

2 Unknown order. A stepwise procedure to determine significant lags

In the following we assume that $\{X_t\}$ is given by (1.1) with $\{e_t\}$ i.i.d., and e_t independent of $\{X_s, s < t\}$ for $t \in N$ or $t \in Z$, where N is the positive integers and Z the set of all integers. For the case $t \in N$, an initial condition $\{X_{-p+1}, \ldots, X_0\}$ has to be specified. Moreover, we assume that g is such that a stationary solution $\{X_t\}$ of (1.1) exists. Some sufficient conditions for this and the stronger property of geometric ergodicity are given in Auestad and Tjøstheim (1990). Geometric ergodicity is of interest since it means that $\{X_t\}$ is strongly mixing with an exponential mixing rate, and this in turn means that asymptotic distributions can in general be found for estimators based on $\{X_t\}$.

One of the main difficulties of the identification problem is the <u>combined</u> identification of the order p and the shape of g.

If p is known then g is given by

$$g(x_1, \ldots, x_p) = E(X_t \mid X_{t-1} = x_1, \ldots, X_{t-p} = x_p).$$

An obvious and well known idea is to estimate g nonparametrically by estimating the conditional expectation. Using kernel estimators the conditional mean

$$m^{(1,\ldots,j)}(x_1, \ldots, x_j) = E(X_t \mid X_{t-1} = x_1, \ldots, X_{t-j} = x_j) \tag{2.1}$$

is estimated from observations $X_1, \ldots, X_n$ by

$$m_n^{(1,\ldots,j)}(x_1, \ldots, x_j) = \frac{\dfrac{1}{n-j} \displaystyle\sum_{t=j}^{n-1} X_{t+1} \prod_{i=1}^{j} K\{h_n^{-1}(x_i - X_{t+1-i})\}}{\dfrac{1}{n-j+1} \displaystyle\sum_{t=j}^{n} \prod_{i=1}^{j} K\{h_n^{-1}(x_i - X_{t+1-i})\}}, \tag{2.2}$$

where, in order to simplify, we have used a product kernel, and where in our simulations K was taken to be standard Gaussian, i.e. $K(x) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}x^2)$. The parameter

$h_n = h_n(j)$ refers to the bandwidth. It determines the degree of smoothing. We have used the same bandwidth in each direction.

Under conditions of stationarity, geometric ergodicity of $\{X_t\}$, smoothness of g and the density function s of $\{e_t\}$ and under mild conditions on h_n given in Robinson (1983) (see also Auestad and Tjøstheim 1990) we have $m_n^{(1,\ldots,j)}(x_1,\ldots,x_j) \xrightarrow{P} m^{(1,\ldots,j)}(x_1,\ldots,x_j)$ and $m_n^{(1,\ldots,j)}(x_1,\ldots,x_j)$ is asymptotically normal such that

$$\{n(h_n)^j\}^{\frac{1}{2}} \left\{ m_n^{(1,\ldots,j)}(x_1,\ldots,x_j) - m^{(1,\ldots,j)}(x_1,\ldots,x_j) \right\}$$

$$\xrightarrow{d} \mathcal{N}\left(0, \frac{v(x_1,\ldots,x_j)}{f(x_1,\ldots,x_j)} \left\{ \int K^2(z)dz \right\}^j \right), \tag{2.3}$$

where $v(x_1,\ldots,x_j) = \mathrm{var}(X_t \mid X_{t-1} = x_1,\ldots,X_{t-j} = x_j)$ and $f(x_1,\ldots,x_j)$ is the joint density function of $(X_{t-1},\ldots,X_{t-j})$. It should be noted that the asymptotic results holds irrespective of whether $j < p$ or $j \geq p$.

The difficult problem which arises in practice is that it is not clear which order p should be chosen for a given data set. In the linear case this obstacle can be attacked through an Akaike-type criterion. In the nonparametric situation, as far as we can tell, the problem has not been much discussed in the literature. For regressions a somewhat similar problem has been analysed by Collomb (1983) and Härdle (1984).

Incidentally, in the modelling of real data the notion of "true order" is of questionable value, and the order produced by the algorithm should only be taken as an indication of a sensible order to use. If the datalength is increased, it may in some cases be reasonable to increase the order.

A tentative solution in the general nonparametric situation was presented in Auestad and Tjøstheim (1990). We will briefly review this procedure as it will be used as a basis in the current paper for selecting the most significant lags. At the outset it should be emphasized that the theoretical difficulties are large, and very little has been obtained in the direction of rigorous asymptotic results. However, the procedure seems to work fairly well on the models tested by us.

We will assume that $\{X_t\}$ is given as a Markov chain of order p as in (1.1) with $\{e_t\}$ being i.i.d., and $E(e_t) = 0$ and $E(e_t^2) = \sigma^2$. It is reasonable to base our work on estimated residuals of order j:

$$\hat{\sigma}_j^2 = \frac{1}{n-j} \sum_{t=j+1}^{n} \{X_t - m_n^{(1,\ldots,j)}(X_{t-1},\ldots,X_{t-j})\}^2$$

with $m_n^{(1,\ldots,j)}$ given as in (2.2). The obvious idea is that $\hat{\sigma}_j^2$ should flatten out for $j \geq p$, since then $\sigma_j^2 \stackrel{\text{def}}{=} E[\{X_t - m^{(1,\ldots,j)}(X_{t-1},\ldots,X_{t-j})\}^2] = \sigma^2$.

The estimated $m^{(1,\ldots,j)}$ and in turn $\hat{\sigma}_j^2$ is critically dependent of the bandwidth h_n. For a Gaussian kernel it can be shown that as $h_n \to 0$, then $\hat{\sigma}_j^2 \to 0$, while as $h_n \to \infty$ (smoothing everything flat), $\hat{\sigma}_j^2 \to (n-j)^{-1} \sum_t (X_t - \overline{X})^2$. It is thus necessary to introduce a penalty factor containing h_n. In Auestad and Tjøstheim (1990), reasoning in a rather heuristic way, in analogy with the Akaike (1969,1970) FPE criterion we formulated a nonparametric FPE criterion as ($\mathrm{FPE}_3(j)$ in Auestad and Tjøstheim 1990):

$$\widehat{\mathrm{FPE}}(j) = \hat{\sigma}_j^2 \frac{1 + n^{-1}(h_n)^{-j} \left\{\int K^2(z)dz\right\}^j B_j}{1 - n^{-1}(h_n)^{-j} \left\{\int K^2(z)dz\right\}^j B_j}$$

$$\stackrel{\text{def}}{=} \hat{\sigma}_j^2 \frac{1 + C_j}{1 - C_j}, \tag{2.4}$$

where B_j is describing the dynamic range of the data in j-dimensional space. This criterion arises from an approximate evaluation of

$$E\left[\{Y_t - m_{n,X}^{(1,\ldots,j)}(Y_{t-1}, \ldots, Y_{t-j})\}^2\right],$$

where the predictor $m_{n,X}^{(1,\ldots,j)}$ of a model of order j computed from $\{X_t\}$ is applied to an independent realization $\{Y_t\}$ generated by the same stochastic mechanism as $\{X_t\}$.

In our experiments the dynamic range was approximated by $B_j = A^j \prod_{k=0}^{j-1} \hat{\sigma}_k$, where the factor A is chosen so that $A\hat{\sigma}_0 \approx X_{\max} - X_{\min}$ and where $\hat{\sigma}_0$ is the estimated standard deviation of the raw data. If there is a strong dependence (correlation) in the model, then once the first variable in $(X_{t-1}, \ldots, X_{t-j})$ is fixed, subsequent variables can only vary according to the residual variance, and this is what is sought expressed by the above formula. As can be expected for highly dependent processes, this approach gives better results than using $B_j = (X_{\max} - X_{\min})^j = A^j \hat{\sigma}_0^j$, which would be reasonable in the i.i.d. case. Still another possibility would be to use some sort of an enveloping volume like the convex hull of the data vectors, but such algorithms are computationally very time consuming.

With the above criterion $h_n = h_n(j)$ and p are determined by minimization in two steps of $\widehat{\mathrm{FPE}}(j)$. First we find h_n by minimization of $\widehat{\mathrm{FPE}}(j)$ for a fixed j, and subsequently the order by minimization over j. The procedure was subjected to a limited set of tests in Auestad and Tjøstheim (1990) with a third order model as the highest order tried, and with j ranging between 0 and 6. In a large majority of the cases the right order was picked by the criterion, but no doubt examples can be constructed where difficulties will arise.

Perhaps the largest objection to the above methodology is that it is rather limited which orders can be tried, since for a moderately large j, an excessive number of datapoints will be required to compute $m_n^{(1,\ldots,j)}(x_1, \ldots, x_j)$ with decent accuracy (although it is only necessary to evaluate it at the observational points in order to compute $\widehat{\mathrm{FPE}}(j)$.). On the other hand observed time series often contain large lags due to e.g. seasonality. This calls for a stepwise procedure, where only the most significant lags are included. Such an approach has been used for parametric models (see Thanoon 1990, and references therein), but would be especially worthwhile for nonparametric models due to the data requirements and the computational problem just alluded to. In this paper we make a preliminary contribution to this problem based on a modification of the $\widehat{\mathrm{FPE}}(j)$ criterion of (2.4).

We now assume that $\{X_t\}$ is given by

$$X_t = g(X_{t-i_1}, \ldots, X_{t-i_d}) + e_t$$

where d is the number of regressor variables in the model and $1 \le i_1 < \cdots < i_d$. The aim is to single out the lags $X_{t-i_1}, \ldots, X_{t-i_d}$ among $X_{t-1}, \ldots, X_{t-R}$, where R is a preset upper limit for the order. Again the analysis is based on conditional moments, in this case on

$$m^{(j_1,\ldots,j_k)}(x_{j_1},\ldots,x_{j_k}) = E(X_t \mid X_{t-j_1} = x_{j_1},\ldots,X_{t-j_k} = x_{j_k})$$

With obvious changes the same asymptotic theory as described by (2.3) is valid. The selection of lags is based on

$$\widehat{\mathrm{FPE}}(j_1,\ldots,j_k) = \hat{\sigma}^2(j_1,\ldots,j_k)\frac{1+C(j_1,\ldots,j_k)}{1-C(j_1,\ldots,j_k)}$$

in analogy with (2.4) although the range factor B_j is computed in a slightly different way. The algorithm is perhaps best explained by simply going through it for k =1 and 2.

For k =1 we have

$$\hat{\sigma}^2(j_1) = \frac{1}{n-j_1}\sum_{t=j_1+1}^{n}\{X_t - m_n^{(j_1)}(X_{t-j_1})\}^2$$

and

$$C(j_1) = n^{-1}\{h_n(j_1)\}^{-1}\{\int K^2(z)dz\}A\hat{\sigma}_0$$

where, as in Auestad and Tjøstheim (1990), we have $A = 7.0$ for $n = 500$. Here j_1 varies between 1 and R. For each value of j_1, the bandwidth $h_n(j_1)$ minimizing $\widehat{\mathrm{FPE}}(j_1)$ was chosen, and then the lag j_1^* minimizing $\{\widehat{\mathrm{FPE}}(j_1), 1 \le j_1 \le R\}$ was selected.

Next j_1^* is fixed and we search for j_2^* by looking at $\{\widehat{\mathrm{FPE}}(j_1^*, j_2), 1 \le j_2 \le R,\ \ j_2 \ne j_1^*\}$. The penalty factor in this case is given by

$$C(j_1^*, j_2) = n^{-1}\{h_n(j_1^*, j_2)\}^{-2}\{\int K^2(z)dz\}^2 A^2\hat{\sigma}_0\hat{\sigma}(j_1^* - j_2),$$

The factor $\hat{\sigma}^2(j_1^* - j_2)$ estimates the residual variance $E\{X_{t-j_2} - E(X_{t-j_2} \mid X_{t-j_1^*})\}^2$ and signifies an important difference from Auestad and Tjøstheim (1990), where we simply used estimated residual variance for a fitted second order model. We believe the above procedure to be better. For each j_2 an optimal smoothing is determined by minimizing $\hat{\sigma}^2(j_1^*, j_2)\{1+C(j_1^*, j_2)\}\{1-C(j_1^*, j_2)\}^{-1}$, and then j_2^* is determined by minimizing over j_2.

The subsequent lags $j_3^*,\ldots,j_d^*$ are determined similarly and by ordering $j_1^*,\ldots,j_d^*$ we obtain $i_1,\ldots,i_d$. The stopping point d is found by minimizing $\widehat{\mathrm{FPE}}(j_1^*,\ldots,j_r^*)$ over $r \le s$, where s is a preset number giving the maximum number of regressors allowed in the model. In practice, s has to be quite low to avoid the problems already mentioned pertaining to data requirements and excessive computations. By analogy with Auestad and Tjøstheim (1990) the residual variance σ^2 may be estimated from

$$\hat{\sigma}^2 = \frac{\hat{\sigma}^2(i_1,\ldots,i_d)}{1-C(i_1,\ldots,i_d)}$$

It is seen from the way the algorithm is constructed that once a certain lag has been entered into the model it cannot be eliminated again. This is a somewhat unfortunate feature of the algorithm, and it is known from cases of stepwise regression that it may lead to wrong results in the parametric case. The reasons for setting up the algorithm in this

manner in our case is purely considerations of computer time. All combinations are not tried and we might miss the right combination of lags, but for the experiments conducted so far promising results have been obtained.

2.1 SIMULATION EXPERIMENTS

We have tested the three models

1. Linear autoregressive
$$X_t = -0.5X_{t-6} + 0.5X_{t-10} + e_t$$

2. Threshold autoregressive
$$X_t = \begin{cases} -0.5X_{t-6} & +0.5X_{t-10} + e_t \quad X_{t-6} \leq 0 \\ & 0.8X_{t-10} + e_t \quad X_{t-6} > 0 \end{cases} \tag{2.5}$$

3. Exponential autoregressive
$$X_t = \{0.4 - 2.0\exp(-50X_{t-6}^2)\}X_{t-6} + \{0.5 - 0.5\exp(-50X_{t-10}^2)\}X_{t-10} + e_t,$$

In all cases $\{e_t\}$ was taken to be Gaussian with $E(e_t) = 0$ and $\mathrm{SD}(e_t) = 0.1$, and the same realizations of $\{e_t\}$ were fed into all of the 3 models. The random number generator G05DDF (NAG-library) was used. Both a sample size of 250 and 500 were considered. For all experiments we chose R=15 and s=4, i.e. we search over all lags between 1 and 15, but only 4 significant lags are allowed. For each model and sample size, only 5 independent realizations were tested, since the computations are very time consuming.

The results are listed in Tables 1a)—f). For each case we have displayed average values over the five realizations for the four most significant lags. It should be noted that these lags varied somewhat from model to model and from one realization to another. However for all 30 realizations the FPE criterion correctly picked out the number of significant lags, i.e. 2. Moreover, for 28 of the realizations lags 6 and 10 were correctly selected as the two significant lags. For two of the realizations of the threshold model with $n = 250$, lags 10 and 11 and 10 and 9 were selected with lag 6 picked in the third and fourth place, respectively. The estimated residual variance with two significant lags included is seen to be close to its correct value of 0.01.

The estimated standard errors computed over 5 realizations for the numbers in Table 1 were in the range 0.0006 to 0.0016 for $n = 500$ and 0.0011 to 0.0029 for $n = 250$.

More experiments are needed to determine how complex functions g can be handled by the above procedure.

3 Projections

In Auestad and Tjøstheim (1990) $m(x) = E(X_t \mid X_{t-1} = x)$ was used to distinguish between various first order nonlinear models. For higher order models the first order conditional expectation $m(x)$ is inadequate. A natural generalization is $m^{(k)}(x) = E(X_t \mid X_{t-k} = x)$, but this is not quite satisfactory either. For a zero-mean Gaussian model, for example, $E(X_t \mid X_{t-k} = x) = \mathrm{corr}(X_t, X_{t-k})x$, and even for the linear model in (2.5) we are

not able to pick up the right lags with correct coefficients. Actually, it is easy to show for this model that $m^{(k)}(x) \neq 0$ for $k \neq 6, 10$ and that $m^{(6)}(x) \neq -0.5x$ and $m^{(10)}(x) \neq 0.5x$. Similar problems arise for the general nonlinear model.

A better characterization of the data is obtained by using $m^{(1,\ldots,j)}(x_1, \ldots, x_j)$, but this is difficult to visualize for $j > 2$. For $j = 2$ we get surfaces in space, and Robinson (1983) has estimated $m^{(j_1,j_2)}(x_{j_1}, x_{j_2})$ for various lags j_1 and j_2 for the sunspot data.

3.1 DEFINITION AND SOME SIMPLE PROPERTIES

For a time series $\{X_t\}$ the k'th order projection is defined (when it exists) by:

$$P_k^j(x) = E\{E(X_t \mid X_{t-1}, \ldots, X_{t-k} = x, \ldots, X_{t-j})\}$$

with $1 \leq k \leq j$. For a model of type (1.1) we have

$$P_k^j(x) = E\{g(X_{t-1}, \ldots, X_{t-k} = x, \ldots, X_{t-j})\}$$

for $j \geq p$. The projections can be combined with the stepwise procedure of Section 2, and we define

$$P_{i_k}^{i_1,\ldots,i_d}(x) = E\{E(X_t \mid X_{t-i_1}, \ldots, X_{t-i_k} = x, \ldots, X_{t-i_d})\} \tag{3.1}$$

for $1 \leq k \leq d$. Clearly $P_{i_k}^{i_1,\ldots,i_d}(x) = P_{i_k}^{i_d}(x)$ in the preceding notation, but using (3.1) simplifies the computational effort.

It should be noted that if $\{X_t\}$ has only d significant lags so that $X_t = g(X_{t-i_1}, \ldots, X_{t-i_d}) + e_t$, then $P_i^j(x) =$ constant for $i \neq i_k$, $1 \leq k \leq d$, and $j \geq i_d$. Thus projections can be used to recognize significant lags, and their use as supplement to the stepwise procedure will be demonstrated in the next subsection. Compared to the stepwise procedure, projections are far less time consuming.

Projections are especially worthwhile for additive models; i.e. models where g of (1.1) is given by

$$g(x_1, \ldots, x_p) = \sum_{j=1}^{p} g_j(x_j)$$

where one or more of the g_j's may be zero. For a recent treatment of additive models in a regression context we refer to Buja et al (1989). For an additive model it follows immediately that

$$P_k^j(x) = g_k(x) + \mu_k \quad , j \geq p,$$

where $\mu_k = \sum_{i \neq k} E\{g_i(X_t)\} = E(X_t) - E\{g_k(X_t)\}$. Hence

$$\sum_{k=1}^{p} P_k^j(X_{t-k}) = \sum_{k=1}^{p} g_k(X_{t-k}) + (p-1)E(X_t) \tag{3.2}$$

and thus $g(X_{t-1}, \ldots, X_{t-p})$ can be estimated by

$$\sum_{k=1}^{p} \hat{P}_k^j(X_{t-k}) - (p-1)\overline{X}$$

and similarly for a model with g given by $g(X_{t-i_1}, \ldots, X_{t-i_d})$. In both cases, for a given sample size, this type of additivity enables us to estimate higher order models than for a general g. A method of estimation for P_k^j is outlined in the next subsection.

The expression (3.2) can also be used to test for additivity since the residuals $X_t - \sum_{k=1}^{p} \hat{P}_k^j(X_{t-k}) + (p-1)\overline{X}$ may be compared to the residuals $X_t - m_n^{(1,\ldots,j)}(X_{t-1}, \ldots, X_{t-j})$. We will return to details in a later publication.

Of the examples in (2.5), 1 and 3 are additive, whereas example 2 is not. In general the threshold model is non-additive, and this creates some complications using projections for recognizing it.

For the model

$$X_t = \begin{cases} a_1 X_{t-2} + e_t, & X_{t-1} \leq 0 \\ a_2 X_{t-2} + e_t, & X_{t-1} > 0 \end{cases}.$$

we have

$$P_1^2(x) = \begin{cases} a_1 E(X_{t-2}), & x \leq 0 \\ a_2 E(X_{t-2}), & x > 0. \end{cases}$$

and

$$P_2^2(x) = \{a_1 P(X_t \leq 0) + a_2 P(X_t > 0)\}x$$

and we will fail to recognize this as a threshold model if a_1 and a_2 are such that $EX_t \approx 0$. We also note that irrespective of the values of a_1 and a_2, $P_2^2(x)$ will come out as a straight line. In this case $m^{(k)}(x)$ may contain additional information, but as we will see in the next subsection this estimator has its own difficulties as already hinted at for the linear model.

3.2 ESTIMATION AND SOME EXAMPLES

In the sequel we will sometimes omit superscripts on the projections. We take as our point of departure the estimate (2.2) and form an estimate of $P_k^j(x)$ simply by averaging, i.e.

$$\hat{P}_k^j(x) = \frac{1}{n-j+1} \sum_{t=j+1}^{n+1} m_n^{(1,\ldots,j)}(X_{t-1}, \ldots, X_{t-k} = x, \ldots, X_{t-j})$$

The estimate $\hat{P}_{i_k}^{i_1,\ldots,i_d}(x)$ is computed likewise from $m_n^{(i_1,\ldots,i_d)}(X_{t-i_1}, \ldots, X_{t-i_k} = x, \ldots, X_{t-i_d})$. Applying mixing and ergodic theorems it is possible to find conditions under which $\hat{P}_k^j(x) \xrightarrow{P} P_k^j(x)$ and such that $\hat{P}_k^j(x)$ is asymptotically normal. In principle the approach of Auestad and Tjøstheim (1990) combined with conditioning on $(X_{t-1}, \ldots, X_{t-j})$ can be used to

evaluate bias and standard error in the asymptotic distribution. We illustrate this for the bias. We have

$$E\left\{\hat{P}_{i_k}(x)\right\} =$$

$$\frac{1}{n-j+1}\sum_{t=j+1}^{n+1} E\left[E\{m_n^{(i_1,\ldots,i_d)}(X_{t-i_1},\ldots,X_{t-i_k}=x,\ldots,X_{t-i_d}) \mid X_{t-i_1},\ldots,X_{t-i_d}\}\right]$$

and reasoning as in Auestad and Tjøstheim (1990) letting $\underline{x}_d = (x_{i_1},\ldots,x_{i_d})$ we can show that in the asymptotic distribution

$$E\{m_n^{(i_1,\ldots,i_d)}(\underline{x}_d)\} =$$

$$m^{(i_1,\ldots,i_d)}(\underline{x}_d) + \frac{1}{2}h_n^2 f^{-1}(\underline{x}_d)\left\{2\sum_{j=1}^{d}\frac{\partial m^{(i_1,\ldots,i_d)}(\underline{x}_d)}{\partial x_{i_j}}\frac{\partial f(\underline{x}_d)}{\partial x_{i_j}} + f(\underline{x}_d)\sum_{j=1}^{d}\frac{\partial^2 m^{(i_1,\ldots,i_d)}(\underline{x}_d)}{\partial x_{i_j}^2}\right\}$$

$$+ \text{ higher order terms}$$

from which $E\{\hat{P}_{i_k}(x)\}$ can by obtained by integrating out x_{i_j} for $j \neq k$. In the linear Gaussian case an explicit expression for the dominating term can be obtained. It turns out that it is linear, so that it does not create serious problem in an identification context. For a general model of type 1 in (2.5) we get

$$E\{\hat{P}_6^{6,10}(x)\} \sim a_6 x - h_n^2\{(1-\rho^2)\mathrm{var}(X_t)\}^{-1}(a_6 - \rho a_{10})x \tag{3.3}$$

where $\rho = \mathrm{corr}(X_{t-6}, X_{t-10})$ and with a similar expression being true for $E\{\hat{P}_{10}^{6,10}(x)\}$. The standard deviation is much more difficult, since a singularity occurs when trying to integrate out x_{i_j} for $j \neq k$.

We have done a number of simulation experiments. In Figs. 1—3 are shown the estimated $\hat{P}_{i_k}(x)$ from 10 independent realizations for the four most significant lags, as determined from the procedure of Section 2. All of the three models defined in (2.5) are included, and we have used a smoothing of $h = 0.03$ which is quite close to the smoothing for $m_n(x)$ used in Auestad and Tjøstheim (1990). This is not unreasonable, since $\hat{P}_k^j(x)$ essentially is a one-dimensional quantity.

For the additive models 1 and 3 the estimates are as predicted from the theory in Section 3.1. The bias and standard error are comparable with that of $m_n(x)$, and for the linear model it is consistent with (3.3) which gives a negligible contribution in this case. It is seen that nonsignificant lags are clearly indicated; i.e. $\hat{P}_{i_k}(x) \approx$ constant for $i_k \neq 6, 10$. Hence projections can be used as a supplement to the stepwise procedure of Section 2, at least for additive models.

For real data only one realization is available. It may be difficult to work out asymptotic theory in the general case, but bootstrapping may give an idea about sampling errors. As a start in this direction, in Fig. 4 we have displayed 10 bootstrap simulations based on one realization of the linear model in (2.5). Clearly, the bootstrap simulations will depend strongly on the original realization they are based on. If this is skewed, the whole picture

will be skewed. The basic realization used in Fig. 4 is not particularly well-behaved. Nevertheless, a comparison with Fig. 1 indicates that bootstrapping can be used to say something sensible about errors, but more extensive investigations are necessary. Possibly, the bootstrap can also be used to determine the degree of smoothing.

On Fig. 5 are displayed some examples of $m^{(k)}(x) = E(X_t \mid X_{t-k} = x)$ ($h_n = 0.05$ is used). It is seen that it can not be decided which lags are significant from this figure.

The threshold model 2 is displayed in Figs. 2a)—d). The projections $\hat{P}_{i_k}(x)$ in this case takes out lags 6 and 10 as significant lags and 6 as the threshold lag. For this model

$$P_6(x) = \begin{cases} -0.5x + 0.5E(X_t) & x \leq 0 \\ 0.8E(X_t) & x > 0 \end{cases}$$

and

$$P_{10}(x) = -0.5E\{X_t I(X_t \leq 0)\} + \{0.5P(X_t \leq 0) + 0.8P(X_t > 0)\}x$$

where $I(\cdot)$ is an indicator variable. The same tendency is seen by using $m_n^{(k)}(x)$, but in this case, as for the linear case, some of the non-significant lags are non-constant too.

It is not difficult to construct threshold models where <u>both</u> methods fail to recognize the threshold character. One example is

$$X_t = \begin{cases} -0.5X_{t-6} + 0.5X_{t-10} + e_t & X_{t-1} \leq 0 \\ 0.8X_{t-10} + e_t & X_{t-1} > 0. \end{cases}$$

3.3 THE WOLF SUNSPOT NUMBERS

The Wolf sunspot numbers have been examined in a number of investigations. Various number of datapoints have been used. We have used the data set listed in Tong (1983) starting in the year 1700 and ending in 1979, altogether 280 data points.

A review of the analysis of the data up to 1977 is given in the paper by Morris (1977). The fitting of nonlinear models is discussed among others by Priestley (1981), Robinson (1983), Subba Rao and Gabr (1984) and Tong (1983). Robinson uses a nonparametric model and plots the surface $m_n^{(i_1,i_2)}(x_{i_1}, x_{i_2})$ for lags picked by a linear procedure devised by Morris (1977). These lags are 1, 2 and 9. Using our stepwise procedure, again allowing only 4 lags, we found lags 1, 2, 3 and 9 to be the most significant, and in that sequence. The results are given in Table 2. It is seen that the minimum of the FPE occurs at the 4th significant lag, and it may be reasonable to include more lags. This is done in fact for the nonlinear models fitted by Tong. The estimated residual variance is given in the last column of Table 2. It is lower than that of a 10-parameter linear model, but higher than a bilinear model based on the 4 first lags of $\{X_t\}$ and $\{e_t\}$ and with 20 parameters. Both of these models are presented in Priestley (1981, p.882).

There is a clear evidence of nonlinearity and its projected form is shown on Fig. 6. Since no test of additivity has been performed one has to be a little careful with the interpretation of these figures. Nonlinearity has also been inferred by other authors using other means. A more detailed analysis of this and other data sets will be presented elsewhere.

4 Further prospects

The methods proposed in Auestad and Tjøstheim (1990) and in this paper, suggest a fairly general approach to time series analysis with a first phase of nonparametric fitting, possibly followed by a parametric model if a suitable linear or nonlinear model can be identified at the first stage. A number of tasks and possibilities for further studies exist. Among them are: Multiple models, additivity and tests of additivity, bootstrap confidence bands, other asymptotic results, considerations of needed sample size, computational problems and parallel processing.

References

Akaike, H. (1969). Fitting autoregressions for predictions. *Ann. Inst. Statist. Math.* **21**, 243–7.

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203–17.

Auestad, B. and Tjøstheim, D. (1990). Identification of nonlinear time series: First order characterization and order determination. *Biometrika,* to appear.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17**, 435–555.

Collomb, G. (1983). From non parametric regression to non parametric prediction: Survey of the mean square error and original results on the predictogram, in *Specifying Statistical Models. From Parametric to Non-parametric, Using Bayesian or Non-Bayesian Approaches* (J.P. Florens, M. Mouchart, J.P. Raoult, L. Simar and A.F.M. Smith eds.). Lecture Notes in Statistics, **16**, 182–204. New York: Springer.

Haggan, V., Heravi, S.M. and Priestley, M.B. (1984). A study of the application of state-dependent models in non-linear time series analysis. *J. Time Series Anal.* **5**, 69–102.

Härdle, W. (1984). How to determine the bandwidth of some nonlinear smoothers in practice, in *Robust and Nonlinear Time series Analysis* (J. Franke, W. Härdle and D. Martin eds.). Lecture Notes in Statistics, **26**, 163– 84. Berlin: Springer.

Morris, J. (1977). Forecasting the sunspot cycle. *J. Roy. Statist. Soc. Ser. A*, **140**, 437–47.

Priestley, M.B. (1981). *Spectral Analysis and Time Series.* London: Academic Press.

Robinson, P.M. (1983). Non-parametric estimation for time series models. *J. Time Series Anal.* **4**, 185–208.

Subba Rao, T. and Gabr, M.M. (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models.* Lecture Notes in Statistics, **24**. New York: Springer.

Thanoon, B.Y. (1990). Subset threshold autoregression with applications. *J. Time Series Anal.* **1**, 75–87.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis.* Lecture Notes in Statistics, **21**. New York: Springer.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach, Monograph.* Oxford University Press, to appear.

k	$\widehat{\mathrm{FPE}}_{(k)}$	$\hat{\sigma}^2$
0	0.03154	0.03154
1	0.01544	0.01516
2	0.01017	0.00971
3	0.01064	0.00988
4	0.01120	0.01032

Table 1a.

k	$\widehat{\mathrm{FPE}}_{(k)}$	$\hat{\sigma}^2$
0	0.02244	0.02244
1	0.01094	0.01073
2	0.01013	0.00964
3	0.01061	0.00960
4	0.01126	0.01038

Table 1b.

k	$\widehat{\mathrm{FPE}}_{(k)}$	$\hat{\sigma}^2$
0	0.01619	0.01609
1	0.01197	0.01146
2	0.01018	0.00942
3	0.01137	0.01024
4	0.01247	0.01141

Table 1c.

k	$\widehat{\mathrm{FPE}}_{(k)}$	$\hat{\sigma}^2$
0	0.03361	0.03361
1	0.01635	0.01594
2	0.01117	0.01048
3	0.01177	0.01082
4	0.01243	0.01138

Table 1d.

k	$\widehat{\mathrm{FPE}}_{(k)}$	$\hat{\sigma}^2$
0	0.02310	0.02310
1	0.01100	0.01055
2	0.01031	0.00959
3	0.01092	0.00978
4	0.01157	0.01055

Table 1e.

k	$\widehat{\mathrm{FPE}}_{(k)}$	$\hat{\sigma}^2$
0	0.01689	0.01689
1	0.01228	0.01177
2	0.01075	0.00975
3	0.01224	0.01072
4	0.01338	0.01202

Table 1f.

Table 1: $\widehat{\mathrm{FPE}}$ values and estimated residual variance $\hat{\sigma}^2 = \hat{\sigma}^2(j_1,\ldots,j_k)\{1 - C(j_1,\ldots,j_k)\}^{-1}$, for the models in (2.5). k is the number of lags in the model and $\widehat{\mathrm{FPE}}_{(k)} = \widehat{\mathrm{FPE}}(j_1,\ldots,j_k)$. 1a and 1d: Linear AR model, 1b and 1e: Threshold model and 1c and 1f: Exponential AR model. Each subtable 1a)—c) is based on 5 independent realizations each consisting of $n = 500$ observations. The subtables 1d)—f) are based on 5 independent realizations each consisting of $n = 250$ observations.

k	$\widehat{\mathrm{FPE}}_{(k)}$	$\hat{\sigma}^2$
0	1495.60	1495.60
1	527.68	515.05
2	249.48	207.97
3	232.75	211.65
4	197.99	180.63

Table 2.

Table 2: $\widehat{\mathrm{FPE}}$ values and estimated residual variance $\hat{\sigma}^2 = \hat{\sigma}^2(j_1,\ldots,j_k)\{1 - C(j_1,\ldots,j_k)\}^{-1}$, for the Wolf sunspot data. k is the number of lags in the model and $\widehat{\mathrm{FPE}}_{(k)} = \widehat{\mathrm{FPE}}(j_1,\ldots,j_k)$. The lags entered into the model in the order: 1, 2, 3 and 9.

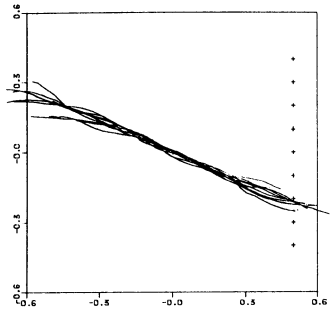


Figure 1a.

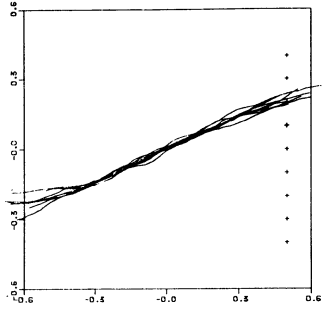


Figure 1b.

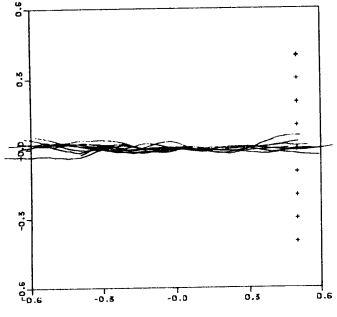


Figure 1c.

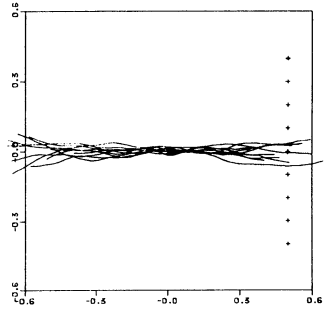


Figure 1d.

Figure 1: $\hat{P}_{i_k}(x)$ plotted against x for model 1 in (2.5); $i_k = 6, 10, 13, 15$ in figures 1a, 1b, 1c and 1d, resp. $P_{i_1}(x), \ldots, P_{i_4}(x)$ are estimated from 10 independent realizations each consisting of $n = 500$ observations.

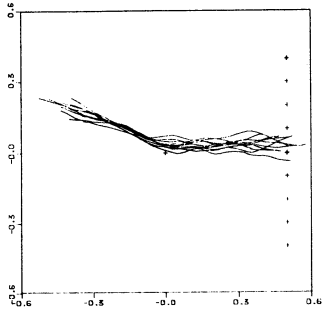


Figure 2a.

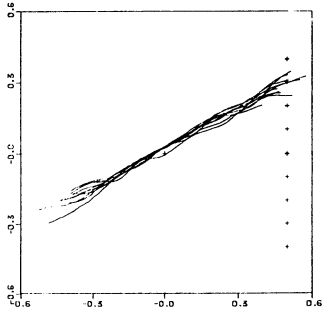


Figure 2b.

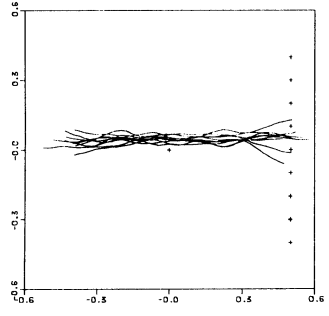


Figure 2c.

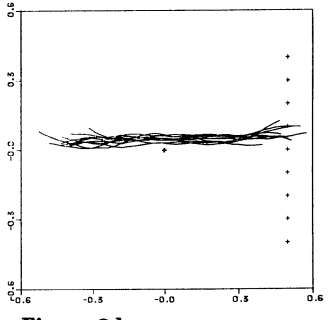


Figure 2d.

Figure 2: $\hat{P}_{i_k}(x)$ plotted against x for model 2 in (2.5); $i_k = 6, 10, 11, 15$ in figures 2a, 2b, 2c and 2d, resp. $P_{i_1}(x), \ldots, P_{i_4}(x)$ are estimated from 10 independent realizations each consisting of $n = 500$ observations.

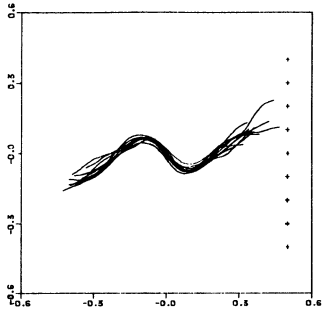


Figure 3a.

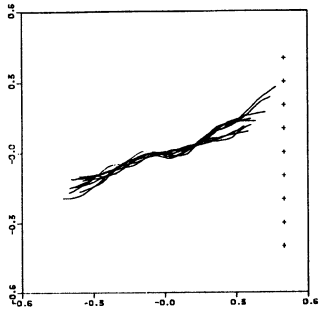


Figure 3b.

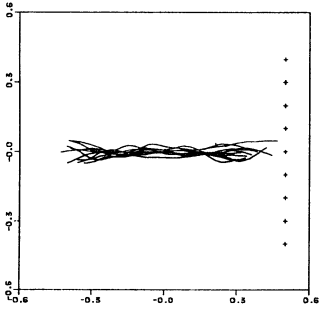


Figure 3c.

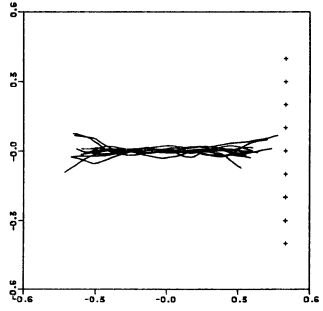


Figure 3d.

Figure 3: $\hat{P}_{i_k}(x)$ plotted against x for model 3 in (2.5); $i_k = 6, 10, 11, 15$ in figures 3a, 3b, 3c and 3d, resp. $P_{i_1}(x), \ldots, P_{i_4}(x)$ are estimated from 10 independent realisations each consisting of $n = 500$ observations.

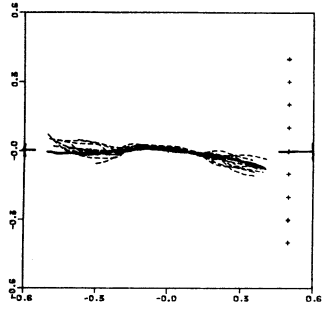


Figure 4a.

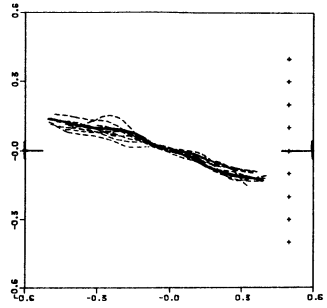


Figure 4b.

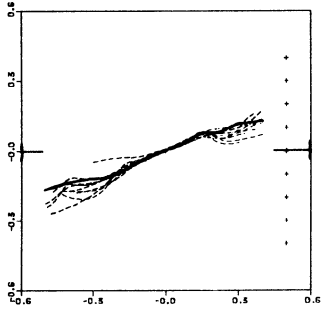


Figure 4c.

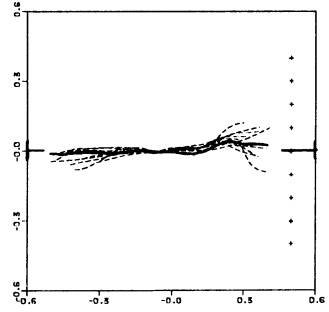


Figure 4d.

Figure 4: Bootstrap for the model 1 in (2.5); — : The estimate of $P_{i_k}(x)$ based on the realization $\{X_t\}$, - - - : The estimates of $P_{i_k}(x)$ based on the bootstrap realizations. $i_k = 4, 6, 10, 12$ in figures 4a, 4b, 4c and 4d, resp.

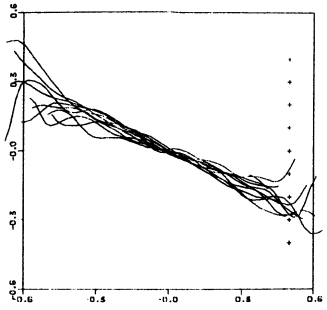


Figure 5a.

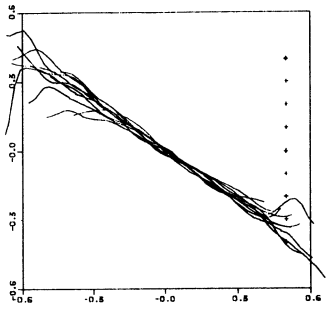


Figure 5b.

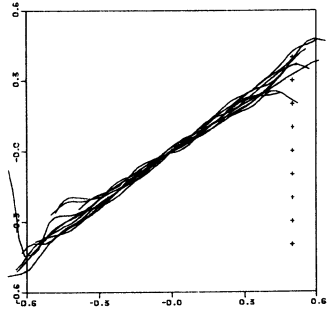


Figure 5c.

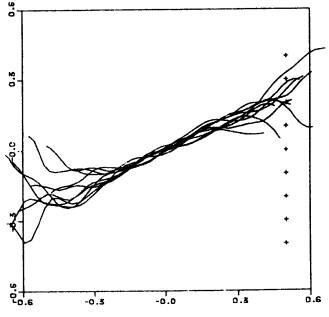


Figure 5d.

Figure 5: $m_n^{(k)}(x)$ plotted against x for model 1 in (2.5); $k = 4, 6, 10, 12$ in figures 5a, 5b, 5c and 5d, resp. $m_n^{(4)}(x), m_n^{(6)}(x), m_n^{(10)}(x)$ and $m_n^{(12)}(x)$ are estimated from 10 independent realizations each consisting of $n = 500$ observations.

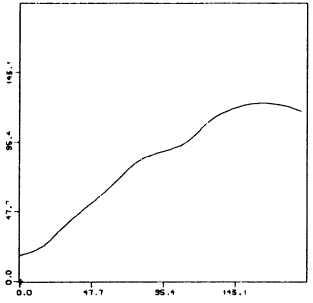


Figure 6a.

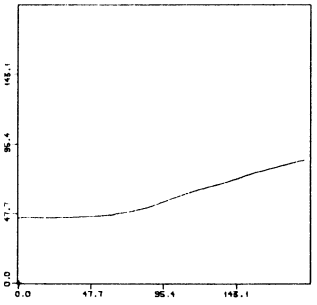


Figure 6b.

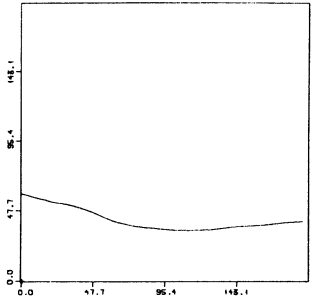


Figure 6c.

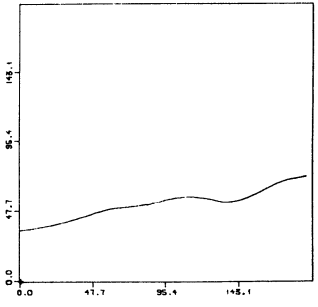


Figure 6d.

Figure 6: $\hat{P}_{i_k}(x)$ plotted against x for the sunspot data; $i_k = 1, 2, 3, 9$ in figures 6a, 6b, 6c and 6d, resp. A smoothing of $h_n = 11.0$ was used.

MODELIZATION, NONPARAMETRIC ESTIMATION AND PREDICTION FOR CONTINUOUS TIME PROCESSES

denis BOSQ
Université Pierre et Marie Curie
Laboratoire de Statistique
Théorique et Appliquée
4, place Jussieu
75252 Paris cedex 05
France

ABSTRACT. In order to predict a continuous time process on an entire time-interval, we introduce and study a Hilbertian autoregressive process X_t, $t=0, \pm 1, \pm 2, \ldots$

First, we present consistent nonparametric estimators for the covariance operator C of X_t and the cross estimator covariance operator of (X_t, X_{t+1}). Then we construct an estimator ρ_n of the autocorrelation operator, by projecting the data on a finite dimensional subspace v_{k_n}, generated by eigenvectors of C. If the eigenvectors are unknown, we construct a preliminary estimator of v_{k_n}. Under mild regularity conditions, we show that the predictor based upon ρ_n converges in probability. When the process (X_t) satisfies an additional mixing condition, we obtain almost sure convergence.

Rates of convergence are given as well as numerical examples.

1. INTRODUCTION

Consider a continuous time process observed on a time-interval $Y = (Y_u, u \in \mathbb{R})$. We seek information on the evolution of Y in the future. From a practical point of view, it is more convenient to predict the process behavior on an entire time-interval rather than at specific points. Therefore, the use of functional spaces techniques seems to be well adapted to this problem. These being nonparametric by themselves, it seems rather heavy to introduce a non parametric model for observations, lying in a functional space, even if some results can be obtained in that direction (see, e.g., GRENANDER (1981), BOSQ-DELECROIX (1985)).

AMS Classification and Key Words : 62M - Prediction, autoregressive process, Hilbert space, continuous time process, mixing.

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 509–529.
© 1991 *Kluwer Academic Publishers.*

Here we prefer making use of a simple linear model which can modelize various situations. We consider a real separable Hilbert space H equined with its norm $\|.\|$ and its scalar product $<.,.>$. We define a first order H-autoregressive process (ARH(1)) as a sequence (X_t) of H-valued random variables such that

$$X_t = \rho(X_{t-1}) + \varepsilon_t \quad ; \; t = 0, \pm 1, \pm 2, \ldots, \tag{1}$$

where (ε_t) is an H-white noise (i.e., the ε_t s are H-valued i.i.d. r.v. such that $0 < E\|\varepsilon_t\|^2 < +\infty$, $E\varepsilon_t = 0$) and ρ is a symmetric compact linear operator on H with $\|\rho\|_{\mathcal{L}} < 1$. (Here $\|.\|_{\mathcal{L}}$ denotes the norm for linear bounded operators on H). In the sequel, we make the additional assumption $E\|X_0\|^4 < +\infty$.

Equation (1) has an almost surely unique strictly stationary solution given by

$$X_t = \sum_{j=0}^{\infty} \rho^j(\varepsilon_{t-j}) \quad ; \; t = 0, \pm 1, \pm 2, \ldots, \tag{2}$$

where the series converge a.s. and in L^2_H (See, e.g., BOSQ (1990) for a survey of properties of ARH(1) processes).

Now, with $X_1, \ldots, X_n$ denoting observations of an ARH(1), our purpose is to estimate $\rho(X_n)$, i.e. the best predictor of X_{n+1} .

First, we present consistent estimators for the covariance operator C of X_0 and the cross covariance operators D and D^* of (X_0, X_1). Rates of convergence are given for these estimators.

The estimation of ρ is a difficult problem even if the invertibility of C implies that $\rho = D\,C^{-1} = C^{-1}\,D^*$. This is due to the fact that C^{-1} is not bounded when H is infinite dimensional.

We construct here an estimator ρ_n of ρ by projecting the data on $\mathcal{V}_{k_n}$, a finite dimensional subspace of H . $\mathcal{V}_{k_n}$ is generated by eigenvectors of C when they are known, or by their estimates otherwise. This construction makes $\rho_n(X_n)$ a natural predictor for X_{n+1} .

Under mild regularity conditions, we will show in the sequel that

$$n^{\eta}\|\rho_n(X_n) - \rho(X_n)\| \xrightarrow{p} 0$$

for a suitable choice of $\eta > 0$.

Under a strong mixing condition we prove a.s. convergence. It should be noticed that some of these results remain valid for a general Hilbertian strictly stationary process.

Our results can be applied to the prediction of a locally square integrable stationary (or stationary increments) continuous time process. If $(Y_u, u \in U)$ is observed, the time set U is divided in n δ-intervals which generate L^2-valued variables $X_1, \ldots, X_n$. The above results provide a predictor of (Y_u) on a δ-interval. Note that one can choose disjoint, adjacent or overlapping intervals.

The contents of this paper are as follows. In Section 2, we study the consistency of the empirical covariance operators for an ARH(1) process. In Section 3, we state

consistency results for eigenvectors and eigenvalues of C . Section 4 is devoted to the method of prediction described above. In Ssection 5, we present some simulations and an application to an experimental data set. Finally, Section 6 contains the proofs of the main results.

The estimation of the covariance operator in the i.i.d. case has been considered by DAUXOIS-FINE-POUSSE (1979), GRENANDER (1981), ANTONIADIS-BEDER (1989). Our results extend these investigations in a more general setting and also give rates of convergence.

2. CONSISTENCY OF THE EMPIRICAL COVARIANCE OPERATORS

The **covariance operator** of X_0 is defined by

$$C(x) = E(< X_0, x > X_0) \quad , x \in H \quad .$$

C is an element of the real separable Hilbert space $\mathcal{S}$ spanned by the Hilbert-Schmidt operators on H . Note that the $\mathcal{S}$-norm satisfies

$$\|s\|_{\mathcal{L}} \leq \|s\|_{\mathcal{S}} \quad , \quad s \in \mathcal{S} \quad .$$

A natural estimator of C is given by the empirical covariance operator defined by

$$C_n(x) = \frac{1}{n} \sum_{t=1}^{n} < X_t, x > X_t \quad , \quad x \in H \quad .$$

C_n is a finite rank symmetric random operator which defines a $\mathcal{S}$-valued r.v.

The convergence of C_n towards C is a "law of large numbers" type result.

Proposition 2.1

$$E\|C_n - C\|_{\mathcal{S}}^2 \leq A\, n^{-1} \, , \tag{3}$$

where $A = E\|X_0\|^4 \left[1 + 4\|\rho\|_{\mathcal{L}}^2 \left(1 - \|\rho\|_{\mathcal{L}}^2\right)^{-1}\right]$.

Furthermore,

$$\|C_n - C\|_{\mathcal{S}} \to 0 \quad \text{a.s.}$$

Let us now define the **cross covariance operators** D and D^* of (X_0, X_1) by

$$D(x) = E(< X_0, x > X_1) \quad , \quad x \in H$$

and

$$D^*(x) = E(< X_1, x > X_0) \quad , \quad x \in H \quad .$$

Their empirical estimators are, respectively,

$$D_n(x) = \frac{1}{n-1} \sum_{t=1}^{n-1} < X_t, x > X_{t+1} \ , \qquad x \in H$$

and

$$D_n^*(x) = \frac{1}{n-1} \sum_{t=1}^{n-1} < X_{t+1}, x > X_t \ , \qquad x \in H$$

Then we have :

Proposition 2.2

For each $n \geq 2$,

$$E\|D_n - D\|_{\mathcal{S}}^2 = E\|D_n^* - D^*\|_{\mathcal{S}}^2 \leq B(n-1)^{-1} \ , \tag{4}$$

where

$$B = 2E\|X_0\|^2 \ E\|\varepsilon_0\|^2 + 2E\|X_0\|^4 \left[1 + \frac{4\|\rho\|_{\mathcal{L}}^2}{1 - \|\rho\|_{\mathcal{L}}^2}\right] ,$$

and

$$\|D_n - D\|_{\mathcal{S}} = \|D_n^* - D^*\|_{\mathcal{S}} \to 0 \quad \text{a. s..}$$

- The strong mixing case

We introduce now a strong mixing condition which allows to obtain exponential rates of convergence for the above estimates.

Let $(v_j, j \geq 1)$ be a complete orthonormal system of eigenvectors of C , $(\lambda_j, j \geq 1)$ whith $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ the corresponding eigenvalues. $\mathcal{V}_k$ denotes the linear space generated by $v_1, \ldots, v_k$ and π^k the orthogonal projection on $\mathcal{V}_k$. Then (X_t) is said to be H-strong mixing if

$$\text{(HSM)} \qquad \alpha(n) = \sup_k \sup_{\substack{A \in \mathcal{F}_{-\infty}^t(\mathcal{V}_k) \\ B \in \mathcal{F}_{t+n}^{\infty}(\mathcal{V}_k)}} |P(A \cap B) - P(A)P(B)| \underset{n \to \infty}{\to} 0 \ ,$$

where $\mathcal{F}_a^b(\mathcal{V}_k)$ is the σ-algebra generated by the r.v.'s $\pi^k(X_a), \ldots, \pi^k(X_b)$.

We have then the following proposition :

Proposition 2.3

If $\|X_0\| \leq d$ a.s. , if (X_t) is geometrically H-strong mixing and if

$$\lambda_j = a\, r^j \quad ; \; a > 0 \; , \; 0 < r < 1 \quad ; \; j = 1,2,\ldots,$$

then, given $0 < \gamma < \frac{1}{4}$, we have, for each $\varepsilon > 0$ and $n \geq n_\gamma$,

$$P[\|C_n - C\|_{\mathcal{S}} > \varepsilon] \leq 2n^\gamma \exp\left[-\frac{\varepsilon\, n^\gamma}{10\, d^2\, e\, \sqrt{2}}\right] + \frac{2ad^2 r}{\varepsilon^2 (1-\sqrt{r})^2} \exp\left[-\mathrm{Log}\, \frac{1}{r}\, n^\gamma\right] \; ; \tag{5}$$

the same bound is valid for $P[\|D_n - D\|_{\mathcal{S}} > \varepsilon]$

substituying $\frac{n^\gamma}{d^2}$ by $\frac{(n-1)^\gamma}{d^2 + \|D\|_{\mathcal{S}}}$ in the first exponential term

and n^γ by $(n-1)^\gamma$ in the second (6)

3. ESTIMATION OF EIGENVALUES AND EIGENVECTORS OF C

Let $\lambda_{1n} \geq \lambda_{2n} \geq \ldots \geq \lambda_{nn} \geq 0 = \lambda_{n+1\,n} = \lambda_{n+2\,n} = \ldots$ be the eigenvalues of C_n and let v_{jn}, $j = 1,2,\ldots$ be a complete orthonormal system of eigenvectors of C_n such that

$$C_n\, v_{jn} = \lambda_{jn}\, v_{jn} \qquad j = 1,2,\ldots \quad .$$

Then λ_{jn} and v_{jn} are natural estimates of λ_j and v_j .
The next lemma provides useful inequalities.

Lemma 3.1

For each positive integer j, we have

$$|\lambda_{jn} - \lambda_j| \leq \|C_n - C\|_{\mathcal{L}} \quad . \tag{7}$$

Furtheremore, if $\lambda_1 > \lambda_2$, then

$$\|v_{1n} - v_1'\| \leq a_1 \|C_n - C\|_{\mathcal{L}} \tag{8}$$

where $a_1 = 2\sqrt{2}\, (\lambda_1 - \lambda_2)^{-1}$;

and, if, for some $j > 1$, $\lambda_j > \lambda_{j+1}$ then

$$\|v_{jn} - v_j'\| \leq a_j\, \|C_n - C\|_{\mathcal{L}} \; , \tag{9}$$

where $a_j = 2\sqrt{2}[\min(\lambda_{j-1} - \lambda_j \; , \; \lambda_j - \lambda_{j+1})]^{-1}$.

Here $v_j' = (\mathrm{sg} < v_{jn}, v_j >)\, v_j$, $j \geq 1$, where $\mathrm{sg}x = 1$ is $x \geq 0$ and $\mathrm{sg}x = -1$ if $x < 0$.

Using Propositions 2.1, 2.3 and Lemma 3.1, we obtain the following consistency result :

Proposition 3.2 We have

$$E\left[\sup_{j\geq 1} |\lambda_{jn} - \lambda_j|^2\right] \leq A\, n^{-1} ,$$

$$\sup_{j\geq 1} |\lambda_{jn} - \lambda_j| \to 0 \quad \text{a.s.} ,$$

and under the assumptions of Proposition 2.3 the rate of convergence is $O(\exp(-C\, n^{\gamma}))$, $0 < \gamma < \frac{1}{4}$.

Proposition 3.3

For each j such that $\lambda_{j-1} > \lambda_j > \lambda_{j+1}$,

$$E\|v_{jn} - v'_j\|^2 \leq a_j\, A\, n^{-1} ,$$

and

$$\|v_{jn} - v'_j\| \to 0 \quad \text{a.s.}$$

Furthermore, if $\lambda_j = a\, r^j$ and $k_n = o(\text{Log}n)$ (resp. $\lambda_j = a\, j^{\beta}$, $\beta > 1$ and $k_n = O(n^{1/(2\beta+2)})$), then

$$E\left[\sup_{1\leq j\leq k_n} \|v_{jn} - v'_j\|^2\right] \to 0 .$$

Finally, under the assumptions of Proposition 2.3 and if $k_n = o(\text{Log}n)$, then

$$\sup_{1\leq j\leq k_n} \|v_{jn} - v'_j\| \xrightarrow[\text{a.s.}]{} 0 ,$$

and the rate of convergence is $O(\exp(-C\, n^{\gamma}))$, $0 < \gamma < \frac{1}{4}$.

Remarks

1. If $\lambda_{j-a} > \lambda_{j-a+1} = \cdots = \lambda_{j+b-1} > \lambda_{j+b}$, then

$$\left\|v_{jn} - \pi_{j-a+1}^{j+b-1}(v_{jn})\right\| \leq 2\sqrt{2}\, \min(\lambda_{j-a}-\lambda_j, \lambda_j-\lambda_{j+b})\, \|C_n - C\|_{\mathcal{L}} ,$$

where π_{j-a+1}^{j+b-1} denotes the orthogonal projection on the linear space generated by $v_{j-a+1}, \ldots, v_{j+b-1}$.

2. If the v_j's are known, alternative estimates of the λ_j's are

$$\lambda'_{jn} = \frac{1}{n} \sum_{t=1}^{n} \langle X_t, v_j \rangle^2 , \quad j = 1,2,\ldots .$$

These estimates are unbiased and consistent.

4. PREDICTION

a) Known v_j's

Let (k_n) be a sequence of integers which tends to infinity. We consider the observed r.v.'s

$$Y_{tn} = \pi^{k_n}(X_t) \quad ; \quad t = 1,\dots,n ,$$

their covariance operators

$$G_n = \pi^{k_n} C \pi^{k_n} , \quad H_n = \pi^{k_n} D \pi^{k_n} , \quad H_n^* = \pi^{k_n} D^* \pi^{k_n} ,$$

and their empirical covariance operators

$$\Gamma_n = \pi^{k_n} C_n \pi^{k_n} , \quad \Delta_n = \pi^{k_n} D_n \pi^{k_n} , \quad \Delta_n^* = \pi^{k_n} D_n^* \pi^{k_n} .$$

In order to construct an estimator of ρ, we make the following assumptions :

$[(A_1)$ $\lambda_j > 0$ for each j

(A_2) For each n , the determinant of

$$\left(\sum_{t=1}^{n} < X_t, v_i > < X_t, v_j >\right)_{1 \leq i,j \leq k_n}$$

does not vanish with probability one.

Under A_1 and A_2, G_n and Γ_n have inverses on $\mathcal{V}_{k_n}$.
We then consider an estimator of ρ

$$\rho_n = \Gamma_n^{-1} \Delta_n^* \qquad \text{(a.s.)}$$

and the associated predictor $\rho_n(X_n)$.

The following lemmas lead to the consistency for $\rho_n(X_n)$.

Lemma 4.1.
Under A_1 and A_2, we have, for each $\varepsilon > 0$,

$$P\|\rho_n - \pi^{k_n}\rho\pi^{k_n}\|_{\mathcal{L}} > \varepsilon) \leq \frac{C_\varepsilon}{n\,\lambda_{k_n}^4} , \tag{10}$$

where

$$C_\varepsilon = \frac{4\,B\,\lambda_1^4}{\varepsilon} + \left(A + \frac{4B}{\varepsilon^2}\right)\lambda_1^2 + \frac{2A}{\varepsilon}\left(1 + \lambda_1\sqrt{\frac{\varepsilon}{2}}\right)^2 . \tag{11}$$

Lemma 4.2. We have

$$E\|\rho(X_n) - (\pi^{k_n}\rho)(X_n)\|^2 = \sum_{k_n+1}^{\infty} < (\rho C \rho)\, v_j\, ,\, v_j > \leq \sum_{k_n+1}^{\infty} \lambda_j \, . \tag{12}$$

Proposition 4.3.

If k_n and $n\,\lambda_{k_n}^4$ both tend to infinity, then

$$\|\rho_n(X_n) - \rho(X_n)\| \overset{P}{\to} 0 \, .$$

In fact, for $\varepsilon > 0$,

$$P(\|\rho_n(X_n) - \rho(X_n)\| > \varepsilon) \leq \frac{C_\varepsilon}{n\,\lambda_{k_n}^4} + \frac{4}{\varepsilon^2} \sum_{k_n+1}^{\infty} \lambda_j \, . \tag{13}$$

Discussion : If $\lambda_j \geq a\, r^j > 0$, $j = 1,2,\ldots,$ then the conditions $k_n \to +\infty$ and $k_n = o(\mathrm{Log}n)$ entail (12). In the special case where $\lambda_j = a\, r^j$, r being known, the optimal $k_n = \left[\frac{1}{5}\,\mathrm{Log}\,\frac{1}{r}\,\mathrm{Log}n\right]+1$ gives

$$\frac{n^{1/10}}{\mathrm{LogLog}n}\,\|\rho_n(X_n) - \rho(X_n)\| \overset{P}{\to} 0 \quad \text{as } n \to \infty \, . \tag{14}$$

Note that in all cases, $r_n = \frac{\lambda_{2n}}{\lambda_{1n}}$ is a consistent estimate for r . A possible practical choice for k_n is given by $\left[\frac{1}{5}\,\mathrm{Log}\,\frac{\lambda_{1n}}{\lambda_{2n}}\,.\,\mathrm{Log}n\right]+1$.

Finally, if $\lambda_j = \frac{c}{j^\gamma}$, $c > 0$, $\gamma > 1$; $j = 1,2,\ldots$,

the asymptotically best choice of k_n is $k_n = [n^{1(5\gamma-1)}]$ which yields

$$\frac{n^{\frac{1}{10}\left(\frac{\gamma-1}{\gamma-(1/5)}\right)}}{\mathrm{LogLog}n}\,\|\rho_n(X_n) - \rho(X_n)\| \overset{P}{\to} 0 \, . \tag{15}$$

b) Unknown v_j's

Let $\hat{\pi}^{k_n}$ be the orthogonal projection on $\hat{\mathcal{V}}_{k_n}$, i.e. the linear space generated by $v_{1n},\ldots,v_{k_n n}$. Putting $\hat{\Gamma}_n = \hat{\pi}^{k_n} C_n \hat{\pi}^{k_n}$, $\hat{\Delta}_n = \hat{\pi}^{k_n} D_n \hat{\pi}^{k_n}$, $\hat{\Delta}_n^* = \hat{\pi}^{k_n} D_n^* \hat{\pi}^{k}$, we take

$\hat{\rho}_n = \hat{\Gamma}_n^{-1}\,\hat{\Delta}_n^{*}$ as an estimate for ρ and $\hat{\rho}_n(X_n)$ as a predictor for X_{n+1} ; the existence of $\hat{\Gamma}_n^{-1}$ being ensured by the assumption

$$(A_3) \qquad P(\lambda_{1n} > \lambda_{2n} > \ldots > \lambda_{nn} > 0) = 1 ,$$

which is fulfilled in all usual cases.

We will also use the assumption that

$$(A_1') \qquad \lambda_1 > \lambda_2 > \ldots > 0 .$$

Lemma 4.4

Assume that A_1', A_2, A_3 hold. Then

$$\|\hat{\rho}_n - \rho_n\| \leq \alpha_n \|C_n - C\| + \beta_n \|D_n - D\| , \tag{16}$$

where

$$\alpha_n = \|D\| \left(\lambda_{k_n}^{-1} \lambda_{k_n n}^{-1} + 2 \sum_{1}^{k_n} a_j \left(1+\lambda_{k_n}^{-1}\right) \right) ,$$

and

$$\beta_n = \lambda_{k_n n}^{-1} .$$

Using the above bound, we obtain readily :

Proposition 4.5

If $\lambda_j = a\, r^j > 0$, and if A_2, A_3 are fulfilled, then for $k_n = o(\mathrm{Log}\, n)$

$$\|\hat{\rho}_n - \rho_n\| \overset{p}{\to} 0 \quad \text{and} \quad \|\hat{\rho}_n(X_n) - \rho(X_n)\| \overset{p}{\to} 0 .$$

Discussion : As in the above results, we have

$$\frac{n^{1/10}}{\mathrm{Log Log}\, n} \|\hat{\rho}_n(X_n) - \rho(X_n)\| \overset{p}{\to} 0 , \tag{17}$$

for $\lambda_j = a\, r^j$ (with r known) and $k_n = \left[\frac{1}{5} \mathrm{Log}\, \frac{1}{r} \mathrm{Log}\, n\right] + 1$.

On the other hand, for $\lambda_j = \frac{c}{j^{\lambda}}$, $c > 0$, $\gamma > 1$ and by taking $k_n = [n^{1/(5\gamma+3)}]$, we obtain

$$\frac{n^{\frac{1}{10}\left(\frac{\gamma-1}{\gamma+(3/5)}\right)}}{\mathrm{Log Log}\, n} \|\rho_n(X_n) - \rho(X_n)\| \overset{p}{\to} 0 , \tag{18}$$

which should be compare with (15).

c) The strong mixing case

Proposition 4.6

We make the following assumptions :

a) (X_t) is bounded and geometrically H-strong mixing ;

b) $\lambda_j = a\, r^j$; $a > 0$, $0 < r < 1$; $j = 1,2,\ldots$;

c) A_2, A_3

d) ρ is an Hilbert-Schmidt operator .

Then, if $k_n = o(\mathrm{Log}\, n)$, we have

$$\|\rho_n - \rho\|_{\mathcal{L}} \xrightarrow{a.s.} 0 \ , \quad \|\hat{\rho}_n - \rho\|_{\mathcal{L}} \xrightarrow{a.s.} 0 \ , \tag{19}$$

and

$$\|\rho_n(X_n) - \rho(X_n)\| \xrightarrow{a.s.} 0 \ , \quad \|\hat{\rho}_n(X_n) - \rho(X_n)\| \xrightarrow{a.s.} 0 \ . \tag{20}$$

Rate of convergence : The proof shows that for any $0 < \gamma < \frac{1}{4}$

$$P(\|\rho_n(X_n) - \rho(X_n)\| > \varepsilon) \leq c_1 \exp\left(-c_2\, r^{2k_n}\, n^{\gamma}\right) , \tag{21}$$

and

$$P(\|\hat{\rho}_n(X_n) - \rho(X_n)\| > \varepsilon) \leq c_3 \exp\left(-c_4\, r^{2k_n}\, n^{\gamma}\right) , \tag{22}$$

where the c_i's are positive constants.

Thus, the rate of convergence is $O\left(\exp\left(-c\, n^{\frac{1}{4} - \delta}\right)\right)$ for some $\delta > 0$

In the special case that $\rho v_j = b s^j v_j$, $j=1,2,\ldots$, where b is a constant, it can be proved that $n^{\delta}\|\hat{\rho}_n(X_n) - \rho(X_n)\| \xrightarrow{a.s.} 0$ for any $0 < \delta < \frac{1}{4}\frac{\mathrm{Log}\, s}{\mathrm{Log}\, r}$. (23)

5. NUMERICAL RESULTS

We present here some simulation results together with estimates based on experimental data.

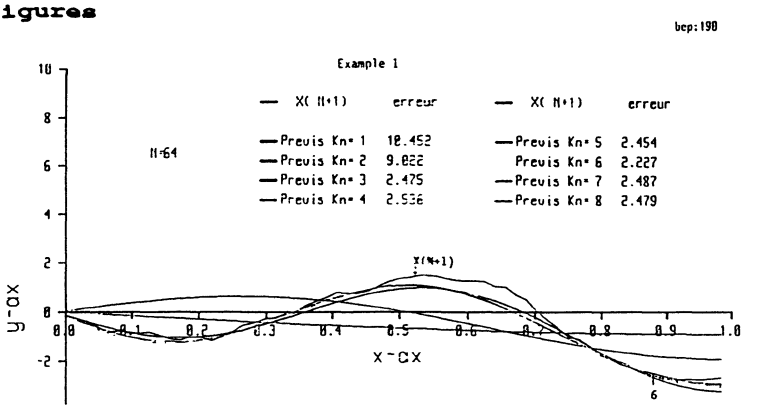
The Hilbert space H is $L^2[0,1]$ and ε_t is the function $u \to W_{t+u} - W_t$, $u \in [0,1]$ where (W_s) is a Wiener process. In Examples 1,3 the v_j's are assumed to be known, in Examples 2,4 and 5 they are considered as unknown.

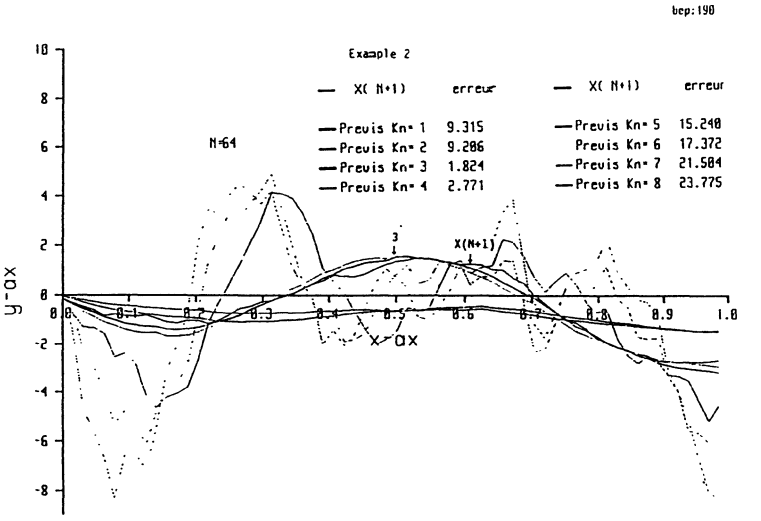
Finally, Example 6 considers the evolution of French train passengers in 1980 using French train traffic data from 1964 to 1979.

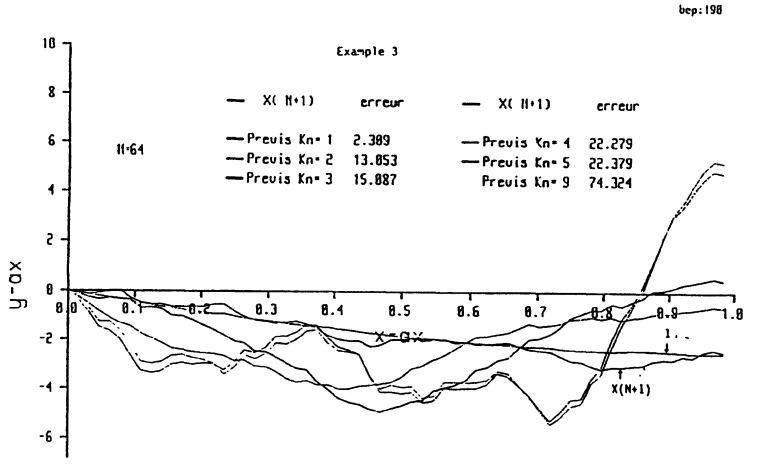
In general, the simulations results can be considered as very good. In Example 8, results are also satisfactory in spite of the fact that size is small.

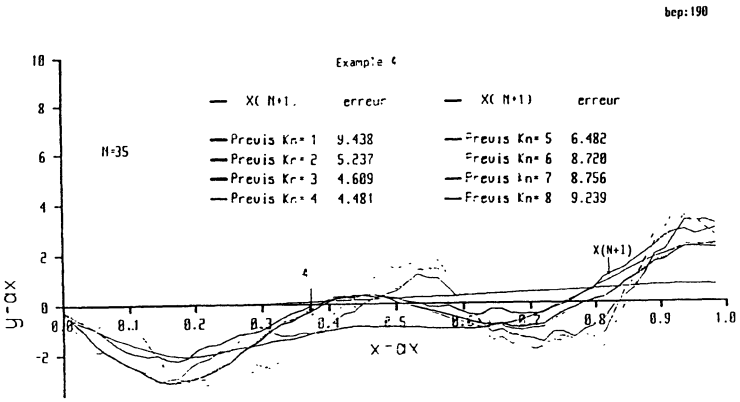
Details about these examples and the choice of k_n are to be found in PUMO (1990).

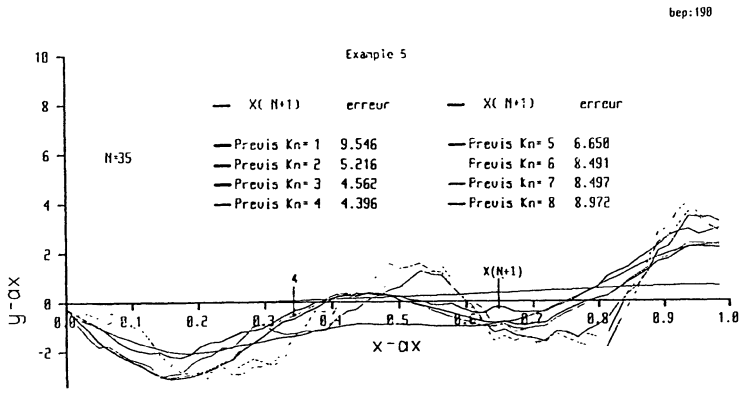
Figures

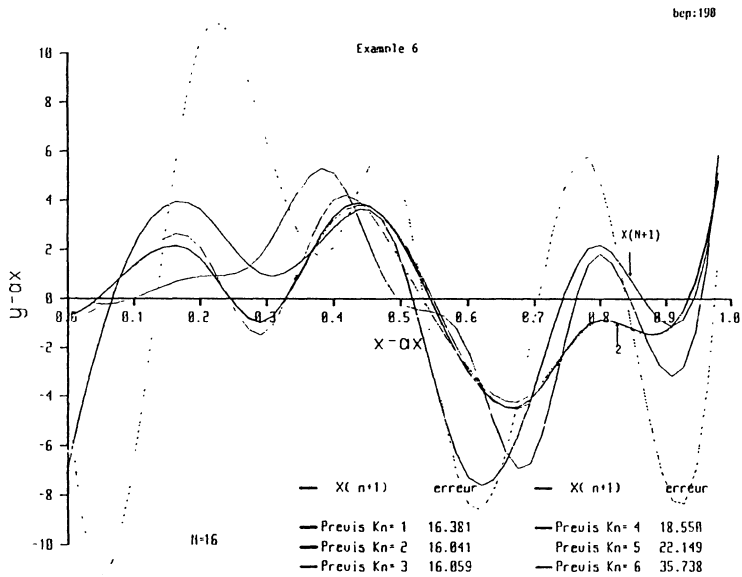

bep:190
Example 1
X(N+1) erreur
X(N+1) erreur
Previs Kn= 1 10.452
Previs Kn= 2 9.022
Previs Kn= 3 2.475
Previs Kn= 4 2.536
Previs Kn= 5 2.454
Previs Kn= 6 2.227
Previs Kn= 7 2.487
Previs Kn= 8 2.479
N=64
X(N+1)
6
y-ox
x-ox


bep:190
Example 2
X(N+1) erreur
X(N+1) erreur
Previs Kn= 1 9.315
Previs Kn= 2 9.206
Previs Kn= 3 1.824
Previs Kn= 4 2.771
Previs Kn= 5 15.240
Previs Kn= 6 17.372
Previs Kn= 7 21.584
Previs Kn= 8 23.775
N=64
3
X(N+1)
y-ox
x-ox


bep:190
Example 3
X(N+1) erreur
X(N+1) erreur
Previs Kn= 1 2.309
Previs Kn= 2 13.053
Previs Kn= 3 15.087
Previs Kn= 4 22.279
Previs Kn= 5 22.379
Previs Kn= 9 74.324
N=64
X(N+1)
y-ox
x-ox


bep:190
Example 4
X(N+1.
erreur
X(N+1)
erreur
N=35
Previs Kn= 1 9.438
Previs Kn= 2 5.237
Previs Kn= 3 4.609
Previs Kn= 4 4.481
Previs Kn= 5 6.482
Previs Kn= 6 8.720
Previs Kn= 7 8.756
Previs Kn= 8 9.239
X(N+1)
y-ax
x-ax


bep:190
Example 5
X(N+1)
erreur
X(N+1)
erreur
N=35
Previs Kn= 1 9.546
Previs Kn= 2 5.216
Previs Kn= 3 4.562
Previs Kn= 4 4.396
Previs Kn= 5 6.650
Previs Kn= 6 8.491
Previs Kn= 7 8.497
Previs Kn= 8 8.972
X(N+1)
y-ax
x-ax


bep:190
Example 6
X(N+1)
y-ax
x-ax
X(n+1)
erreur
X(n+1)
erreur
N=16
Previs Kn= 1 16.381
Previs Kn= 2 16.041
Previs Kn= 3 16.059
Previs Kn= 4 18.558
Previs Kn= 5 22.149
Previs Kn= 6 35.738

6. PROOFS

Proof of Proposition 2.1

Consider the $\mathcal{S}$ valued r.v.

$$Z_t = \langle X_t , . \rangle X_t - C \quad , t = 1,2,\dots .$$

Since $E\|Z_t\|^2_{\mathcal{S}} = E\|X_0\|^4$, we have

$$E\|Z_n+\dots+Z_{n+p-1}\|^2 = pE\|X_0\|^4 + 2\sum_{h=1}^{p-1} (p-h)E\langle Z_0,Z_h\rangle_{\mathcal{S}} . \qquad (24)$$

In order to get a bound for $E\langle Z_0,Z_h\rangle_{\mathcal{S}}$, we write

$$E\langle Z_0,Z_h\rangle_{\mathcal{S}} = \sum_{j,\ell} E(\langle X_0,e_j\rangle\langle X_0,e_\ell\rangle\langle X_h,e_j\rangle\langle X_h,e_\ell\rangle)$$
$$- \sum_{j,\ell} \langle C\, e_j,e_\ell\rangle^2 , \qquad (25)$$

where (e_j) is a complete orthonormal system in H .

Using $X_\ell = \varepsilon_h + \rho(\varepsilon_{h-1}) + \dots + \rho^{h-1}(\varepsilon_1) + \rho^h(X_0)$

and replacing X_ℓ by its value in (24), it yields after some calculations

$$E\langle Z_0,Z_h\rangle_{\mathcal{S}} = \sum_{j,\ell} E(\langle X_0,e_j\rangle\langle X_0,e_\ell\rangle\langle X_0,\rho^h(e_j)\rangle\langle X_0,\rho^h(e_\ell)\rangle$$

$$- \sum_{j,\ell} \langle C\, e_j,e_\ell\rangle\langle(\rho^h C\, \rho^h)\, e_j,e_\ell\rangle .$$

We now specify (e_j) by choosing eigenvectors of ρ of the form :

$$\rho\, e_j = \alpha_j\, e_j \; ; \; j = 1,2,\dots \quad |\alpha_1| \geq |\alpha_2| \geq \dots$$

So we obtain

$$E\langle Z_0,Z_h\rangle_{\mathcal{S}} = \sum_{j,\ell} \alpha_j^h\, \alpha_\ell^h\, E(\langle X_0,e_j\rangle^2\langle X_0,e_\ell\rangle^2)$$

$$- \sum_{j,\ell} \alpha_j^h\, \alpha_\ell^h \langle C\, e_j,e_\ell\rangle^2 .$$

It follows that

$$|E\langle Z_0,Z_h\rangle_{\mathcal{S}}| \leq 2\|\rho\|^{2h}_{\mathcal{L}}\, E\|X_0\|^4 .$$

By using these expressions in (24) we obtain

$$E\|Z_n + \dots + Z_{n+p-1}\|^2_{\mathcal{S}} \leq p\left[E\|X_0\|^4 + \frac{4E\|X_0\|^4\, \|\rho\|^2_{\mathcal{L}}}{1-\|\rho\|^2_{\mathcal{L}}}\right] , \qquad (26)$$

which implies (3).

Furthermore, using (26) and adapting the usual proof of the law of large numbers in the real case (see, e.g. DOOB (1953) pages 492-493) we obtain readily a.s. convergence. ∎

Proof of Proposition 2.2 Let

$$Z'_t = \langle X_t , . \rangle X_{t+1} - D \quad \text{for} \quad t = 1,2,\dots .$$

By (1), and since $D = \rho C$, we have

$$Z'_t = \langle X_t , . \rangle \varepsilon_{t+1} + \rho[\langle X_t , . \rangle X_t - C] .$$

Hence

$$\overline{Z'_n} := \frac{Z'_1 + \dots + Z'_{n-1}}{n-1} = \frac{1}{n-1} \sum_{t=1}^{n-1} \langle X_t , . \rangle \varepsilon_{t+1} + \rho(C_{n-1} - C) . \tag{27}$$

By choosing again a complete orthogonal system in H, it is easy to show that

$$E\left\|\overline{Z'_n}\right\|_{\mathcal{B}}^2 \leq \frac{2\, E\|X_0\|^2\, E\|\varepsilon_0\|^2}{n-1} + 2\, \|\rho\|_{\mathcal{L}}^2\, E\|C_{n-1} - C\|^2 .$$

Using (3), we then obtain (4) .

Now, using (27), we get

$$\|Z'_n\|_{\mathcal{B}} \leq \left\|\frac{1}{n-1} \sum_{t=1}^{n-1} \langle X_t , . \rangle \varepsilon_{t+1}\right\|_{\mathcal{B}} + \|\rho\|_{\mathcal{B}} = \|C_{n-1} - C\|_{\mathcal{B}} .$$

By Proposition 2.1., we have $\|C_{n-1} - C\|_{\mathcal{B}} \xrightarrow{a.s.} 0$ and we can apply the same SLLN as obove since the r.v. $\langle X_t , . \rangle \varepsilon'_{t+1}$s are orthogonal in $L^2_{\mathcal{B}}$.

Proof of Proposition 2.3.

Let $(s_{\ell k})$ be a complete orthogonal system in $\mathcal{B}$, defined by

$$s_{\ell k}(.) = \langle . , v_\ell \rangle v_k \quad ; \ \ell, k = 1,2,\dots .$$

For a given integer ν_n , consider the decomposition

$$\|C_n - C\|_{\mathcal{B}}^2 = \sum_{1 \leq \ell, k \leq \nu_n} (\langle C_n, s_{\ell k} \rangle_{\mathcal{B}} - C, s_{\ell k} \rangle_{\mathcal{B}})^2$$

$$+ \sum_{\ell, k > \nu_n} (\langle C_n, s_{\ell k} \rangle_{\mathcal{B}} - \langle C, s_{\ell k} \rangle_{\mathcal{B}})^2$$

$$=: A_n + B_n .$$

We have

$$P(\|C_n - C\|_{\mathcal{B}} > \varepsilon) \leq \sum_{1 \leq \ell, k \leq \nu_n} P\left(|\langle C_n, s_{\ell k} \rangle_{\mathcal{B}} - \langle C, s_{\ell k} \rangle_{\mathcal{B}}| > \frac{\varepsilon}{\nu_n \sqrt{2}}\right)$$

$$+ \frac{2}{\varepsilon^2} E\, B_n .$$

By using an exponential inequality, due to CARBON (1983), we

find for $\gamma' \in \left]0,\frac{1}{4}\right[$ and $n \geq n_{\gamma'}$,

$$\left(|< C_n-C,\ s_{\ell k} >| > \frac{\varepsilon}{\nu_n\sqrt{2}}\right) \leq 2\ \exp\left(-\ \frac{\varepsilon}{10d^2e\sqrt{2}}\ \frac{n^{\gamma'}}{\nu_n}\right) \quad . \tag{28}$$

On the other hand

$$\begin{aligned} EB_n &\leq \sum_{\ell,k>\nu_n} E[<X_t,v_\ell><X_t,v_k> - E(<X_t,v_\ell><X_t,v_\ell><X_t,v_k>)]^2 \\ &\leq \sum_{\ell,k>\nu_n} E(<X_t,v_\ell>^2\ <X_t,v_k>^2) \\ &\leq \sum_{\ell,k>\nu_n} (E<X_t,v_\ell>^4)^{1/2}(E<X_t,v_k>^4)^{1/2} \\ &\leq d^2 \sum_{\ell,k>\nu_n} (E<X_t,v_\ell>^2)^{1/2}(E<X_t,v_k>^2)^{1/2} \\ &\leq d^2\left(\sum_{j>\nu_n} \lambda_j^{1/2}\right)^2 \\ &\leq \frac{a\ d^2\ r}{(1-\sqrt{r})^2}\ r^{\nu_n} \quad . \end{aligned} \tag{29}$$

Finally, by using (28), (29), and choosing $\nu_n = \left[n^{\frac{\gamma}{2}}\right] + 1$, we obtain (5) .

We omit the proof of (6) which is very similar. ∎

Proof of Lemma 3.1

Since C and C_n are symmetric compact operators, (7) follows from Corollary 2.3, p.31, of GOHBERG-KREJN (1971). We next turn to the proof of (9). First, we have

$$C\ v_{jn} - \lambda_j\ v_{jn} = (C - C_n)\ v_{jn} + (\lambda_{jn} - \lambda_j)\ v_{jn}\ .$$

Thus, by (7) :

$$\|C\ v_{jn} - \lambda_j\ v_{jn}\| \leq 2\|C_n - C\| \quad . \tag{30}$$

Now, if $j > 1$,

$$\begin{aligned} \|C\ v_{jn} - \lambda_j\ v_{jn}\|^2 &= \sum_{\ell=1}^{\infty} (<C\ v_{jn},v_\ell> - <\lambda_j\ v_{jn},v_\ell>) \\ &= (<v_{jn},v_j> - sg<v_{jn},v_j>)^2 + \sum_{\ell\neq j} <v_{jn},v_\ell>^2 \\ &\leq (1-|<v_{jn},v_j>|^2) + \sum_{\ell\neq j} <v_{jn},v_\ell>^2 \end{aligned}$$

$$\leq 2 \sum_{\ell \neq j} \langle v_{jn}, v_\ell \rangle^2 \tag{32}$$

Now, by using (30), (31), (32) we get (9).

The proof of (8) is very similar and therefore omitted.

Proof of Proposition 3.2 : It follows easily from Proposition 2.1. and (7). ∎

Proof of Proposition 3.3 : Is a consequence of Propositions 2.1., 2.3 and inequalities (8) and (9). ∎

Proof of Lemma 4.1

We start with the inequality

$$\left\| \rho_n - \pi^{k_n} \rho \pi^{k_n} \right\| \leq \left\| \Gamma_n^{-1} - G_n^{-1} \right\| \left\| \Delta_n^* \right\| + \left\| G_n^{-1} \right\| \left\| \Delta_n^* - H_n^* \right\| ,$$

where $\|.\|$ denotes $\|.\|_{\mathcal{L}}$, and we estimate each summand separately.

First, note that $\left\| G_n^{-1} \right\| = \lambda_{k_n}^{-1}$. Therefore if $\|\Gamma_n - G_n\| < \lambda_{k_n}$, then $1 - \left\| G_n^{-1} \right\| \|\Gamma_n - G_n\| > 0$ an by a well known inequality for linear operators, we obtain

$$\left\| \Gamma_n^{-1} - G_n^{-1} \right\| \leq \frac{\left\| G_n^{-1} \right\|^2 \|\Gamma_n - G_n\|}{1 - \left\| G_n^{-1} \right\| \|\Gamma_n - G_n\|} \quad \text{a.s.} \tag{33}$$

(see, e.g., DUNFORD-SCHWARTZ (1958), p. 585).

Note that, by (3),

$$\begin{aligned} P(\|\Gamma_n - G_n\| \geq \lambda_{k_n}) &\leq P(\|C_n - C\|_{\mathcal{B}} \geq \lambda_{k_n}) \\ &\leq A\, n^{-1} \lambda_{k_n}^{-2} \quad . \end{aligned} \tag{34}$$

Using now (33), we see that, for $n > 0$, $E_n = \left\{ \left\| \Gamma_n^{-1} - G_n^{-1} \right\| > \eta,\ \|\Gamma_n - G_n\| < \lambda_{k_n} \right\}$

implies $$\frac{\lambda_{k_n}^{-2} \|\Gamma_n - G_n\|}{1 - \lambda_{k_n}^{-1} \|\Gamma_n - G_n\|} \geq \eta ,$$

which in turn implies

$$\|\Gamma_n - G_n\| > \lambda_{k_n}^2 \eta \left(1 + \lambda_{k_n} \eta\right)^{-1} \quad .$$

Thus, using again (3) ,

$$P(E_n) \leq \frac{E\|\Gamma_n - G_n\|^2 \left(1 + \lambda_{k_n} \eta\right)^2}{\lambda_{k_n}^4 \eta^2}$$

$$\leq \frac{A\left(1+\lambda_{k_n}\eta\right)^2}{\eta^2}\,\frac{1}{n\,\lambda_{k_n}^4} \quad . \tag{35}$$

Finally, using (34) and (35), we obtain

$$P\left(\left\|\Gamma_n^{-1} - G_n^{-1}\right\| > \eta\right) \leq \frac{A\left(1+\lambda_{k_n}\eta\right)^2}{\eta^2}\,\frac{1}{n\,\lambda_{k_n}^4} + \frac{A}{n\,\lambda_{k_n}^2} \quad . \tag{36}$$

On the other hand, by (4), we get

$$P\left(\left\|\Delta_n^*\right\| > \|D^*\|+1\right) \leq P(\|D_n-D\| > \|D\|+1) \leq \frac{2B}{n} \quad , \tag{37}$$

and

$$P\left(\left\|G_n^{-1}\right\|\,\left\|\Delta_n^* - H_n^*\right\| > \eta(\|D^*\|+1)\right) \leq \frac{2B}{n\,\lambda_{k_n}^2\,\eta^2\,(\|D\|+1)^2} \quad . \tag{38}$$

Combining (36), (37), and (38), we obtain a bound for $P\left(\left\|\rho_n-\pi^{k_n}\rho\,\pi^{k_n}\right\| > 2\eta(\|D\|+1)\right)$. By taking $\eta = \frac{\varepsilon}{2(\|D\|+1)}$, we get the desired result. ∎

Proof of Lemma 4.2

Note that

$$\begin{aligned} E < \rho(X_n),\, v_j >^2 &= E < X_n\,,\, \rho(v_j) >^2 \\ &= < C[\rho(v_j)],\, \rho(v_j) > \\ &= < (\rho\, C\, \rho)(v_j), v_j > . \end{aligned}$$

Hence the first result is proved by the equality

$$E\left\|\rho(X_n) - \left(\pi^{k_n}\rho\right)(X_n)\right\|^2 = \sum_{j=k_n+1}^{\infty} E< \rho(X_n), v_j >^2 .$$

On the other hand it is easy to see that

$$C = \rho\, C\, \rho + C_\varepsilon \ ,$$

where C_ε denotes the covariance operator of ε. Thus,

$$\begin{aligned} \lambda_j = < C\, v_j, v_j > &= <(\rho\, C\, \rho)(v_j), v_j > + < C_\varepsilon\, v_j, v_j > \\ &\geq <(\rho\, C\, \rho)(v_j, v_j)> , \end{aligned}$$

which implies (12). ∎

Proof of Lemma 4.3

Note that

$$\begin{aligned} \|\rho_n(X_n)-\rho(X_n)\| \leq &\left\|\rho_n-\pi^{k_n}\rho\,\pi^{k_n}\right\|\,\|X_n\| \\ &+ \left\|\pi^{k_n}\rho\right\|\,\left\|\pi^{k_n}(X_n) - X_n\right\| \\ &+ \left\|\left(\pi^{k_n}\rho\right)(X_n) - \rho(X_n)\right\| \quad . \end{aligned}$$

By Lemmas 4.1. and 4.2., the first and third summands tend to zero in probability. For the second one, it suffices to note that

$$P\left(\left\|\pi^{k_n}(X_n) - X_n\right\| > \eta\right) \leq \frac{1}{\eta^2} \sum_{k_n+1}^{\infty} \lambda_j \quad . \quad \blacksquare$$

Proof of Lemma 4.4

Put $r_n = G_n^{-1} H_n^*$ and consider

$$\hat{\rho}_n - r_n = \hat{\Gamma}_n^{-1}\left(\hat{\Delta}_n^* - H_n^*\right) + \left(\hat{\Gamma}_n^{-1} - G_n^{-1}\right) H_n^* \quad .$$

We clearly have

$$\|\hat{\rho}_n - r_n\| \leq \lambda_{k_n,n}^{-1} \|D_n - D\| + \|D\| \frac{\|C_n - C\|}{\lambda_{k_n} \lambda_{K_n,n}} \quad . \qquad (40)$$

Consider now

$$r_n - \rho_n = G_n^{-1} \left(H_n^* - \Delta_n^*\right) + \left(G_n^{-1} - \Gamma_n^{-1}\right) \Delta_n^* \quad .$$

We have

$$\|r_n - \rho_n\| \leq \lambda_{k_n}^{-1} \, 2\|D\| \left\|\pi^{k_n} - \hat{\pi}^{k_n}\right\| + 2 \|D\| \sum_1^{k_n} \left\|v_{j_n} - v_j'\right\| \quad .$$

Using (8) and (9), we obtain,

$$\|r_n - \rho_n\| \leq \|D\| \left(\sum_1^{k_n} a_j\right) \left(1 + \lambda_{k_n}^{-1}\right) \|C_n - C\| \ , \qquad (41)$$

and therefore (40) and (41) imply (16). $\blacksquare$

Proof of Proposition 4.5

By Proposition 4.3, we need only show that $\|\hat{\rho}_n - \rho_n\| \xrightarrow{P} 0$. For this purpose, we use (16).

It is easy to check that $\sum_{j=1}^{k_n} a_j = O\left(r^{-k_n}\right)$. Define now $F_n = \left\{\frac{\lambda_{k_n}}{2} < \lambda_{k_n,n} < \frac{3\,\lambda_{k_n}}{2}\right\}$. Then, by (3), we have

$$P\left(F_n^c\right) = P\left\{\left|\lambda_{k_n} - \lambda_{k_n,n}\right| > \frac{\lambda_{k_n}}{2}\right\} \leq P\left(\|C_n - C\| > \frac{\lambda_{k_n}}{2}\right)$$

$$\leq \frac{4\,A}{n\,\lambda_{k_n}^2} \quad . \qquad (42)$$

Therefore by (16) and (42) and for each $\varepsilon > 0$,

$$P(\|\hat{\rho}_n-\rho_n\| > \varepsilon) \leq P(\|\hat{\rho}_n-\rho_n\| > \varepsilon,\ F_n) + P\left(F_n^c\right)$$

$$\leq \frac{c_1(\varepsilon)}{n\ r^{2k_n}} + \frac{c_2(\varepsilon)}{n\ r^{4k_n}} + \frac{c_3(\varepsilon)}{n\ r^{2k_n}} \quad , \qquad (43)$$

where the $c_i(\varepsilon)$ $i=1,2,3$ are strictly positive constants, and the conclusions follow easily. ∎

Proof of Proposition 4.6

The proof is essentially the same as that of Propositions 4.3. and 4.5, with the exception that we use (5) and (6) instead of (3) and (4).

We have, in the first place

$$P(E_n) \leq 2n^{\gamma} \exp\left(- \frac{a^2\ r^{2k_n}\ \eta\ n^{\gamma}}{\left(1+\lambda_{k_n}\ \eta\right)\ 10\ d^2\ e\ \sqrt{2}}\right)$$

$$+ \frac{2ad^2 r\left(1+\lambda_{k_n}\ \eta\right)^{-2}}{\lambda_{k_n}^4\ \eta^2\ (1-\sqrt{r})^2} \exp\left(-\mathrm{Log}\ \frac{1}{r}\ n^{\gamma}\right)$$

(where we recall that $E_n = \left\{\left\|\Gamma_n^{-1} - G_n^{-1}\right\| > \eta,\ \|\Gamma_n - G_n\| < \lambda_{k_n}\right\}$).

Following the proof of Lemma 4.1. and by using (6), it is easy to see that $\left\|\rho_n - \pi^{k_n}\ \rho\ \pi^{k_n}\right\| \xrightarrow{a.s.} 0$. (44)

Using the Hilbert-Schmidt property of ρ, we get after some calculations

$$\left\|\pi^{k_n}\ \rho\ \pi^{k_n} - \pi^{k_n}\ \rho\right\|_{\mathcal{S}}^2 \leq \sum_{j > k_n} \|\rho\ v_j\|^2 \to 0 \ , \qquad (45)$$

and

$$\left\|\rho - \pi^{k_n}\ \rho\right\|_{\mathcal{S}} \leq \sum_{j > k_n} \|\rho\ v_j\|^2 \to 0 \ . \qquad (46)$$

From (44), (45), (46) we deduce

$$\|\rho_n - \rho\| \xrightarrow{a.s.} 0 \ .$$

On the other hand, the same argument as those in the proof of Lemma 4.4. and Proposition 4.5. leads to

$$P\left(F_n^c\right) \leq 2\ n^{\gamma}\ \exp\left(-\ \frac{a^2\ r^{k_n}\ n^{\gamma}}{10\ d^2\ e\ \sqrt{2}}\right)$$

$$+\ \frac{4\ a\ d^2\ r}{\lambda_{k_n}^2\ (1-\sqrt{r})}\ \exp\left(-\mathrm{Log}\ \frac{1}{r}\ n^{\gamma}\right)\ .$$

Assuming that F_n holds, we have the bounds

$$\alpha_n \leq b\ r^{-2k_n}\ ,\quad \beta_n \leq b\ r^{-k_n}\ ,$$

where b is a constant.

Therefore, by (16)

$$P(\|\hat{\rho}_n - \rho_n\| > \varepsilon\ ,\ F_n) \leq P\left(b\ r^{-k_n}\|D_n - D\| > \frac{\varepsilon}{2}\right)$$

$$+\ P\left(b\ r^{-2k_n}\|C_n - C\| > \frac{\varepsilon}{2}\right)$$

and the conclusions follow easily from (5) and (6). ∎

Acknowledgement : I would like to thank Professor P. Deheuvels and Professor G. Roussas for helpful comments.

References

[1] ANTONIADIS A. - BEDER J.H. - (1989) - Joint estimation of the mean and the covariance of a Banach valued Gaussian vector - Statistics 20, 1, 77-93.

[2] BOSQ D. - (1989) - Propriétés des opérateurs de covariance empiriques d'un processus stationnaire hilbertien - C.R. Acad. Sci. Paris, I, 309, 873-875.

[3] BOSQ D. - (1990) - Modèle autorégressif hilbertien. Application à la prévision d'un processus à temps continu sur un intervalle. Preprint.

[4] BOSQ D. - DELECROIX M. (1985) - Nonparametric prediction of a Hilbert space valued random variable - Stochastic Process. Appl., 19, 271-280.

[5] CARBON M. (1983) - Inégalité de Bernstein pour les processus fortement mélangeants non nécessairement stationnaires - C.R. Acad. Sci. Paris I, 297, 303-306.

[6] DAUXOIS J. - POUSSE A. - ROMAIN Y. (1982) - Asymptotic theory for the principal component analysis of a vector random function : some applications to statistical inference. J. Multivariate Anal. 12, 136-154.

[7] DEHLING H. - (1983) - Limit theorem for sums of weakly dependent Banach space valued random variables. Z. Wahrscheinlichkeitstheorie und verw. Gebiete, 63, 393-432.

[8] DOOB J. (1953) - Stochastic Processes - Wiley, New-York.
[9] DUNFORD N. - SCHWARTZ J.T. (1958) - Linear operators, I, Wiley, New-York.
[10] GOHBERG I.C. - KREIN M.G. (1971) - Opérateurs linéaires non auto-adjoints dans un espace Hilbertien - Dunod.
[11] GRENANDER U. (1981) - Abstract inference - Wiley, New-York.
[12] PUMO B. (1990) - Etude pratique du problème de l'estimation et de la prévision d'un processus autorégressif d'ordre 1 . Preprint, Université Pierre et Marie Curie, Paris.

ESTIMATION OF CHAOTIC DYNAMIC SYSTEMS WITH CONTROL VARIABLES

ALEXANDER A. GEORGIEV
Department of Biostatistics, Epidemiology and Systems Science
Medical University of South Carolina
Charleston, SC 29425-2503
U.S.A.

ABSTRACT. In this paper we develop some nonparametric models for chaotic dynamic systems with control variables. The theory of chaotic dynamic systems suggests that complicated or apparently stochastic phenomenon can be explained through a simple nonlinear model. There is physically a large class of nonlinear systems with input or control variables. An important task in many practical situations is to create reasonable techniques for modeling chaotic dynamic systems on the basis of data without any assumptions about the structure of the system. In recent years, various parametric methods have been developed for estimation of chaotic systems. The purpose of this paper is to survey selectively some results in the area of nonlinear system estimation and to present one application of kernel nonparametric techniques to real data set.

1. Introduction

Chaos is the irregular behavior of simple equations, and irregular behavior is ubiquitous in nature. A primary motivation for studying chaos is this: Given an observation of irregular behavior, is there a simple explanation that can account for it? And if so, how simple? There is a growing consensus that a useful understanding of the physical world will require more than finally uncovering the fundamental laws of physics. Simple systems, which obey simple laws, can nonetheless exhibit exotic and unexpected behavior. Nature is filled with surprises that turn out to be direct consequences of Newton's laws.

Physically, a dynamic system is anything that moves. Mathematically, a dynamic system is defined by a state space R^d that describes the instantaneous states available to the system and evolution operator ϕ_t that tells how the state of the system changes in time. One usually thinks of ϕ_t as the physics of the system. An element $y \in R^d$ of the state space specifies the current state of the system, and d characterizes the number of degrees of freedom in the system. For a particle on space, y might represent the three coordinates of the particle's position, and the three coordinates of its momentum, for a total of d=6 degrees of freedom. The evolution operator is a family of functions $\phi_t : R^d \to R^d$ that map the current state of the system into its future state at time t units later. The operator ϕ_t satisfies $\phi_0(y) = y$ and $\phi_{t+s}(y) = \phi_t [\phi_s(y)]$. The dynamic system is nonlinear if, in general,

$$\phi_t (c_1 y' + c_2 y'') \neq c_1 \phi_t (y') + c_2 \phi_t (y'').$$

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 531–540.
© 1991 *Kluwer Academic Publishers.*

The function ϕ_t can be defined either as a discrete map or in terms of a set of ordinary difference or differential equations. Partial differential equations and differential delay equations have also been studied (in these last two cases, d is infinite). In the rest of this paper we will limit our analysis only to the difference equations. One possible example of chaotic dynamic system with control variable is the Hénon map (Hénon, 1976)

$$y_{n+1} = 1.0 - a\, y_n^2 + u_n ,$$

$$u_{n+1} = b\, y_n ,$$

with a = 1.4 and b= 0.3. The reader is referred to Gleick's (1987) best-seller *Chaos* as the easiest and undoubtedly the most enjoyable book from which to start learning about chaos, and the dynamic personalities of scientists in the field. Reviews of chaos includes Eckmann and Ruelle (1985), Shaw (1981), Crutchfield *et al.* (1986), Ott (1981), and Holden (1986). A recommended text with many references is Schuster (1988). See also Abraham *et al.* (1984) for an early review of experimental observations of chaos.

The theory of chaotic dynamic systems suggests that complicated or apparently stochastic phenomenon can be explained through a simple nonlinear model. The main problem is how to determine the class of the nonlinear model. In recent years, various parametric methods have been developed for estimation of the nonlinear systems. Normally, the nonlinear parametric models are useful in the analysis of the chaotic systems for a particular set of data, but fail for others. This is because the choice of the parametric nonlinear model is only an approximation of the nature. The problem becomes more complicated if you consider modeling of chaotic systems with control variables. This type of system is more frequently used in real-life problems.

Since the analysis of chaotic controlled systems is often dependent on finding a mathematical description that defines the relationship between the system input and output, systems estimation is of fundamental importance in chaotic systems. Most systems encountered in practice are nonlinear to some extent, and although it may be possible to represent systems that are only *'mildly'* nonlinear or which operate over a restricted range by a linear model, in general, nonlinear processes can only be adequately characterized over their whole operating range by a nonlinear model.

There are mainly three different approaches for discrete-time estimating of nonlinear systems and these are based on functional series expansions, block oriented systems, and nonlinear difference equations. The functional series methods can be used to represent a very broad class of nonlinear systems, but this is achieved at the expense of creating a large parameter set and of the excessive computational effort. Block structured systems represent a much smaller class of systems than functional series. The model building for block systems can only be applied if the system has the known structure. We restrict our considerations only to the nonlinear difference equation models.

Our goal is to develop nonparametric models for chaotic systems with control variables on the basis of very little *a priori* information. The typical assumption is that no structures or functional forms of the system characteristics are known. The main idea is "to let the data speak for themselves" without forcing too much structure limitation. In the

past few years, several studies have been done to employ nonparametric estimation in system modeling.

There are generally two nonparametric approaches to system modeling: modeling in terms of unparametrized characteristics such as the impulse response or frequency response, and modeling of systems by techniques derived from nonparametric estimates of autoregression or regression functions. This last time domain approach has been attracting more attention. The reader is referred to Georgiev (1984a, 1984b, 1985, 1989a, 1989b), Greblicki and Pawlak (1986, 1987, 1989a, 1989b), Krzyzak (1990), Rutkowski (1982, 1984, 1985), Yakowitz (1985, 1987, 1989) and references there.

In the present study, we report selectively a few results about the class of chaotic systems which can be estimated by nonparametric methodology and the application of nonparametric kernel techniques for estimating a real system (a turbogenerator set). We will use later the terms modeling and estimation of systems as a synonyms.

2. Stochastic Models for Controlled Chaotic Systems

Consider a chaotic system described by the following difference equation of order p

$$y(t+p) = m[y(t+p-1), ..., y(t), u(t+p-1), ..., u(t)] \ . \tag{2.1}$$

There are at last two ways for generalization of the model (2.1) to the stochastic case. First, assume that the input sequence (control variables u(t)) is a realization of some stochastic process. We assume that inputs are independent, identically distributed random variables (i.i.d.r.v.). We obtain an output sequence, which will be also a realization of stochastic process. Let all stochastic processes be defined on the same probability space (Ω, A, P). The dynamics of the system is represented by the following difference equation

$$Y(t+p) = m[Y(t+p-1), ..., Y(t), U(t+p-1), ..., U(t)] \ , \qquad t=1,2, ..., n. \tag{2.2}$$

We will denote by capital letters all random variables.

Second, in many situations in the practice, we can observe that the system and/or the measurements of the input-output sequence are distributed into unknown ways. These disturbances are not measurable and we do not know exactly their influence on the system. Therefore, the next step of generalization of the recursive equation (2.2) is to introduce an additive noise, i.e.

$$Y(t+p) = m[Y(t+p-1), ..., Y(t), U(t+p-1), ..., U(t)] + Z(t+p-1), \qquad t = 1,2,...,n \ . \tag{2.3}$$

The noise Z(t) is taken additively to facilitate the analysis.

We assume the disturbances (or system noise) Z(t) form an innovations sequence. Given an increasing sequence of σ-algebras $\{A_t\}$ in the probability space (Ω, A, P), a sequence $\{Z(t)\}$ of random variables adapted to $\{A_t\}$ is said to be an *innovations sequence* if

$$E[Z(t) \mid A_{t-1}] = 0 \quad \text{for all t.}$$

The main object in estimating the chaotic system represented by (2.3) is to determine the behavior of the output Y(t+p) given {Y(t+p-1), ..., Y(t), U(t+p-1), ..., U(t)} for some p > 0. The stochastic model representation (2.3) was introduced for the systems of order one (p = 1) by Georgiev (1984a) in the context of kernel modeling. He studied first a kernel model for the nonlinear function m[·] using the conditional mean of Y(t+1) given {Y(t), U(t)}.

We see that, considering different stochastic interpretations of the input and output sequences, the chaotic system is transformed mathematically into a stochastic nonlinear input-output model (2.3). This model is the generalization to the nonlinear case of the model

$$Y(t+p) = a_1 Y(t+p-1) + \ldots + a_p Y(t) + b_1 U(t\ p-1) + \ldots + b_p U(t) + Z(t)\ .$$

The above linear model, usually called the ARMAX (AutoRegressive Moving Average with eXogenous inputs) model or *controlled* ARMA model, is the standard model used in estimation of linear dynamic systems (see Ljung (1987) and Priestley (1981)). The nonlinear model (2.3) for chaotic systems can be called the NARMAX (Nonlinear ARMAX) model. It is a model that can describe a wide class of chaotic systems encountered in the nature.

3. Characterization of Stochastic Models

In this section, we will be concerned with the probabilistic properties of the input-output stochastic processes associated with the stochastic chaotic system (2.3). We will look for the class of nonlinear functions m[·] for which the nonparametric estimation methodology is applicable. In addition, we show that the distribution of the innovation sequence {Z(t)} is also important for the properties of the system (2.3). We will consider only first order systems for the ease of exposition. With more complicated notation all of the results below can be generalized for the p-order systems.

The nonparametric estimation of discrete-time processes, described by NARMAX equation

$$Y(t+1) = m[Y(t), U(t)] + Z(t)\ , \tag{3.1}$$

is a natural generalization of the nonparametric estimation of a regression function. Normally, we are dealing only with a *single* realization of the input-output stochastic process, i.e. {U(t), Y(t); t ≥ 0}. When we estimate regression function (static system, or memoryless system), the outputs {Y(t), t ≥ 0} form a sequence of independent random variables. In the NARMAX equation (3.1), the random disturbance (system noise) {Z(t), t ≥ 0} is independent of the sequence {U(t), Y(t), t ≥ 0} having the following properties

$$\left.\begin{array}{ll} E\,Z(t) = 0 & \text{for all } t \geq 0\ , \\ \multicolumn{2}{l}{Z(t) \text{ and } Z(t') \text{ are independent, identically distributed,}} \\ \multicolumn{2}{l}{\text{random variables (i.i.d.r.v) for } t \neq t', \text{ all } t, t' \geq 0\ .} \end{array}\right\} \qquad (3.2)$$

When we are dealing with estimation of dynamic systems, the output $Y(t+1)$ depends upon the previous output $Y(t)$ and the input $U(t)$. Since we use only a single realization of the process $\{U(t), Y(t); t \geq 0\}$, the following question arises:

Problem:
For what class of NARMAX equations (3.1) and distributions of the system noise (3.2) can we estimate the chaotic system on the basis of single realization of the input-output process?

Result: (Georgiev, 1989a)
Sufficient conditions for estimation of the chaotic system (3.1) on the basis of single realization of the input-output process are that there exists a constant $c > 0$, such that:

a) $$\begin{array}{l} E\,|\,m[y, U(t)]\,| + E\,|\,Z(t)\,| < \infty\ , \text{ for } |y| \leq c\ ; \\ E\,|\,m[y, U(t)]\,| \leq |y| - E\,|\,Z(t)\,|\ , \text{ for } |y| > c\ ; \end{array} \qquad (3.3)$$

b) $m(y,u)$ is continuous on the y for all u. (3.4)

Example:
Consider the following controlled chaotic system

$$Y(t+1) = \{\psi + \pi \exp(-\gamma\, Y^2(t))\}\, Y(t) + U(t) + Z(t)\ , \quad \gamma \geq 0\ , \qquad (3.5)$$

where the input $U(t)$ and the noise $Z(t)$ have a density function with infinite support and symmetric about zero. $\{U(t)\}$ and $\{Z(t)\}$ form sequences of i.i.d.r.v. The reader may check that the system (3.5) satisfies the conditions of the above result.

The interesting fact, it should be noted, is that the conditions (3.3) and (3.4) are closely related with the stability and chaotic properties of the system.

4. Nonparametric Estimation of Controlled Chaotic Systems

The nonparametric estimate of the unknown function $m(.,.)$ was first introduced and theoretically developed by Georgiev (1984a). A brief description of the algorithm is added below for better insight. To spare notation, the version with $p = 1$ (first order) is presented.

Assume $m(.,.)$ is constant in the neighborhood A of a fixed point (y_0, u_0). Let $J = \{i: (Y(i), U(i)) \in A,\ 1 \leq i \leq n\}$. The estimate of m is the average of $Y(i+1)$, where $(Y(i), U(i)) \in A$. Define $|J| = \text{card } J$. We have

$$\sum_{i \in J} Y(i+1)/\,|J| = \sum_{i=1}^{n} Y(i+1) I_A(Y(i), U(i)) \Big/ \sum_{i=1}^{n} I_A(Y(i), U(i)),$$

where $I_A(.)$ is an indicator function. When m is not constant, we take

$$A = A_n = \{(y, u): ((y_0 - y)/h_n, (u_0 - u)/h_n) \in B\},$$

where B is a small neighborhood of 0 and h_n is a sequence of positive numbers which tend to zero, as $n \to \infty$. The estimate we obtain in this case is

$$m_n(y_0, u_0) = \sum_{i=1}^{n} Y(i+1) K((y_0 - Y(i))/h_n, (u_0 - U(i))/h_n) \Big/ \sum_{i=1}^{n} K((y_0 - Y(i))/h_n, (u_0 - U(i))/h_n), \tag{4.1}$$

where $K = I_A$. There is no need for the K (so-called kernel) to be a set indicator; one may take K as a probability density function, bounded on R^2. In our experimental study, we use Gaussian kernel defined on R^8 (p = 4).

5. Experimental Results

5.1 THE NONLINEAR SYSTEM AND DATA ACQUISITION

The efficiency of proposed kernel method is demonstrated by estimating a turbogenerator set. The turbogenerator set (Unbehauen, 1985) consists of an air pressure turbine driving a synchronous generator. Two output signals, the frequency Y_1 and phase voltage Y_2, are excited by two input signals, the field current U_1 of the synchronous generator, and the pressure U_2 of the turbine. The generator supplies a balanced load. The preliminary developments of the turbogenerator set indicate a strong nonlinear dynamic behavior of the plant.

To generate the data series used for modeling, the following operation point was adjusted: field current U_1 = 0.4A, actuating valve U_2 = 60%, frequency Y_1 = 800 Hz, phase voltage Y_2 = 11 V. The plant is extended by a process interface including the actuators and measurement devices. The inputs are white uniform distributed noises and are generated by the process computer through digital-analog converters in the ranges $\Delta U_1 = \pm 0.3$ A, and $\Delta U_2 = 0\%$. The measurements of the dynamic outputs are in the ranges $\Delta Y_1 = \pm 200$ Hz, $\Delta Y_2 = \pm 10$ V, and are logged via analog-digital conversion to the process computer. All input and output dynamic signals are normalized to the range $\pm$ 1 for the computation convenience. Two data series of 500 measuring values are generated, where the sampling rate is 0.1 sec., with different bandwidth of input signals.

5.2 RESULTS

In the present experimental study, only single-input-single-output characteristics are developed. More specifically, the model for the U_1/Y_2 (respectively field current/voltage) was estimated.

For this case two data series were generated. One series (Fig. 1) was applied for estimation, using the estimate (4.1), and one (Fig. 2) for validation of the estimate. The input signals of the estimation and validation data sets consist of white uniformly distributed noise with amplitudes $\Delta U_1 = \pm 1$ ($\pm$ 0.3 A), but different bandwidth. The U_2 is held constant at the operation point.

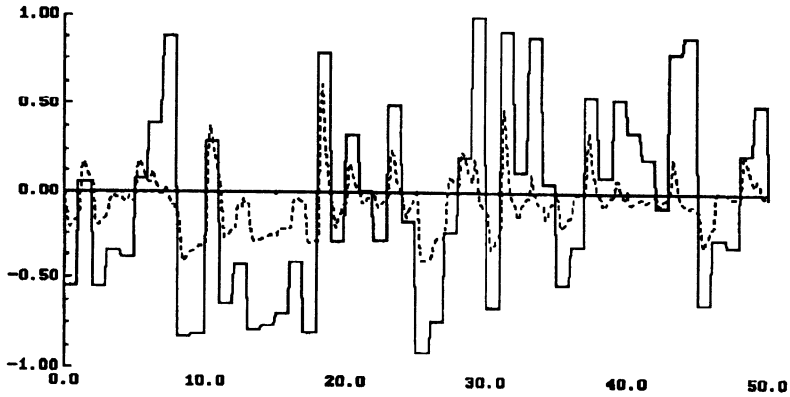


Fig. 1. Estimation data: —— input, ---- output.

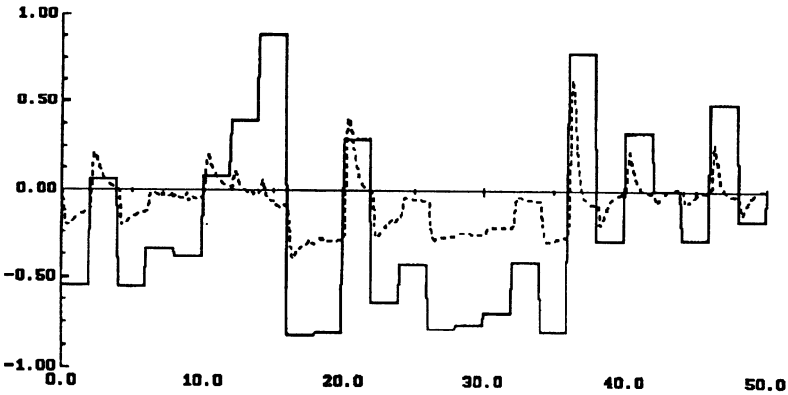


Fig. 2. Validation data: —— input, ---- output.

To develop the efficiency of the nonparametric kernel estimate of dynamic relationship U_1/Y_2, we estimate m by m_n (4.1) and estimation data set (Fig. 1). The kernel is Gaussian, the order p = 4 and the h_n = 0.2. The fitted model m_n is then used as one-step-ahead predictor for validation data set (Fig. 2). A comparison between the output obtained by the predictor and true output is shown in Fig. 3.

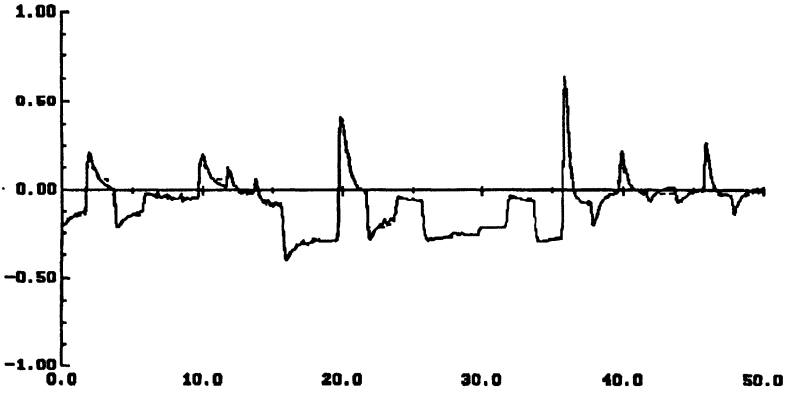


Fig. 3. Comparison: — system's response, --- model's response.

The difference between predicted and true voltage Y_2 is shown in Fig. 4.

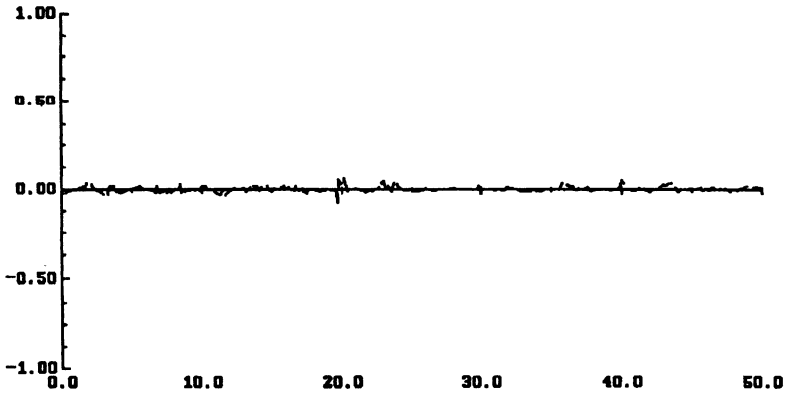


Fig. 4. The difference.

6. Conclusions

We present the usefulness of the nonparametric methodology in the analysis of chaotic systems with control variables. The fields for application are diverse, including ecology, population dynamics, engineering, etc. A characterization of the estimation problem of nonlinear stochastic systems is given. The estimation of single-input-single-output dynamic relationships in turbogenerator set, as a case study of controlled chaotic systems, has illustrated prediction accuracy of the nonparametric estimation approach introduced in Georgiev (1984a). The main reason for this enormous improvement, with comparison to the parameter models, is that the dynamic nonlinearity of the system was significantly better represented by the nonparametric algorithm.

References

Abraham, N.B., Gollub, J.P., and Swinney, H.L. (1984) "Testing nonlinear dynamics", *Physica*, **11D**, 252.

Crutchfield, J.P., Farmer, J.D., Packard, N.H., and Shaw, S. (1986) "Chaos", *Sci. Am.*, **255**, 46.

Eckmann, J.-P. and Ruelle, D. (1985) "Ergodic theory of chaos and strange attractors", *Rev. Mod. Phys*, **57**, 617.

Georgiev, A.A. (1984a) "Nonparametric system identification by kernel methods", *IEEE Trans. Autom. Control*, **29**, 356-358.

Georgiev, A.A. (1984b) "A nonparametric algorithm for identification of linear dynamic SISO systems of unknown order", *Syst. Control Letters*, **4**, 273-280.

Georgiev, A.A. (1985) "Identification of spatially variable parameters in linear distributed systems via kernel methods: A convergence study", in H.A. Barker, and P.C. Young (eds.), *Identification and System Parameter Estimation 1985*, Pergamon Press, Oxford-New York, pp. 775-779.

Georgiev, A.A. (1989a) *Modeling of Discrete-Time Processes from Stochastic Input-Output Data: A Kernel Approach*, In preparation.

Georgiev, A.A. (1989b) "On the identification of time-varying dynamic systems by nonparametric procedures", *IEEE Trans. Automat. Control*, **34**, 223-225.

Gleick, J. (1987) *Chaos: Making a New Science*, Viking, New York.

Greblicki, W. and Pawlak, M. (1986) "Identification of discrete Hammerstein systems using kernel regression estimates", *IEEE Trans. Automat. Control*, **31**, 74-77.

Greblicki, W. and Pawlak, M. (1987) "Hammerstein system identification by non-parametric regression estimation", *Int. J. Control*, **45**, 343-354.

Greblicki, W. and Pawlak, M. (1989a) "Nonparametric identification of Hammerstein systems ", *IEEE Trans. Informat. Theory*, **35**, 409-418.

Greblicki, W. and Pawlak, M. (1989b) "Recursive nonparametric identification of Hammerstein systems ", *J. Franklin Instit.*, **326**, 461-481.

Krzyzak, A. (1990) "On the estimation of a class of nonlinear systems by the kernel regression estimate", *IEEE Trans. Informat. Theory*, **36**, 141-152.

Ljung, L. (1987) *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs N.J..

Hénon, M. (1976) "A two-dimensional mapping with a strange attractor", *Commun. Math. Phys.*, **50**, 69.

Holden, A.V., ed. (1986) *Chaos*, Princeton Univ. Press, Princeton N.J..

Ott, E. (1981) "Strange attractors and chaotic motions of dynamical systems", *Rev. Mod. Phys.*, **53**, 655.

Priestley, M.B. (1981) *Spectral Analysis and Time Series*, Academic Press, New York.
Rutkowski, L. (1982) "On-line identification of time-varying systems by nonparametric techniques", *IEEE Trans. Automat. Control*, **27**, 228-230.
Rutkowski, L. (1984) "On nonparametric identification with prediction of time-varying systems", *IEEE Trans. Automat. Control*, **29**, 58-60.
Rutkowski, L. (1985) "Nonparametric identification of quasi-stationary systems," *Syst. Control Letters*, **6**, 33-35.
Schuster, H.G. (1988) *Deterministic Chaos: An Introduction,* VCH (West Germany), Weinheim.
Shaw, R.S. (1981) "Strange attractors, chaotic behavior, and information flow", *Z. Naturforsch.*, **36a**, 80.
Unbehauen, H. (1985) "A laboratory experiment for teaching system identification and adaptive control", in H.A. Barker, and P.C. Young (eds.), *Identification and System Parameter Estimation 1985*, Pergamon Press, Oxford-New York, pp. 583-587.
Yakowitz, S. (1985) "Markov flow models and the flood warning problem", *Water Resources Research*, **21**, 81-88.
Yakowitz, S. (1987) "Nearest neighbor methods for time series analysis", *J. Time Series Anal.*, **8**, 235-247.
Yakowitz, S. (1989) "Nonparametric density and regression estimation for markov sequences without mixing assumptions", *J. Multivariate Anal.*, **30**, 124-136.

NONPARAMETRIC ESTIMATION OF A CLASS OF NONLINEAR TIME SERIES MODELS

M. PAWLAK
Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, MB, Canada R3T 2N2

W. GREBLICKI
Institute of Engineering Cybernetics
Technical University of Wroclaw
Poland

ABSTRACT. The problem of estimation of nonlinear time series models which are a composition of nonlinear elements and linear stochastic processes is considered. The compositions studied include the cascade and parallel connections. The problem of nonparametric estimation of underlying nonlinearities is examined. It is resolved by solving Fredholm's integral equations of the second kind arising in the estimation problem. As a result, the nonparametric orthogonal series estimates of nonlinearities are derived and their asymptotic as well as some small sample properties are established.

1. Introduction

Much time series analysis is carried out within the framework of linear models. Such models have been widely and successfully used in describing and predicting time series behaviour [6]. There is no particular reason, however, to assume that the time series observed in the physical process is generated linearly. In fact, the linearity assumption can be regarded as a first-order approximation of stochastic discrete time phenomena.

This paper will discuss the problem of nonparametric estimation of a certain class of nonlinear time series models. They are a composition of nonlinear elements and linear time invariant stochastic processes. In particular, we study the cascade and parallel connections. Such models have been extensively employed in several fields of engineering and biology [17], [20].

This paper concerns itself with the following two nonlinear time series models

$$Y_n = \sum_{i=0}^{\infty} g_i \, \varphi(U_{n-i}) + Z_n \tag{1}$$

$$Y_n = \sum_{i=0}^{\infty} g_i \, \varphi(U_{n-i}) + \sum_{j=0}^{\infty} t_j \, \psi(V_{n-j}) + Z_n \, , \tag{2}$$

where the values Y_i, U_i, V_i, $i = 1, ..., n$ are given, Z_n is an unknown stationary noise

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 541–552.
© 1991 *Kluwer Academic Publishers.*

with mean 0 and variance σ^2. The sequences $\{g_i\}$, $\{t_j\}$ represent characteristics of linear filters and $\varphi(u)$, $\psi(u)$ are some unknown measurable functions. Both $\{g_i\}$, $\{t_j\}$ and $\varphi(u)$, $\psi(u)$ must be estimated from the data record, i.e. from $(U_1, Y_1), \ldots, (U_n, Y_n)$ for the model in (1), and $(U_1, V_1, Y_1) \ldots, (U_n, V_n, Y_n)$ for the model in (2).

The model in (1) can be viewed as a cascade connection of two subsystems, where the first is nonlinear with the characteristic $\varphi(u)$, while the second is a linear asymptotically stable dynamic subsystem with the characteristic $\{g_i\}$. For $g_0 = 1$, $g_i = 0$, $i \geq 1$, (1) reduces to the classical regression case, $Y_n = \varphi(U_n) + Z_n$, [8]. The model in (2), in turn, is a parallel connection of two cascades given in (1). For $g_0 = t_0 = 1$, $g_i = t_i = 0$, $i \geq 1$, (2) becomes the additive regression model $Y_n = \varphi(U_n) + \psi(V_n) + Z_n$, [24]. The aim of the paper is to devise a nonparametric regression estimation technique for recovering the nonlinear characteristics of the models.

It should be noted that the characteristics of the linear subsystems can be estimated, independently on the estimation technique used for the nonlinear transformations.

To demonstrate the point, let us observe that for the model in (1)

$$\mathrm{cov}(Y_{n+s}, U_n) = c \sum_{i=0}^{\infty} g_i r_{s-i}, \; c = \mathrm{cov}(U_0, \varphi(U_0)) / \mathrm{var}(U_0), \; r_n = \mathrm{cov}(U_k, U_{k+n}) .$$

Using this identity and assuming that $\{U_n\}$ is a stationary Gaussian process Brillinger [5] has constructed a consistent estimate of the transfer function of the linear subsystem. He has not considered, however, the problem of estimation of the nonlinearity $\varphi(u)$. A similar analysis can be carried out for the model in (2), [22]. Nevertheless, we will assume throughout the paper that $\{g_i\}$ and $\{t_i\}$ are known.

In Section 2 the cascade model is investigated, whereas Section 3 briefly discusses analogous issues for the parallel connection.

2. Estimation of the Cascade Model

2.1 PRELIMINARIES

We assume that $\sum_{i=0}^{\infty} |g_i| < \infty$. The input process $\{U_n\}$ is strictly stationary with mean 0 and a symmetric marginal density f(u). The processes $\{U_n\}$ and $\{Z_n\}$ are mutually independent. The nonlinear characteristic φ is unknown and it is assumed, without loss of generality, to be odd. Let also $\mathbf{E}\,\varphi^2(U_0) < \infty$. These conditions imply that $\mathbf{E}\,Y_n = 0$ and $\mathbf{E}\,Y_n^2 \leq \mathbf{E}\,\varphi^2(U_0) \left(\sum_{i=0}^{\infty} |g_i| \right)^2$. Thus, the above restrictions are totally independent of the estimation method presented in this paper. They simply make the problem well posed in the sense that the output Y_n is a random variable. Note also that $\{Y_n\}$ is a stationary process. Clearly, Y_n's are dependent

even if the input is white noise. This is caused by the fact that the model embraces a dynamic subsystem which makes Y_n and Y_m dependent.

Some authors [3], [7], [9], [15], [23], [27] have investigated the behaviour of nonparametric curve estimates under some mixing type dependence conditions. The process $\{Y_n\}$ in (1), however, need not exhibit such a property. In fact, let $Y_n = \frac{1}{2} Y_{n-1} + \varphi(U_n)$, i.e., $Y_n = \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^i \varphi(U_{n-i})$, where $\varphi(x) = 1$ for $x > 0$ and 0 otherwise, and $\{U_n\}$ is a white Gaussian process. Then, using results of [1], [2] one can easily demonstrate that the $\{Y_n\}$ process is not strong mixing.

The problem of nonparametric estimation of $\varphi(u)$ in (1) has been recently examined by a number of authors [10], [11], [12], [18]. Assuming that the input process $[U_n\}$ is white noise it was observed [10] that $\varphi(u) = \mathbf{E}\{Y_n \mid U_n = u\} = m(u)$, say. That is, the $\varphi(u)$ can be derived from nonparametric regression estimation techniques [8], [25]. In [10], [11], [12], [18] the kernel regression nonparametric estimate has been applied and various asymptotic as well as small sample properties have been established.

In this paper we face a more difficult problem since the input process is not assumed to be white noise but colour. Then, clearly, $\varphi(u) \neq \mathbf{E}\{Y_n \mid U_n = u\}$. The relationship between the regression functions $m(u)$ and the characteristic $\varphi(u)$ is now in the form of a Fredholm's integral equation of the second kind. Hence, it is easy to see from (1) that

$$m(u) = \varphi(u) + \int K(u,z)\, \varphi(z)\, dz \ , \tag{3}$$

where $K(u,z) = \sum_{i=1}^{\infty} g_i\, f_i(u,z)/f(u)$, $f_i(u,z)$ is the bivariate density of (U_n, U_{n+i}) and $m(u)$ is the regression function of Y_n on U_n.

Assuming that $f_i(u,z)$ may be expressed in a diagonal expansion of some orthonormal functions [19] we give an explicit solution of the equation.

Then, assuming additionally that the process is Markov, a consistent estimate of $\varphi(u)$ is constructed.

2.2 INTEGRAL EQUATION

It is well known [19] that if $f_i(x,y)$ satisfies the following weak integrability condition:

$$\iint \frac{f_i^2(x,y)}{f(x)\, f(y)}\, dx\, dy < \infty \tag{4}$$

then $f_i(x,y)$ can be written in the diagonal series expansion [19]

$$f_i(x,y) = f(x)\, f(y) \left[1 + \sum_{t=1}^{\infty} \rho_t(i)\, h_t(x)\, h_t(y) \right] . \tag{5}$$

Here, the set $\{h_i(x)\}$ is a sequence of functions being orthonormal and complete with

respect to $f(x)$ with $h_0(x) = 1$. Moreover, $\rho_t(i)=E\{h_t(U_0)\, h_t(U_i)\}$. Clearly, $|\rho_t(i)| \le 1$. By means of (5) we can rewrite the equation in (3) as

$$w(u) = \alpha(u) + \int R(u,z)\, \alpha(z)\, dz \ , \tag{6}$$

where $w(u) = m(u)\sqrt{f(u)}$, $\alpha(u) = \varphi(u)\sqrt{f(u)}$, and

$$R(u,z) = \sqrt{f(u)\, f(z)} \sum_{r=1}^{\infty} \tau_r\, h_r(u)\, h_r(z)\ ,\ \tau_r = \sum_{i=1}^{\infty} \rho_r(i)\, g_i\ .$$

The kernel function $R(u, z)$ has some important properties. First, it is a symmetric function, moreover $\iint R^2(u,z)\, du\, dz = \sum_{r=1}^{\infty} \tau_r^2$. Let us assume that the latter series is convergent, i.e., that the kernel is quadratically integrable. Note that for a broad class of stochastic processes $\rho_t(i) = \rho^{ti}$ where $|\rho| < 1$, [19], [26], and then this assumption is automatically satisfied.

Finally, observing that $\int w^2(u)\, du < \infty$, one can conclude that the integral equation in (6) possesses a unique, quadratically integrable solution, see [28]. Furthermore, noting that the kernel has eigenvalues τ_i and corresponding eigen functions $\sqrt{f(x)}\, h_i(x)$, $i = 1, 2, \ldots$, and using the result of [28, p. 115], one can get the following solution of (6)

$$\varphi(u) = m(u) - \sum_{t=1}^{\infty} \frac{\tau_t}{1+\tau_t}\, b_t\, h_t(u)\ , \tag{7}$$

where $b_t = \int m(u)\, h_t(u)\, f(u)\, du = E\{Y_n\, h_t(U_n)\}$.
In this way, we have shown

<u>Theorem 1.</u> Let $\sum_{i=0}^{\infty} |g_i| < \infty$ and $E\, \varphi^2(U_0) < \infty$.

If (4) and $\sum_{r=1}^{\infty} \left(\sum_{i=1}^{\infty} \rho_r(i)\, g_i \right)^2 < \infty$

are satisfied, then the integral equation in (3) has one solution of the class L_2 given by the formula in (7).

The solution given in (7) allows us to construct various estimates of $\varphi(u)$. It requires only to estimate the regression function $m(u)$ from the sample $(U_1, Y_1), \ldots, (U_n, Y_n)$ and then to truncate the second term on the right-hand side of (7) with b_t replaced by $\frac{1}{n} \sum_{i=1}^{n} Y_i\, h_t(U_i)$.

In the next subsection we define an orthogonal series estimate of $\varphi(u)$ and show its consistency.

2.3 ESTIMATION ALGORITHM AND CONSISTENCY

Multiplying the equation (7) scalarly by $h_s(u)\, f(u)$, we get

$$c_s = (1+\tau_s)^{-1}\, b_s \quad , \tag{8}$$

where $c_s = \int \varphi(u)\, h_s(u)\, f(u)\, du$ is the sth coefficient of the expansion of $\varphi(u)$ in the $\{h_t(u)\}$ series.

Because of this and introducing some smoothing by omitting the infinite sum $\sum_{r=1}^{\infty} c_r\, h_r(u)$ after some finite number, N, of terms, we may define

$$\hat{\varphi}(u) = \sum_{r=1}^{N} \hat{c}_r\, h_r(u) \quad , \tag{9}$$

where $\hat{c}_r = (1+\tau_r)^{-1}\, \hat{b}_r$, $\hat{b}_r = \frac{1}{n} \sum_{i=1}^{n} Y_i\, h_r(U_i)$.

The estimate in (9) is closely related to orthogonal series estimates of density and regression functions studied in the literature [8], [13], [14], [16]. However, in these references a simple case of inference from independent data has been mainly examined. Let us now consider the asymptotic properties of $\hat{\varphi}(u)$.

Let the mean integrated squared error $\mathrm{MISE}(N) = \mathbf{E} \int \left(\hat{\varphi}(u) - \varphi(u)\right)^2 f(u)\, du$ be chosen as a discrepancy measure between $\hat{\varphi}$ and φ. To prove the consistency result, we assume that $\{U_n\}$ is a Markov sequence. Then $\rho_t(i) = \rho_t^i$ [26] and moreover for many bivariate densities $\rho_t = \rho^t$, $|\rho| < 1$, [26]. Thus, without loss of generality, let us assume that $\{U_n\}$ is a stationary Gauss-Markov process with the marginal density f(u) that of N(0,1). The bivariate density of (U_n, U_{n+i}) has an expansion as in (5) with $\rho_t(i) = \rho^{ti}$, $\rho = \mathrm{corr}(U_0, U_1)$ and $h_r(u) = H_r(u)/\sqrt{r!}$ being the rth orthonormal (with respect to f) Hermite polynomial. To establish consistency, we need the following:

<u>Lemma</u>. Let $\{U_t\}$ be a Gauss-Markov process, with $\mathrm{corr}\,(U_0, U_1) = \rho$.
Let $\alpha(u)$ and $\gamma(u)$ be Borel measurable functions. Then for $a < b < c < d$

$$|\mathrm{cov}\,(\alpha(U_d)\, \gamma(U_c), \alpha(U_b)\, \gamma(U_a))|$$

$$\leq \left(\sup_u |\gamma(u)|\right)^2 \mathbf{E}\, \alpha^2(U_0) \sum_{t=1}^{\infty} |\rho|^{t\,(c-b)} \quad .$$

The proof easily results from (5) and the fact that $f(u_a, u_b, u_c, u_d) = f(u_d|u_c)\, f(u_c|u_b)\, f(u_a, u_b)$.

We are now in a position to prove

<u>Theorem 2</u>. Let all conditions of Theorem 1 be satisfied. Let $|\rho| < 1$ and let $\varphi(u)$ be a bounded function. If

$$N(n) \to \infty \qquad \text{and} \qquad \frac{1}{n}\sum_{r=1}^{N(n)}\left(1+\sum_{i=1}^{\infty} g_i\,\rho^{ri}\right)^{-2} \to 0$$

then

$$\mathrm{MISE}(N(n)) \to 0 \text{ as } n \to \infty .$$

<u>Proof</u>. Applying Parseval's formula we get

$$\mathrm{MISE}(N) = \sum_{r=1}^{N} \mathrm{var}(\hat{c}_r) + \sum_{r=N+1}^{\infty} c_r^2 .$$

It is clear that it is sufficient to consider $\mathrm{var}\,(\hat{b}_r)$.
Due to the stationarity, we get

$$\begin{aligned}\mathrm{var}(\hat{b}_r) &= \frac{1}{n}\mathrm{var}\,(Y_1\, h_r(U_1)) \\ &\quad + \frac{2}{n}\sum_{s=2}^{n-1}\left(1-\frac{s}{n}\right)\mathrm{cov}\,(Y_s\, h_r\,(U_s),\, Y_1\, h_r\,(U_1)) \qquad (10)\\ &= A + B, \text{ say.}\end{aligned}$$

Applying the Cauchy-Schwartz inequality, it is easy to see that

$$nA \le \overline{\varphi}^2\left(\sum_{i=0}^{\infty}|g_i|\right)^2 + \sigma^2 \text{ , where } \overline{\varphi} = \sup_x |\varphi(x)| \text{ and } \sigma^2 = \mathrm{var}\,(Z_0) .$$

Now, let us take the term B in (10) into consideration. It is plain that

$$\begin{aligned}\mathrm{cov}\,(Y_s\, h_r\,(U_s),\, Y_1\, h_r\,(U_1)) &= \mathrm{cov}\,(h_r(U_s)\,\varphi(U_s),\, h_r\,(U_1)\,\varphi(U_1))\\ &\quad + \sum_{j=1}^{\infty} g_j\,\mathrm{cov}\,(h_r\,(U_s)\,\varphi\,(U_s),\; h_r(U_1)\,\varphi(U_{1-j}))\\ &\quad + \sum_{i=1}^{\infty} g_i\,\mathrm{cov}\,(h_r\,(U_s)\,\varphi\,(U_{s-i}),\; h_r(U_1)\,\varphi(U_1))\\ &\quad + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} g_i\, g_j\,\mathrm{cov}\,(h_r\,(U_s)\,\varphi\,(U_{s-i}),\; h_r(U_1)\,\varphi(U_{1-j})) . \qquad (11)\end{aligned}$$

Clearly, the first term in this decomposition is not greater than $\overline{\varphi}^2 \sum_{t=1}^{\infty} |\rho|^{t(s-1)}$.

Let us take the last term in (11) into account.
For the $1 - j < 1 < s - i < s$ case the Lemma can be applied yielding the upper bound

$$\overline{\varphi}^2 \sum_{i=1}^{\infty}|g_i| \sum_{t=1}^{\infty}|\rho|^{t(s-i-1)} \sum_{j=1}^{\infty}|g_j| .$$

The cases $i = s - 1$ and $i > s - 1$ can be evaluated by the Cauchy-Schwartz inequality giving upper bounds $\overline{\varphi}^2\, g_{s-1} \sum_{j=1}^{\infty}|g_j|$ and $\overline{\varphi}^2 \sum_{i=s-2}^{\infty}|g_i| \sum_{j=1}^{\infty}|g_j|$, respectively.

Since the third term in (11) can be examined in a similar way, while the second

one is equal to zero, we have shown that $nB \to \text{const}$ as $n \to \infty$. Thus, the proof has been completed.

Corollary. If $\{U_n\}$ is white noise (i.e., $\rho \equiv 0$), then one can show that

$$\sum_{r=1}^{N} \operatorname{var}(\hat{c}_r) = \left(\sigma^2 + E\,\varphi^2(U_0) \sum_{i=1}^{\infty} g_i^2\right)\frac{N}{n}$$
$$+ \frac{1}{n}\sum_{r=1}^{N} E\left\{\varphi^2(U_0)\, h_r^2(U_0)\right\} - \frac{1}{n}\sum_{r=1}^{N} c_r^2 .$$

Hence,

$$\frac{n}{N(n)} \int \operatorname{var}\left(\hat{\varphi}(x)\right) f(x)\, dx \to \sigma^2 + E\,\varphi^2(U_0) \sum_{i=1}^{\infty} g_i^2 .$$

It should be observed that no boundedness condition on φ is required here.

Let us now evaluate the rate of convergence.

Let $\beta_p(x;\varphi) = (x - \frac{d}{dx})^p \varphi(x)$. It has been shown, see [13] and references therein, that if $\beta_p(x;\varphi)$ and φ are square integrable, then

$$|c_r| \le |d_{r+p}| / (r+1)^{p/2} ,$$

where d_r is the rth coefficient of the expansion of $\beta_p(x;\varphi)$ in the Hermite series.

Then, clearly $\sum_{r=N+1}^{\infty} c_r^2 = O(N^{-p})$.

By this result and recalling the proof of Theorem 1, we have

Theorem 3. Let all conditions of Theorem 2 be satisfied. Let, moreover, $\beta_p(x;\varphi)$, $p \ge 1$, be square integrable with respect to the weight function $f(x)$.

If $N(n) \sim n^{1/(p+1)}$ then $\text{MISE}(N(n)) \cong O(n^{-p/(p+1)})$.

To support the asymptotic theory let us consider an example illustrating the effect of correlated data on the performance of the estimate.

To this end, let $Y_n = a\,Y_{n-1} + \varphi(U_n) + Z_n$, where $Z_n = N(0, 0.05)$, $|a| < 1$ and

$$\varphi(u) = \begin{cases} 2u & |u| \le 0.5 \\ 2(1-u) & 0.5 < |u| \le 1 \\ 0 & \text{otherwise} \end{cases}$$

The process $\{U_n\}$ is generated by $U_n = \rho U_{n-1} + (1-\rho^2)^{1/2}\varepsilon_n$, $\varepsilon_n = N(0, 1)$, $|\rho| < 1$. Figure 1 plots the MISE(N) (x 100) versus a(for $\rho = 0$) and ρ(for $a = 0$). The sample size was set to $n = 40$ and $N = 11$ (this is the N minimizing MISE(N) for $a = 0$, $\rho = 0.5$).

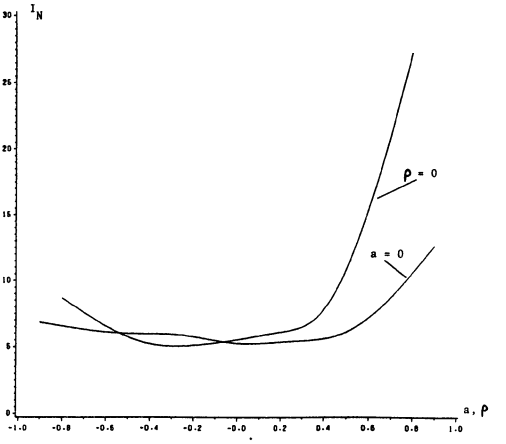


Fig. 1. MISE(N) versus a and ρ, n = 40, N = 11 .

Apparently, the estimate is more vulnerable to the model dynamics than to the correlation of the input signal. Furthermore, the degradation of estimate performance for positively correlated data (a, ρ > 0) is also revealed. This is illustrated in Fig. 2 where MISE(N), N = 11, as a function of (ρ, a), $|a| \leq 1$, $|\rho| \leq 1$ is plotted.

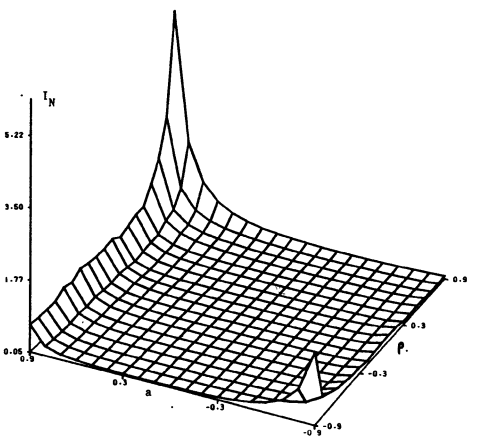


Fig. 2. MISE(N) versus (ρ,a), N = 11, n = 40.

2.4 AUTOMATIC SELECTION OF SMOOTHING PARAMETER

Clearly, the parameter N controls the level of smoothing applied to the data. Its automatic selection can be worked out similarily as in the density estimation case, see [16] and references cited therein. This is, particularly, feasible in the case when $\{U_n\}$ is a white process. Then, $\mathrm{var}(\hat{c}_r) = \frac{1}{n}\mathrm{var}(Y_1\, h_r(U_1))$ and the variance in this expression can be estimated by

$$\eta_r = (n - 1 + \alpha)^{-1} \sum_{i=1}^{n} (Y_i h_r(U_i) - \hat{c}_r)^2 .$$

The constant $\alpha \in (-n + 1, 1)$ reflects the correlation of the data and should be chosen by the user.
Thus, the following estimate

$$\hat{\Delta}_r = \eta_r/n - \tilde{c}_r ,$$

of Δ_r = MISE (r) - MISE (r - 1) = var $(\hat{c}_r) - c_r^2$ can be proposed.
Here $\tilde{c}_r = 2(n(n-1))^{-1} \sum\sum_{i<j} Y_i Y_j h_r(U_i) h_r(U_j)$.
This yields the following version of the inclusion rule [16]:

Include all successive terms which pass the test $\hat{\Delta}_r < 0$.

Typically, it is required to truncate the estimate whenever the rule fails for a priori determined number of consecutive terms. This prevents the algorithm from getting stuck in a local minima.
Fig. 3 shows the typical realization of the process

$$\xi_k = \sum_{r=1}^{k} \hat{\Delta}_r \text{ for } 1 \le k \le 15, \alpha = -n/2, n = 40, a = 0.5, \rho = 0.0,$$

and the other data as in Fig. 1.
Note that ξ_k estimates MISE(k) - MISE(0).

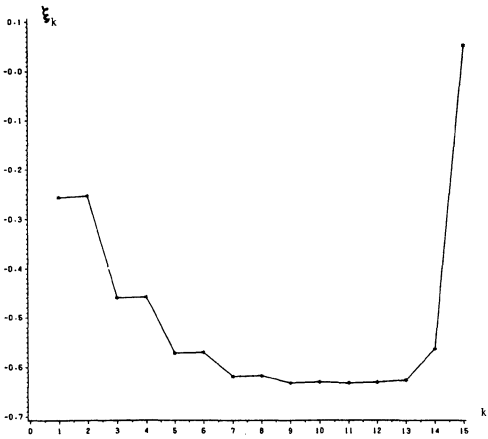


Fig. 3. The process $\{\xi_k\}$, a = 0.5, ρ = 0, n = 40.

The optimal N minimizing MISE(N) in this case is 11. It is seen that the inclusion rule is able to detect a minimum point assuming that the optimization process is able to skip over one local minimum.
The problem of consistency of this scheme is left for further research.

3. Estimation of the Parallel Model

Let us now consider briefly the additive model defined in (2). The estimation of characteristics $\varphi(u)$ and $\psi(u)$ can be basically carried out in the similar fashion as for the cascade model. The correlation structure of the additive connection, however, can be more complex yielding some tedious analysis. To demonstrate the point, let us assume that the process (U_n, V_n) is generated by the following linear stochastic equations

$$U_n = \Delta U_{n-1} + \left(1 - \Delta^2\right)^{1/2} \varepsilon_n \ ,$$
$$V_n = \tau V_{n-1} + \left(1 - \tau^2\right)^{1/2} \epsilon_n \ ,$$

where $\varepsilon_n, \epsilon_n$ are N(0, 1) white noise processes with $\mathrm{cov}(\varepsilon_n, \epsilon_m) = b\,\delta_{nm}$, $|\Delta| < 1$, $|\tau| < 1$, $|b| \le 1$.

The integral equation in (3) is now replaced by the system of integral equations.

$$\begin{cases} E\{Y_n|U_n = u\} = m_1(u) = \varphi(u) + \int K(u,z)\,\varphi(z)\,dz + \int L(u,z)\,\psi(z)\,dz \\ E\{Y_n|V_n = v\} = m_2(v) = \psi(v) + \int O(v,z)\,\psi(z)\,dz + \int P(v,z)\,\varphi(z)\,dz \end{cases} , \qquad (12)$$

where $K(u,z) = \sum_{i=1}^{\infty} g_i\, f_i(z|u)$, $f_i(z|u)$ the conditional density of U_i on U_0 ,

$L(u,z) = \sum_{j=0}^{\infty} t_j\, \eta_j(z|u)$, $\eta_j(z|u)$ the conditional density of V_j on U_0 ,

$P(u,z) = \sum_{i=0}^{\infty} g_i\, \xi_i(z|v)$, $\xi_i(z|v)$ the conditional density of U_i on V_0 ,

$O(v,z) = \sum_{i=1}^{\infty} t_i\, g_i(z|v)$, $g_i(z|v)$ the conditional density of V_i on V_0.

The system in (12) can be solved using the technique worked out in Section 2.2.
This technique and some algebra yield the following solutions for $\varphi(u)$ and $\psi(v)$

$$\varphi(u) \cong \sum_{s=1}^{\infty} a_s\, h_s(u) \ , \quad \psi(v) \cong \sum_{s=1}^{\infty} b_s\, h_s(v) \ ,$$

where $a_s = \theta_s/G(\Delta^s) - b_s\, \rho^s\, T(\Delta^s)/G(\Delta^s)$,

$$b_s = \left(\beta_s\, G(\Delta^s) - \rho^s\, G(\tau^s)\, \theta_s\right) / \left(T(\tau^s)\, G(\Delta^s) - \rho^{2s}\, T(\Delta^s)\, G(\tau^s)\right) .$$

Here $\theta_s = E\{Y_n\, h_s(U_n)\}$, $\beta_s = E\{Y_n\, h_s(V_n)\}$, $\rho = b\left((1 - \Delta^2)(1 - \tau^2)\right)^{1/2} / (1 - \Delta\tau)$.

Moreover, $G(z) = \sum_{i=0}^{\infty} g_i\, z^i$ and $T(z) = \sum_{j=0}^{\infty} t_j\, z^j$ are transfer functions of linear subsystems $\{g_j\}$ and $\{t_j\}$, respectively.

The estimates of $\varphi(u)$ and $\psi(v)$ can be easily defined as

$$\hat{\varphi}(u) = \sum_{r=1}^{N} \hat{a}_r\, h_r(u), \quad \hat{\psi}(v) = \sum_{r=1}^{M} \hat{b}_r\, h_r(v),$$

where $\hat{a}_r$ and $\hat{b}_r$ are defined as a_r and b_r with θ_r, β_r replaced by

$$\hat{\theta}_r = \frac{1}{n} \sum_{i=1}^{n} Y_i\, h_r(U_i), \quad \hat{\beta}_r = \frac{1}{n} \sum_{i=1}^{n} Y_i\, h_r(V_i), \text{ respectively.}$$

Here N and M define smoothing parameters.

Asymptotic properties of these estimates can be established, in principle, in the same way as that in Theorem 2, although it requires some tedious but straightforward algebra, see [21], [22].

Finally, let us note that the approach proposed above to recover the components of the additive model may be considered as an alternative strategy to those proposed in [4] and [24]. Our method, however, is confined to a certain class of input signals. On the other hand, we deal with dependent data, while in [4] and [24] the iid case is mostly examined.

Acknowledgment

This work was supported by NSERC Grant A8131 and by Grant RP.I.02.

References

[1] Andrews, D.W.K. (1984), 'Non-strong mixing autoregressive processes', J. Appl. Prob., 21, 930-934.

[2] Andrews, D.W.K. (1985), 'A nearly independent, but non-strong mixing, triangular array', J. Appl. Prob., 22, 729-731.

[3] Bierens, H.T. (1983), 'Uniform consistency of kernel estimators of a regression function under generalized conditions', J. Amer. Statist. Assoc., 78, 699-707.

[4] Breiman, L. and Friedman, J.H. (1985), 'Estimating optimal transformations for multiple regression and correlation', J. Amer. Statist. Assoc., 80, 580-597.

[5] Brillinger, D.R. (1977), 'The identification of a particular nonlinear timer series system', Biometrika, 64, 509-515.

[6] Brillinger, D.R. (1980), Time Series, Data Analysis and Theory, Holden-Day, Inc., San Francisco, Expanded Edition.

[7] Collomb, G. and Härdle, W. (1985), 'Strong uniform convergence rates in robust nonparametric time series analysis and prediction: kernel regression estimation from dependent observations', Stochastic Process.Appl. , 23, 77-81.

[8] Eubank, R.L. (1980), Spline Smoothing and Nonparametric Regression, Marcel Dekker, New York.

[9] Georgiev, A. (1984), 'Nonparametric system identification by kernel methods', IEEE Trans. Automat. Control, AC-29, 356-358.

[10] Greblicki, W. and Pawlak, M. (1986), 'Identification of discrete Hammerstein systems using kernel regression estimates', IEEE Trans. Automat. Control, AC-31, 74-77.

[11] Greblicki, W. and Pawlak, M. (1989), 'Recursive nonparametric identification of Hammerstein systems', J. of Franklin Institute, 326, 461-481.

[12] Greblicki, W. and Pawlak, M. (1989), 'Nonparametric identification of Hammerstein systems', IEEE Trans. Inform. Theory, 35, 409-418.

[13] Greblicki, W. and Pawlak, M. (1984), 'Hermite series estimates of a probability

density and its derivatives', J. Multivariate Anal. 15, 176-182.
[14] Greblicki, W. and Pawlak, M. (1985), 'Fourier and Hermite series estimates of regression functions', Ann. Inst. Statist. Math., 37, 443-459.
[15] Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989), Nonparametric Curve Estimation from Time Series, Springer-Verlag, Berlin.
[16] Hall, P. (1987), 'Cross-validation and the smoothing of orthogonal series density estimators', J. Multivariate Anal., 21, 189-206.
[17] Hunter, I.W. and Korenberg, M.J. (1986), 'The identification of nonlinear biological systems: Wiener and Hammerstein cascade models', Biol. Cybern., 55, 135-144.,
[18] Krzyzak, A. (1990), 'On estimation of a class of nonlinear systems by the kernel regression estimate', IEEE Trans. Inform. Theory, IT-36, 141-152.
[19] Lancaster, H.O. (1958), 'The structure of bivariate distributions', Ann. Math. Statist., 29, 719-736.
[20] O'Leary, D.P. (1985), "Identification of sensory systems and neural systems', IFAC Symposium on Identification and System Parameter Estimation, Pergamon Press, 77-84.
[21] Pawlak, M. (1988), 'Identification of a class of parallel systems', The 26th Allerton Conf. in Communication, Control, and Computing, I, 366-367.
[22] Pawlak, M. (1990), 'Estimation of a class of additive nonlinear time series models', Manuscript in preparation.
[23] Robinson, P.M. (1986), 'On the consistency and finite-sample properties on nonparametric kernel time series regression, autoregression and density estimators', Ann. Inst. Statist. Math., Part A, 38, 539-549.
[24] Stone, C.T. (1985), 'Additive regression and other nonparametric models', Ann. Statist., 13, 689-705.
[25] Stone, C.J. (1977), 'Consistent nonparametric regression', Ann. Statist., 5, 595-645.
[26] Wise, G.L. and Thomas, J.B. (1975), 'A characterization of Markov sequences', J. Franklin Institute, 229, 269-278.
[27] Yakowitz, S. (1985), 'Nonparametric density estimation prediction and regression for Markov sequences', J. Amer. Statist. Assoc., 80, 215-221.
[28] Tricomi, F.G. (1985), Integral Equations, Dover, New York.

SEMIPARAMETRIC AND NONPARAMETRIC INFERENCE FROM IRREGULAR OBSERVATIONS ON CONTINUOUS TIME STOCHASTIC PROCESSES

J.M. STOYANOV
Institute of Mathematics
Academy of Sciences
Sofia
Bulgaria

and

P.M. ROBINSON
Department of Economics
London School of Economics
Houghton Street
London WC2A 2AE UK

ABSTRACT. A continuous time stochastic process contains unknown, nonparametric and/or semiparametric characteristics which are estimated by means of discrete and ordered data obtained at random times. Two specific models are considered in detail. Estimates are suggested and their asymptotic behaviour studied.

I. Introduction

This paper deals with stochastic models defined as the pair (ξ,W) where the first component $\xi = (\xi_t,\ t \in R^1)$ is a continuous time stochastic process and the second component $W = \{\tau_1,\tau_2,\ldots\}$ consists of an increasing sequence of random times. Both components ξ and W can contain parametric and/or nonparametric characteristics. A main assumption is that we can observe the process ξ only at the times $t \in W$. Hence all the available data are discrete. Using these discrete observations we wish to make inference about unknown characteristics of the model (ξ,W). Several practical phenomena, e.g. from economics, the natural sciences, medicine, technology etc., can be adequately described by continuous–discrete stochastic models of the type (ξ,W), see e.g. Parzen (1984). Earlier results can be found in Brillinger (1973), Le Breton (1976), Robinson (1977),(1980), Masry (1983), Stoyanov (1984), Nguyen (1985), Vysotskii (1986), Genon–Catalot (1990), Deniau, Oppenheim and Viano (1986) and Prakasa Rao (1988).

Among the variety of estimation problems for continuous–discrete stochastic models we confine ourselves to two specific models: in one ξ is parametric, in the other ξ is nonparametric. In both, W is nonparametric. In Section 2 the problem is to estimate the mean of τ_i W by observations of a diffusion process ξ with given constant coefficients. Here the model for (ξ,W) could be called semiparametric. Section 3 is devoted to estimation of a nonparametric autocovariance function of a weakly stationary process ξ using observations recorded at times associated with a point process. Here the model for (ξ,W) is nonparametric. In both cases we suggest explicit estimates of the unknown characteristics. Both estimates are based on "ordered" data. This means that for each natural N we have at our disposal the data $\tilde{D}_N = \{X_1,\ldots,X_N\}$ where $X_1 = \xi_{\tau_1},\ldots,X_n = \xi_{\tau_N}$, i.e. we know the observations of the process ξ in the order of their appearance but we do not know the times $\tau_1,\ldots,\tau_N$ themselves. (Cases in which the τ_i are also observed are important, as are ones when, by contrast, even the order of observations is unknown; see Robinson (1977), Stoyanov (1984), Vysotskii (1986) and Prakasa Rao (1988).) We present two propositions which establish the asymptotic properties of the desired estimators, sketching their proofs. Finally some concluding remarks are given.

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 553–558.
© 1991 *Kluwer Academic Publishers.*

2. Semiparametric Model

Suppose the continuous–discrete model (ξ, W) has the following specific structure:

(I) The process $\xi = (\xi_t,\ t \geq 0)$ is defined by

$$\xi_t = at + \sigma W_t$$

where $W = (W_t,\ t \geq 0)$ is a standard Wiener process and a and σ are given constants, $a \in R^1$, $\sigma > 0$. This means that ξ is a simple diffusion process with a constant drift coefficient a and a constant diffusion coefficient σ^2.

(II) The set of times for observations $W = \{\tau_1, \tau_2, \ldots\}$ is a renewal type process, i.e. $\tau_1, \tau_2 - \tau_1, \tau_3 - \tau_2, \ldots$ are positive independent and identically distributed random variables. Denote by F the distribution function of τ_1 (and of any of $\tau_j - \tau_{j-1}$).

Now, according to the procedure described in the introduction, the process ξ is observed only at the times $t \in W$. Suppose that F is nonparametric, that is not a given function of finitely many unknown parameters, but we are only interested in estimating $m_1 = E\tau_1 = \int_0^\infty t dF(t)$, assuming also the finiteness of the second order moment $m_2 = E(\tau_1^2)$.

Thus the problem is to find "good" estimates of m_1 on the basis of the ordered data $\tilde{D}_N = \{X_1, \ldots, X_N\}$ with $X_1 = \xi_{\tau_1}, \ldots, X_N = \xi_{\tau_N}$.

The last of our assumptions it that the random set W and the process ξ are independent.

Now we are in a position to formulate the following result where by definition we have $\tau_0 = 0$ and $X_0 = 0$.

PROPOSITION 1. If $a \neq 0$ let

$$\tilde{m}_{1N} = \frac{X_N}{a_N}$$

while if $a = 0$

$$\tilde{m}^*_{1N} = \frac{X_N^2}{\sigma^2 N}$$

and

$$\tilde{m}^{**}_{1N} = \frac{1}{\sigma^2 N} \sum_{j=1}^{N} (X_j - X_{j-1})^2.$$

Then each of the statistics $\tilde{m}_{1N}$, $\tilde{m}^*_{1N}$ and $\tilde{m}^{**}_{1N}$ is an unbiased and weakly consistent estimate of the unknown parameter m_1.

<u>Sketch of proof</u>. The unbiasedness property of $\tilde{m}_{1N}$, $\tilde{m}^*_{1N}$ and $\tilde{m}^{**}_{1N}$ is based on the Wald identities for the Wiener process: $EW_t = 0$ and $E(W_t^2) = E\tau$ which hold for any integrable random time τ adapted with W, see Liptser and Shiryaev (1978). Hence, if $a \neq 0$, we have $X_N = a\tau_N + \sigma W_{\tau_N}$ and obviously $EX_N = aNm_1$, or equivalently $E\tilde{m}_{1N} = m_1$. Further, if $a = 0$, then $X_N = \sigma W_{\tau_N}$ and $E(X_N^2) = \sigma^2 E(W^2_{\tau_N}) = \sigma^2 E\tau_N = \sigma^2 N m_1$, i.e. $E\tilde{m}^*_{1N} = m_1$. Finally, $X_j - X_{j-1} = \sigma(W_{\tau_j} - W_{\tau_{j-1}})$ and by the strong Markov property of the Wiener process we easily find that

$$E\{(X_j-X_{j-1})^2\} = \sigma^2 E\{(W_{\tau_j}-W_{\tau_{j-1}})^2\} = \sigma^2 E(\tau_j-\tau_{j-1}) = \sigma^2 m_1 .$$

Thus $E\tilde{m}^{**}_{1N} = m_1$. Therefore each of $\tilde{m}_{1N}$, $\tilde{m}^{*}_{1N}$ and $\tilde{m}^{**}_{1N}$ is an unbiased estimate of m_1.

The consistency property of $\tilde{m}_{1N}$ (and analogously of $\tilde{m}^{*}_{1N}$ and $\tilde{m}_{1N}$), i.e. the relation $\tilde{m}_{1N} \to_p m_1$ as $N \to \infty$, is a consequence of the weak law of large numbers valid for the Wiener process W even if we consider the values of W at random times. Of course, we arrive at the same conclusion directly calculating the quantity $E\{(\tilde{m}_{1N}-m_1)^2\}$ (analogously $E\{(\tilde{m}^{*}_{1N}-m_1)^2\}$ and $E\{(\tilde{m}^{**}_{1N}-m_1)^2\}$) and then applying the Chebyshev inequality.

Note in addition that under the assumption $E(\tau_1^4) < \infty$ we can prove the following stronger result: $\tilde{m}_{1N} \xrightarrow{a.s.} m_1$ asd $N \to \infty$, i.e. $\tilde{m}_{1N}$ is a strongly consistent estimate of m_1 (analogously for $\tilde{m}^{*}_{1N}$ and $\tilde{m}^{**}_{1N}$). For this, we have first to calculate $E\{(\tilde{m}_{1N}-m_1)^4\}$ using the known expressions for $E(W_\tau^3)$ and $E(W_\tau^4)$ also called Wald identities of a higher order, see again Liptser and Shiryaev (1978). The last step is to show that $\Sigma_{N=0}^{\infty} E\{(\tilde{m}_{1N}-m_1)^4\} < \infty$ which, by the Borel–Cantelli lemma implies the a.s. convergence of $\tilde{m}_{1N}$ to m_1 as $N \to \infty$.

Remarks. (i) The stimates can be shown under suitable calculations to be asymptotically equally as efficient. (ii) We can similarly discuss the estimation of m_2, or of both m_1 and m_2 simultaneously.

3. Nonparametric Model

We redefine (ξ, W) as follows.

Denote by P the class of continuous time stochastic processes $\xi = (\xi_t, t \in R^1)$ satisfying conditions (1) – (4) listed below.

(1) ξ is real valued, and $E(\xi_1^4) < \infty$.
(2) ξ is weakly stationary with zero mean and autocovariance function $C(t) = E\{\xi_s \xi_{s+t}\}$.
(3) ξ is stationary to order 4 in that the expectations $E\{\xi_h \xi_{h+t_1} \xi_{h+t_1+t_2} \xi_{h+t_1+t_2+t_3}\}$ does not depend on h, and thus is a function, say $H(t_1,t_2,t_3)$, of t_1, t_2 and t_3 only. It is assumed that $H(0,t,0)$ is continuous in t.
(4) There is a constant M_1 such that

$$\int_0^\infty \left| H(t_1,t,t_3) - C(t_1)C(t_3) \right| dt \leq M_1 .$$

Further, let the set of times for observations $W = \{\tau_1, \tau_2, \ldots,\}$ be a renewal process, i.e. $\tau_1, \tau_2-\tau_1, \tau_3-\tau_2, \ldots$ are positive independent random variables whose common distribution function is denoted by F. We assume also that the following conditions (5)–(6) on W are satisfied.

(5) The distribution function F is absolutely continuous and its density $f = F'$ is continuous. The random variable τ_1 has a finite first order moment: $m_1 = E\tau_1 = \int_0^\infty t f(t) dt < \infty$.

(6) Denote $F_K = F * \ldots * F$ (the K–th convolution of F with itself) and let $U(t) = \Sigma_{K=1}^{\infty} F_K(t)$, $t \in R^1$. The existence of U can be derived by using Laplace transforms, see Feller (1971). The continuity of $f = F'$ implies that

$U(t)$, $t \in R^1$ is differentiable.

We also assume:

(7) In ξ and W are independent.

(8) $Q(t) = C(t)u(t)$ is twice differentiable with a finite second derivative.

It is supposed that the positive function $u(t) = U'(t)$. $t \in R^1$ is bounded, i.e. for some constant M_2 we have $u(t) \leq M_2$ for all t.

We wish to estimate $C(t)$, which is not parameterized, on the basis, as before, of the ordered data $\tilde{D}_N = \{X_1,\ldots,X_N\}$ where $X_j = \xi_{\tau_j}$, $j = 1,\ldots,M$.

Before formulating our result we introduce also the following function:

$$g_N(t) = \begin{cases} (1/2)\sqrt{N}, & \text{if } |t| \leq 1/\sqrt{N} \\ 0, & \text{if } |t| > 1/\sqrt{N}. \end{cases}$$

<u>PROPOSITION 2</u>. Let

$$\tilde{C}_N(t) = \frac{1}{Nu(t)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X_i X_j g_N((j-i)m_1 - t).$$

Then for each t we have:

(i) $\lim_{N\to\infty} E\tilde{C}_N(t) = C(t)$

and

(ii) $\lim_{N\to\infty} \operatorname{Var} \tilde{C}_N(t) = 0,$

i.e. $\tilde{C}_N(t)$ is an asymptotically unbiased and L^2–consistent estimate of the unknown covariance function $C(t)$.

<u>Sketch of proof</u>. In order to establish the validity of relation (i) we proceed as follows.

<u>Step 1</u>. Using the properties of conditional expectations and the stationarity of ξ we find the following expression for $E\tilde{C}_N(t)$:

$$E\tilde{C}_N(t) = \frac{1}{Nu(t)} \sum_{\kappa=1}^{N-1} (N-\kappa) \int_0^\infty C(s)dF_\kappa(s) g_N(\kappa m_1 - t).$$

In particular, we have used also the fact that if $j-i = \kappa$, then the random variable $\tau_j - \tau_i$ has a distribution function equal to F_K.

For the next steps we rewrite $E\tilde{C}_M(t)$ as

$$E\tilde{C}_N(t) = I_N(t) - J_N(t)$$

where

$$I_N(t) = \frac{1}{u(t)} \sum_{\kappa=1}^{N-1} g_N(\kappa m_1 - t) \int_0^\infty C(s)dF_\kappa(s),$$

$$J_N(t) = \frac{1}{Nu(t)} \sum_{\kappa=1}^{N-1} g_N(\kappa m_1 - t)\, \kappa \int_0^\infty C(s)dF_\kappa(s).$$

<u>Step 2</u>. The conditions imposed on f and C allow us to deduce that for each t and large N we have

$$|J_N(t)| = O(1/N) .$$

Step 3. Consider the quantity $I_N(t)$. Since $g_N(v) = 0$ for any v which is outside the interval $[t-1/\sqrt{N},\ t+1/\sqrt{N}]$ and $d(\Sigma_{K=1}^{\infty} F_K(s)) = dU(s) = u(s)ds$ we find after transformation that

$$I_N(t) = \frac{1}{u(t)} \int_0^{\infty} Q(s) g_N(s-t) ds + \Delta_N(t) ,$$

where $Q(s) = u(s)C(s)$ and $|\Delta_N(t)| = 0(1/N)$ for large N and each t.

Step 4. Write down the Taylor expansion of the function Q(s) in a neighbourhood of the point t with a remainder term of the form $0((s-t)^2)$ and put this expansion in $I_N(t)$.

Step 5. Combine Steps 1–4 and take the limit as $N \to \infty$ thus getting the desired relation (i).

The proof of relation (ii) is based on the analysis of $E\{\tilde{C}_N^2(t)\}$. The idea is to represent $E\{\tilde{C}_N^2(t)\}$ as a suitable sum of terms which have to be estimated or transformed. Conditions (1)–(8) are essential for this analysis. Similar arguments to those above are used.

4. Final Remarks

The models treated above can be further studied. Other asymptotic properties can be established. When, for example, ξ is Gaussian and/or the set W consists of the consecutive times of jumps of a homogeneous Poisson process, stronger results can be proved. More general models can be considered. In connection with this, another line of research in Robinson (1980) can be further pursued. Robinson (1980) found that if ξ has rational spectral density, $\tilde{D}_N$ can be an autoregressive moving average sequence, and also showed that the irregular sampling can avoid the identification problem inherent in equally–spaced sampling of continuous processes.

Acknowledgement

The first author's research was carried out while he was a Science and Engineering Council (SERC) Visiting Fellow (reference number: GR/872734) at the London School of Economics. The second author's research is funded by the Economic and Social Research Council (ESRC) reference number: R000231441.

References

Brillinger, D.R. (1973) 'Estimation of the mean of a stationary time series by sampling', Journal of Applied Probability 10, 419–431.

Deniau, C., Oppenheim, G., and Viano, C. (1986) 'Time random sampling', Preprint No.19, Université, Paris-Sud.

Feller, W. (1971) An Introduction to Probability Theory and its Applications, Vol.2, (2nd ed.), John Wiley & Sons, New York.

Genon-Catalot, V. (1990) 'Maximum contrast estimation for diffusion processes from discrete observations', Statistics 21, 99–116.

Le Breton, A. (1976) 'On continuous and discrete sampling for parameter estimation in diffusion type processes', Math. Progr. Studies 5, 124–144.

Liptser, R., and Shiryaev, A. (1978) Statistics of Random Processes, II Springer-Verlag, New York.

Masry, R. (1983) 'Nonparametric covariance estimation from irregularly spaced data', Advances in Applied Probaility 15, 113–132.

Nguyen, H.T. (1985) 'On point process sampling in continuous-time model', Pub. Inst. Statist. 30, University, Paris, 73–95.

Parzen, E. (1984)(Ed.) Time Series Analysis of Irregularly Observed Data, Lecture Notes in Statistics 25. Springer-Verlag, Berlin.

Prakasa Rao, B.L.S. (1988) 'Statistical inference from sampled data for stochastic processes', Contemporary Mathematics (AMS) 80, 249–284.

Robinson, P.M. (1977) 'Estimation of a time series model from unequally spaced data', Stochastic Processes and their Applications 6, 9–24.

Robinson, P.M. (1980) 'Continuous model fitting from discrete data', in D.R. Brillinger and G.C. Tiao (eds.) Directions in Time Series, Hayward (CA), pp.263–278.

Stoyanov, J. (1984) 'Problems of estimation in continuous-discrete stochastic models', in M. Iosifescu (ed.), Proceedings of the 7th Brasov Conference on Probability Theory, Academiei, Bucharest, pp.363–373.

Stoyanov, J., and Vladeva, D. (1982) 'Estimation of the parameter of a Poisson random window by observations of a diffusion process', Proceedings of the 11th Conference of the Union of Bulgarian Mathematics, Academia, Sofia, pp.431–435.

Vysotskii, V.I. (1986) 'Estimation of the correlation function of a stationary process observed at random times', Controlled Queueing Systems, No.4, University of Tomsk, Tomsk, pp.14–21.

VII. Various Topics

COMPLEXITY REGULARIZATION WITH APPLICATION TO ARTIFICIAL NEURAL NETWORKS

ANDREW R. BARRON
University of Illinois
Departments of Statistics and
Electrical & Computer Engineering
725 South Wright Street
Champaign, Illinois 61820 USA

ABSTRACT. Concepts of universal data compression lead to minimum description-length criteria for parsimonious statistical model selection for the estimation of functions. In this paper we define general complexity regularization criteria and establish bounds on the statistical risk of the estimated functions. These bounds establish consistency, yield rates of convergence, and demonstrate the near asymptotic optimality of the model selection criterion in both parametric and nonparametric cases. A fundamental role is played by an index of resolvability that quantifies the tradeoff between complexity and accuracy of candidate models. Applications are given to polynomial regression and artificial neural networks.

1. Introduction

In the context of statistical estimation of functions, a list of parametric models is provided from which one is to be automatically selected from the data. These lists of parametric models have the property that for essentially any target function there exist a convergent sequence of approximating functions on the list. In the absence of knowledge of the true function, the problem is to estimate an appropriate size model that provides a good tradeoff between the approximation error, which is best for larger models, and the parameter estimation error, which is best for smaller models.

In this paper we present an index of resolvability that defines the best tradeoff between approximation error and the complexity of models. A complexity regularization criterion is defined which adds an information-theoretic complexity penalty to the empirical loss, and the estimator is taken which achieves the optimum value of this criterion. It is shown that the statistical risk of the estimated function is bounded by the index of resolvability.

The methods and theory we present here are closely related to minimum complexity density estimation as developed in Barron and Cover (1990). The difference is that here we do not necessarily stick to the minimum description-length paradigm from Rissanen (1983,1984) and Barron (1985). In that paradigm the model selection criterion is required to correspond to a total description length for the sample. That paradigm works quite well for density estimation, using the fact that the best codes

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 561–576.
© 1991 *Kluwer Academic Publishers.*

for a given distribution have length equal to minus the logarithm of the density. The complexity penalty then serves as the length of a preample that describes what estimated distribution is used in the description of that sample. However, for other curve fitting problems we may be interested in estimating, say, a regression function using a penalized squared-error criterion. If the distribution of the errors is unknown (and in particular not Gaussian), then the squared error term in the criterion does not correspond to the length of an accurate code for the data. Nevertheless, regularization of the estimator by the addition of a complexity penalty is still fruitful in this context, despite the lack of a complete justification on grounds of optimal data compression.

A Bayesian formulation of the model selection problem as in Schwarz (1978) yields test statistics that are similar to criteria motivated by the minimum description-length principle. Although closely related, the complexity regularization criterion for the estimation of functions does not necessarily correspond to a Bayesian solution. Again, the reason is that the choice of an empirical loss function is not constrained to correspond to the log-likelihood for a model of the error distribution.

Other statistical performance criteria have been analyzed for their asymptotic statistical properties. In particular, Shibata (1981) has demonstrated asymptotic optimality properties of the Akaike's final prediction error and Akaike's *AIC* criterion in the case of Gaussian errors. For more general error distributions, Li (1987) demonstrates asymptotic optimality properties of Mallows' C_p criterion, cross-validation, and generalized cross-validation. Li's analysis applies only to linear models and he requires a condition that indirectly restrict the number of candidate models that may be considered. The analysis that we give here is similar in spirit to the work of Shibata and Li.

An advantage of the complexity regularization criterion is that it is amenable to statistical analysis for essentially any list of candidate functions. In particular, linearity of the parametric models need not be assumed. Also, no restriction need be placed on the number of candidate models that is considered for each sample size. In essence the effect of the number of candidate models is automatically accounted for in the complexity penalty. To permit this increased generality, our bounds do not yield asymptotic optimality in the sense of Shibata for the nonparametric case. Nevertheless, the bounds do show that the statistical risk is within a logarithmic factor of the optimum. Moreover, our results are seen to be applicable both for parametrically represented functions and for nonparametric functions, without prior knowledge of whether the true function is in the list of parametric models.

Below we give two examples to which the convergence theory applies. The theory is not restricted to these cases. These examples provide lists of parametric families of functions $f_M(x,\theta)$ for $x \in R^d$ and $\theta \in R^m$, where the dimension m depends on the model M. As a concession to computational feasibility, the size of the list of models may be restricted to depend on the sample size n, as long as the restrictions are relaxed as $n \to \infty$. Statistical and information-theoretical considerations show that there is no essential loss to restrict the coordinates of θ to a grid of points spaced at width equal to a small multiple of $1/\sqrt{n}$, where n is the sample size. Because of the discretization of the parameters, the list of models leads to a countable list Γ_n of candidate functions.

For each candidate function f, a complexity $C_n(f)$ is assigned. In general, these numbers $C_n(f)$ are arbitrary subject to a summability restriction on $2^{-C_n(f)}$. This restriction means that $C_n(f)$ may be interpreted as a codelength for the description of f, or, what turns out to be operationally equivalent but motivationally quite distinct, $2^{-C_n(f)}$ may be interpreted as a prior probability of f. For the examples we give, the complexities may be chosen to take on the following form

$$C_n(f_M(\cdot,\theta)) = \frac{m}{2}\log n + c_\theta + c_M, \tag{1}$$

where the logarithm is taken with base 2. Here $(m/2)\log n + c_\theta$ may be interpreted as a codelength for the parameter θ, or 2^{-c_θ} as a prior density function for the parameter. Likewise c_M may be interpreted as a codelength for the model M, or 2^{-c_M} as a prior probability, where $\sum_M 2^{-c_M} \leq 1$. The term c_M indirectly accounts for the effect of the number of candidate models. When there are many competing models, this term help to avoid overfit problems due to selection bias.

The examples we study have the property that essentially every function f^* is well approximated by functions in the list Γ_n. In particular, for every L^2 function with compact domain in R^d, there is a sequence of functions f_n in Γ_n, such that the L^2 approximation error converges to zero as $n \to \infty$. A consequence is that the index of resolvability $R_n(f^*) = \min_{f\in\Gamma_n}(\|f-f^*\|^2 + C_n(f)/n)$, which quantifies the tradeoff between the accuracy and complexity of the best approximations, is seen to converge to zero. The speed at which it converges depends on how well tailored the list Γ_n is to the function f^*. If f^* is a member of one of the candidate families then $R_n(f^*) = O(\log n)/n$. In other smooth cases, $R_n(f^*)$ converges at rate given by a fractional power of $(\log n)/n$. The theory we develop here gives conditions such that the statistical risk of estimated functions converges at rate bounded by $R_n(f^*)$.

EXAMPLE 1.1: MULTIVARIATE BASIS FUNCTION SELECTION

Let $B_k(x)$, $x \in R^d$, $k \in \{0,1,2,\ldots\}^d$ denote a collection of basis functions for multivariate functional expansion. In particular, we may take basis functions for polynomials, splines, or trigonometric series. For example, $B_k(x) = x_1^{k_1}\cdots x_d^{k_d}$ is the traditional basis for polynomials. Finite sets M of multi-indices k yield candidate models

$$f_M(x,\theta) = \sum_{k\in M}\theta_k B_k(x). \tag{2}$$

The lists of candidate models consist of collections of such f_M. In theory one may take the collection of all finite subsets M, and use the complexity regularization criterion to choose a statistically suitable subset. However, in practice it is more feasible to impose sequences of bounds on the maximal degree and the interaction order of the basis functions, and perhaps to restrict to hierarchical sets of terms, thereby yielding a sequence of lists Γ_n of candidate functions. Theoretically, the sequence of bounds should be fairly large (and tend to infinity as $n \to \infty$), so that we do not interfere with the desired approximation properties. The complexity regularization criterion makes the final choice of model within the possibly restricted class.

There are various statistical contexts in which such basis function expansions may be used, including least squares regression (see Cox (1988), Friedman (1990)),

maximum likelihood estimation of the log-density (see Stone (1990), Barron and Sheu (1988)), and logistic regression for nonparametric discrimination. As we shall see, complexity regularization and indices of resolvability are defined for each of these contexts.

Suppose the target function f^* is in the Sobolev space of functions on $[0,1]^d$ for which all partial derivatives up to order r are square integrable. In the contexts of regression and log-density estimation using polynomials, splines, or trigonometric series, the index of resolvability converges at rate

$$R_n(f^*) = O\left(\frac{\log n}{n}\right)^{2r/(2r+d)}. \tag{3}$$

This is shown for $d=1$ in Barron and Cover (1990) and extends by the same method to $d>1$ using multivariate approximation results in Sheu (1989).

These bounds on the index of resolvability then apply, using the main results from Section 4, to show that the statistical risk of the estimated function converges to zero at rate $O(\log n)/n$ in the case that the true function is in one of the candidate families and at rate $O(\log n/n)^{2r/(2r+d)}$ when the function is in the Sobolev space. Again these rates are obtained without apriori knowledge of the family or of the smoothness class.

EXAMPLE 1.2: ARTIFICIAL NEURAL NETWORKS

Feedforward artificial neural networks are a relatively flexible parametric family of functions $f(x,\theta)$ on R^d, which are expressed as the mathematical composition of fixed non-linear functions of one variable and parameterized linear functions of several variables. The most popular non-linear function in this context is the logistic $h(u) = 1/(1+e^{-u})$ for $-\infty < u < \infty$, although polynomial nonlinearities have also had some practical success. The references Farlow (1984), Rumelhart, et. al. (1986), Anderson and Rosenfeld (1988), Barron and Barron (1988), Lippmann (1987), and Lee and Lippmann (1990) provide a starting point on network methods of empirical modeling. Despite the name "neural," these networks are intended for general purpose approximation and estimation of functions; they are connected to biological networks only through a loose analogy. Artificial neural networks are intended as competitors to other methods of nonparametric function estimation, particularly in high-dimensional contexts in which the networks have an advantage of dimensionality reduction by composition. That is, if the true response is a function of many variables, that depends on these variables through the composition of lower-dimensional smooth functions, then network methods may be better suited than traditional series expansions which suffer from an exponentially large the number of candidate terms as a function of the dimension.

The architecture of an artificial neural network is the specification of the mathematical composition scheme. The layers of the network refer to the depth of composition. A basic network architecture is the multiple layer networks in which the output of the functions on one layer are inputs to every function on the next layer. Thus

$$f(x,\theta) = \sum_{m=1}^{M} \alpha_m g_m(x,\beta_m), \tag{4}$$

where each parameterized function $g(x,\beta)$ takes the form

$$g(x,\beta) = h(\beta^T z), \tag{5}$$

with $z = (1,z_1,\ldots,z_K)$ for some K, and $\beta \in R^{K+1}$. These elemental functions (5) are called units, nodes, or artificial neurons. In the case of one layer of non-linear nodes, the $z_k = x_k$ are original input variables. In the case of two-layer networks, each intermediate input z_k is itself the output of a another parameterized function of the form (5). This process is repeated for any additional layers of the network. The parameter vector θ consists of the linear parameters α and the parameters β from all of the nodes in the network.

Cybenko (1989) demonstrated an essential approximation-theoretic property of families of networks with the logistic non-linearity (or other bounded non-linearities with distinct left and right limits). Namely, the set of such networks with one layer of non-linear nodes is dense in the space of all bounded continuous functions with compact domain in R_d. It follows that classes of multiple layer networks are also dense. Similar results show the density of networks in the space of L^2 functions with compact domain and show that the density of networks holds for unbounded non-linearities such as polynomials, see Barron (1989). Very little is known about the rates of approximation of networks. One result in this direction, based on the work of Jones (1990), is that if a sequences of of network functions exist for which $\| f - f^* \|$ tends to zero and if the sum of the absolute values of the α parameters is bounded, then there exists network functions f_M of the form (4) with integrated squared error bounded by $\| f_M - f^* \|^2 \leq O(1/M)$.

Without appropriate statistical control on the size of an artificial neural network and a control on the number of estimated parameters, artificial neural networks are subject to the same overfit problems (inadequacy of the generalization of the predictions to new observation) that are associated with other parameter intensive models, such as the models in Example 1.1 above. To realize the potential benefits of neural network approximations, it is important to have a statistically accurate model selection criterion. Moreover, the criterion should be demonstrably applicable to network models that can be highly nonlinear functions of the parameters. These considerations were motivating factors in the development and analysis of the complexity regularization criterion.

The statistical convergence theory we present shows that artificial neural network estimated from data achieve the best rates of approximation, as measured by the index of resolvability for the given class of networks. This convergence theory does not necessarily hold for arbitrary neural network training algorithms; it applies to network estimators that optimize the complexity regularization criterion.

Results on consistency of statistically estimated networks were announced in Barron and Barron (1988). See also, White (1990) for consistency results for networks based on the method of sieves. The bounds on the statistical risk of functions estimated by complexity regularization that are proved in the present paper were announced in the neural network context in Barron (1989).

2. Loss Functions for Functional Estimation

Before giving our convergence result for functions estimated by complexity regularization, we address the issue of the choice of the loss function and the measure of statistical risk.

Let $(X_i,Y_i)_{i=1}^n$ be independent observations drawn from the unknown joint distribution of random variables X,Y, where the support of X is in $\mathbf{R}^d$. Here X is the vector of explanatory variables and Y is the response variable. Functions $f(X)$ are used in predicting the response. The error incurred by a prediction is measured by a distortion function $d(Y,f(X))$. (Informally d is sometimes referred to as a loss function, although in a strict decision-theoretic sense it is not, since it depends on the random variables X,Y). The empirical loss is $(1/n)\sum_{i=1}^n d(Y_i,f(X_i))$. Functions f_n are typically estimated by minimizing the empirical loss over a class of candidate functions, or by minimizing the empirical loss with a penalty added for the complexity of the functions.

We let f^* be a function which minimizes $E(d(Y,f(X)))$ over all measurable functions on $\mathbf{R}^d$. For the distortion functions we investigate, such a optimum function $f^*(x)$ exists and is related explicitly to the conditional distribution of Y given $X = x$. When a function f is used in place of the optimum function f^* the regret is measured by the difference between the expected distortions

$$r(f,f^*) = E(d(Y,f(X))) - E(d(Y,f^*(X))). \tag{6}$$

In our formulation, the quantity $r(f,f^*)$ is the theoretical loss incurred by the function f when the true best function is f^*. This loss function $r(f,f^*)$ depends on the function f over its whole domain (the support of the distribution of X) and not just on its values at one or several points. It quantifies the ability of an estimate to generalize, on the average, to new data from the distribution of X,Y. The statistical risk is the expected value of the loss for a given estimator f_n,

$$\text{risk} = E(r(f_n,f^*)). \tag{7}$$

More generally, we define $d^* = \inf_f E(d(Y,f(X)))$, $r(f) = E(d(Y,f(X))) - d^*$, and risk $= E(r(f_n))$, all of which are defined even if there does not exist an optimal function f^*. This $r(f)$ is the same as the loss function $r(f,f^*)$ in the case that an optimal function $f^*(x)$ exists.

The most common choice for the distortion function $d(Y,f(X))$ is the squared error $(Y-f(X))^2$, although other choices can also be handled in the theory, including the absolute error $|Y-f(X))|$. In the case of a dichotomous random variable Y, assumed to take values of either $+1$ or -1, reasonable choices include, in addition to the squared error, the zero-one distortion function $1_{\{Y \neq \operatorname{sgn}(f(X))\}}$, and the logistic distortion function $-Yf(X) + \log(e^{f(X)} + e^{-f(X)})$, which is the same, except for a linear rescaling as the distortion function used in traditional logistic regression.

A general class of distortion functions are those which take the form $d(Y,f(X)) = -\log p(Y \mid f(X))$ where $p(y \mid f(x))$ models the conditional density of Y given X. In particular, the squared error distortion corresponds to a Gaussian conditional density and the logistic distortion corresponds to the Bernoulli model written in exponential form $p(y \mid x) = e^{yf(x)}/(e^{f(x)} + e^{-f(x)})$, $y = \pm 1$, where $f(x)$ models

one-half the log-odds ratio in favor of class $+1$ versus class -1. For the estimation of the density function for the random vector X, we use $d(Y, f(X)) = -\log f(X)$ and require that $f(x)$ is non-negative and integrates to one with respect to a specified dominating measure.

For the squared error distortion, the loss function reduces to an integrated squared error

$$r(f, f^*) = E((f(X) - f^*(X))^2) \tag{8}$$

and $f^*(x) = E[Y \mid X = x]$ is the conditional mean of Y given $X = x$. In particular, for the dichotomous case with squared error loss, $f^*(x)$ is the difference between the conditional probabilities of class $+1$ and class -1. For the zero-one loss, $r(f, f^*)$ is the difference between the probability of error based on f and the Bayes optimal probability of error.

For distortion functions based on a family of conditional densities, f^* is the choice which makes the conditional density $p(y \mid f^*(x))$ closest to the true conditional density in the relative entropy sense. If the family of conditional densities is correctly specified, i.e. if it includes the true conditional density, then $r(f, f^*)$ is the average relative entropy distance between $p(\cdot \mid f^*(x))$ and $p(\cdot \mid f(x))$.

$$r(f, f^*) = E \log \frac{p(Y \mid f^*(X))}{p(Y \mid f(X))}. \tag{9}$$

In this context, a role will also be played by the squared Hellinger distance,

$$d_H^2(f(x), f^*(x)) = \int \left(\sqrt{p(y \mid f(x))} - \sqrt{p(y \mid f^*(x))}\right)^2 \mu(dy), \tag{10}$$

where μ is the measure assumed to dominate the family of conditional distributions for Y given X.

3. Complexity Regularization and the Index of Resolvability

We have already indicated that the complexity regularization criterion is motivated in certain cases by the minimum description-length principle. Indeed, these criteria are seen to coincide when the distortion function d is taken to be equal to minus the logarithm of likelihood and when the parameter λ in the definition of the criterion is set to equal one. Here we motivate the criterion in the context of uniform bounds on the statistical risk of estimators.

There are two contributions to the risk $E(r(f_n, f^*))$ of estimators based on the optimization of an empirical loss: namely, the approximation error $r(f, f^*)$ achieved by functions f in the given class as an approximation to the desired function f^*, and the estimation error, which is due to the discrepancy between the empirical and theoretical loss. By techniques in Vapnik (1982), Devroye (1988), or Haussler (1989), this discrepancy between empirical and theoretical averages can be shown to be uniformly bounded by $O(\sqrt{C_n/n})$, in probability, for families of functions of complexity bounded by C_n. (Essentially, C_n is taken there as the logarithm of the number of functions required to approximate functions in the class to within a prescribed accuracy.) By generalizations of these techniques, it is shown that the estimation error

can be bounded in probability by a multiple of $\sqrt{C_n(f)/n}$ for arbitrary bounded distortion functions and by $r(f,f^*) + C_n(f)/n$ for the squared error and for the likelihood-based distortion functions, uniformly for all candidate functions f. Here, instead of requiring a uniform complexity bound, we permit unbounded complexities $C_n(f)$ that may depend on the candidate functions. (These "complexities" $C_n(f)$ are arbitrary numbers satisfying a summability requirement, as given in equation (15) below, in accordance with an information-theoretic interpretation.) In the absence of constraints on the form or number of candidate functions, the empirical loss is not necessarily uniformly accurate for all candidate functions. The complexity penalties used here is the smallest penalty we are able to impose to force the criterion to be accurate at least for the functions which achieve the minimum criterion value.

Thus we are led to complexity regularization criteria and to corresponding indices of approximation. Depending on whether bounds of order $\sqrt{C_n(f)/n}$ or $C_n(f)/n$ arise in controlling the estimation error, we add the appropriate complexity penalty to the empirical loss to define the criterion for function estimation, for in this way it seen that the minimizer of the empirical criterion has a performance essentially as good as that achievable by the theoretical analog of the criterion.

Definition: Given a collection Γ_n of functions, numbers $C_n(f)$, $f \in \Gamma_n$, satisfying the summability condition (15), and a positive constant λ, the method of *complexity regularization* selects the function to minimize one of the following two criteria,

$$\frac{1}{n}\sum_{i=1}^{n} d(Y_i,f(X_i)) + \lambda(\frac{1}{n}C_n(f))^{1/2} \tag{11}$$

or

$$\frac{1}{n}\sum_{i=1}^{n} d(Y_i,f(X_i)) + \lambda\frac{1}{n}C_n(f). \tag{12}$$

The theoretical analog of these criteria leads to the following indices of resolvability, which quantify the tradeoff between the complexity and accuracy of approximations to f^*,

$$R_n^{(1)}(f^*) = \min_{f\in\Gamma_n}(r(f,f^*) + \lambda(\frac{1}{n}C_n(f))^{1/2}), \tag{13}$$

and

$$R_n^{(2)}(f^*) = \min_{f\in\Gamma_n}(r(f,f^*) + \lambda\frac{1}{n}C_n(f)). \tag{14}$$

The latter quantity is the index of resolvability introduced in an information-theoretic context in Barron and Cover (1990).

The requirement that is imposed on the numbers $C_n(f)$ is the following summability condition:

$$\sum_{f\in\Gamma_n} 2^{-C_n(f)} \le s, \tag{15}$$

for some finite constant s. There is an information-theoretic interpretation of the summability condition with $s = 1$: this is the Kraft-McMillan inequality which is

necessary and sufficient for the existence of uniquely decodable binary codes, with codelengths $C_n(f)$, for $f \in \Gamma_n$. We shall also require that $C_n(f) \geq l$ for all f and all n, for some positive constant l. This also is automatically satisfied, with $l = 1$ in the case of binary codes for a set Γ_n with more than one function. A trivial consequence of the latter requirement is that $R_n^{(1)}$ is not of order smaller than $1/\sqrt{n}$ and $R_n^{(2)}$ is not of order smaller than $1/n$.

4. Bounds on the Statistical Risk

Now we present the main results on statistical convergence properties of functions using a complexity regularization criterion. Let

$$f_n^{(1)} = \underset{f \in \Gamma_n}{\operatorname{argmin}} \left(\frac{1}{n}\sum_{i=1}^{n} d(Y_i, f(X_i)) + \lambda\left(\frac{1}{n}C_n(f)\right)^{1/2}\right), \tag{16}$$

and

$$f_n^{(2)} = \underset{f \in \Gamma_n}{\operatorname{argmin}} \left(\frac{1}{n}\sum_{i=1}^{n} d(Y_i, f(X_i)) + \lambda\frac{1}{n}C_n(f)\right). \tag{17}$$

The first estimator is used for distortion functions such as the zero-one distortion and the absolute error for which the ideal rate of convergence of the risk would be close to $1/\sqrt{n}$. The second estimator is used for squared error and log-likelihood based distortions for which close to $1/n$ would be the ideal rate.

Our main result, which we now give, shows that the statistical rate of convergence of functions estimated by complexity regularization is bounded by the index of resolvability.

The result requires in some cases that $d(Y, f(X))$ be almost surely bounded. This is forced by a constraint on the support of Y and by clipping the functions $f(X)$, or by explicit choice of a bounded distortion function. For distortions based on the log-likelihood, with a correctly specified family of conditional densities, no boundedness of the distortion is required.

Convergence Theorem for Complexity Regularization: *Assume that the indices of approximation $R_n^{(1)}(f^*)$ and $R_n^{(2)}(f^*)$ tend to zero as $n \to \infty$. If the range of $d(Y, f(X))$ for every f in Γ_n is in a fixed interval of length b, and if $\lambda > b/\sqrt{2 \log e}$ in the definition of the complexity regularized estimator $f_n^{(1)}$, then the statistical risk of the estimator converges to zero at rate bounded by $R_n^{(1)}(f^*)$, i.e.,*

$$E(r(f_n^{(1)}, f^*)) \leq O(R_n^{(1)}(f^*)). \tag{18}$$

Indeed, for all $n \geq 1$,

$$E(r(f_n^{(1)}, f^*)) \leq R_n^{(1)}(f^*) + \frac{c_0}{\sqrt{n}}, \tag{19}$$

where $c_0 = (s+1)b\sqrt{\pi/2}$.

For the squared error distortion function, if the support of Y and the range of each function $f(X)$ is in a known interval of length b, then with $\lambda \geq 5b^2/3 \log e$ in the definition of the estimator $f_n^{(2)}$, the mean squared error converges to zero at rate bounded

by $R_n^{(2)}(f^*)$, *i.e.,*

$$E((f_n^{(2)}(X) - f^*(X))^2) \leq O(R_n^{(2)}(f^*)). \tag{20}$$

If the distortion function is $d(Y, f(X)) = -\log p(Y \mid f(X))$ *where the true conditional density is* $p(Y \mid f^*(X))$, *then for all* $\lambda > 1$ *in the definition of the estimator* $f_n^{(2)}$, *the expected squared Hellinger distance between the conditional densities converges at rate bounded by the index of resolvability* $R_n^{(2)}(f^*)$, *i.e.,*

$$E(d_H^2(f_n^{(2)}(X), f^*(X))) \leq O(R_n^{(2)}(f^*)). \tag{21}$$

The L^1 distance, which takes the form $\int |p(y \mid f(x)) - p(y \mid f^*(x))| \mu(dy)$, is known to not be greater than twice the Hellinger distance. Therefore, a consequence of (21) is that the expected square of the L^1 distance also converges at rate bounded by $R_n^{(2)}(f^*)$.

For the Gaussian error case, the Hellinger distance can be evaluated and lower bounded as in Barron and Cover (1990). It is seen that for any $c > 0$, the risk $E(\min((f_n(X) - f^*(X))^2, c))$ converges to zero at rate bounded by $R_n^{(2)}(f^*)$.

5. Proof of the Convergence Theorem for Complexity Regularization

For simplicity in the proof we assume, without loss of generality, that the complexities are converted to base e. This means that the summability condition becomes $\sum_{f \in \Gamma_n} e^{-C_n(f)} \leq s$. Accordingly, for the purpose of the proof, the logarithms in the theorem are now interpreted as base e instead of base 2.

For the first two conclusions of the theorem, the proof uses a bound on the probability of the event that $r(f_n, f^*) > t$ in terms of a sum of probabilities of related events for each $f \in \Gamma_n$, to which inequalities of Hoeffding and Bernstein can be applied. The bounds on the probabilities are then integrated for $t > 0$ to obtain the indicated bounds on the risk. The proof of the third conclusion is based in part on results in Barron and Cover (1990) which uses inequalities of Chernoff.

The inequality of Hoeffding (1963, Theorem 2) states that if $(U_i)_{i=1}^n$ are independent random variables taking values in intervals of length b, then the distribution of the sample average $\overline{U} = (1/n)\sum_{i=1}^n U_i$ has the following exponential bound for all $\varepsilon > 0$,

$$P\{\overline{U} - E\overline{U} \geq \varepsilon\} \leq e^{-2n\varepsilon^2/b^2}. \tag{22}$$

The Bernstein-type inequalities make direct use of the variance as well as the expected values of the random variables. To state the Bernstein inequality, let U_i be independent random variables that satisfy the moment condition that for some $h > 0$,

$$E|U_i - EU_i|^k \leq \frac{\mathrm{var}(U_i)}{2} k!\, h^{k-2} \tag{23}$$

for $k \geq 2$, $i = 1, \ldots, n$. This is satisfied in particular if $|U_i - EU_i| \leq M$ with $h = M/3$. Bernstein's inequality states that for $t > 0$,

$$P\{\overline{U} - E\overline{U} \geq t\sigma_{\overline{U}}\} \leq \exp\{-t^2/(2 + 2ht/n\sigma_{\overline{U}})\} \tag{24}$$

see Craig (1933), Bennett (1962).

5.1. PROOF OF THE THEOREM: BOUNDED LOSS FUNCTION CASE

We examine the theoretical and empirical loss,

$$r(f) = Ed(Y, f(X)) - d^* \tag{25}$$

and

$$\hat{r}_n(f) = \frac{1}{n}\sum_{i=1}^{n} d(Y_i, f(X_i)) - d^*, \tag{26}$$

where we have subtracted the constant $d^* = \inf_f Ed(Y, f(X))$. By Hoeffding's inequality and the union of events bound, for any $\varepsilon_n(f) > 0$,

$$r(f) - \hat{r}_n(f) < \varepsilon_n(f) \quad \text{for all } f \in \Gamma_n, \tag{27}$$

except in an event of probability not greater than $\sum_{f \in \Gamma_n} e^{-2n(\varepsilon_n(f)/b)^2}$. Given $\delta > 0$, we choose $\varepsilon_n(f)$ such that $2n(\varepsilon_n(f)/b)^2 = C_n(f) + \ln 1/\delta$, to obtain that

$$r(f) - \hat{r}_n(f) < \frac{b}{\sqrt{2}}\left(\frac{C_n(f)}{n} + \frac{\ln 1/\delta}{n}\right)^{1/2} \quad \text{for all } f \in \Gamma_n \tag{28}$$

except in a set of probability not greater $\delta \sum_{f \in \Gamma_n} e^{-C_n(f)}$, which by the assumption on the complexities $C_n(f)$ is not greater than $s\delta$.

For the estimator f_n defined to achieve the minimum value of $\hat{r}_n(f) + \lambda(C_n(f)/n)^{1/2}$ and for $\lambda > b/\sqrt{2}$, we have that the following bounds hold on the loss $r(f_n)$, except in a set of probability not greater than $s\delta$,

$$\begin{aligned} r(f_n) &< \hat{r}_n(f_n) + \lambda\left(\frac{C_n(f_n)}{n}\right)^{1/2} + \frac{b}{\sqrt{2}}\left(\frac{\ln 1/\delta}{n}\right)^{1/2} \\ &\le \hat{r}_n(f_n^*) + \lambda\left(\frac{C_n(f_n^*)}{n}\right)^{1/2} + \frac{b}{\sqrt{2}}\left(\frac{\ln 1/\delta}{n}\right)^{1/2}. \end{aligned} \tag{29}$$

Taking f_n^* to be a function that achieves the best resolvability, that is, a function minimizing $r(f) + \lambda(C_n(f)/n)^{1/2}$, and applying Hoeffding's inequality once more, to get that $\hat{r}_n(f_n^*) \le r(f_n^*) + \lambda((\ln 1/\delta)/n)^{1/2}$ except in a set of probability not greater than δ, we obtain that

$$r(f_n) < r(f_n^*) + \lambda\left(\frac{C_n(f_n^*)}{n}\right)^{1/2} + 2\frac{b}{\sqrt{2}}\left(\frac{\ln 1/\delta}{n}\right)^{1/2}, \tag{30}$$

except in a set of probability not greater than $(s+1)\delta$. This shows that the loss of the estimator is bounded in terms of the index of resolvability,

$$r(f_n) \le R_n^* + O\left(\frac{1}{n}\right)^{1/2}, \tag{31}$$

in probability, where the index of resolvability is

$$R_n^* = \min_{f \in \Gamma_n}(r(f) + \lambda(C_n(f)/n)^{1/2}). \tag{32}$$

In particular, setting $\delta = e^{-(1/2)nt^2/b^2}$ for $t > 0$, we have

$$P\{r(f_n) \geq R_n^* + t\} \leq (s+1)e^{-(1/2)nt^2/b^2}. \tag{33}$$

Integrating for $0 < t < \infty$, yields,

$$\begin{aligned} E\, r(f_n) - R_n^* &\leq \int_0^\infty P\{r(f_n) - R_n^* \geq t\}\, dt \\ &\leq (s+1)\int_0^\infty e^{-(1/2)nt^2/b^2} dt \\ &= (s+1)\frac{b\sqrt{\pi/2}}{\sqrt{n}} \\ &= \frac{c_0}{\sqrt{n}}. \end{aligned} \tag{34}$$

Thus for all $n \geq 1$, the risk of the estimator is bounded by

$$E\, r(f_n) \leq R_n^* + \frac{c_0}{\sqrt{n}}. \tag{35}$$

This completes the proof of the bounds on the statistical risk of the complexity regularization estimator for bounded loss functions.

5.2. PROOF OF THE THEOREM: SQUARED ERROR CASE

For the squared error loss function, the ideal rate of convergence is close to $1/n$ instead of $1/\sqrt{n}$. For this case we use the criterion and the index of resolvability with penalty $\lambda C_n(f)/n$ instead of $\lambda(C_n(f)/n)^{1/2}$ and instead of Hoeffding's inequality we use a variant of Bernstein's inequality.

Direct use of Bernstein's inequality for our purposes is workable but cumbersome. We find it easier to use the following inequality, that Craig (1933) develops in his proof of Bernstein's inequality. If U_i are independent random variables satisfying Bernstein's moment condition, then

$$P\{\overline{U} - E\overline{U} \geq \frac{\tau}{n\varepsilon} + \frac{n\varepsilon \operatorname{var}(\overline{U})}{2(1-c)}\} \leq \exp\{-\tau\}, \tag{36}$$

for $0 < \varepsilon h \leq c < 1$ and $\tau > 0$.

Now to treat the complexity regularization estimator with the squared error distortion function, denote the difference in empirical loss at f and f^* by

$$\begin{aligned} \hat{r}_n(f, f^*) &= \frac{1}{n}\sum_{i=1}^n (Y_i - f(X_i))^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - f^*(X_i))^2 \\ &= -\frac{1}{n}\sum_{i=1}^n U_i, \end{aligned} \tag{37}$$

where

$$U_i = -(Y_i - f(X_i))^2 + (Y_i - f^*(X_i))^2.$$

Under the assumption that $f(X_i)$ and Y_i take values in a fixed interval of length b, we have that Bernstein's condition is satisfied with $h = 2b^2/3$. Next we bound the variance of U_i. To this end, we expand the square in the definition of U_i to get

$$U_i = 2(Y_i - f^*(X_i))(f(X_i) - f^*(X_i)) - (f(X_i) - f^*(X_i))^2. \tag{38}$$

The covariance between these two terms is zero. The expected square of the first term is $4E((\sigma^2_{Y|X})(f(X)-f^*(X))^2)$ which is not greater than $b^2E(f(X)-f^*(X))^2$ (since the variance of a distribution concentrated on an interval of length b is not greater than the variance $b^2/4$ achieved by the distribution that places mass 1/2 at each endpoint). The expected square of the second term is also bounded by $b^2E(f(X)-f^*(X))^2$. Together these bounds yield

$$\mathrm{var}(U_i) \le 2b^2 r(f, f^*). \tag{39}$$

It follows that $n\ \mathrm{var}(\overline{U}) \le 2b^2 r(f, f^*)$. Now apply the union of events bound and the Bernstein-type inequality with $\tau = C_n(f) + \ln 1/\delta$ and $\varepsilon = 1/\lambda$, to obtain that

$$r(f, f^*) - \hat{r}_n(f, f^*) < \lambda \frac{C_n(f)}{n} + \frac{b^2}{\lambda(1-c)} r(f, f^*) + \lambda \frac{\ln 1/\delta}{n} \tag{40}$$

for all f in Γ_n, except in an event of probability not greater than $s\delta$. Set $c = \varepsilon h = 2b^2/(3\lambda)$. The assumption that $\lambda > 5b^2/3$ implies that $\alpha = b^2/(\lambda(1-c)) < 1$. We collect the terms involving $r(f, f^*)$ on the left side, and evaluate at the complexity regularization estimator f_n to obtain

$$\begin{aligned}(1-\alpha) r(f_n, f^*) &\le \hat{r}_n(f_n, f^*) + \lambda \frac{C_n(f_n)}{n} + \lambda \frac{\ln 1/\delta}{n} \\ &\le \hat{r}_n(f_n^*, f^*) + \lambda \frac{C_n(f_n^*)}{n} + \lambda \frac{\ln 1/\delta}{n},\end{aligned} \tag{41}$$

where f_n^* is a function that achieves the best resolvability. Applying the Bernstein-type inequality once more, but now with $\tau = \ln 1/\delta$, to get that $\hat{r}_n(f_n^*, f^*) \le r(f_n^*, f^*) + \alpha r(f_n^*, f^*) + \lambda(1/n)\ln 1/\delta$ except in an event of probability not greater that δ, we obtain that

$$(1-\alpha) r(f_n, f^*) \le (1+\alpha) r(f_n^*, f^*) + \lambda \frac{C_n(f_n^*)}{n} + + 2\lambda \frac{\ln 1/\delta}{n}, \tag{42}$$

except in an event of probability not greater than $(s+1)\delta$. Dividing through by $(1-\alpha)$, and using the definition of the index of resolvability, we have that

$$r(f_n, f^*) \le \frac{1+\alpha}{1-\alpha} R_n(f^*) + 2\lambda \frac{\ln 1/\delta}{n}, \tag{43}$$

except in an event of probability not greater than $(s+1)\delta$. Setting $\delta = e^{-nt/(2\lambda)}$ and integrating the probability for $0 < t < \infty$ as in the previous case, we conclude that for all $n \ge 1$,

$$E(\, r(f_n, f^*)) \le \frac{1+\alpha}{1-\alpha} R_n(f^*) + \frac{2\lambda}{n}. \tag{44}$$

This completes the proof of the bound on the mean squared error of the complexity regularization estimator.

5.3. PROOF OF THE THEOREM: LIKELIHOOD CASE

The third conclusion of the Theorem is essentially given in Barron and Cover (1990). There it is shown that the squared Hellinger distance converges in probability at rate

given by the index of resolvability. Here we complete the reasoning to show that the expected squared Hellinger distance also converges at the same rate.

Let $P = P_{Y|f^*(X)}P_X$ be the true distribution of X, Y, let $P_n = P_{Y|f_n(X)}P_X$ be the joint distribution obtained by plugging in the minimum complexity estimator f_n, and let $P_n^* = P_{Y|f_n^*(X)}P_X$ be the joint distribution obtained by plugging in a function f_n^* that achieves the best resolvability. Using inequalities of Chernoff, exponential bounds are derived in Barron and Cover (1990) for the P_n^* probability that the ratio of the squared Hellinger distance (between f_n and f_n^*) and the index of resolvability is greater than an arbitrary positive constant. In particular,

$$P_n^*\{(1-\alpha)\frac{d_H^2(f_n,f_n^*)}{R_n(f^*)} > c\} \le s\, e^{-cnR_n(f^*)}e^{\lambda C_n(f_n^*)}, \tag{45}$$

where $\alpha = 1/\lambda$. Applying the Lemma with $U = (1-\alpha)d_H^2(f_n,f_n^*)/R_n(f^*)$, with $r_1 = nR_n(f^*)/2$ and $r_2 = \lambda C_n(f_n^*)$, it follows that

$$\begin{aligned}(1-\alpha)\frac{E(d_H^2(f_n,f_n^*))}{R_n(f^*)} &\le s\int_0^\infty e^{-cnR_n(f^*)/2}dc + \frac{nr(f_n^*,f^*) + \lambda C_n(f_n^*)}{nR_n(f^*)/2} + \frac{e^{-1}}{nR_n(f^*)/2} \\ &\le \frac{s + e^{-1}}{nR_n(f^*)/2} + 2 \\ &\le 2\frac{s + e^{-1}}{l} + 2,\end{aligned} \tag{46}$$

where we have used the fact that, in the present context, $r(f_n^*,f^*)$ is the relative entropy distance between P and P_n^*. Also, in the last line we used $nR_n(f^*) \ge C_n(f_n^*) \ge l$. We conclude that the expected squared Hellinger distance between f_n and f_n^* is bounded by a constant times the index of resolvability, for all $n \ge 1$,

$$E(d_H^2(f_n,f_n^*)) \le c_1 R_n(f^*), \tag{47}$$

where the constant is $c_1 = 2(1+(s+e^{-1})/l)/(1-\alpha)$. Using the triangle inequality to get $d_H^2(f_n,f^*) \le 2(d_H^2(f_n,f_n^*) + d_H^2(f^*,f_n^*))$ and then using the fact that the squared Hellinger distance $d_H^2(f^*,f_n^*)$ is bounded by the relative entropy $r(f_n^*,f^*)$ which in turn is less than $R_n(f^*)$, we obtain

$$E(d_H^2(f_n,f^*)) \le 2(c_1 + 1)R_n(f^*). \tag{48}$$

This completes the proof of the Theorem.

Lemma: For any nonnegative random variable U, any pair of distributions P and Q, and constants $r_1 > 0$, and $r_2 \ge 0$,

$$E(U) \le e^{-r_2}\int_0^\infty Q\{U > c\}e^{r_1 c}dc + \frac{D(P\,||\,Q) + r_2 + e^{-1}}{r_1}, \tag{49}$$

where the expectation is with respect to P, and $D(P\,||\,Q)$ is the relative entropy distance between P and Q.

Proof of the Lemma: The inequality is trivial if the relative entropy distance is infinite. Now suppose $D(P \| Q)$ is finite, so there is a density ratio dP/dQ. By a simple calculation as in Barron and Cover (1990, Lemma 2),

$$P\{U > c\} \leq Q\{U > c\}e^{r_1 c - r_2} + P\{\frac{1}{r_1}(r_2 + \log \frac{dP}{dQ}) > c\}. \tag{50}$$

Integrating for $c > 0$ then gives

$$E(U) \leq 2^{-r_2}\int Q\{U > c\}e^{cr_1}dc + \frac{1}{r_1}E(r_2 + \ln \frac{dP}{dQ})^+. \tag{51}$$

Using the fact that $E(\ln dP/dQ)^+ \leq D(P \| Q) + e^{-1}$ completes the proof of the Lemma.

6. Acknowledgement

This work is supported by an Office of Naval Research Contract N00014-89-J-1811.

7. References

Anderson, James A. and Rosenfeld, Edward (1988) *Neurocomputing: Foundations of Research*, MIT press.

Barron, Andrew R. (1985) "Logically smooth density estimation," Ph.D. dissertation, Department of Electrical Engineering, Stanford University.

Barron, Andrew R. (1989) "Statistical properties of artificial neural networks," *Proc. 28th Conference on Decision and Control*, IEEE, New York.

Barron, Andrew R. and Barron, Roger L. (1988) "Statistical learning networks: a unifying view," *Computing Science and Statistics: Proc. 20th Symp. Interface.*, Edward Wegman, editor, Amer. Statist. Assoc., Washington, DC., 192-203.

Barron, Andrew R. and Cover, Thomas M. (1990) "Minimum complexity density estimation," To appear in *IEEE Trans. Inform. Theory.*

Barron, Andrew R. and Sheu, Chyong-Hwa (1988) "Approximation of density functions by sequences of exponential families," To appear in *Ann. Statist.*

Bennett, George (1962) "Probability inequalities for the sum of independent random variables," *J. Amer. Statist. Assoc.*, **57**, 33-45.

Cox, Dennis D. (1988) "Approximation of least squares regression on nested subspaces," *Ann. Statist.*, **18**, 713-732.

Craig, Cecil C. (1933) "On the Tchebychef inequality of Bernstein" *Ann. Math. Statist.*, **4**, 94-102.

Cybenko, George (1989) "Approximations by superpositions of sigmoidal functions," *Math. Control, Signals, Systems*, **2**, 303-314.

Devroye, Luc (1988) "Automatic pattern recognition: a study of the probability of error," *IEEE Trans. Pattern Anal. Mach. Intelligence*, **10**, 530-543.

Farlow, Stanley J. (1984) "Self Organizing Methods in Modeling: GMDH Type Algorithms," Marcel Dekker, New York.

Friedman, Jerome H. (1990) "Multivariate adaptive regression splines (with discussion." To appear in the *Ann. Statist.*

Haussler, David (1989) "Generalizing the PAC model for neural net and other learning applications," Computer Research Laboratory, Technical Report 89-30, University of California, Santa Cruz.

Hoeffding, W. (1963) "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, **58**, 13-30.

Jones, Lee (1990). "A simple lemma on greedy approximations in Hilbert space and convergence rates for projection pursuit regression and neural network training," To appear in IAnn. Statist..

Lee, Yuchun and Lippmann, Richard P. (1990) "Practical characteristics of neural network and conventional pattern classifiers on artificial and speech problems," *Advances in Neural Information Processing Systems 2*, David S. Touretzky, editor, Morgan Kauffmann Publishers, San Mateo, CA.

Li, Ker-Chau (1987) "Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: discrete index set. *Ann. Statist.*, **15**, 958-975.

Lippmann, Richard P. (1987) "An introduction to computing with neural nets," *IEEE Communications Magazine*, **4**, 4-22.

Rissanen, Jorma (1983) "A universal prior for integers and estimation by minimum description length," *Ann. Statist.*, **11**, 416-431.

Rissanen, Jorma (1984) "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, **30**, 629-636.

Rumelhart, David E., McClelland, James L., et. al. (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press.

Schwarz, Gideon (1978) "Estimating the dimension of a model," *Ann. Statist.*, **6**, 461-464.

Sheu, Chyong-Hwa (1989) "Density estimation with Kullback-Leibler loss," Ph.D. Thesis, Department of Statistics, University of Illinois, Champaign, Illinois.

Shibata, Ritei (1981) "An optimal selection of regression variables," *Biometrika*, **68**, 45-54.

Vapnik, V. N. (1982) *Estimation of Dependences Based on Empirical Data*, Springer Verlag, New York.

White, Halbert (1990) "Connectionists nonparametric regression: multilayer feedforward networks can learn arbitrary mappings," To appear in *Neural Networks*.

DESIGNING PREDICTION BANDS

RUDOLF BERAN
University of California
Berkeley, California 94720 USA

ABSTRACT. This article develops four principles for the design of good prediction bands for a random process. The issues addressed include: appropriate asymptotic convergence of conditional and unconditional coverage probabilities; probability centering of prediction bands; controlling dispersion of conditional coverage probabilities; and increasing the rate of convergence of unconditional coverage probabilities. Examples illustrate the design issues and a proposed bootstrap construction for good prediction bands.

1. Introduction

Prediction bands for a random process will be discussed in the following setting. An observed learning sample Y_n and a potentially observable variable X have a joint distribution $P_{\theta,n}$. The unknown parameter θ lies in a parameter space Θ. To be predicted is the potentially observable random process $Z = \{Z(u,X): u \in U\}$, where the index set U may be *infinite* and the function $Z(\cdot,\cdot)$ is specified. In this article, both Θ and U are metric spaces.

The treatment will emphasize the case where the process Z is real-valued, the index set U is infinite, and the aim is to devise a good one-sided prediction band for Z. By careful extension of the index set U and of the function $Z(\cdot,\cdot)$, the analysis for this one-sided case also applies to two-sided prediction bands and to multivariate prediction regions. These possibilities will be illustrated through examples.

Let $z = \{z(u): u \in U\}$ denote a generic possible value of the random process Z. Define the one-sided prediction band for Z

$$D_n = \{z: z(u) \le c_n(u) \text{ for every } u \in U\}, \qquad (1.1)$$

where the critical values $c_n(u)$ depend on the learning sample Y_n. Clearly, D_n is equivalent to simultaneously asserting the one-sided prediction intervals

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 577–586.
© 1991 *Kluwer Academic Publishers.*

$$D_{n,u} = \{z(u): z(u) \le c_{n(u)}\}, \qquad u \in U \tag{1.2}$$

for the respective process values $Z(u)$.

The coverage probability of prediction band D_n or of each prediction interval $D_{n,u}$ can be assessed conditionally, given the learning sample, or unconditionally. Let $P_\theta(\cdot|Y_n)$ denote the conditional distribution of X given Y_n. The *conditional coverage probability* of D_n given Y_n is

$$CP(D_n|Y_n,\theta) = P_\theta(Z \in D_n|Y_n). \tag{1.3}$$

The *unconditional coverage probability* of D_n is

$$\begin{aligned} CP(D_n|\theta) &= P_{\theta,n}(Z \in D_n|Y_n) \\ &= E_\theta CP(D_n|Y_n,\theta) \end{aligned} \tag{1.4}$$

where the expectation is calculated with respect to the distribution $Q_{\theta,n}$ of Y_n. The notations $CP(D_{n,u}|Y_n,\theta)$ and $CP(D_{n,u}|\theta)$ similarly denote the conditional and unconditional coverage probabilities of $D_{n,u}$. Conditional coverage probabilities have been emphasized by Box and Jenkins (1976) in prediction of time series and by writers on tolerance regions (cf. Guttman 1970). Unconditional coverage probabilities have been supported by Cox and Hinkley (1974), among others. Additional background references on prediction regions are cited in Beran (1990a) and Carroll and Ruppert (1989).

On what basis should we choose the critical values $\{c_n(u): u \in U\}$ in the definition (1.1) of D_n? Suppose the real-valued random variables $\{Z(u,\cdot): u \in U\}$ measure logically similar attributes. In addition, suppose that X and Y_n are independent. Two natural design requirements on the critical values $\{c_n(u)\}$ are then:

Design Principle 1. Choose the $\{c_n(u)\}$ so that

$$CP(D_n|Y_n,\theta) \to \alpha \text{ in } Q_{\theta,n} \text{ probability} \tag{1.5}$$

as n increases. The value of the constant $\alpha \in (0,1)$ is prechosen.

Design Principle 2. Having satisfied design principle 1, choose the $\{c_n(u)\}$ so that

$$\sup_u |CP(D_{n,u}|Y_n,\theta) - \beta(\alpha,\theta| \to 0 \text{ in } Q_{\theta,n} \text{ probability} \tag{1.6}$$

for some constant $\beta(\alpha,\theta)$ which does not depend on u but can depend on α and θ.

The first design principle controls the asymptotic simultaneous coverage probability of prediction band D_n, both conditionally and unconditionally. The second design principle controls the *relative* asymptotic conditional and unconditional coverage probabilities of the $\{D_{n,u}\}$. On the one hand, since the $\{Z(u,\cdot)\}$ are assumed to be logically similar, it is only fair to require that the $\{D_{n,u}\}$ have equal coverage probabilities. From this viewpoint, design principle 2 has been called conditional asymptotic *balance* of the prediction band D_n (cf. Beran 1988 and 1990b). On the other hand, for particular choices of the $\{Z(u,\cdot)\}$, design principle 2 serves to *center* the prediction band D_n, in a conditional probability sense. This second interpretation is illustrated in Examples 1 and 2 of Section 2.

When the learning sample Y_n and the future observable X are dependent, the technical formulation of design principle 2 becomes a little more complex, since then (1.6) no longer represents the most general possibility.

Typically, more than one construction of the critical values $\{c_n(u)\}$ will satisfy design principles 1 and 2. On what basis might the choice be narrowed further? Considering the conditional and unconditional errors in the balance and coverage probability of prediction band D_n suggests two futher design goals:

Design Principle 3. Having satisfied design principles 1 and 2, choose the $\{c_n(u)\}$ so as to minimize asymptotically the dispersions of $CP(D_n|Y_n,\theta) - \alpha$ and of the $\{CP(D_{n,u}|Y_n,\theta) - \beta(\alpha,\theta)\}$.

Design Principle 4. Having satisfied design principles 1, 2 and 3, choose the $\{c_n(u)\}$ so as to maximize the rates of convergence to 0 of $CP(D_n|\theta) - \alpha$ and of the $\{CP(D_{n,u}|Y_n,\theta) - \beta_n(\alpha,\theta)\}$ for some $\{\beta_n(\alpha,\theta)\}$ that converge to $\beta(\alpha,\theta)$.

Design principle 4 addresses asymptotic biases in the conditional simultaneous converage probability and balance of D_n while design principle 3 seeks to control the dispersions of the various conditional coverage probabilities.

2. Design Principles 1 and 2

Suppose X and Y_n are independent, with distributions P_θ and $Q_{\theta,n}$ respectively. Let L(U) denote the set of all functions on U which take values in [0,1]. Metrize L(U) with the supremum norm $\|\cdot\|$. Assume without loss of generality that the sample paths of the process Z belong to L(U). This requirement can always be achieved by a strictly monotone increasing transformation of the $\{Z(u,\cdot): u \in U\}$ without changing the structure of D_n as defined by (1.1).

Define the functional

$$B(f,\theta) = P_\theta[Z(u,\cdot) \le f(u) \text{ for every } u \in U, \quad f \in L(U)] \tag{2.1}$$

and the cdf's

$$A_u(x,\theta) = P_\theta[Z(u,\cdot) \le x], \quad u \in U. \tag{2.2}$$

Two further assumptions are used in the analysis of this section:

Assumption A. For every θ, the cdf's $\{A_u(x,\theta): u \in U\}$ are equicontinuous and strictly monotone in x; and the corresponding quantile functions $\{A_u^{-1}(x,\theta): u \in U\}$ are equicontinuous in x.

Define the cdf

$$\begin{aligned} A(x,\theta) &= P_\theta[Z(u,\cdot) \le A_u^{-1}(x,\theta) \text{ for every } u] \\ &= P_\theta[\sup_u A_u\{Z(u,\cdot),\theta\} \le x] \\ &= B[A_{\cdot}^{-1}(x,\theta),\theta]. \end{aligned} \tag{2.3}$$

Assumption B. For every θ, the functional $B(f,\theta)$ is continuous in f and the cdf $A(x,\theta)$ is strictly monotone and continuous in x.

These assumptions yield a simple necessary and sufficient condition for satisfying design principles 1 and 2.

Proposition 1. Suppose Assumptions A and B hold. Then, design principles 1 and 2 are satisfied if and only if

$$\sup_u |c_n(u) - A_u^{-1}[A^{-1}(\alpha,\theta),\theta]| \to 0 \text{ in } Q_{\theta,n} \text{ probability.} \tag{2.4}$$

Remark A. Arguments in Beran (1990b) for the case of finite U can be extended to prove this result.

Remark B. Suppose $\hat\theta_n = \hat\theta_n(Y_n)$ is an estimate of θ such that $\hat\theta_n \to \theta$ in $Q_{\theta,n}$ probability. Then, the plug-in critical values

$$c_n(u) = A_u^{-1}[A^{-1}(\alpha,\hat\theta_n),\hat\theta_n], \quad u \in U \tag{2.5}$$

satisfy design principles 1 and 2, given the conditions for Proposition 1 and equicontinuity of the quantile functions in (2.5).

Example 1. Let the $\{X_i: i \ge 1\}$ be a sequence of i.i.d. r-variate $N(\mu,\Sigma)$ random vectors, the parameter $\theta = (\mu,\Sigma)$ being unknown. Let $U = \{u \in R^r: |u|=1\}$ be the unit ball let $Y_n = (X_1,\ldots,X_n)$ and let $X = X_{n+1}$. Define the process Z on U by $Z(u,X) = u'X$. The aim is to predict Z on the basis of learning sample Y_n.

Let $\hat{\theta}_n = (\bar{X}_n, S_n)$ denote the usual estimate of $\theta = (\mu, \Sigma)$, let Φ denote the standard normal cdf, and write $\sigma_u^2 = u'\Sigma u$, $s_{n,u}^2 = u'S_n u$. Evidently

$$A_u(x,\theta) = \Phi[x-u'\mu)/\sigma_u] \tag{2.6}$$

$$A(x,\theta) = \chi_r\{[\Phi^{-1}(x)]^2\},$$

where χ_r is the cdf of the chi-squared distribution with r degrees of freedom. From Proposition 1 and Remark B, the one-sided prediction band

$$D_n = \{z : z(u) \le u'\bar{X}_n + s_{n,u}\{\chi_r^{-1}(\alpha)\}^{1/2} \quad \text{for every } u \in U\} \tag{2.7}$$

for the process $Z(u,\cdot)$ satisfies design principles 1 and 2.

By the defilnition $Z(u,X) = u'X$, the set D_n can be rewritten equivalently as a multivariate prediction region for X:

$$D_n = \{x : (x - \bar{X}_n)'S_n^{-1}(x - \bar{X}_n) \le \chi_r^{-1}(\alpha)\}. \tag{2.8}$$

An arbitary tangent hyperplane to prediction ellipsoid (2.8) divides R^r into two halfspaces. The conditional probability, given Y_n, that X lies in the halfspace *not* containing the ellipsoid is asymptotically the same, whatever the choice of tangent hyperplane. In this sense, prediction ellipsoid (2.8) for X is probability centered; and this centering property is equivalent here to satisfying design principle 2.

Example 2. Replace the normality assumption in Example 1 with the weaker assumption that each X_i has an unknown cdf F whose density exists and is strictly positive over R^r. Estimate $\theta = F$ by $\hat{\theta}_n = \hat{F}_n$, the empirical cdf of the learning sample. Then $A_u(\cdot,\hat{F}_n)$ is the empirical cdf on the $\{u'X_i : 1 \le i \le n\}$ for every unit vector u; and $A_u(\cdot,\hat{F}_n)$ is the empirical cdf of the $\{\sup_{|u|=1} A_u(u'X_i,\hat{F}_n) : 1 \le i \le n\}$ or

equivalently of the $\{\sup_{|u|=1} n^{-1}\text{rank}(u'X_i) : 1 \le i \le n\}$.

The applicability of Proposition 1 and Remark B to this example can be established. Consequently, prediction band D_n, defined by (1.1) with critical values

$$c_n(u) = A_u^{-1}[A^{-1}(\alpha,\hat{F}_n),\hat{F}_n], \quad u \in U, \tag{2.9}$$

satisfies design principles 1 and 2. Viewed as a prediction region for the vector X, D_n is a convex set that is probability centered, asymptotically in the sense of Example 1.

Example 3. The discussion of gait analysis by Olshen, Biden, Wyatt and Sutherland (1989) suggests the following simplified model. Let the $\{X_i : i \geq 1\}$ be i.i.d Gaussian processes on [1,2], with unknown mean function $\mu(\cdot)$ and covariance kernel $\Sigma(\cdot,\cdot)$. The parameter θ is the pair $\left(\mu(\cdot), \Sigma(\cdot,\cdot)\right)$. On the basis of the learning sample $Y_n = (X_1,\ldots,X_n)$, the aim is to predict the process

$$Z(u,X) = \begin{cases} X(u) & \text{if } 1 \leq u \leq 2 \\ -X(-u) & \text{if } -1 \leq u \leq -1 \end{cases} \tag{2.10}$$

where $X = X_{n+1}$. Thus $U = [-2,-1] \cup [1,2]$.

Let $\hat{\theta}_n = (\bar{X}_n, S_n)$ denote the usual estimate of θ and write $\sigma^2(u) = \Sigma(u,u)$, $s_n^2(u) = S_n(u,u)$ for u in [0,1]. In this example,

$$A_u(x,\theta) = \begin{cases} \Phi[\sigma^{-1}(u)\{x - \mu(u)\}], & 1 \leq u \leq 2 \\ \Phi[\sigma^{-1}(-u)\{x - \mu(-u)\}], & -2 \leq u \leq -1 \end{cases} \tag{2.11}$$

and

$$A(x,\theta) = \Pr[\sup_{u \in U} \Phi\{\sigma^{-1}(u)|Z(u)-\mu(u)|\} \leq z]. \tag{2.12}$$

By Proposition 1 and Remark B, the critical values

$$c_n(u) = \begin{cases} \bar{X}_n(u) + s_n(u)\Phi^{-1}[A^{-1}(\alpha,\hat{\theta}_n)], & 1 \leq u \leq 2 \\ -\bar{X}_n(u) + s_n(-u)\Phi^{-1}[A^{-1}(\alpha,\hat{\theta}_n)], & -2 \leq u \leq -1 \end{cases} \tag{2.13}$$

satisfy design principle 1 and 2. The one-sided prediction band for $Z(\cdot,\cdot)$, defined by (1.1) and (2.13), is equivalent to a two-sided prediction band for the process X:

$$D_n = \{x : |x(u) - \bar{X}_n(u)| \leq s_n(u)\Phi^{-1}[A^{-1}(\alpha,\hat{\theta}_n)], \quad 1 \leq u \leq 2\}. \tag{2.14}$$

This band for X is asymptotically probability centered at every u in [0,1], because design principle 2 is satisfied in its construction.

3. Design Principles 3 and 4

Having satisfied design principles 1 and 2 through (2.4), we seek to control the dispersion and bias of the conditional coverage probabilities of D_n and the $\{D_{n,u}\}$. Define on the unit interval, the cdf

$$\tilde{A}(x,\theta,t) = B[A^{-1}(x,t)\theta] \tag{3.1}$$

and the functions

$$C_u(x,\theta,t) = A_u[A_u^{-1}\{A^{-1}(x,t),t\},\theta]$$

$$C(x,\theta,t) = \tilde{A}[A^{-1}\{A^{-1}(x,t),\theta,t]. \tag{3.2}$$

When the critical values $\{c_n(u)\}$ of D_n are given by (2.5),

$$CP(D_{n,u}|Y_n,\theta) = C_u(\alpha,\theta,\hat{\theta}_n)$$

$$CP(D_n|Y_n,\theta) = C(\alpha,\theta,\hat{\theta}_n). \tag{3.3}$$

For simplicity, suppose the parameter θ is real-valued. Impose the following assumption on the estimate $\{\hat{\theta}_n\}$:

Assumption C. The asymptotic distribution of $\{n^{1/2}(\hat{\theta}_n-\theta)\}$ as n increases is $N(0,\sigma^2(\theta))$ for some finite $\sigma^2(\theta)$. There exist functions $\{a_j(\theta)\colon j=1,2\}$ such that

$$E_\theta(\hat{\theta}_n-\theta)^j = n^{-1}a_j(\theta) + O(n^{-1}) \tag{3.4}$$

for $j = 1,2$.

In what follows, notation like $C^{(i,j,k)}(x,\theta,t)$ represents the partial derivative $\partial^{i+j+k}C(x,\theta,t)/\partial x^i\partial\theta^j\partial t^k$ and $C_u^{(i,j,k)}(x,\theta,t)$ denotes the appropraite norm derivative of $\{C_u(x,\theta,t)\colon u \in U\}$ as an element of L(U).

Proposition 2. Suppose Assumption C holds, the critical values $\{c_n(u)\}$ are given by (2.5),

$$\tau_u^2(\theta) = [C_u^{(0,0,1)}(\alpha,\theta,\theta)]^2\sigma^2(\theta),$$

$$\tau(\theta) = [C^{(0,0,1)}(\alpha,\theta,\theta)]^2\sigma^2(\theta) \tag{3.5}$$

and

$$b_u(\alpha,\theta) = \sum_{j=1}^{2} a_j(\theta)C_u^{(0,0,j)}(\alpha,\theta,\theta)/j!$$

$$b(\alpha,\theta) = \sum_{j=1}^{2} a_j(\theta)C_u^{(0,0,j)}(\alpha,\theta,\theta)/j!\ . \tag{3.6}$$

Then, under regularity condition,

$$n^{1/2}[CP(D_{n,u}|Y_n,\theta) - A^{-1}(\alpha,\theta)] \Rightarrow \tau_u(\theta)W$$

$$n^{1/2}[CP(D_n|Y_n,\theta) - \alpha] \Rightarrow \tau(\theta)W \tag{3.7}$$

where W is a standard normal random variable and the first weak convergence occurs as random elements of L(U). Moreover,

$$CP(D_{n,u}|\theta) = A^{-1}(\alpha,\theta) + n^{-1}b_u(\alpha,\theta) + O(n^{-2})$$

$$CP(D_n|\theta) = \alpha + n^{-1}b(\alpha,\theta) + O(n^{-2}) \tag{3.8}$$

the first convergence occurring as elements of L(U).
Remark C. Proposition 2 extends to infinite U some results in Beran (1990b). To satisfy design principle 3, it follows from (3.7) that the estimate $\{\hat{\theta}_n\}$ should be asymptotically efficient. On the other hand, (3.8) is not the best taht we can achieve toward satisfying design principle 4; this point is demonstrated by Proposition 3 below.

Let

$$S_n(u) = A[A_u\{Z(u,\cdot),\hat{\theta}_n\},\hat{\theta}_n] \tag{3.9}$$

and define the cdf's

$$H_{n,u}(x,\theta) = P_{\theta,n}[S_n(u) \le x],\ u \in U$$

$$H_n(x,\theta) = P_{\theta,n}[\sup_u H_{n,u}(S_n(u),\theta) \le x]. \tag{3.10}$$

Observe that $H_{n,u}(\alpha,\theta)$ is precisely $CP(D_{n,u}|\theta)$ when the critical values are given by (2.5)

Proposition 3. Suppose that the critical values $\{c_n(u)\}$ are given by

$$c_n(u) = A_u^{-1}[A^{-1}\{H_{n,u}^{-1}[H_n^{-1}(\alpha,\hat{\theta}_n),\hat{\theta}_n],\hat{\theta}_n\},\hat{\theta}_n] \tag{3.11}$$

and that the other assumptions of Proposition 2 hold. Then, under regularity conditions, the conclusions (3.7) are still valid. However,

$$CP(D_{n,u}|\theta) = H_n^{-1}(\alpha,\theta) + O(n^{-2})$$

$$CP(D_n|\theta) = \alpha + O(n^{-2}) \tag{3.12}$$

instead of (3.8), the first convergence ocurring as elements of L(U).

Remark D. The critical values defined in (3.11) differ from those in (2.5) by $O_p(n^{-1})$. This small perturbation does not matter for the purpose of achieving design principles 1,2 and 3, but does help toward satisfying design principle 4.

Remark E. Calculating the cdf's $A(\cdot,\theta)$ and $\{A_u(\cdot,\theta)\}$ may require a bootstrap algorithm (cf. Beran 1988, 1990b). The supremum over the infinite set U may be approximated by a systematic search algorithm or by random search over U (cf. Beran and Millar 1987, Beran 1988). Computing the cdf's $H(\cdot,\theta)$ and $\{H_{n,u}(\cdot,\theta)\}$ usually requries another round of bootstrap approximation. These comments apply to the calculation of the critical values (2.5) and (3.11).

Example 1 (continued). Using the refined critical values (3.11) in this example replaces the chi-squared quantile in (2.7) with the αth quantile of the F distribution with r and n-r degrees of freedom. Consequently, the $\{CP(D_{n,u}|\theta)\}$ are equal and $CP(D_n|\theta) = \alpha$, both results being exact. The construction of Proposition 3 has found the classical prediction pivot, thereby satisfying design principle 4. Since $\hat{\theta}_n = (\bar{X}_n, S_n)$ is asymptotically efficient and Proposition 2 is easily extended to vector parameters, design principle 3 is also satisfied by the classical pivotal construction of D_n.

Example 2 (continued). An analog of Proposition 2 can be proved for this example, although θ is function-valued. However, Proposition 3 does not hold: the refined critical values (3.11) fail to produce improved rates of convergence for the errors in the unconditional balance and overall coverage probability of D_n. A good resolution to this difficulty remains to be found.

Example 3 (continued). Though the parameter θ is function-valued in this example too, analogs of Proposition 2 and 3 can be proved by essentially similar arguments.

Acknowledgement

This research was supported in part by NSF Grant DMS-8701426.

References

Beran, R. (1988). 'Balanced simultaneous confidence sets'. J. Amer. Statist. Assoc. **83**, 679-697.

Beran, R. (1990a). 'Calibrating prediction regions'. J. Amer. Statist. Assoc. **85**, to appear.

Beran, R. (1990b). 'Simultaneous prediction regions'. Unpublished preprint.

Beran, R. and Millar, P.W. (1987). 'Stochastic estimation and testing'. Ann. Statist. **15**, 1131-1154.

Box, G.E.P. and Jenkins, G.M. (1976). 'Time Series Analysis: Forecasting and Control', revised edition. Holden-Day, Oakland.

Carroll, R.J. and Ruppert, D. (1989). 'Prediction and tolerance intervals with transformation and/or weighting'. Unpublished preprint.

Cox, D.R. and Hinkley, D.V. (1974). 'Theoretical Statistics'. Chapman Hall, London.

Guttman, I. (1970). 'Statistical Tolerance Regions: Classical and Bayesian'. Griffin, London.

Olshen, R.A., Biden, E.N., Wyatt, M.P. and Sutherland, D.H. (1989). 'Gait analysis and the bootstrap'. Ann. Stataist. **71**, 1419-1440.

ANALYSIS OF OBSERVATIONAL STUDIES FROM THE POINT OF VIEW OF NONPARAMETRIC REGRESSION

P.K. BHATTACHARYA
University of California
Davis, California 95616
USA

ABSTRACT. Pair-matching is used in two-group observational studies to remove bias due to imbalance in covariate distribution between groups. In this paper, the traditional pair-matching technique is modified by comparing the responses in one group with the nonparametric regression estimates of the responses in the other group at the same covariate values. Properties of such estimates are derived, taking into consideration the possibility of treatment- covariate interaction.

1. Introduction

Many studies are conducted by government agencies, public health organizations and social science researchers for the purpose of estimating or testing significance of the effects of various conditions on a response variate Y, based on random samples from populations in which these conditions prevail. In practically all situations, the variate Y is influenced by one or more covariates X in addition to the conditions whose effects are of primary interest, and these covariates often follow different distributions in different groups. Such studies have been called observational studies by Cochran (1983), as opposed to experimental studies which are controlled by matching subjects with respect to covariates to form "blocks" and then randomly assigning the subjects in each block to the different conditions.

Observational studies usually involve human populations and the investigator is restricted to collecting data as they exist. It is clear that the purpose of an observational study is not served by a comparison of the mean Y-values in the different groups due to bias entering through covariates. To remove this bias, one must compare Y- values in different groups after making suitable adjustments for the covariates X. In classical analysis of covariance, such adjustments are made by fitting parallel linear regressions of Y on X in the different groups and then estimating the effects of the conditions from the differences between Y- intercepts of these lines. However, such estimates would be meaningless if the regressions are

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 587–602.
© 1991 *Kluwer Academic Publishers.*

non-linear, non-parallel and the covariates are distributed differently in the different groups.

A more robust approach to adjustment for covariate effect invloves matching, which is based on the idea of smooth nonparametric regression of Y on X. Various matching techniques have been proposed in the literature. For description, comparative studies and critical appraisal of some of these matching procedures, we refer to Rubin (1973), Cochran and Rubin (1973), McKinlay (1977) and Rosenbaum (1989).

In this paper, we consider a comparison between the effects of two conditions which we call "treatments", on a continuous response Y in the presence of a continuous covariate X, which is one-dimensional in our discussion for simplicity of notation. Suppose the data consists of two random samples (X_{1j},Y_{1j}), $j=1,\ldots,m$ and (X_{2j},Y_{2j}), $j=1,\ldots,n$ from the two treatment groups. We assume the following model:

$$Y_{1j} = \mu(X_{1j}) + Z_{1j}\ ,\ Y_{2j} = \mu(X_{2j}) + \theta(X_{2j}) + Z_{2j}, \qquad (1)$$

where $X_{11},\ldots,X_{1m}$ are iid as X_1 having pdf f_1, $X_{21},\ldots,X_{2n}$ are iid as X_2 having pdf f_2, $Z_{11},\ldots,Z_{1m}$ are iid as Z_1 having mean 0 and variance σ_1^2, $Z_{21},\ldots,Z_{2n}$ are iid as Z_2 having mean 0 and variance σ_2^2, all the random variables $\{X_{1j}\}$, $\{X_{2j}\}$, $\{Z_{1j}\}$, $\{Z_{2j}\}$ being mutually independent. In this model, $\theta(x)$ is the local treatment effect at the covariate value of x. If the function $\theta(\cdot)$ is a constant, i.e., if the regressions in the two groups are parallel, then the constant value of $\theta(\cdot)$ is the treatment effect; otherwise, one can only think of a treatment main effect in the presence of treatment-covariate interaction.

In Section 2, we introduce an estimate of treatment main effect by modifying the well-known pair-matching technique, using kernel estimates of nonparametric regression in one group evaluated at each of the X-values of the other groups. In Section 3, we show that under regularity conditions, this estimate converges to a certain version of treatment main effect at the rate of $1/\sqrt{N}$ as the total sample size $N \to \infty$ and is asymptotically normal. We conclude with a few remarks in Section 4. Some technical details of the proofs in Section 3 are given in an Appendix.

2. Pair-Matching and its Nonparametric Regression Analogue

Pair-matching is a very commonly used matching procedure which aims at removing bias due to X by matching each unit in the smaller of two groups with a unit in the larger group having approximate agreement in X, and then comparing the Y- values of the matched units. Although the basic idea of pair-matching is appealing, it has some obvious drawbacks such as lack of clear-cut algorithm and loss of

data due to unmatched units in the larger group. These difficulties and the issue of appropriate statistical inference from pair- matched data when there is treatment- covariate interaction have been discussed by McKinlay (1977).

If we look at pair-matching from the point of view of smooth nonparametric regression, it seems natural to modify the traditional procedure by comparing each Y-value in one group by the estimated regression in the other group evaluated at the same X rather than an actual Y- value with approximately the same X. This leads to the consideration of

$$T_N = n_2^{-1} \sum_{k=1}^{n_2} [Y_{2j} - \hat{\mu}_1(X_{2k})], \tag{2}$$

where $\hat{\mu}_1(x)$ is the kernel estimate of the regression of Y on X in the first group, evaluated at X = x, defined by

$$\hat{\mu}_1(x) = \{(n_1 b_N)^{-1} \sum_{j=1}^{n_1} 1(|X_{1j}-x| \le b_N/2)Y_{1j}\}/\hat{f}_1(x),$$

$$\hat{f}_1(x) = (n_1 b_N)^{-1} \sum_{j=1}^{n_1} 1(|X_{1j}-x| \le b_N/2), \tag{3}$$

the bandwidth $b_N \downarrow 0$ at an appropriate rate.

In the next section we show that T_N estimates the treatment effect when there is no treatment-covariate interaction and it estimates a version of treatment main effect,

$$\theta = E[\theta(X_2)] = \int \theta(x) f_2(x)\, dx$$

when there is interaction.

3. Asymptotics of T_N

Of the following regularity conditions, A1 is heavily used in order to ensure that the reciprocal of $\hat{f}_1(\cdot)$ appearing in T_N, converges to the reciprocal of $f_1(\cdot)$ at $X_{21},\ldots,X_{2n_2}$ in an appropriate manner. Had it not been for the restriction already imposed by A1, conditions A2 and A3 could serve their purpose in slightly weaker forms.

A1. The pdf f_1 vanishes outside a finite interval I_1 and is bounded away from 0 on I_1. Moreover, f_1 and its first two derivatives are bounded.

A2. The pdf f_2 vanishes outside a finite interval $I_2 \subset I_1$ on which f_2 and its first two derivatives are bounded.

A3. The regression μ and its first two derivatives are bounded.

A4. The local treatment effect $\theta(x)$ has finite second moment in the second group, i.e., $E[\theta^2(X_2)] = \int \theta^2(x) f_2(x)dx < \infty$.

A5. The residuals Z_1, Z_2 have mean 0 and finite variances $E(Z_i^2) = \sigma_i^2$, $i=1,2$. Moreover, $E(|Z_1|^3) = \beta_3 < \infty$.

A6. $n_1/N \to \lambda$, $n_2/N \to 1-\lambda$, $0 < \lambda < 1$, as $N \to \infty$.

We also require the following conditions on the bandwidth b_N and the sample sizes n_1, n_2 as $N = n_1+n_2 \to \infty$.

A7. $b_N = O(N^{-1/4-\delta})$ with $\delta > 0$ but suitably small. This implies $b_N \downarrow 0$ but $Nb_N \to \infty$.

Note that some of the conditions are asymmetric between the two groups. This is because the estimator T_N itself gives a primary role to the response variate in the samples from the second group and uses the samples from the first group by nonparametric regression estimates. This asymmetry will be discussed further in Section 4.

We now state the main result of this paper.

THEOREM 1. Under conditions A1-A7,

$$\sqrt{N}(T_N-\theta) \xrightarrow{d} \mathcal{N}(0,\ \sigma_0^2 + \lambda^{-1}\sigma_1^2\gamma_1 + (1-\lambda)^{-1}\gamma_2),$$

where

$$\begin{aligned} \theta &= E[\theta(X_2)]\ , \quad \sigma_0^2 = \lambda^{-1}\sigma_1^2 + (1-\lambda)^{-1}\sigma_2^2\ , \\ \gamma_1 &= E[f_2(X_2)/f_1(X_2)] - 1, \\ \gamma_2 &= E[\Delta^2(X_2)] = E[\{\theta(X_2) - \theta\}^2] = E[\theta^2(X_2)] - \theta^2. \end{aligned} \tag{4}$$

To prove Theorem 1, use the expressions for Y_{1j} and Y_{2j} given by (1) to write T_N given by (2) and (3) in the form

$$\begin{aligned} T_N &= n_2^{-1} \sum_{k=1}^{n_2} \theta(X_{2k}) + n_2^{-1} \sum_{k=1}^{n_2} Z_{2k} \\ &+ (n_1 n_2 b_N)^{-1} \sum_{k=1}^{n_2} \left[\sum_{j=1}^{n_1} w_N(X_{1j},\ X_{2k})\{\mu(X_{2k}) - \mu(X_{1j})\}\right]/\hat{f}_1(X_{2k}) \\ &- (n_1 n_2 b_N)^{-1} \sum_{k=1}^{n_2} \left[\sum_{j=1}^{n_1} w_N(X_{1j}, X_{2k}) Z_{1j}\right]/\hat{f}_1(X_{2k}), \end{aligned}$$

where

$$w_N(x_1,x_2) = 1(|x_1-x_2| \le b_N/2).$$

Now let $V_{1j} = (X_{1j},Z_{1j})$, $V_{2k} = (X_{2k},Z_{2k})$. Then $\sqrt{N}\,(T_N - \theta)$ is expressed as

$$\sqrt{N}(T_N-\theta) = \sqrt{N}(n_1n_2)^{-1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_2} g_N(V_{1j},V_{2k}) + R_{N1} + R_{N2}, \tag{5}$$

where the first term is a U-statistic with

$$g_N(V_{1j},V_{2k}) = -\,b_N^{-1}w_N(X_{1j},X_{2k})Z_{1j}/f_1(X_{2k}) + Z_{2k} + \{\theta(X_{2k}) - \theta\}, \tag{6}$$

and the remainder terms are

$$R_{N1} = \sqrt{N}(n_1n_2b_N)^{-1}\sum_{k=1}^{n_2}\Big[\sum_{j=1}^{n_1} w_N(X_{1j},X_{2k})\{\mu(X_{2k}) - \mu(X_{1j})\}\Big]/\hat{f}_1(X_{2k}),$$

$$R_{N2} = \sqrt{N}(n_1n_2b_N)^{-1}\sum_{k=1}^{n_2}\Big[\sum_{j=1}^{n_1} w_N(X_{1j},X_{2k})Z_{1j}\Big]\cdot[1/f_1(X_{2k}) - 1/\hat{f}_1(X_{2k})]. \tag{7}$$

Asymptotic normality of $\sqrt{N}(T_N-\theta)$ will be proved by handling the asymptotics of the first term on the right-hand side of (5) by a suitable generalization of the standard U-statistic technique which takes into account the dependence of the kernel g_N on the sample size N, and by showing that the remainder terms R_{N1} and R_{N2} are $o_p(1)$.

For the U-statistic part of the argument, we need the following lemma, while the negligibility of the remainder terms is proved in the Appendix.

LEMMA 1. Let $S_N = (n_1n_2)^{-1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_2} g_N(V_{1j},V_{2k})$ be a two-sample U-statistic based on independent sequences of iid random vectors $\{V_{1j}\}$, $\{V_{2k}\}$.

Suppose that $n_1/N \to \lambda \in (0,1)$ and the kernel g_N satisfies the following conditions as $N = n_1 + n_2 \to \infty$:

B1. There exist $h_1(V_{11})$, $h_2(V_{21})$ with finite variances ζ_1,ζ_2,

respectively, such that

$$\varphi_{iN}(V_{i1}) = E[g_N(V_{11},V_{21})|V_{i1}] = h_i(V_{i1}) + \varepsilon_{iN}(V_{i1}),$$

where

$$E[\varepsilon^2_{iN}(V_{i1})] = o(N^{-1}),\ i=1,2.$$

B2. $\mathrm{Var}[g_N(V_{11},V_{21})] = o(N)$.

Then as $N \to \infty$,

(a) $\alpha_N = E[g_N(V_{11},V_{21})] = \alpha + o(N^{-1/2})$, where $\alpha = E[h_1(V_{11})]$
$= E[h_2(V_{21})]$,

(b) $\zeta_{iN} = \mathrm{Var}[\varphi_{iN}(V_{i1})] = \zeta_i + o(1)$, $i=1,2$,

(c) $N^{1/2}(S_N-\alpha) \xrightarrow{d} \mathcal{N}(0,\lambda^{-1}\zeta_1 + (1-\lambda)^{-1}\zeta_2)$.

PROOF. This lemma is a special case of Lemma 4 in [1]. ∎

We now use this lemma to prove Theorem 1.

PROOF OF THEOREM 1. We verify conditions B1, B2 of Lemma 1.

The conditional expectation $\varphi_{1N}(V_{11})$ at $V_{11} = (x,z)$ is obtained as

$$\varphi_{1N}(x,z) = E[g_N(V_{11},V_{21})|V_{11} = (x,z)]$$

$$= -\ zb_N^{-1}\ E[1(|X_2-x| \le b_N/2)/f_1(X_2)]$$

$$= -\ z \int_{-1/2}^{1/2} [f_2(x+b_Nu)/f_1(x+b_Nu)]\ du$$

$$= -\ z\ f_2(x)/f_1(x) + z\rho_{1N}(x),$$

where $|\rho_{1N}(x)| \le Cb_N^2$ by regularity conditions, so that

$E[Z_1^2\ \rho^2_{1N}(X_1)] \le C\sigma_1^2\ b_N^4 = o(N^{-1})$, by our choice of $b_N = o(N^{-1/4})$.

We next find the conditional expectation $\varphi_{2N}(V_{21})$ at $V_{21} = (x,z)$:

$$\varphi_{2N}(x,z) = E[g_N(V_{11},V_{21})|V_{21} = (x,z)] = Z_{2k} + \{\theta(X_{2k}) - \theta\}.$$

Thus condition B1 holds with

$$h_1(X_1,Z_1) = - Z_1 f_2(X_1)/f_1(X_1),$$

$$h_2(X_2,Z_2) = Z_2 + \{\theta(X_2) - \theta\} = Z_2 + \Delta(X_2),$$

having common mean $\alpha = 0$ and variances

$$\zeta_1 = E[h_1^2(X_1,Z_1)] = \sigma_1^2\ E[f_2^2(X_1)/f_1^2(X_1)] = \sigma_1^2\ E[f_2(X_2)/f_1(X_2)],$$

$$\zeta_2 = E[h_2^2(X_2,Z_2)] = \sigma_2^2 + E[\Delta^2(X_2)].$$

To verify condition B2, we note that since $Nb_N \to \infty$ as $N \to \infty$, it is enough to show that $E[g_N^2(V_{11},V_{21})] = O(b_N^{-1})$. Since $X_{11},X_{21},Z_{11},Z_{21}$ are mutually independent with $E(Z_{11}) = E(Z_{21}) = 0$, we have

$$E[g_N^2(V_{11},V_{21})]$$

$$= E[b_N^{-2} w_N(X_{11},X_{21}) Z_{11}^2 / f_1^2(X_{21}) + Z_{21}^2 + \Delta^2(X_{21})]$$

$$= b_N^{-1}\sigma_1^2 \int \{E[b_N^{-1} w_N(X_1,x)]/f_1^2(x)\} f_2(x)dx + \sigma_2^2 + E[\Delta^2(X_2)]$$

$$\le b_N^{-1}\ \sigma_1^2\ a^{-2}\ \sup_x E[b_N^{-1} w_N(X_1,x)] + \sigma_2^2 + E[\Delta^2(X_2)],$$

which is $O(b_N^{-1})$, since $f(x) \ge a > 0$ for all x, $E[\Delta^2(X_2)]$

$$\le E[\theta^2(X_2)] < \infty,$$

and

$$E[b_N^{-1} w_N(X_1,x)] = \int_{-1/2}^{1/2} f_1(x + b_N u)du \le M.$$

Thus Lemma 1 applies to the first term on the right-hand side of (5), which is asymptotically normal with mean 0 and variance

$$\lambda^{-1}\zeta_1 + (1-\lambda)^{-1}\zeta_2$$

$$= \lambda^{-1}\sigma_1^2\ E[f_2(X_2)/f_1(X_2)] + (1-\lambda)^{-1}\{\sigma_2^2 + E[\Delta^2(X_2)]\}$$

$$= [\lambda^{-1}\sigma_1^2 + (1-\lambda)^{-1}\sigma_2^2] + \lambda^{-1}\sigma_1^2\ \{E[f_2(X_2)/f_1(X_2)] - 1\}$$

$$+ (1-\lambda)^{-1}\ E[\Delta^2(X_2)]$$

$$= \sigma_0^2 + \lambda^{-1}\ \sigma_1^2\ \gamma_1 + (1-\lambda)^{-1}\gamma_2,$$

where σ_2^2, γ_1 and γ_2 are as in the statement of Theorem 1. The remainder terms R_{N1} and R_{N2} on the right-hand side of (5) are shown to be $o_p(1)$ in the Appendix, and we use that fact to conclude the proof. ∎

4. Remarks

4.1. ESTIMATION OF STANDARD ERROR OF T_N

For this, we need consistent estimators of σ_1^2, σ_2^2, $E[f_2(X_2)/f_1(X_2)]$ and $E[\Delta^2(X_2)]$. Such estimators are constructed as follows. As before, let $w_N(x_1,x_2) = 1(|x_1-x_2| \le b_N/2)$. Then the following facts are direct consequences of our regularity conditions with the additional requirement $E(Z_1^4) < \infty$, and we omit their proofs.

$$A_{Ni} = \binom{n_i}{2}^{-1} b_N^{-1} \sum_{1\le j<k\le n_i} w_N(X_{ij},X_{ik})(Y_{ik}-Y_{ik})^2$$

$$= 2\,\sigma_i^2\|f_i\|^2 + o_p(1), \quad i=1,2,$$

$$B_{Ni} = \binom{n_i}{2}^{-1} b_N^{-1} \sum_{1\le j<k\le n_i} w_N(X_{ij},X_{ik}) = \|f_i\|^2 + o_p(1), \quad i=1,2,$$

$$C_N = n_1\{n_2(n_2-1)\}^{-1} \sum_{k=1}^{n_2} \Big[\sum_{j\ne k=1}^{n_2} w_N(X_{2j},X_{2k}) / \sum_{j=1}^{n_1} w_N(X_{1j},X_{2k})\Big]$$

$$= E[f_2(X_2)/f_1(X_2)] + o_p(1),$$

$$D_N = n_2^{-1} \sum_{k=1}^{n_2} [Y_{2k} - \hat{\mu}_1(X_{2k})]^2 = \sigma_2^2 + E[\theta^2(X_2)] + o_p(1).$$

From these, we have

$$\sigma_{Ni}^2 = A_{Ni}/(2B_{Ni}) = \sigma_i^2 + o_p(1), \; i=1,2,$$

$$\Delta_N = D_N - T_N^2 - \sigma_{N2}^2 = E[\Delta^2(X_2)] + o_p(1).$$

Putting these together, we obtain the following estimator

$$SE(T_N) = [A_{N1}C_N/(2n_1B_{N1}) + (D_N - T_N^2)/n_2]^{1/2}$$

of the standard error of T_N

$$\{(\sigma_0^2 + \lambda^{-1}\sigma_1^2\gamma_1 + (1-\lambda)^{-1}\gamma_2)/N\}^{-1/2}.$$

Using this estimated standard error, we can construct confidence intervals

$$T_N \pm \Phi^{-1}(1-\alpha/2)\; SE(T_N)$$

for θ with confidence coefficient $1-\alpha$, where Φ is the standard normal cdf.

4.2. TREATMENT-COVARIATE ESTIMATION AND ITS ESTIMATE

The quantity $E[\Delta^2(X_2)]$ is the variance of $\theta(X_2)$, and so it can be thought of as a measure of treatment-covariate interaction. A consistent estimator of $E[\Delta^2(X_2)]$ has been mentioned above. Since it is possible for this estimator to take negative values, it would be more appropriate to use

$$\Delta_N^+ = [D_N - T_N^2 - A_{N2}/(2B_{N2})]^+$$

to estimate $E[\Delta^2(X_2)]$. In many situations, a scale-invariant measure of treatment-covariate interaction $E^{1/2}[\Delta^2(X_2)]/|\theta|$ and its estimate $\sqrt{\Delta_N^+}/|T_N|$ may be more meaningful, provided that $\theta \neq 0$.

4.3. ASYMMETRIC NATURE OF PAIR-MATCHING

In pair-matching, each sample in the smaller group is matched with a sample in the larger group by approximate agreement in the covariate. This assigns a special role to the smaller group. The nonparametric regression analogue of pair- matching discussed here, also assigns such a special role to one of the groups. The estimator T_N given by (2) and (3) assigns this role to the second sample which is not necessarily the smaller of the two samples. As a consequence, T_N estimates the version $E[\theta(X_2)]$ of the treatment main effect. By reversing the roles of the two samples, we would be estimating $E[\theta(X_1)]$. For example, if we are comparing Y = diastolic blood pressure between Group 1 consisting of non- smokers and Group 2 consisting of smokers, making adjustment for the covariate X = age, then

$E[\theta(X_1)]$ = increase in average diastolic blood pressure if all the non-smokers were smoking, and

$E[\theta(X_2)]$ = decrease in average diastolic blood pressure if all the smokers were not smoking.

In this context, $E[\theta(X_1)]$ and $E[\theta(X_2)]$ would be different since $\theta(x)$ is not a constant in x and the age distribution is different in the two groups. Proper interpretation of the estimate of treatment effect is, therefore, an important matter in analysis of observational studies. In [1], a class of estimators has been developed in which the two samples have symmetric roles.

When no treatment-covariate interaction is suspected and the covariate distributions in the two groups have the same support, we can calculate two estimators

$$T_{N1} = n_1^{-1} \sum_{j=1}^{n_1} [\hat{\mu}_2(X_{1j}) - Y_{1j}] \ , \ T_{N2} = n_2^{-1} \sum_{k=1}^{n_2} [Y_{2k} - \hat{\mu}_1(X_{2k})],$$

where $\hat{\mu}_i$ is the nonparametric estimator of $E(Y|X=x)$ in the ith group, $i=1,2$. Theorem 1 extends in a straightforward manner to yield the asymptotic bivariate normality of $\sqrt{N}\ (T_{N1} - \theta,\ T_{N2} - \theta)$ with mean $(0,0)$, variances $\sigma_0^2 + \psi_1$, $\sigma_0^2 + \psi_2$ and covariance σ_0^2, where

$$\psi_1 = (1-\lambda)^{-1}\sigma_2^2\ \{E[f_1(X_1)/f_2(X_1)] - 1\},$$

$$\psi_2 = \lambda^{-1}\sigma_1^2\ \{E[f_2(X_2)/f_1(X_2)] - 1\},$$

The best linear combination of T_{N1} and T_{N2} is

$$T_N^* = (\psi_1^{-1}T_{N1} + \psi_2^{-1}T_{N2})/(\psi_1^{-1} + \psi_2^{-1}),$$

and

$$\sqrt{N}(T_N^* - \theta) \xrightarrow{d} \mathcal{N}(0,\ \sigma_0^2 + \psi_1\psi_2/(\psi_1 + \psi_2)).$$

4.4. EFFECT OF UNKNOWN REGRESSION ON THE STANDARD ERROR OF T_N

Suppose there is no treatment-covariate interaction. In such a case, if the regression μ were known, then we would estimate the constant treatment effect θ by

$$\hat{\theta}_N = n_2^{-1} \sum_{k=1}^{n_2} [Y_{2k} - \mu(X_{2k})] - n_1^{-1} \sum_{j=1}^{n_1} [Y_{1j} - \mu(X_{1j})],$$

and

$$\sqrt{N}\ \ (\hat{\theta}_N - \theta) \xrightarrow{d} \mathcal{N}(0,\sigma_0^2),\ \sigma_0^2 = \lambda^{-1}\sigma_1^2 + (1-\lambda)^{-1}\sigma_2^2.$$

On the other hand, the asymptotic variance of $\sqrt{N}\ (T_N - \theta)$ is $\sigma_0^2 + \lambda^{-1}\sigma_1^2\gamma_1$ where $\gamma_1 = E[f_2(X_2)/f_1(X_2)] - 1 = 0$ if $f_1 = f_2$ and $\gamma_1 > 0$, otherwise, since

$$E[f_2(X_2)/f_1(X_2)] > 1/E[f_1(X_2)/f_2(X_2)] = 1/\int_{\{x:f_2(x)>0\}} f_1(x)dx \geq 1,$$

by Jensen's inequality when $f_1 \neq f_2$ on a set of positive Lebesgue measure. Thus in the absence of treatment-covariate interaction, the

nonparametric regression analogue of pair-matching makes adjustment for the covariate at no cost when $f_1 = f_2$ and at the cost of inflating the asymptotic variance by a factor of $1 + \gamma_1\sigma_1^2 / (\lambda\ \sigma_2^2)$ otherwise.

Appendix

The remainder terms R_{N1}, R_{N2} given in (7) are shown to be $o_p(1)$ in the following lemmas. Let M denote an upper bound for $f_1, f_2, |f_1'|$, $|f_2'|$, $|f_1''|, |f_2''|$ as well as $|\mu|$, $|\mu'|$, $|\mu''|$, and let a denote a positive lower bound for f_1. The existence of a and M are assured by the regularity conditions. We shall also let C, C_1, C_2, etc., denote generic constants determined only by a and M unless otherwise specified.

LEMMA 2. If $0 < t_N < 1/M$ and $b_N^2/t_N = o(1)$, then for large N,

$$\sup_x P[|1/\hat{f}_1(x) - 1/f_1(x)| \geq t_N] \leq 2 \exp[-CN\ b_N t_N^2],\ C > 0.$$

PROOF. Since $t_N f_1(x) \leq t_N M < 1$ for all x, $|1/\hat{f}_1(x) - 1/f_1(x)| \geq t_N$ implies

$$|\hat{f}_1(x) - f_1(x)| \geq A_N(x) = t_N f_1^2(x)/\{1 + t_N f_1(x)\}.$$

Moreover, letting $\pi_{1N}(x) = P(|X_1 - x| \leq b_N/2)$, we have $E[\hat{f}_1(x)]$

$$= b_N^{-1}\pi_{1N}(x)$$

and $|b_N^{-1}\pi_{1N}(x) - f_1(x)| = |\int_{-1/2}^{1/2} f_1(x+b_N u)du - f_1(x)| \leq Mb_N^2/24.$

Hence

$$P[|1/\hat{f}_1(x) - 1/f_1(x)| \geq t_N]$$

$$\leq P[|\hat{f}_1(x) - b_N^{-1}\pi_{1N}(x)| \geq \frac{1}{2} A_N(x)]$$

$$+ 1(|b_N^{-1}\pi_{1N}(x) - f_1(x)| \geq \frac{1}{2} A_N(x))$$

$$\leq P[|n_1^{-1} \sum_{j=1}^{n_1} w_N(X_{1j}, x) - \pi_{1N}(x)| \geq \frac{1}{2} A_N(x)]$$

$$+ 1(Mb_N^2 \geq 12\ A_N(x)).$$

The second term vanishes for large N, because $t_N f_1(x) < 1$ and $f_1^2(x) > a$ imply $12\ A_N(x) \geq 6t_N$, which becomes $< Mb_N^2$ since $b_N^2/t_N = o(1)$.

In the first term, $\{w_N(X_{1j},x)\}$ are independent Bernoulli variables with probability $\pi_{1N}(x)$, and by Bernstein's inequality (see Uspensky [9], p.205),

$$P[|n_1^{-1} \sum_{j=1}^{n_1} w_N(X_{1j},x) - \pi_{1N}(x)| \geq \frac{1}{2} b_N A_N(x)]$$

$$\leq 2 \exp[- \frac{1}{4} n_1 b_N A_N^2(x)\{2b_N^{-1}\pi_{1N}(x) + A_N(x)\}^{-1}].$$

The lemma is now proved by using $2b_N^{-1}\pi_{1N}(x) + A_N(x) \leq 3f_1(x)$, $A_N^2(x)/f_1(x) \geq a^3 t_N^2/4$ and $n_1/N \to \lambda > 0$. ∎

LEMMA 3. If $b_N^3/t_N = o(1)$, then for large N,

$$\sup_x P[|n_1^{-1} \sum_{j=1}^{n_1} w_N(X_{1j},x)\{\mu(X_{1j}) - \mu(x)\}| \geq t_N]$$

$$\leq 2 \exp[-CNt_N^2/b_N^2], \ C > 0.$$

PROOF. Let $\mu_N(x) = E[w_N(X_1,x)\{\mu(X_1) - \mu(x)\}]$. By regularity conditions,

$$|\mu_N(x)| = b_N|\int_{-1/2}^{1/2} \{\mu(x+b_N u) - \mu(x)\}f_1(x+b_N u)du| \leq M^2 b_N^3/4.$$

Moreover,

$$|w_N(X_{1j},x)\{\mu(X_{1j}) - \mu(x)\}| \leq \max_{|u|\leq 1/2} |b_N u\mu'(x+b_N u)|$$

$$\leq Mb_N/2.$$

Using Hoeffding's inequality (Hoeffding (1963)), we now have

$$P[|n_1^{-1} \sum_{j=1}^{n_1} w_N(X_{1j}, x)\{\mu(X_{1j}) - \mu(x)\}| \geq t_N]$$

$$\leq P[|n_1^{-1} \sum_{j=1}^{n_1} w_N(X_{1j}, x)\{\mu(X_{1j}) - \mu(x)\} - \mu_N(x)| \geq t_N/2]$$

$$+ 1 \ (|\mu_N(x)| \geq t_N/2)$$

$$\leq 2 \exp[-CNt_N^2/b_N^2] + 1(b_N^3/t_N \geq 2/M^2).$$

The second term vanishes for large N since $b_N^3/t_N = o(1)$, and the lemma is proved. ∎

LEMMA 4. $\sup_x P\left[\left|n_1^{-1/2} \sum_{j=1}^{n_1} w_N(X_{1j},x)\, Z_{1j}\right| \geq t_N\right]$

$$\leq \exp\,[-C_1 N b_N^2] + C_2(b_N^{1/2}/t_N)\,\exp[-C_3 t_N^2/b_N]$$

$$+ C_4(\sigma_1,\ \beta_3)N^{-1/2}\, b_N t_N^{-3},$$

where C_1, C_2, C_3 depend only on a and M, and C_4 depends on a, M, σ_1 and $\beta_3 = E(|Z_1|^3)$.

PROOF. $P\left[\left|n_1^{-1/2} \sum_{j=1}^{n_1} w_N(X_{1j},\ x)\, Z_{1j} \geq t_N\right]\right.$

$$= \sum_{r=0}^{n_1} P\left[\sum_{j=1}^{n_1} w_N(X_{1j},x) = r\right]\cdot P\left[\left|\sum_{j=1}^{r} Z_{1j}\right| \geq n_1^{1/2} t_N\right]$$

$$\leq P\left[n_1^{-1} \sum_{j=1}^{n_1} \{w_N(X_{1j},\ x) - \pi_{1N}(x)\} > \delta\ \pi_{1N}(x)\right]$$

$$+ \max_{r\leq(1+\delta)n_1\pi_{1N}(x)} P\left[\left|r^{-1/2}\sigma_1^{-1} \sum_{j=1}^{r} Z_{1j}\right| \geq (n_1/r)^{1/2} t_N/\sigma_1\right]$$

with a suitable $\delta > 0$. By Hoeffding's inequality (Hoeffding (1963)), the first term is

$$\leq \exp[-2n_1\delta^2\pi_{1N}^2(x)] \leq \exp[-C_1 N b_N^2],$$

using $\pi_{1N}(x) \geq ab_N/2$ as seen above, and by Theorem 2, Chapter V, page 73 of Esseen (1945) with third moment condition on Z_1, the second term is

$$\leq 2[1- \Phi(t_N\sigma_1^{-1}\{(1+\delta)\pi_{1N}(x)\}^{-1/2})]$$
$$+ C(\sigma_1,\beta_3)(1+\delta)\pi_{1N}(x)n_1^{-1/2}t_N^3.$$

The lemma is proved by using

$$1-\Phi(x) \leq x^{-1}\Phi'(x) \text{ and } \pi_{1N}(x) \leq 2M\ b_N$$

in the last expression. ∎

The above three lemmas provide probability inequalities for deviations of

$$\xi_N(x) = 1/f_1(x) - 1/\hat{f}_1(x),$$

$$\eta_N(x) = n_1^{-1} \sum_{j=1}^{n_1} w_N(X_{1j}, x)\{\mu(X_{1j}) - \mu(x)\},$$

$$\zeta_N(x) = n_1^{-1/2} \sum_{j=1}^{n_1} w_N(X_{1j}, x)\, Z_{1j}.$$

Moreover, using Lemma 1 and proceeding as in the proof of Theorem 1, we can verify that

$$Q_N = N^{1/2}(n_1 n_2 b_N)^{-1} \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} w_N(X_{1j}, X_{2k})\{\mu(X_{2k})$$

$$- \mu(X_{1j})\}/f_1(X_{2k}) = o_p(1).$$

To show that the remainder terms in (7) are $o_p(1)$, we note that

$$|R_{N1}| = |Q_N + N^{1/2} b_N^{-1} n_2^{-1} \sum_{k=1}^{n_2} \xi_N(X_{2k})\eta_N(X_{2k})| \tag{8a}$$

$$\le |Q_N| + N^{1/2} b_N^{-1} \max_{1\le k\le n_2} |\xi_N(X_{2k})| \max_{1\le k\le n_2} |\eta_N(X_{2k})|,$$

$$|R_{N2}| = |(N/n_1)^{1/2} b_N^{-1} n_2^{-1} \sum_{k=1}^{n_2} \xi_N(X_{2k})\zeta_N(X_{2k})|$$

$$\le (N/n_1)^{1/2} b_N^{-1} \max_{1\le k\le n_2} |\xi_N(X_{2k})| \cdot \max_{1\le k\le n_2} |\zeta_N(X_{2k})|. \tag{8b}$$

Lemmas 2, 3 and 4 are now used to prove the desired results.

LEMMA 5. $R_{N1} = o_p(1)$ as $N \to \infty$.

PROOF. By (8a), for $\varepsilon > 0$,

$$P[|R_{N1}| \ge \varepsilon] \le P[|Q_N| \ge \varepsilon/2]$$

$$+ n_2 \max_x P[|\xi_N(x)| \ge u_{1N}] + n_2 \max_x P[|\eta_N(x)| \ge v_{1N}], \tag{9}$$

where $u_{1N}v_{1N} = N^{-1/2} b_N \varepsilon/2$. Use $b_N = N^{-(1/4)-\delta}$ with δ positive and suitably small, and take $v_{1N} = N^{-7/16}$, $u_{1N} = N^{-(5/16)-\delta}(\varepsilon/2)$. Now b_N^2/u_{1N} and b_N^3/v_{1N} are both $o(1)$, so on the right-hand side of (9),

the second term tends to 0 by Lemma 2 and the third term tends to 0 by Lemma 3,
as $N \to \infty$. Since $Q_N = o_p(1)$, the proof is complete. ∎

LEMMA 6. $R_{N2} = o_p(1)$ as $N \to \infty$.

PROOF. By 8(b), for $\varepsilon > 0$,

$$P[|R_{N2}| \geq \varepsilon] \leq n_2 \max_x P[|\xi_N(x)| \geq u_{2N}] + \eta_2 \max_x P[|\zeta_N(x)| \geq v_{2N}] , \tag{10}$$

where $u_{2N}v_{2N} = (n_1/N)^{1/2} b_N \varepsilon$. Take $v_{2N} = N^{1/12}$ and u_{2N}

$= (n_1/N)^{1/2} N^{-1/3-\delta} \varepsilon$. Now on the right-hand side of (10), as $N \to \infty$, the first term tends to 0 by Lemma 2 since $b_N^2/u_{2N} = o(1)$, and the second term tends to 0 by Lemma 4, and that completes the proof. ∎

Acknowledgement

This research was done with partial support from NSF Grant No. DMS 89-04422.

References

Bhattacharya, P.K. (1989) 'Estimation of treatment main effect and treatment-covariate interaction in observational studies using bandwidth-matching'. Tech. Report No. 188, Division of Statistics, University of California, Davis.

Cochran, W.G. (1983) Planning and Analysis of Observational Studies. Wiley.

Cochran, W.G. and Rubin, D.B. (1973) 'Controlling bias in observational studies: A review'. Sankhyā Ser. A 35, 417-446.

Esseen, C.G. (1945) 'Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law'. Acta Math. 77, 1-125.

Hoeffding, W. (1963) 'Probability inequalities for sums of bounded random variables'. J. Amer. Statist. Assoc., 58, 13-30.

McKinlay, S.M. (1977) 'Pair-matching — a reappraisal of a popular technique'. Biometrics 33, 725-735.

Rosenbaum, P.R. (1989) 'Optimal matching for observational studies'. J. Amer. Statist. Assoc., 84, 1024-1032.

Rubin, D.B. (1973) 'Matching to remove bias in observational studies'. Biometrics 29, 159-183.

Uspensky, J.V. (1937) Introduction to Mathematical Probability. McGraw-Hill.

SOME ISSUES IN CROSS-VALIDATION

PRABIR BURMAN
University of California
Davis, California 95616
USA

and

DEBORAH NOLAN
University of California
Berkeley, California 94720
USA

ABSTRACT. Leave-one-out cross-validation (i.e., ordinary cross-validation) and variants such as v-fold cross-validation are often used for model selection problems with independent observations. Extensions and generalizations of these methods for dependent observations are discussed in this note. It is shown that ordinary cross-validation can be used for certain types of dependence. A new type of cross-validation, which is an off-shoot of both ordinary cross-validation and v-fold cross-validation, is proposed here for model selection problems when the data is generated by a stationary process.

1. Introduction

Cross-validation (Stone (1974), Geisser (1975)) is acclaimed as a method for estimating prediction error in regression and classification problems, and in recent years it has received much attention for its model selection ability in the nonparametric setting (Härdle and Marron (1985)). In this note, we review the technique of cross-validation and discuss two variations: v-fold cross-validation and h-block cross-validation. The first variant is useful because it is not as computationally intensive as ordinary cross-validation (see Burman (1989)). The second handles the broader setting where observations are not independent.

In the next section, we formulate the prediction-error problem, provide both parametric and nonparametric examples of the problem, and explain how ordinary cross-validation and v-fold cross-validation estimate the prediction error by reusing the sample, all within the i.i.d. setting. Section 3 makes a heuristic argument for why the technique of leave-one-out cross-validation can provide a good

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 603–612.
© 1991 *Kluwer Academic Publishers.*

estimate of prediction error. We argue in the regression setting that this estimate is good when there is an underlying martingale-like structure in the errors. This suggests ordinary cross-validation applies to more models than those with independent structures. In Section 4, we present a new type of cross-validation that is suitable for general stationary processes. This new cross-validation, which we call h-block cross-validation, makes use of the ideas behind v-fold cross-validation. The discussion throughout this note will be nontechnical.

2. Formulation of the prediction-error problem

Let $Z_1,\ldots,Z_n$ be independent and identically distributed (i.i.d.) random vectors on R^d with distribution function (d.f.) F. Write F_n for the empirical d.f. based on the observations $Z_1,\ldots,Z_n$. The problem is to estimate a quantity such as

$$s_n = \int T(z,F_n)dF(z) \text{ or } E(s_n).$$

Typically, T measures the prediction error of an estimator based on F_n, and s_n is the expected error incurred by using this estimator to predict a future observation. Before proceeding any further let us give a few examples of T and s_n. Though our examples all involve squared error loss, our formulation and can handle Kullback-Leibler information and L^1 loss.

Example 1. Consider the model

$$Y_i = \mu(X_i) + \varepsilon_i,\ i=1,\ldots,n.,$$

where $E(\varepsilon_i|X_i) = 0$. The regression function μ is unknown and has to be estimated.

Here $Z_i = (Y_i,X_i)$, $i=1,\ldots,n$. Assume μ is estimated using splines with k knots or a polynomial of degree k, and then the parameters are determined by the least squares method. Take an $n+1^{st}$ observation (Y_{n+1}, X_{n+1}) independent of (Y_i,X_i), $i=1,\ldots,n$. If we call the fitted regression function $\hat{\mu}_k$, then the error (expected quadratic loss) from predicting Y_{n+1} by $\hat{\mu}_k(X_{n+1})$ is

$$\begin{aligned} s_n &= E[\{Y_{n+1} - \hat{\mu}_k(X_{n+1})\}^2 | (X_i,Y_i),\ i=1,\ldots,n] \\ &= \int \{y-\hat{\mu}_k(x)\}^2 dF(y,x). \end{aligned}$$

Here, $T(z,F_n) = \{y-\hat{\mu}_k(x)\}^2$ with $z = (y,x)$.

Example 2. Consider the regression model,

$$Y_i = \Sigma_{1\le t\le M_n}\theta_t X_{ti} + \varepsilon_i,\ i=1,\ldots,n.$$

In some problems of practical interest M_n/n is not small. If only

first k of the independent variables are retained and their coefficients are estimated by the method of least squares, then the prediction error is

$$s_n = \int (y - \Sigma_{1\le t\le k}\hat{\theta}_t x_t)^2 dF(y,x_1,\ldots,x_k).$$

Here, $z = (y,x_1,\ldots,x_k)$, $Z_i = (Y_i,X_{1i},\ldots,X_{ki})$, $i=1,\ldots n$ and

$$T(z,F_n) = (y-\Sigma_{1\le t\le k}\hat{\theta}_t x_t)^2.$$

Example 3. Breiman and Friedman (1985) considered transformations $h(Y)$, $\varphi_1(X_1),\ldots,\varphi_p(X_p)$ of variables Y, $X_1,\ldots,X_p$ with the constraint that these transformed random variables have zero means and finite variances and $Var(h(Y))=1$. Transformations h^*, $\varphi_1{}^*,\ldots,\varphi_p{}^*$ are called optimal if they minimize $E[h(Y) - \varphi_1(X_1)-\cdots-\varphi_p(X_p)]^2$. Breiman and Friedman showed that optimal transformations exist. If we consider spline approximations of these transformations, based on n observations, then a quantity of interest is

$$s_n = \int [\hat{h}(y)-\hat{\varphi}_1(x_1)-\cdots-\hat{\varphi}_p(x_p)]^2 dF(y,x_1,\ldots,x_p)$$

where $\hat{h}$, $\hat{\varphi}_1,\ldots,\hat{\varphi}_p$ are estimated spline transformations.

In all three examples, the goal is to minimize s_n over k, each k corresponds to a lower-dimensional model. Since F is unknown the problem becomes one of finding an adequate estimator of s_n.

Consider the problem of estimating s_n using ordinary cross-validation and a variant called v-fold cross-validation. Other methods of estimating s_n include the bootstrap, repeated learning-testing (Burman (1989)), partitioned cross-validation (Marron (1988)), generalized cross-validation (Wahba (1985)). These are not discussed here.

Leave-one-out cross-validation (i.e., ordinary cross-validation).

Let $F_{n,-i}$ be the empirical d.f. based on the n-1 observations excluding Z_i. Then the ordinary cross-validated estimate of s_n is $CV_n = n^{-1}\Sigma_{1\le i\le n}T(Z_i,F_{n,-i})$.

v-fold cross-validation.

Divide the data randomly into v groups of as near equal size as possible. If the sizes of the groups are $m_1,\ldots,m_v$, then $m_\alpha \simeq n/v$, $\alpha=1,\ldots,v$. For $\alpha=1,\ldots,v$, let $F_{n\alpha}$ be the empirical d.f. based on the $n-m_\alpha$ observations not in group α and $\bar{F}_{n\alpha}$ be the empirical d.f. constructed from the m_α observations in group α. Then the v-fold cross-validated estimate of s_n is

$$CV_{nv} = \Sigma_{1\le\alpha\le v}(m_\alpha/n)\int T(z,F_{n\alpha})d\bar{F}_{n\alpha}(z).$$

From a theoretical point of view, CV_{nv} estimates s_n fairly well if $v \to \infty$. Unfortunately, for small v (say 3 or 4) CV_{nv} may not be a good estimate of s_n since $E\{CV_{nv}-s_n\} \approx c_o(v-1)^{-1}n^{-1}$ where c_o is a constant that depends on T and F (see Burman (1989)). However, it is for small v that this version of cross-validation is useful, because it can greatly reduce the amount of computation required to estimate s_n.

Corrected v-fold cross-validation.

As pointed out above, CV_{nv} may not be a good estimate of s_n. To remedy this, a correction term is needed. The corrected v-fold cross-validated estimate does approximate s_n well.

$$CV^*_{nv} = CV_{nv} - \Sigma_{1\le\alpha\le v}(m_\alpha/n)\int T(z,F_{n\alpha})dF_n(z) + \int T(z,F_n)dF_n(z).$$

The estimate CV^*_{nv} is better than the uncorrected v-fold cross-validated estimate CV_{nv} because $E\{CV^*_{nv} - s_n\} \approx (v-1)^{-1}c_1n^{-2}$ for some constant c_1 depending on T and F (see Burman (1989)).

3. How does ordinary cross-validation work?

We now heuristically argue to find out what makes ordinary cross-validation work. Consider the regression example (Example 1 in the last section):

$$Y_i = \mu(X_i)+\varepsilon_i,\ i=1,\ldots,n,$$

where $E(\varepsilon_1|X_1) = 0$ and $E(\varepsilon_1^2|X_1=x) = \sigma^2(x)$.

Let $\hat{\mu}$ be an estimate of μ. Then it can be shown (Burman and Nolan (1990)) that

$$E[CV_n - s_n] \approx E[\Sigma_{1\le i\le n}\ \sigma^2(X_i)W(X_i,X_i)/n^2 - \Sigma_{1\le i,j\le n}\ \varepsilon_i\varepsilon_jW(X_i,X_j)/n^2]$$

where W is a deterministic weight function. If

$$E(\varepsilon_i\varepsilon_j|X_1,\ldots,X_j) = 0 \text{ for all } i<j \qquad (*)$$

then

$$E[CV_n - s_n] \approx 0.$$

The property (*) is martingale-like. It suggests that ordinary cross-validation estimates s_n well for a broad class of processes, those which satisfy (*). Burman and Nolan (1990) explored this point in detail for Markov chains.

In Burman and Nolan (1990) we considered the problem of estimating the one step prediction function $\mu(\cdot) = E(X_2|X_1 = \cdot)$ on the basis of n+1 observations $X_1,\ldots,X_{n+1}$ from a stationary Markov chain. If we

write $Y_i = X_{i+1}$ and $\varepsilon_i = X_{i+1} - \mu(X_i)$, $i=1,\ldots,n$, then it is fairly easy to see that condition (*) holds in this case and hence, under a few additional conditions, cross-validation should still work here.

Consider spline estimates of μ with k knots and choose the value of k which minimizes

$$CV_n(k) = n^{-1}\Sigma_{1\le i\le n}\{X_{i+1} - \hat{\mu}_{k,-i}(X_i)\}^2$$

where $\hat{\mu}_{k,-i}$ is the spline estimate of μ with k knots, excluding the ith observation. If $\hat{k}$ is the index at which CV_n is minimized, then declare $\hat{\mu}_{\hat{k}}$ as the estimate of μ. For this example, Burman and Nolan (1990) show that if μ is bounded and the marginal density of X_1 is bounded away from zero, then under appropriate mixing conditions (φ-mixing) the following Shibata type optimality result holds for any $0 < \delta < 1$:

$$\int(\hat{\mu}_{\hat{k}} - \mu)^2 dF/\inf_{1\le k\le n^{1-\delta}}\int(\hat{\mu}_k - \mu)^2 dF \to 1, \text{ in probability.}$$

Here is a numerical example. We generated observations from the model:

$$X_{i+1} = .5 + .25\sin(2\pi X_i) + \varepsilon_i,\ i=1,\ldots,100,$$

where ε_i are i.i.d. Uniform(-.25,.25). We fit quadratic splines with knots at the sample quantiles. Ordinary cross-validation led to the selection of only three knots. A figure given at the end of this paper has a scatter diagram of X_i vs. X_{i+1}. The true and the cross-validated curves are also shown. Notice that the fitted curve is remarkably close to the true curve except perhaps near endpoints.

4. Cross-validation for the stationary case

Now assume that the n observations $Z_1,\ldots,Z_n$ are from a d-dimensional strictly stationary process satisfying some mixing condition. The recent book by Györfi et al. (1989) is a good source for many nonparametric problems for dependent data. As before assume that the marginal d.f. of Z_1 is F and that F_n is the empirical d.f.. The goal once again is to estimate a quantity of the form:

$$s_n = \int T(z,F_n)dF(z) \text{ or } E(s_n).$$

We provide few examples of T and s_n.

Example 4 (Autoregressive model fitting).

Let $X_1,\ldots,X_n$ be n observations from a stationary process. Fit an autoregressive model of order k where the parameters are estimated by the least squares method. Assume that $X_1',\ldots,X_n'$ is another process which has the same distribution as $X_1,\ldots,X_n$ but is independent of it. Following Akaike (1970), the error in predicting X_{k+1}' by

$\hat{\theta}_o + \hat{\theta}_1 X_k' + \cdots + \hat{\theta}_k X_1'$ is

$$s_n = \int (x_{k+1} - \hat{\theta}_o - \hat{\theta}_1 x_k - \cdots - \hat{\theta}_k x_1)^2 dF(x_{k+1}, \ldots, x_1).$$

Here $Z_i = (X_{i+k+1}, X_{i+k}, \ldots, X_{i+1})$ and $T(z, F_n) = (x_{k+1} - \hat{\theta}_o - \hat{\theta}_1 x_k - \cdots - \hat{\theta}_k x_1)^2$ with $z = (x_{k+1}, \ldots, x_1)$.

Example 5 (Additive prediction).

Let $X_1, \ldots, X_n$ be n observations from a stationary process. Generalizing the ideas in Hastie and Tibshirani (1986) or Stone (1985), we may consider the problem of predicting X_i by an additive model in $X_{i-1}, \ldots, X_{i-d}$ of the form $\mu_0 + \mu_1(X_{i-1}) + \cdots + \mu_d(X_{i-d})$. In his Ph.D. thesis, Chen (1990) has considered a number of such nonparametric estimators in time series problems. If we fit splines or polynomials for $\mu_1, \ldots, \mu_d$ and estimate them by the least squares method, then as in Example 4 the prediction error for predicting X_{d+1}' from $\hat{\mu}_0 + \hat{\mu}_1(X_d') + \cdots + \hat{\mu}_d(X_1')$ is given by

$$s_n = \int [x_{d+1} - \hat{\mu}_o - \hat{\mu}_1(x_d) - \cdots - \hat{\mu}_d(x_1)]^2 \, dF(x_{d+1}, \ldots, x_1)$$

where F is the distribution function of $(X_{d+1}, X_d, \ldots, X_1)$. Here

$$Z_i = (X_{i+d+1}, X_{i+d}, \ldots, X_{i+1}) \text{ and}$$

$$T(z, F_n) = [x_{d+1} - \hat{\mu}_o - \hat{\mu}_1(x_d) - \cdots - \hat{\mu}_d(x_1)]^2$$

with $z = (x_{d+1}, \ldots, x_1)$.

Example 6 Consider the regression model.

$$Y_i = \mu(X_i) + \varepsilon_i, \quad i=1, \ldots, n,$$

where $\{Z_i = (X_i, Y_i)\}$ is assumed to be stationary. The problem here is to estimate the regression function μ. Once again, if we fit a polynomial of degree k (or splines with k knots) and estimate the parameters by the method of least squares, then the quantity of interest is

$$s_n = \int [y - \hat{\mu}_k(x)]^2 dF(y, x)$$

where $\hat{\mu}_k$ is the estimate of μ.

How do we estimate s_n in these examples? Intuitively speaking, it is quite clear that ordinary cross-validation may easily run into trouble because of the dependent structure present in the model. One way out of this trouble is as follows. For each i, we obtain estimates after deleting h observations on both sides of Z_i and then test this estimate on Z_i. The average of these n tested values will be an estimate of s_n. This proposal also appears in Chu (1989), Györfi et al (1989). Call this version of cross-validation h-block

cross-validation. Both show, under a variety of conditions, that this method "works" for optimal model selection in nonparametric curve estimation problems. Their choice of h requires: $h/n \to 0$; knowledge of the smoothness of the unknown curve; and knowledge of the amount of dependence of the stationary process. Their arguments are tied to the nonparametric curve fitting situation.

Here we consider the case where h is allowed to be a fixed fraction of n. We believe that deleting a fixed fraction of the sample is useful in long range dependence. As in the case of v-fold cross-validation for i.i.d. case, certain correction terms are needed to make this work. Since the discussion here is nontechnical, we will try to motivate the main steps. Detailed arguments will appear in Burman, Chow and Nolan (1990). Let h be a positive integer. For $i=1,\ldots,n$, let F_{ni} be the following estimator of F

$$F_{ni}(A) = \Sigma_{1\le j\le n}\, w_{ij} I(Z_j \varepsilon A), \quad A \text{ any borel set in } R^d,$$

where $\{w_{ij}: 1\le i,j\le n\}$ is a nonnegative double array of numbers satisfying

$$w_{ij} = 0 \text{ whenever } |i-j| \le h$$

$$\Sigma_{1\le i\le n} w_{ij} = 1 \quad \text{for all } j=1,\ldots,n.$$

The first condition is imposed to make sure that the learning set $\{Z_j: |j-i| > h\}$ and the test set $\{Z_i\}$ are almost independent. The second condition is a technical condition needed to ensure that

$$n^{-1}\Sigma_{1\le i\le n} F_{ni} = F_n.$$

Many choices of $\{w_{ij}\}$ satisfy these constraints; this issue will not be discussed. The h-block cross-validated estimate of s_n is

$$CV_{nh} = n^{-1}\Sigma_{1\le i\le n}\, T(Z_i, F_{ni}).$$

Heuristic calculations show that

$$E\{CV_{nh} - s_n\} = n^{-1}\Sigma_{1\le i\le n}\, E\{T(Z_i,F_{ni}) - \int T(z,F_{ni})dF(z)\}$$
$$+ n^{-1}\Sigma_{1\le i\le n}\, E\{\int T(z,F_{ni})dF(z) - \int T(z,F_n)dF(z)\}.$$

In the last expression, the first term is negligible if h is large, whereas the second term is roughly $c_2 h n^{-2}(1-2h/n)^{-1}$ where c_2 is a constant that depends on T and the probability distribution of the process. Typically c_2 increases with k, the number of parameters being estimated. So the second term is not necessarily small if h/n or k is not small. To tackle this problem, we have introduced a correction term. The corrected h-block cross-validated estimate of s_n is

$$CV^*_{nh} = CV_{nh} - n^{-1}\Sigma_{1\le i\le n} \int T(z,F_{ni})dF_n(z) + \int T(z,F_n)dF_n(z).$$

This corrected h-block cross-validated estimate works fairly well. It turns out that

$$E\{CV^*_{nh} - s_n\} \approx c_3 h n^{-3}(1-2h/n).$$

where c_3 depends on T and the probability distribution of the process.

The following table summarizes some of the findings from a simulation study to appear in Burman, Chow and Nolan (1990). We generated a sample of size n=36 observations from a mean zero stationary Gaussian process with unit variance whose autocorrelations of order one, two etc. are 0, .6, 0, $(.6)^2$, 0, $(.6)^3$, 0,.... We fitted a autoregressive model of order 1 as in Example 4. We calculated the means and the standard deviations of s_n, CV_{nh} and CV^*_{nh} for various values of (2h/n)100% (the percentage of deleted observations). The simulation shows that ordinary cross-validation does not perform well. It also shows that corrected h-block cross-validation CV^*_{nh} does better than CV_{nh}; the improvement is not only in the mean but also in the standard deviation.

The weights $\{w_{ij}\}$ used here are as follows:

for $1 \le j \le h$, $w_{ij}=1/(n-j-h)$ when $|i-j| > h$ and zero otherwise,

for $h < j < n-h$, $w_{ij}=1/(n-2h-1)$ when $|i-j| > h$ and zero otherwise,

for $n-h \le j \le n$, $w_{ij}=1/(j-h-1)$ when $|i-j| > h$ and zero otherwise.

The simulation results presented here are based on 1000 repetitions. Note that h/n=0 corresponds to the case of ordinary cross-validation.

Table: Estimated values of the mean and the standard deviation of h-block cross-validated estimates of s_n

$E(s_n) = 1.20$, $s.d.(s_n)=.23$

	Percentage of deleted observations					
	0	20%	30%	40%	50%	60%
$E(CV_{nh})$	.93	1.22	1.25	1.31	1.40	1.46
$E(CV^*_{nh})$	.92	1.18	1.19	1.20	1.21	1.22
$s.d.(CV_{nh})$	.32	.52	.55	.64	.75	.86
$s.d.(CV^*_{nh})$	.32	.49	.51	.56	.61	.66

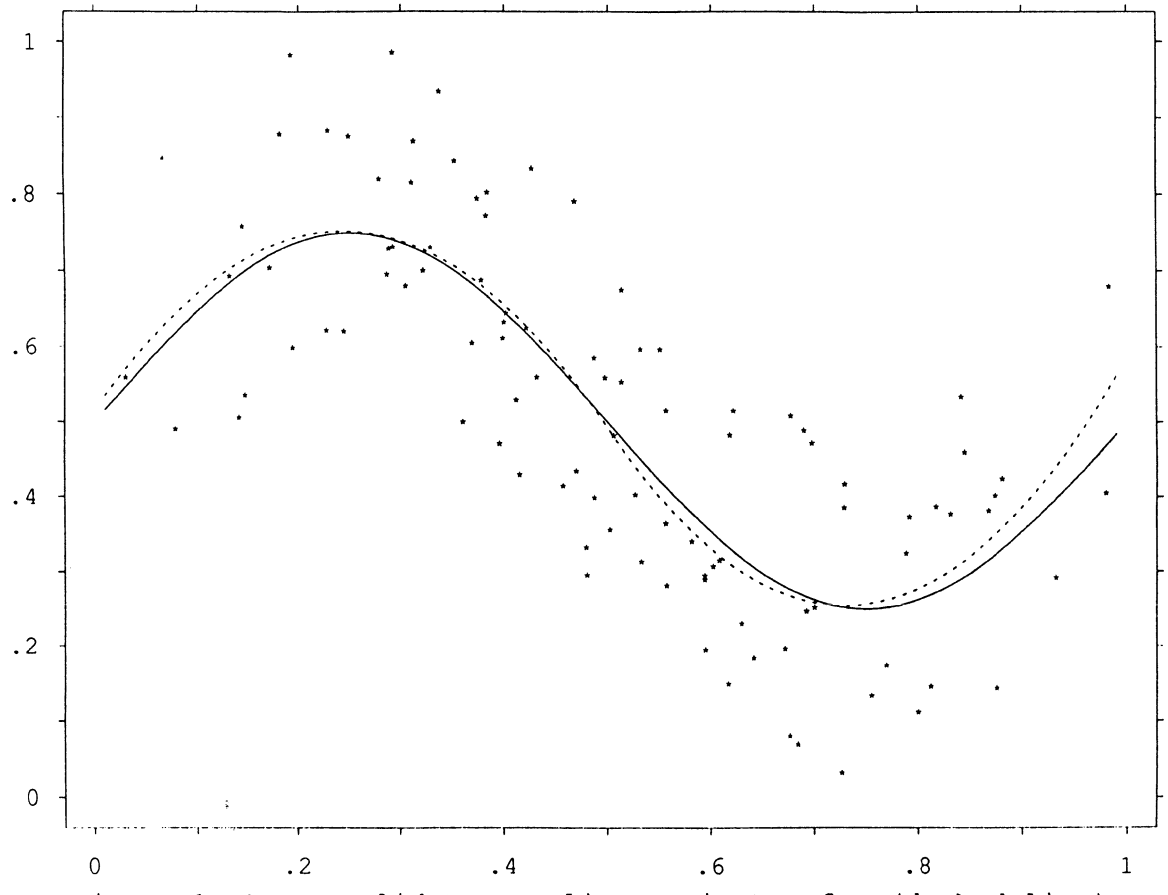


Figure 1. Cross-validatory spline estimate of μ (dashed line)

References

Akaike, H. (1970) 'Statistical predictor identification'. Ann. Inst. Statist. Math., 22, 203-217.

Breiman, L. and Friedman, J. (1985) 'Estimating optimal transformations for regression and correlation'. J. Amer. Statist. Assoc., 80, 580-619.

Burman, P. (1989) 'A comparative study of ordinary cross-validation, v-fold cross-validation and repeated learning-testing methods'. Biometrika, 76, 503-514.

Burman, P. and Nolan, D. (1990) 'Data dependent estimation of prediction functions'. Unpublished manuscript.

Burman, P., Chow, E. and Nolan, D. (1990) 'A modified cross-validation method for dependent observations'. In preparation.

Chen, R. (1990) 'Nonlinear time series analysis'. Ph.D. dissertation, Department of Statistics, Carnegie-Mellon University.

Chu, C.K. (1989) 'Some results in nonparametric regression'. Ph.D. dissertation, Department of Statistics, University of North Carolina, Chapel Hill.

Geisser,S. (1975) 'A predicitive sample reuse method with applications'. J. Amer. Statist. Assoc., 70, 320-328.

Györfi, L., Härdle. W., Sarda, P. and Vieu, P. (1989) 'Nonparametric curve estimation from time series', Springer-Verlag, New York.

Härdle, W. and Marron, J.S. (1985) 'Optimal bandwidth selection in nonparametric regression function estimation'. Ann. Statist., 13, 1465-1481.

Hastie, T. and Tibshirani, R. (1986) 'Generalized addition models (with discussion)'. Statist. Sci., 1, 297-318.

Marron, J.S. (1988) 'Automatic smoothing parameter selection: A survey'. Empirical Economics, 13, 187-208.

Stone, M. (1974) 'Cross-validatory choice and the assessment of statistical predictions' (with discussion). J. Roy. Statist. Soc. Ser. B, 36, 111-133.

Stone, C. (1985) 'Additive regression and other nonparametric models'. Ann. Statist., 13, 689-705.

Wahba, G. (1985) 'A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem'. Ann. Statist., 13, 1378-1402.

NONPARAMETRIC FUNCTION ESTIMATION INVOLVING ERRORS-IN-VARIABLES

Jianqing **FAN**
Department of Statistics
University of North Carolina
Chapel Hill, N.C. 27599

Young K. **TRUONG**
Department of Biostatistics
University of North Carolina
Chapel Hill, N.C. 27599

Yonghua **WANG**
Department of Statistics & Biostatistics
University of North Carolina
Chapel Hill, N.C. 27599
USA

ABSTRACT. We examine the effect of errors in covariates in nonparametric function estimation. These functions include densities, regressions and conditional quantiles. To estimate these functions, we use the idea of deconvoluting kernels in conjunction with the ordinary kernel methods. We also discuss a new class of function estimators based on local polynomials.

1 Introduction

In nonparametric regression analysis, one is interested in analyzing the relationship between the response Y and the covariate X through the regression function $E(Y|X)$. Suppose now X contains another random variable X° so that $X = X^\circ + \varepsilon$—a situation in which the variable of interest (X°) is *measured with error.* An important issue arises as how to recover the association between Y and X°. Note that X° is not directly observable and this complicates the problem of estimating the regression function $E(Y|X^\circ)$ since one will have to "recover" X° from X first. Alternately, it may be worth exploring the association between Y and X° by considering the conditional median function $\text{med}(Y|X^\circ)$, especially in situations involving asymmetric conditional distributions. Note that it is necessary to make this problem identifiable by assuming the "error" ε has a *known* distribution and that it is *independent* of X° and Y.

Given a training set $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution of (X, Y) with $X = X^\circ + \varepsilon$, the problem of estimating the regression function $m(X^\circ) =$

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 613–627.
© 1991 *Kluwer Academic Publishers.*

$E(Y|X^\circ)$ or the conditional median function $m(X^\circ) = \text{med}(Y|X^\circ)$ is called regression analysis with errors-in-variables. It is said to be *parametric* if the regression function is assumed to be a specific function with unknown parameters. See, for example, Armstrong (1985), Stefanski (1985), Stefanski and Carroll(1985), Prentice (1986), Whittemore and Keller (1986) and Fuller (1987). In this approach, there are no formal ways of verifying the appropriateness of the regression model. Moreover, the parameters are estimated by maximizing the likelihood equation which is usually very complicated.

To overcome the above difficulties in parametric analysis, the present paper adopts the nonparametric approach by estimating the regression function directly. In the absence of measurement errors, there is now a huge literature on regression function estimation for studying the structures between Y and X. In situation with measurement errors, it is necessary to modify these methods since regression is used to study the effect of X° on Y, not X on Y. To achieve this, we must know how to recover X° from X. This operation is called *deconvolution* and it was first used in density estimation involving measurement errors.

This paper is outlined as follows. Section 2 begins with the problem of density estimation based on deconvoluted kernel method, since it is really the building block for estimating other functions. It then considers the problem of estimating nonparametric regression functions such as conditional mean and conditional median. Section 3 reviews some optimal properties possessed by these estimators. Section 4 discusses generalizations of these kernel methods to design-adaptive method based on local polynomials. Numerical examples are given in Section 5. Section 6 contains some concluding remarks.

2 Methods

2.1 Deconvoluted kernel density estimators

Set $X = X^\circ + \varepsilon$ and let $X_1, \ldots, X_n$ denote a training sample from the distribution of X. Assume that the error ε has a non-vanishing characteristic function $\phi_\varepsilon(t) \neq 0$. Denote the characteristic functions of X and X° by $\phi_X(\cdot)$ and $\phi_{X^\circ}(\cdot)$, respectively. By Fourier inversion, the density function of X° is given by

$$f_{X^\circ}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itx)\phi_{X^\circ}(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itx)\frac{\phi_X(t)}{\phi_\varepsilon(t)}\,dt.$$

Thus the problem of estimating the density function of X° can be reduced to the estimation of the function

$$\phi_X(t) = \int_{-\infty}^{\infty} \exp(itx) f_X(x)\, dx,$$

which, in turn is a density estimation problem based on X. To this end, let $K(\cdot)$ denote a kernel function and h_n be a bandwidth. Suppose the density function $f_X(\cdot)$ of X (no errors) is now estimated by the usual kernel method:

$$\hat{f}(x) = \frac{1}{nh_n} \sum_1^n K\left(\frac{x - X_j}{h_n}\right).$$

Then we arrive at an estimator of $\phi_X(\cdot)$:

$$\begin{aligned}
\hat{\phi}_X(t) &= \int_{-\infty}^{\infty} \exp(itx) \hat{f}(x)\, dx \\
&= \int_{-\infty}^{\infty} \exp(itx) \frac{1}{nh_n} \sum_1^n K\left(\frac{x - X_j}{h_n}\right) dx \\
&= \frac{1}{n} \sum_1^n \exp(itX_j) \int_{-\infty}^{\infty} \exp(ituh_n) K(u)\, du \\
&= \hat{\phi}_n(t) \phi_K(th_n),
\end{aligned}$$

where $\phi_K(\cdot)$ is the Fourier transform of $K(\cdot)$ and $\hat{\phi}_n(t) = n^{-1} \sum_1^n e^{itX_j}$. This leads to a natural estimator of the density function of X°:

$$\begin{aligned}
\hat{f}_n(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx) \frac{\hat{\phi}_X(t)}{\phi_\varepsilon(t)}\, dt \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx) \phi_K(th_n) \frac{\hat{\phi}_n(t)}{\phi_\varepsilon(t)}\, dt. \qquad (2.1)
\end{aligned}$$

It is useful to note that the above estimator can be rewritten in the kernel form:

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_1^n K_n\left(\frac{x - X_j}{h_n}\right), \qquad (2.2)$$

where the kernel (called deconvoluted kernel) is given by

$$K_n(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx) \frac{\phi_K(t)}{\phi_\varepsilon(t/h_n)}\, dt. \qquad (2.3)$$

This deconvoluted kernel density estimator was considered by Stefanski and Carroll (1990) and Zhang (1990). Optimal properties are established by Fan (1991a). See also references given therein.

2.2 Deconvoluted kernel regression function estimators

Given a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution of (X, Y), we first review the kernel estimator of the regression function $m(x) = E(Y|X = x)$. As before, let $K(\cdot)$ denote a kernel function and h_n be a bandwidth. Then the kernel estimator is the weighted average given by (Nadaraya (1964), Watson (1964))

$$\hat{m}_n(x) = \frac{\sum_j Y_j K\left(\frac{x-X_j}{h_n}\right)}{\sum_i K\left(\frac{x-X_i}{h_n}\right)}. \tag{2.4}$$

Now suppose $X = X^\circ + \varepsilon$ and consider the problem of estimating the regression function $E(Y|X^\circ)$. Since the deconvoluted kernel (2.3) has the effect of separating X° from X in constructing the correct neighborhood around a given x, hence the kernel estimator (2.4) should be modified by replacing $K(\cdot)$ with $K_n(\cdot)$ to account for this effect. Based on this idea, Fan and Truong (1990) proposed the following estimator for $E(Y|X^\circ = x)$:

$$\hat{m}_n(x) = \frac{\sum_j Y_j K_n\left(\frac{x-X_j}{h_n}\right)}{\sum_i K_n\left(\frac{x-X_i}{h_n}\right)}. \tag{2.5}$$

This estimator has many interesting optimal properties depending on the error distributions, which will be discussed with more details in Section 3.

2.3 Deconvoluted kernel median function estimators

Now consider the problem of estimating the conditional median function $m(x) = \text{med}(Y|X^\circ)$. We first consider the problem of estimating the conditional distribution in the absence of measurement errors. A kernel estimator of $F(y|X = x) = P(Y \leq y|X = x)$ is given by

$$F_n(y|x) = \frac{\sum_j K\left(\frac{x-X_j}{h_n}\right) 1(Y_j \leq y)}{\sum_i K\left(\frac{x-X_i}{h_n}\right)}. \tag{2.6}$$

From this, one obtains the following conditional median estimator [see Truong (1989)]:

$$m_n(x) = F_n^{-1}(1/2|x). \tag{2.7}$$

To deal with the case involving measurement errors, we simply replace the kernel function $K(\cdot)$ by $K_n(\cdot)$ given in (2.3). This yields the deconvoluted kernel estimator of $F(y|X^\circ = x)$:

$$\hat{F}_n(y|x) = \frac{\sum_j K_n\left(\frac{x-X_j}{h_n}\right) 1(Y_j \le y)}{\sum_i K_n\left(\frac{x-X_i}{h_n}\right)}. \tag{2.8}$$

To estimate $m(x) = \text{med}(Y|X^\circ = x)$, we propose

$$\hat{m}_n(x) = \hat{F}_n^{-1}(1/2|x). \tag{2.9}$$

More generally, it is easy to modify this approach to cover the problem of estimating conditional quantiles. Let $0 < p < 1$. The pth conditional quantile is defined as $F^{-1}(p|X^\circ = x)$, which is the pth quantile of the conditional distribution $F(\cdot|X^\circ = x)$. To estimate this quantile, we propose the following deconvoluted kernel estimator:

$$\hat{F}_n^{-1}(p|x). \tag{2.10}$$

Since (2.5) and (2.9) are respectively the mean and the median of (2.8), it is expected that the conditional median estimator (2.9) shares the same optimal properties as the regression estimator (2.5).

3 Optimal rates of convergence

Theoretical aspects of the problem on estimation of the density function f_{X° and the regression function $E(Y|X^\circ)$ are given in Fan (1991a, b) and Fan and Truong (1990). These results can be highlighted as follows.

- The estimates (2.1) and (2.5) are optimal in terms of rate of convergence.
- The rates of convergence depend on the smoothness of error distributions, which can be characterized into two categories: ordinary smooth and super smooth. Let ϕ_ε be the characteristic function of the error distribution. The distribution of ε is said to be
  - super smooth of order β: if the function $\phi_\varepsilon(\cdot)$ satisfies

    $$d_0|t|^{\beta_0}\exp(-|t|^\beta/\gamma) \le |\phi_\varepsilon(t)| \le d_1|t|^{\beta_1}\exp(-|t|^\beta/\gamma), \quad \text{as } t \to \infty, \tag{3.1}$$

    where d_0, d_1, β, γ are positive constants and β_0, β_1 are constants;

– ordinary smooth of order β: if the function $\phi_\varepsilon(\cdot)$ satisfies

$$d_0|t|^{-\beta} \leq |\phi_\varepsilon(t)| \leq d_1|t|^{-\beta} \quad \text{as } t \to \infty, \tag{3.2}$$

for positive constants d_0, d_1, β.

For example, super smooth distributions are

$$\begin{cases} N(0,1) & \text{(normal)} \quad \text{with } \beta = 2, \\ \frac{1}{\pi}\frac{1}{1+x^2} & \text{(Cauchy)} \quad \text{with } \beta = 1. \end{cases}$$

Ordinary smooth distributions are

$$\begin{cases} \frac{\alpha^p}{\Gamma(p)} x^{p-1} e^{-\alpha x} & \text{(Gamma)} & \text{with } \beta = p, \\ \frac{1}{2} e^{-|x|} & \text{(double exponential)} & \text{with } \beta = 2\ . \end{cases}$$

The rates of convergence for deconvoluted kernel estimators depend on β, the order of smoothness of the error distribution. They also depend on the smoothness of the regression function $m(x)$ and the marginal density function. For regression functions with bounded k-th derivatives, the following table illustrates the optimal rates of convergence according to the error distribution.

Error distribution	*Rates*	*Error distribution*	*Rates*
$N(0,1)$	$(\log n)^{-k/2}$	Gamma(α, p)	$n^{-k/(2k+2p+1)}$
Cauchy (0,1)	$(\log n)^{-k}$	Double exponential	$n^{-k/(2k+5)}$

For density estimation based on deconvolution, a similar table is provided in Fan (1991a).

- The optimal choice of bandwidth h_n depends also on the error distribution. For the supersmooth error distribution of order β, the optimal bandwidth is $h_n = c(\log n)^{-1/\beta}$ for some constant c depending only on the error distribution and the kernel function. In the ordinary case, the optimal choice of bandwidth is $h_n = dn^{-\frac{1}{2k+2\beta+1}}$, for some constant d.

- Suppose $\min_{x\in[a,b]} f_{X^\circ}(x) > 0$ and that the conditional variance

$$\sigma^2(x) = \text{var}(Y|X^\circ = x) = \sigma^2.$$

Set

$$\hat{\sigma}_n^2 = \sum_1^n (Y_j - \hat{m}_n(X_j))^2 1_{\{a \le X_j \le b\}} / \sum_1^n 1_{\{a \le X_j \le b\}}$$

and

$$\hat{V}_n^2 = \frac{\hat{\sigma}_n^2}{n} \sum_1^n \frac{1}{h_n^2} K_n^2 \left(\frac{x - X_j}{h_n} \right).$$

Then under appropriate conditions on h_n (which depends on the error distribution),

$$\sqrt{n} \frac{\hat{f}_n(x)[\hat{m}_n(x) - m(x)]}{\hat{V}_n} \longrightarrow N(0,1).$$

Note that this result is presented so that confidence intervals about the regression function can be easily obtained. For details, see Fan, Truong and Wang (1990). See also therein for a discussion on how to choose h_n for different type of error distributions. A similar result for density estimation is given Fan (1991b).

4 Extensions: a design-adaptive approach

4.1 Regression functions

As in the case of ordinary kernel regression function estimation, the optimal properties of the deconvoluted kernel regression estimator (2.5) depends on some smoothness condition of the marginal distribution. For example, to show that (2.5) possesses the optimal rates of convergence when the regression function $m(x)$ has a bounded $(k+1)$st derivative, it is also required that the marginal density function has a bounded $(k+1)$st derivative, too. See Fan and Truong (1990). In this section, we propose a different approach that would remove the extra smoothness condition imposed on the marginal distribution.

Note that the deconvoluted kernel regression estimator (2.5) can be viewed as $\hat{m}_n(x) = \hat{a}$ where $\hat{a}$ minimizes L^2-discrepancy:

$$G(a) = \sum_j (Y_j - a)^2 K_n \left(\frac{x - X_j}{h_n} \right),$$

with $K_n(\cdot)$ being the deconvoluted kernel function (2.3). Suppose now the regression function $m(x)$ has a bounded $(k+1)$st derivative. Define a new

estimator by $\hat{m}_n(x) = \hat{a}_0$, where $\hat{a}_0, \ldots, \hat{a}_k$ minimize

$$G(a_0, \ldots, a_k) = \sum_j (Y_j - a_0 - a_1(x - X_j) - \cdots - a_k(x - X_j)^k)^2 K_n\left(\frac{x - X_j}{h_n}\right). \tag{4.1}$$

In fact, in the absence of measurement errors, Stone (1982) established that the estimators defined by (4.1) achieve the optimal rates of convergence without putting smoothness conditions on the marginal density, and Fan (1990) extended this idea to the smooth kernel case which possesses a number of efficient properties. The present approach is inspired by the latter idea based on smooth kernel and it is called design-adaptive since it does not require extra smoothness conditions on the marginal density function for achieving the optimal rates of convergence. In other words, the estimator adapts to both nearly uniform designs and highly clustered designs.

Example 1: **Weighted Linear Regression**

Suppose the regression function $m(x) = E(Y|X^\circ = x)$ has a bounded second derivative. Define $\hat{m}_n(x) = \hat{a}$, where $\hat{a}$, along with $\hat{b}$, minimize

$$G(a, b) = \sum_j (Y_j - a - b(x - X_j))^2 K_n\left(\frac{x - X_j}{h_n}\right).$$

Simple algebra shows

$$\hat{m}_n(x) = \sum_1^n w_{n,j} Y_j / \sum_1^n w_{n,i},$$

where

$$w_{n,j} = K_n\left(\frac{x - X_j}{h_n}\right)(s_{n,2} - (x - X_j)s_{n,1})$$

and

$$s_{n,l} = \sum_{j=1}^n K_n\left(\frac{x - X_j}{h_n}\right)(x - X_j)^l, \qquad l = 1, 2.$$

4.2 Quantile functions

The deconvoluted kernel median function estimator (2.9) can also be viewed as $\hat{m}_n(x) = \hat{a}$ where $\hat{a}$ minimizes L_1-discrepancy:

$$G(a) = \sum_j |Y_j - a| K_n\left(\frac{x - X_j}{h_n}\right),$$

with $K_n(\cdot)$ given by (2.3). Similar to the approach given in the previous section, define a new estimator by $\hat{m}_n(x) = \hat{a}_0$, where $\hat{a}_0, \ldots, \hat{a}_k$ minimize

$$G(a_0, \ldots, a_k) = \sum_j |Y_j - a_0 - a_1(x - X_j) - \cdots - a_k(x - X_j)^k| K_n\left(\frac{x - X_j}{h_n}\right). \tag{4.2}$$

The estimator defined by (4.2) can further be extended as follows. Let $\hat{m}_n(x) = \hat{a}_0$, with $\hat{a}_0, \ldots, \hat{a}_k$ minimizing

$$G(a_0, \ldots, a_k) = \sum_j g(Y_j - a_0 - a_1(x - X_j) - \cdots - a_k(x - X_j)^k) K_n\left(\frac{x - X_j}{h_n}\right), \tag{4.3}$$

where $g(z) = (pz_+ + (1-p)z_-)$, $0 < p < 1$. This leads to the estimation of conditional quantiles in general. Note that, when $p = 1/2$, (4.3) reduces to (4.2).

5 Numerical Results

In this section, we are interested in the estimation of regression function for binary responses. Let Y denote a 0-1 random variable whose distribution depends on another random variable (covariate) X° so that the logistic transformation of its regression function is quadratic (see Figure 2):

$$m(x) = E(Y|X^\circ = x) = P(Y = 1|X^\circ = x) = \frac{\exp(-6(x - 0.5)^2)}{1 + \exp(-6(x - 0.5)^2)}.$$

That is, given $X^\circ = x$, Y takes the value 1 with probability $m(x)$. Now suppose $X^\circ \sim \text{unif}(0, 1)$ and that X° is not available, instead we observe it through $X = X^\circ + \varepsilon$, where ε is a random error with mean 0 and finite variance. Given a random sample from the distribution of (X, Y), we would like to estimate $m(x) = E(Y|X^\circ = x)$. From the theoretical aspect of Section 3, we know that the error distribution plays an important role in determining the sampling behaviors of the deconvoluted kernel estimator. Computationally, this will be addressed in the following examples.

Example 2: **Super Smooth Errors**

Suppose ε has a normal distribution with mean 0 and variance σ_0^2. Let $(X_1, Y_1), \ldots, (X_{200}, Y_{200})$ denote a random sample from the distribution of (X, Y) so that

$$\begin{aligned} X_i &= X_i^\circ + \varepsilon_i, \quad X_i^\circ \sim_{\text{iid}} \text{unif}(0,1), \quad \varepsilon_i \sim_{\text{iid}} N(0, \sigma_0^2), \\ Y_i &= \begin{cases} 1, & \text{with probability } m(X_i^\circ); \\ 0, & \text{with probability } 1 - m(X_i^\circ). \end{cases} \end{aligned} \tag{5.1}$$

Here σ_0 is selected so that $\sigma_0^2/\text{var}(X_i^\circ) = 0.10$. Suppose the kernel $K(\cdot)$ is an inverse triangular density so that its Fourier transform is given by

$$\phi_K(t) = (1 - |t|)_+.$$

By $\phi_\varepsilon(t) = \exp(-\frac{1}{2}\sigma_0^2 t^2)$ and (2.3),

$$K_n(x) = \frac{1}{\pi} \int_0^1 \cos(tx)(1-t) \exp\left(\frac{\sigma_0^2 t^2}{2h_n^2}\right) dt. \tag{5.2}$$

For the estimator (2.5) to achieve the optimal rates of convergence, the bandwidth h_n is chosen so that $h_n = c\sigma_0(\log n)^{-1/2}$ with $c > 1$. See Fan and Truong (1990). Note that the deconvoluted kernel function depends on n. Figure 1 plots these functions for different values of the constant factor c. Deconvoluted estimators with different c are presented in Figure 2.

Example 3: **Ordinary Smooth Errors**

As in (5.1), $(X_1, Y_1), \ldots, (X_{200}, Y_{200})$ denotes a random sample from the distribution of (X, Y) with ε now having a double exponential distribution:

$$f_\varepsilon(z) = \sigma_0^{-1} \exp(-2|z|/\sigma_0).$$

Here σ_0 is chosen so that $\sigma_0^2/\text{var}(X_i^\circ) = 0.10$. The 200 simulated data are plotted in Figure 4. A star " * " indicates an observed datum point.

Note that

$$\phi_\varepsilon(t) = \frac{1}{1 + \frac{1}{4}\sigma_0^2 t^2}.$$

If $K(\cdot)$ is the Gaussian kernel $K(x) = (\sqrt{2\pi})^{-1} \exp(-x^2/2)$, then by (2.3),

$$K_n(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx)\phi_K(t) \left(1 + \frac{\sigma_0^2 t^2}{4h_n^2}\right) dt$$

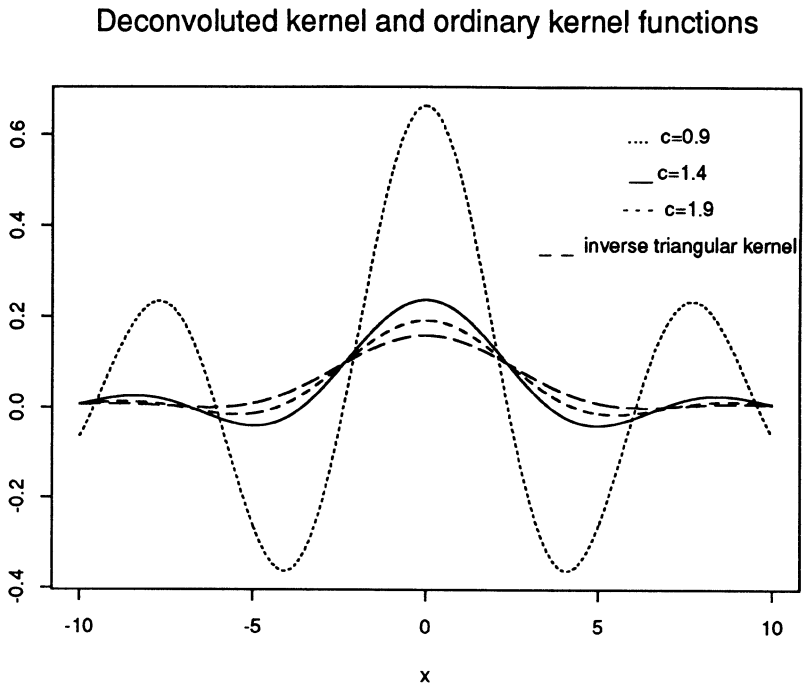


Figure 1

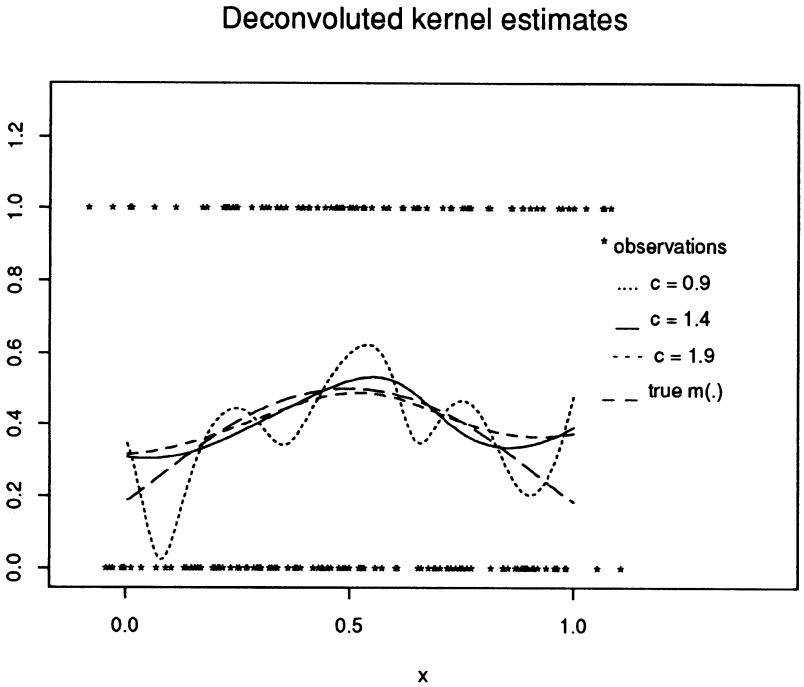


Figure 2

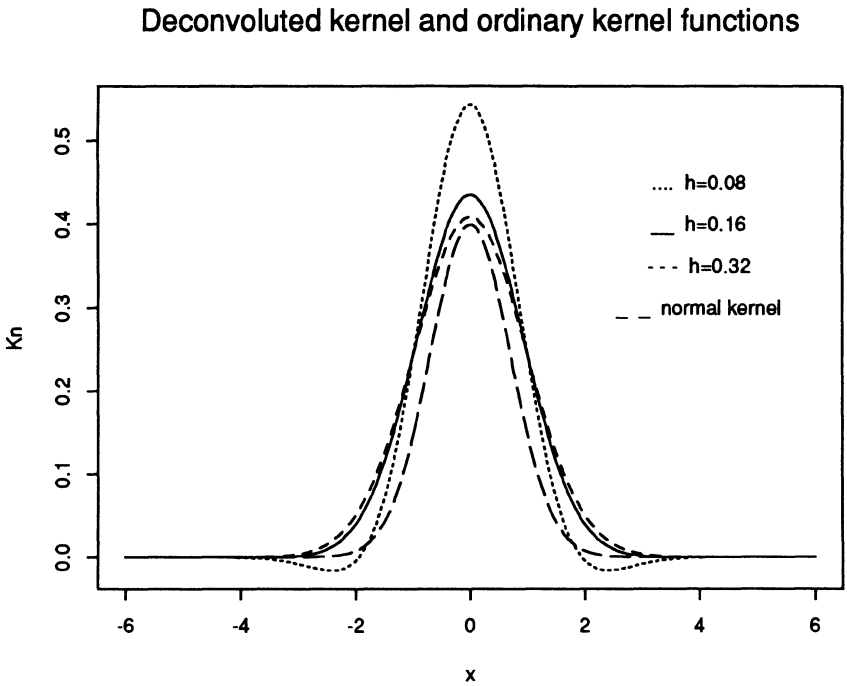


Figure 3

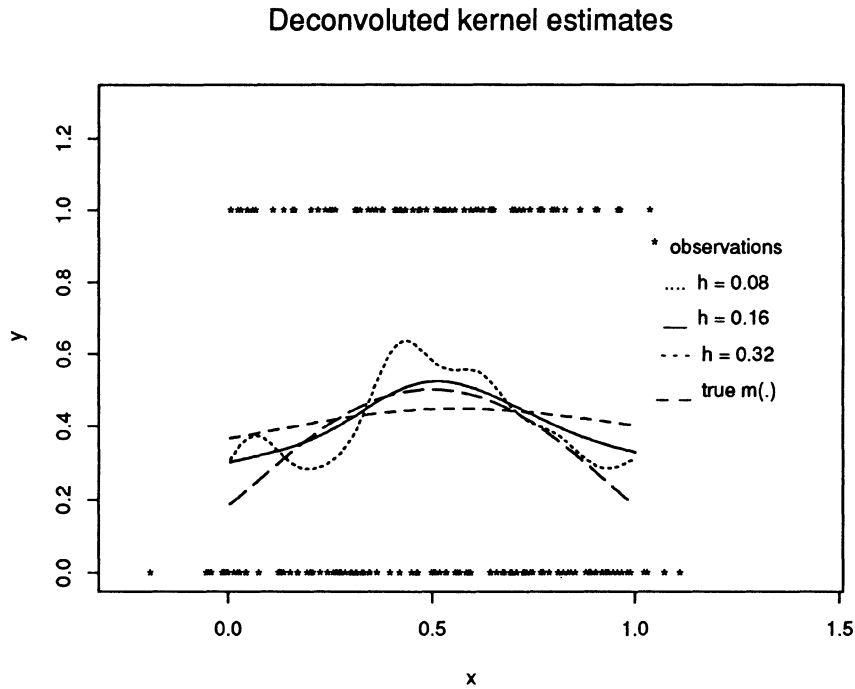


Figure 4

$$
\begin{aligned}
&= K(x) + \frac{\sigma_0^2}{4h_n^2}\frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itx)t^2\phi_K(t)\,dt \\
&= K(x) - \frac{\sigma_0^2}{4h_n^2}K''(x) \\
&= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)\left[1 - \frac{\sigma_0^2}{4h_n^2}(x^2-1)\right].
\end{aligned}
$$

Figure 3 plots the deconvoluted kernel functions and Figure 4 gives the estimators (2.5) with $h_n = 0.08, 0.16, 0.32$. It is clear that the estimate with $h_n = 0.08$ undersmoothes the curve as the estimate is not smooth enough, while the estimate with $h_n = 0.32$ oversmoothes the curve since the estimate does not capture the curvature.

Fan and Truong (1990) showed that the optimal rates of convergence can be achieved by choosing $h_n \sim cn^{-1/9}$ with $c > 0$.

In these examples, the estimate (2.5) is also computed with other choices of bandwidth (not shown here). We observed that the estimates for the super smooth case are very sensitive to the choice of the bandwidth when $h_n = c\sigma_0(\log n)^{-1/2}$ *with* $c < 1$. This is compatible with the theory: for $c < 1$, the variance is so large that the estimate is not even *consistent*. One the other hand, the bandwidth is relatively less sensitive for the ordinary smooth case.

6 Conclusions

In this paper, we discuss methods for estimating nonparametric functions involving errors-in-variables. Density estimation was considered first because it sets the foundation for *deconvolution*. We then address problems on estimating the regression function and conditional quantiles. Extensions of these estimators to design-adaptive ones (i.e. local polynomials) are also given. Although we have not provided proofs, it seems, however, plausible that the design-adaptive estimators would achieve the optimal rates of convergence without requiring extra smoothness conditions on the marginal distribution. There are also some other important open problems:

- the performance of the deconvoluted conditional quantile estimators (2.7) and (2.10);

- the sampling properties of design-adaptive quantile estimators (4.2) and (4.3);
- the amount of noise level $\sqrt{\mathrm{var}(\varepsilon)}$ that would make nonparametric deconvolution feasible.

References

[1] Armstrong, B. (1985). Measurement error in the generalized linear models. *Communications in Statistics — Simulation and Computation*, **14**, 529–544.

[2] Fan, J. (1990). A remedy to regression estimators and nonparametric minimax efficiency. *Institute of Statistics Mimeo Series #2028*, University of North Carolina, Chapel Hill.

[3] Fan, J. (1991a). On the optimal rates of convergence for nonparametric deconvolution problem. To appear in *Ann. Statist.*

[4] Fan, J. (1991b). Asymptotic normality for deconvolving kernel density estimators. To appear in *Sankhyā*.

[5] Fan, J. and Truong, Y. (1990). Nonparametric regression with errors-in-variables. *Institute of Statistics Mimeo Series #2023*, Univ. of North Carolina, Chapel Hill.

[6] Fan, J., Truong, Y. and Wang, Y. (1990). Confidence intervals for nonparametric regression with errors-in-variables. *Manuscript.*

[7] Fuller, W. A. (1987). *Measurement error models.* Wiley, New York.

[8] Nadaraya, E. A. (1964). On estimating regression. *Theor. Probab. Appl.*, **9**, 141–142.

[9] Prentice, R. L. (1986). Binary regression using an extended Beta-Binomial distribution, with discussion of correlation induced by covariate measurement errors. *J. Amer. Statist. Assoc.*, **81**, 321–327.

[10] Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika*, **72**, 583–592.

[11] Stefanski, L. A. and Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Ann. Statist.*, **13**, 1335–1351.

[12] Stefanski, L. A. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, **21**, 169-184.

[13] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, **10**, 1040-1053.

[14] Truong, Y.K. (1989). Asymptotic properties of kernel estimators based on local medians. *Ann. Statist.*, **17**, 606-617.

[15] Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A*, **26**, 359–372.

[16] Whittemore, A. S. and Keller, J. B. (1986). Approximations for errors in variables regression. *J. Amer. Statist. Assoc.*, **83**, 1057-1066.

[17] Zhang, C. H. (1990). Fourier methods for estimating mixing densities and distributions. *Ann. Statist.*, **18**, 806–830.

VIII. Various Topics (Continued)

A CONSISTENT GOODNESS OF FIT TEST BASED ON THE TOTAL VARIATION DISTANCE

László GYÖRFI
Hungarian Academy of Sciences
Technical University of Budapest
Stoczek u.2
H-1111 Budapest, Hungary

Edward C. VAN DER MEULEN
Department of Mathematics
Katholieke Universiteit Leuven
Celestijnenlaan 200B
B-3030 Leuven, Belgium

ABSTRACT. In this paper we propose a consistent test for testing a simple hypothesis versus a composite alternative. The simple hypothesis consists of a continuous distribution ν, while the composite alternative is the set of probability measures μ such that μ is in total variation $T(\mu,\nu)$ at least distance $\delta > 0$ away from ν. For this hypothesis testing problem no uniformly consistent test exists. The proof of the consistency result implies a way to restrict the alternative in order to have uniform consistency, and thus uniform exponential consistency.

1. Introduction

Let $X_1, X_2, \ldots$ be independent and identically distributed realvalued random variables, and consider the problem of testing the simple null hypothesis H_0 that the X_i's are distributed according to some continuous distribution function F_0 with corresponding probability measure ν, versus the composite alternative hypothesis H_1 that the underlying probability measure μ has a total variation distance from ν of at least δ for some $\delta > 0$, i.e.,

$$H_1 = \{\mu : T(\mu,\nu) \geq \delta\} . \tag{1.1}$$

We recall that the total variation $T(\mu,\nu)$ between μ and ν is defined by

$$T(\mu,\nu) = \sup_A |\mu(A) - \nu(A)| , \tag{1.2}$$

where the supremum is taken over *all* measurable sets A. It is well-known that if μ and ν are absolutely continuous with respect to some λ with densities f and g, respectively, then

$$T(\mu,\nu) = \frac{1}{2} \int |f(x) - g(x)| d\lambda(x) .$$

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 631–645.
© 1991 *Kluwer Academic Publishers.*

The above problem corresponds to the more general case of testing the simple null hypothesis H_0^* that ν is arbitrary but fixed, versus the composite alternative hypothesis H_1^* given by

$$\{\mu : d(\mu,\nu) \geq \delta\} ,$$

where $d(\mu,\nu)$ is *some* distance measure. It has been investigated in the literature under which conditions there exist consistent tests for these testing situations. For our purposes we briefly mention the following results in this regard.

Hoeffding and Wolfowitz (1958) showed the positive result that if

$$d(\mu,\nu) = d_\pi(\mu,\nu) = \sum_{A\in\pi} |\mu(A) - \nu(A)|$$

for a *finite* partition π, then uniformly consistent tests for testing H_0^* versus H_1^* do exist. On the other hand, Le Cam (1973) showed (see also Barron (1989)) the negative result that if ν is nondiscrete, then there does *not* exist a uniformly consistent test for testing ν versus $\{\mu : T(\mu,\nu) \geq \delta\}$ for all sufficiently small $\delta > 0$. Le Cam and Schwartz (1960) gave necessary and sufficient conditions for the existence of uniformly consistent tests in a general setting.

Another useful result, obtained by Schwartz (1965) and Le Cam (1973), is that in the case of independent random variables (as we assume here), if the two hypotheses H_0^* and H_1^* are fixed (do not vary with n), then the existence of a uniformly consistent test is equivalent to the existence of a uniformly exponentially consistent (UEC) test.

Since we are assuming from the outset that ν is nondiscrete (F_0 is continuous), we are thus addressing ourselves to a situation where a uniformly consistent test cannot be obtained. It is still of interest, though, to investigate whether reasonably good test procedures can be derived in this case. Indeed, we will show that it is still possible to introduce a test procedure for testing ν versus $\{\mu : T(\mu,\nu) \geq \delta\}$ which is consistent.

Without loss of generality we may assume that, since F_0 is continuous, ν is uniform on [0,1], otherwise we transform each X_i by F_0. Moreover, we can assume that $\mu([0,1]) = 1$, because the result of the transformation by F_0 is within [0,1].

Our test statistic will be based on the L_1-distance between the histogram corresponding to a suitable partition of [0,1], and the uniform density on [0,1]. The advantage of our test procedure is that, besides being relatively easy to carry out, it is consistent against a large class of alternatives. Indeed, the distance measure $T(\mu,\nu)$ corresponds to a fairly strong mode of convergence, and, moreover, the measure μ in (1.1) is completely arbitrary. We now describe our test statistic.

Let $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \ldots, A_{n,m_n}\}$ be a finite partition of [0,1], where the $A_{n,i}$'s are intervals of common length $\nu(A_{n,i}) = h_n$ and $m_n = h_n^{-1}$ is an integer.

Let μ_n stands for the empirical measure for $X_1,\ldots,X_n$, i.e.,

$$\mu_n(A_{n,i}) = \frac{N_{in}(X_1,\ldots,X_n;A_{n,i})}{n} \tag{1.3}$$

with $N_{in}(X_1,\ldots,X_n;A_{n,i})$ being the number of X_j's falling in $A_{n,i}$. We then introduce the following test statistic :

$$J_n = \sum_{i=1}^{m_n} |\mu_n(A_{n,i}) - \nu(A_{n,i})| \ . \tag{1.4}$$

Observe that

$$J_n = \sum_{i=1}^{m_n} \int_{A_{n,i}} \left| \frac{\mu_n(A_{n,i})}{\nu(A_{n,i})} - 1 \right| d\nu(x)$$

$$= \int_0^1 |f_n(x) - g(x)|\, d\nu(x) = 2T(\mu_n,\nu) \tag{1.5}$$

with

$$f_n(x) = \frac{\mu_n(A_{n,i})}{\nu(A_{n,i})} \quad \text{if} \quad x \in A_{n,i} \tag{1.6}$$

and $g(x) = 1$ if $x \in [0,1]$. Thus, J_n is just the L_1-distance between the histogram f_n and the uniform density g. We now turn to the actual hypothesis testing problem for testing H_0 versus H_1 based on this test statistic J_n.

2. The Test Procedure

Consider again the problem of testing H_0 : ν is uniform on [0,1] versus H_1 : $\mu \in \{\mu : T(\mu,\nu) \geq \delta\}$ where $\delta > 0$ is fixed.

First choose the common length $h_n = \nu(A_{n,i})$ of the sets $A_{n,i}$ in the partition $\mathcal{P}_n$ of [0,1] such that

$$\lim_{n\to\infty} h_n = 0 \tag{2.1}$$

and

$$\lim_{n\to\infty} nh_n = \infty . \tag{2.2}$$

Furthermore, choose C* such that

$$0 < C^* < 2\delta .$$

Then, based on the random sample $X_1,\ldots,X_n$, and the selected partition $\mathscr{P}_n$, calculate the test statistic J_n defined in (1.4). Our test procedure for testing H_0 versus H_1 rejects H_0 if

$$J_n \geq C^* , \tag{2.3}$$

and accepts otherwise.

The following theorem analyzes the behavior of this test.
Let

$$\Delta_n = \frac{c}{nh_n} + \left\{\frac{1 - h_n}{nh_n}\frac{2}{\pi}\right\}^{1/2} \tag{2.4}$$

where $c = \dfrac{4\pi c_0}{3\sqrt{3}}$, with c_0 being the constant in the nonuniform version of the Berry–Esseen theorem. Michel (1981) showed that $c_0 \leq 30.54$, hence $c \leq 73.86$. Clearly, (2.1) and (2.2) imply that

$$\lim_{n\to\infty} \Delta_n = 0 .$$

We then have

THEOREM. Under H_0

$$P[J_n \geq C^*] < 2e^{-\frac{n}{8}([C^*-\Delta_n]^+)^2} . \tag{2.5}$$

If $\mu \in H_1$, then

$$\lim_{n\to\infty} P[J_n < C^*] = 0 . \tag{2.6}$$

REMARK 1. It follows from (2.5) that under H_0 the proposed test has a positive type I asymptotic error exponent r_0, satisfying

$$r_0 \triangleq -\limsup_{n\to\infty} \left(\frac{1}{n}\log P[J_n \geq C^* | H_0]\right)$$

$$\geq -\limsup_{n\to\infty} \left(\frac{1}{n}\left(-\frac{n}{8}([C^* - \Delta_n]^+)^2 + \log 2\right)\right)$$

$$= \frac{C^{*2}}{8} > 0 . \tag{2.7}$$

Unfortunately, for testing H_0 versus H_1, there does not exist a uniformly exponentially consistent test, as was shown by Le Cam (1973) and Barron (1989). Therefore, our test cannot have a positive type II asymptotic uniform error exponent, which will be defined by (3.3). From the proof of the Theorem, it will become clear why the convergence of the type II error is not exponentially uniform (or uniform for that matter). In Section 3 we will show how the alternative H_1 can be restricted so that our test becomes uniformly exponentially consistent for testing H_0 against this restricted alternative. In the latter case we will provide a positive lower bound for the type II asymptotic error exponent.

The proof of the Theorem follows from several lemma's.

LEMMA 1. For each $\epsilon > 0$

$$P[|J_n - E(J_n)| > \epsilon] \leq 2e^{-n\epsilon^2/8} , \tag{2.8}$$

regardless whether H_0 or H_1 is true.

Proof. We apply the technique of Devroye (1988). Let $\mathfrak{F}_n$ be the σ-algebra generated by $X_1,\ldots,X_n$, and define

$$Z_{nk} = E(J_n | \mathfrak{F}_k) - E(J_n | \mathfrak{F}_{k-1}) , \tag{2.9}$$

$k = 1,\ldots,n$, with $\mathfrak{F}_0$ being trivial. Then

$$E(Z_{nk} | \mathfrak{F}_{k-1}) = 0 ,$$

so that the sequence $\{(Z_{nk},\mathfrak{F}_k), k = 1,\ldots,n\}$ forms a martingale difference sequence. Moreover,

$$\sum_{k=1}^{n} Z_{nk} = E(J_n|\mathfrak{F}_n) - E(J_n|\mathfrak{F}_0) = J_n - E(J_n) .$$

We will first show that

$$|Z_{nk}| \le \frac{2}{n},$$

and then apply an exponential martingale inequality to obtain (2.8). Put

$$W_{i,m,k} = \frac{1}{n} \sum_{m \le j \le k} (I_{\{X_j \in A_{n,i}\}} - \nu(A_{n,i}))$$

for $1 \le i \le n$, $1 \le m \le k \le n$. Then

$$J_n = \sum_{i=1}^{m_n} |W_{i,1,n}| = \sum_{i=1}^{m_n} |W_{i,1,k-1} + W_{i,k,k} + W_{i,k+1,n}| \tag{2.10}$$

with $W_{i,1,k-1}, W_{i,k,k}, W_{i,k+1,n}$ being independent. We denote the distribution functions of $W_{i,k,k}, W_{i,k+1,n}$, and $W_{i,k,k} + \frac{\nu(A_{n,i})}{n}$ by $F_{1,i}, F_{2,i}$, and $F_{3,i}$, respectively. Then, by (2.9), (2.10), and independence,

$$Z_{nk} = \sum_{i=1}^{m_n} \int |W_{i,1,k-1} + W_{i,k,k} + u| dF_{2,i}(u)$$
$$- \sum_{i=1}^{m_n} \iint |W_{i,1,k-1} + v + u| dF_{1,i}(v) dF_{2,i}(u).$$

Hence

$$|Z_{nk}| \le \sum_{i=1}^{m_n} \iint |W_{i,k,k} - v| dF_{2,i}(u) dF_{1,i}(v) = \sum_{i=1}^{m_n} \int |W_{i,k,k} - v| dF_{1,i}(v)$$

$$= \sum_{i=1}^{m_n} \int \left| W_{i,k,k} + \frac{\nu(A_{n,i})}{n} - w \right| dF_{3,i}(w)$$

$$\leq \sum_{i=1}^{m_n} \int \left[\left| W_{i,k,k} + \frac{\nu(A_{n,i})}{n} \right| + w \right] dF_{3,i}(w) = \frac{2}{n}, \tag{2.11}$$

since

$$\sum_{i=1}^{m_n} \left[W_{i,k,k} + \frac{\nu(A_{n,i})}{n} \right] = \frac{1}{n} \sum_{i=1}^{m_n} I_{\{X_k \in A_{n,i}\}} = \frac{1}{n}$$

and

$$\sum_{i=1}^{m_n} \int w dF_{3,i}(w) = \sum_{i=1}^{m_n} E\left[\frac{1}{n} I_{\{X_k \in A_{n,i}\}} \right] = \frac{1}{n}.$$

Since the Z_{nk}'s are uniformly bounded, we now obtain by an inequality due to Azuma (1967) (see also Stout (1974, pp. 238-239)) that

$$P[|J_n - E(J_n)| > \epsilon] = P\left[\left| \sum_{k=1}^{n} Z_{nk} \right| > \epsilon \right]$$

$$\leq 2e^{-\frac{\epsilon^2}{2n(2/n)^2}} = 2e^{-\frac{n\epsilon^2}{8}},$$

which proves the lemma.

LEMMA 2. Under H_0

$$\left| E(J_n) - \sqrt{\frac{2(1-h_n)}{\pi n h_n}} \right| \leq \frac{c}{n h_n}, \tag{2.12}$$

where c is the constant figuring in (2.4). Hence $E(J_n) \leq \Delta_n$, with Δ_n as defined in (2.4).

Proof. Let

$$S_{n,i} = |\mu_n(A_{n,i}) - \nu(A_{n,i})| ,$$

then $J_n = \sum_{i=1}^{m_n} S_{n,i}$. We rewrite

$$S_{n,i} = \left|\frac{1}{n}\sum_{j=1}^{n} (I_{\{X_j \in A_{n,i}\}} - \nu(A_{n,i}))\right|$$

and denote

$$Y_j = I_{\{X_j \in A_{n,i}\}} - \nu(A_{n,i}) .$$

Then, under H_0, $Y_1,\ldots,Y_n$ are i.i.d. with $E(Y_1) = 0$, $E(Y_1^2) = h_n(1 - h_n)$, and $E(|Y_1|^3) \le E(Y_1^2)$.
By virtue of Lemma 8 of Devroye and Györfi (1985, p. 90), we have that

$$\left|E(S_{n,i}) - \sqrt{\frac{2h_n(1-h_n)}{n\pi}}\right| \le \frac{cE|Y_1|^3}{h_n(1-h_n)n} \le \frac{c}{n}$$

and hence

$$\left|E(J_n) - \sqrt{\frac{2(1-h_n)}{\pi n h_n}}\right| \le \frac{c}{nh_n} ,$$

which proves the lemma.

LEMMA 3. If $h_n \to 0$, $nh_n \to \infty$, and $\mu \in H_1$, then

$$\lim_{n\to\infty} E(J_n) = 2T(\mu,\nu) . \tag{2.13}$$

Proof. Assume that $\mu \in H_1$ is the underlying distribution, μ arbitrary. Then, by Jensen's inequality,

$$E(J_n) = \sum_{i=1}^{m_n} E|\mu_n(A_{n,i}) - \nu(A_{n,i})| \geq \sum_{i=1}^{m_n} |\mu(A_{n,i}) - \nu(A_{n,i})| .$$

On the other hand, by the triangle inequality,

$$J_n \leq \sum_{i=1}^{m_n} |\mu_n(A_{n,i}) - \mu(A_{n,i})| + \sum_{i=1}^{m_n} |\mu(A_{n,i}) - \nu(A_{n,i})| ,$$

so that

$$\sum_{i=1}^{m_n} |\mu(A_{n,i}) - \nu(A_{n,i})| \leq E(J_n) \leq \sum_{i=1}^{m_n} E|\mu_n(A_{n,i}) - \mu(A_{n,i})| + \sum_{i=1}^{m_n} |\mu(A_{n,i}) - \nu(A_{n,i})| .$$

Therefore, Lemma 3 will be proved if we show that

$$\lim_{n\to\infty} \sum_{i=1}^{m_n} |\mu(A_{n,i}) - \nu(A_{n,i})| = 2T(\mu,\nu) \tag{2.14}$$

and

$$\lim_{n\to\infty} \sum_{i=1}^{m_n} E|\mu_n(A_{n,i}) - \mu(A_{n,i})| = 0 . \tag{2.15}$$

We first show (2.15). Indeed,

$$\sum_{i=1}^{m_n} E|\mu_n(A_{n,i}) - \mu(A_{n,i})| \leq \sum_{i=1}^{m_n} \sqrt{E|\mu_n(A_{n,i}) - \mu(A_{n,i})|^2}$$

$$\leq \sum_{i=1}^{m_n} \sqrt{\frac{\mu(A_{n,i})(1-\mu(A_{n,i}))}{n}} \leq \sqrt{\frac{\sum_{i=1}^{m_n} \mu(A_{n,i}) \sum_{i=1}^{m_n} 1}{n}}$$

$$= \frac{1}{\sqrt{nh_n}}, \tag{2.16}$$

so that (2.15) follows from (2.2). In order to prove (2.14) we introduce the notations

$$f_{n1}(x) = \frac{\mu(A_{n,i})}{\mu(A_{n,i}) + \nu(A_{n,i})}, \quad \text{if } x \in A_{n,i}$$

and

$$f_{n2}(x) = \frac{\nu(A_{n,i})}{\mu(A_{n,i}) + \nu(A_{n,i})}, \quad \text{if } x \in A_{n,i}.$$

Clearly,

$$\sum_{i=1}^{m_n} |\mu(A_{n,i}) - \nu(A_{n,i})| = \int |f_{n,1}(x) - f_{n,2}(x)| \, d(\mu + \nu). \tag{2.17}$$

Furthermore,

$$\int |f_1(x) - f_2(x)| \, d(\mu + \nu) - \int |f_{n,1}(x) - f_1(x)| \, d(\mu + \nu)$$

$$- \int |f_{n,2}(x) - f_2(x)| \, d(\mu + \nu) \leq$$

$$\int |f_{n,1}(x) - f_{n,2}(x)| \, d(\mu + \nu) \leq$$

$$\int |f_{n,1}(x) - f_1(x)| \, d(\mu + \nu) + \int |f_{n,2}(x) - f_2(x)| \, d(\mu + \nu)$$

$$+ \int |f_1(x) - f_2(x)| \, d(\mu + \nu),$$

where

$$f_1 = \frac{d\mu}{d(\mu + \nu)} \quad \text{and} \quad f_2 = \frac{d\nu}{d(\mu + \nu)} .$$

Observe that

$$\int |f_1(x) - f_2(x)|\, d(\mu + \nu) = \int \left| \frac{d\mu}{d(\mu+\nu)} - \frac{d\nu}{d(\mu+\nu)} \right| d(\mu + \nu) = 2T(\mu,\nu) . \quad (2.18)$$

From Lebesgue's density theorem (cf. Devroye and Györfi, 1985, Theorem 1, p. 6) and (2.1) we obtain that

$$\lim_{n\to\infty} \int |f_{n,i}(x) - f_i(x)|\, d(\mu + \nu) = 0 \quad , i = 1,2. \quad (2.19)$$

Now (2.14) follows from (2.17), (2.18) and (2.19), which completes the proof.

Proof of the Theorem

Clearly, if $C^* \leq \Delta_n$, then (2.5) holds. Therefore, we may assume in the sequel that $C^* > \Delta_n$. Now applying Lemma 1 and Lemma 2, we obtain that under H_0

$$\begin{aligned} P[J_n \geq C^*] &= P[J_n - E(J_n) \geq C^* - E(J_n)] \\ &\leq P[J_n - E(J_n) \geq C^* - \Delta_n] \\ &= P[|J_n - E(J_n)| \geq [C^* - \Delta_n]^+] \\ &\leq 2e^{-\frac{n}{8}([C^*-\Delta_n]^+)^2} . \end{aligned} \quad (2.20)$$

For proving (2.6), let $\mu \in H_1$ and $0 < \gamma < 1$. Then

$$\begin{aligned} P[J_n < C^*] &= P[J_n - E(J_n) < C^* - E(J_n)] \\ &\leq P[|J_n - E(J_n)| > E(J_n) - C^*] \\ &\leq P[|J_n - E(J_n)| > \gamma(2T(\mu,\nu) - C^*)] + I_{\{E(J_n) - C^* < \gamma(2T(\mu,\nu) - C^*)\}} \\ &\leq P[|J_n - E(J_n)| > \gamma(2\delta - C^*)] + I_{\{2T(\mu,\nu) - E(J_n) > (2\delta - C^*)(1-\gamma)\}} . \end{aligned} \quad (2.21)$$

From Lemma 1 it follows that

$$P[|J_n - E(J_n)| > \gamma(2\delta - C^*)] \leq 2e^{-\frac{n}{8}(\gamma(2\delta - C^*))^2}, \tag{2.22}$$

whereas Lemma 3 implies that

$$\lim_{n\to\infty} I_{\{2T(\mu,\nu)-E(J_n)>(2\delta-C^*)(1-\gamma)\}} = 0 . \tag{2.23}$$

From (2.21), (2.22), and (2.23) it follows that

$$\lim_{n\to\infty} P[J_n < C^*] = 0 . \tag{2.24}$$

REMARK 2. After having given the proof of the Theorem it will be clear why the convergence of the type II error is not uniform. Looking at the proof, we see that the uniformity (in μ) holds in (2.22), but breaks down in (2.23). On its turn, (2.23) is based on Lemma 3, and its proof depends on the limits (2.14) and (2.15). Of these two limits, the convergence in (2.15) is uniform in μ (as can be seen from (2.16)), whereas the convergence in (2.14) is not. In Section 3 we will investigate how we can restrict the alternative H_1 so as to make the convergence uniform.

3. From Consistency to Uniform Consistency

We will now show how, by restricting the alternative, the consistency of our test statistic can be made uniform for all μ in the restricted class. Suppose H_1 is replaced by $H_1' = H_1 \cap K$, where K is a set of probability measures such that on K the convergence in (2.23) is uniform. Then there exists an n_0 such that for $n \geq n_0$ and all $\mu \in H_1'$

$$I_{\{2T(\mu,\nu)-E(J_n)>(2\delta-C^*)(1-\gamma)\}} = 0 . \tag{3.1}$$

Therefore, in this case, for $n \geq n_0$ and all $\mu \in H_1'$, by (2.21) and (2.22)

$$P[J_n < C^*] \leq 2e^{-\frac{n}{8}(\gamma(2\delta-C^*))^2} . \tag{3.2}$$

Hence, defining the type II asymptotic error exponent

$$r_1' \triangleq -\limsup_{n\to\infty} \left(\frac{1}{n}\log(\sup_{\mu\in H_1'} P(J_n < C^*))\right) \tag{3.3}$$

it follows that

$$r_1' \geq \frac{(\gamma(2\delta - C^*))^2}{8}.$$

Since $0 < \gamma < 1$ is arbitrary, we get that

$$r_1' \geq \frac{(2\delta - C^*)^2}{8}. \tag{3.4}$$

This yields the interesting result that if there exists a set K on which the convergence in (2.23) in uniform, then the corresponding type II asymptotic error exponent r_1' is positive, and moreover, there is a positive lower bound on it, which is universal, i.e., is independent of K.

Notice that the minimum of the two lower bounds on the asymptotic error exponents r_0 and r_1' (given by (2.7) and (3.4), respectively) is maximized by $C^* = \delta$, in which case the two bounds become equal to $\delta^2/8$, i.e.,

$$r_0 \geq \frac{\delta^2}{8} \qquad \text{and} \qquad r_1' \geq \frac{\delta^2}{8}.$$

Barron (1989) gave an example of a set K for which r_1' is positive. In his example, K is any set of absolutely continuous measures such that the corresponding density functions are uniformly equicontinuous.

We provide here another example of such a set K. We restrict ourselves, as Barron did, to alternative probability measures μ which have density functions f with respect to the Lebesgue measure ν, but we do not require these densities to be smooth. First, choose any differentiable density function ϕ on [0,1] and let, for $a > 0$,

$$\phi_a(x) = \frac{1}{a}\,\phi\left(\frac{x}{a}\right).$$

Then define

$$V_f = \sup_{a>0} \int |(\phi_a * f)'| \tag{3.5}$$

where $*$ denotes convolution. It can be shown (cf. Devroye and Györfi, 1985, p. 86) that V_f is independent of ϕ. Now, given any constant D, $0 < D < \infty$, define

$$K^{(D)} = \{\mu : V_f < D,\ f = \frac{d\mu}{d\nu}\}. \tag{3.6}$$

We will show that for any D, $0 < D < \infty$, for testing H_0 versus $H_1^{(D)} = H_1 \cap K^{(D)}$, our test described by (2.3) is uniformly exponentially consistent. Let g be the uniform density on [0,1] and, for a given partition $\mathcal{P}_n$ with $\nu(A_{n,i}) = h_n$, let

$$f_n(x) = \frac{\mu_n(A_{n,i})}{\nu(A_{n,i})} \quad \text{if} \quad x \in A_{n,i} ,$$

as in (1.6). Then

$$Ef_n(x) = \frac{\mu(A_{n,i})}{\nu(A_{n,i})} \qquad \text{if} \quad x \in A_{n,i} ,$$

and

$$\sum_{i=1}^{m_n} |\nu(A_{n,i}) - \mu(A_{n,i})| = \int_0^1 |g(x) - Ef_n(x)| d\nu(x) . \tag{3.8}$$

We recall from Remark 2 that it suffices to show that (3.8) converges to $2T(\mu,\nu)$ uniformly in $\mu \in H_1^{(D)}$ as $n \to \infty$. Obviously,

$$\int |f(x) - g(x)| d\nu(x) - \int |f(x) - Ef_n(x)| d\nu(x) \le \int |g(x) - Ef_n(x)| d\nu(x)$$

$$\le \int |f(x) - g(x)| d\nu(x) + \int |f(x) - Ef_n(x)| d\nu(x)$$

and

$$\int |f(x) - g(x)| d\nu(x) = 2T(\mu,\nu) .$$

Therefore, the uniform convergence of (3.8) to $2T(\mu,\nu)$ is proved if the convergence of

$$\int |f(x) - Ef_n(x)| d\nu(x) \tag{3.9}$$

to zero is uniform. However, (3.9) represents the bias of the histogram. Similarly to Lemma 21, p. 118 of Devroye and Györfi (1985), it can be shown that

$$\int |f(x) - Ef_n(x)| d\nu(x) \le h_n V_f . \tag{3.10}$$

Hence, for any $\mu \in H_1^{(D)}$ we have

$$\int |f(x) - Ef_n(x)| d\nu(x) \leq h_n D . \tag{3.10}$$

The uniform convergence to zero now follows from (2.1).

This technique provides an example, different from the one given by Barron, of a class of alternatives which are in total variation at least distance δ away from H_0, against which a uniformly consistent test exists. Indeed, the densities of measures in $H_1^{(D)}$ are not necessarily continuous, in contrast to Barron's example.

4. References

Azuma, K. (1967), "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal*, vol. 37, pp. 357-367.

Barron, A.R. (1989), "Uniformly powerful goodness of fit tests", *Annals of Statistics*, vol. 17, pp. 107-124.

Devroye, L. (1988), "The kernel estimate is relatively stable", *Probability Theory and Related Fields*, vol. 77, pp. 521-536.

Devroye, L. and Györfi, L. (1985), *Nonparametric Density Estimation : The L_1 View*, Wiley, New York.

Hoeffding, W. and Wolfowitz, J. (1958), "Distinguishability of sets of distributions", *Annals of Mathematical Statistics*, 29, pp. 700-718.

Le Cam, L. (1973), "Convergence of estimates under dimensionality restrictions", *Annals of Statistics*, vol. 1, pp. 38-53.

Le Cam, L. and Schwartz, L. (1960), "A necessary and sufficient condition for the existence of consistent estimates", *Annals of Mathematical Statistics*, 31, pp. 40-150.

Michel, R. (1981), "On the constant in the nonuniform version of the Berry-Esseen theorem", *Z. Wahrsch. verw. Gebiete*, 55, pp. 109-117.

Schwartz, L. (1965), "On Bayes' procedures", *Z. Wahrsch. verw. Gebiete*, 4, 10-26.

Stout, W.D. (1974), *Almost Sure Convergence*, Academic Press, New York.

ON A PROBLEM IN SEMIPARAMETRIC ESTIMATION

S. R. JAMMALAMADAKA and X. WAN
Department of Statistics and Applied Probability
University of California
Santa Barbara, CA 93106 U.S.A.

ABSTRACT. The estimation problem in a semiparametric model , namely, the generalized Lehmann alternative model, is considered here. Suppose that two independent samples $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ with d.f.'s F and G, respectively, are observed. Assume that $G(\cdot)=H(F(\cdot);\theta)$, where the form of the function H is known, but F and the parameter θ are unknown. The problem is to estimate θ in the presence of the nuisance function F. We give two methods of estimating the parametric component of the model based on sample quantiles and the Mann-Whitney statistic. The asymptotic variances of these estimators are compared.

1. Introduction

Semiparametric models have become an active research topic in recent years. In such a model, there exist both parametric and nonparametric components. For example, Begun, Hall, Huang and Wellner(1983) study the model in which $X_1,\ldots,X_n$ are i.i.d. with density $f=f(\cdot;\theta,g)$ with respect to Lebesgue measure μ on the real line, where θ is a real number and g belongs to a class of densities sufficiently small that θ is identifiable. There is considerable literature on the well known Lehmann alternatives(Lehmann(1953)). See, for example, Young(1973), Brooks(1974) and Savage(1980).

Fukui and Miura(1988) consider the estimation problem in the following two sample semiparametric model. Let $X_1,\ldots,X_m$ be i.i.d. random variables(r.v.'s) with distribution function(d.f.) F and let $Y_1,\ldots,Y_n$ be i.i.d. r.v.'s with d.f. $G(\cdot) = H(F(\cdot);\theta)$, where $\{H(x;\theta)\colon \theta\in(a,b)\}$ is a known family of d.f.'s on [0,1], while the true value of the parameter θ as well as the function F are unknown. It is desired to estimate θ in the presence of the nuisance function F. This semiparametric model is known as the generalized Lehmann alternative model since Lehmann alternative is a special case of this corresponding to $H(x;\theta) = x^{\theta}$. We now briefly review the work of Fukui and Miura (1988):

Since $G(x) = H(F(x);\theta)$, then $F(x) = H^{-1}(G(x);\theta)$. Let $D_{mn}(\theta)$ denote the Kolmogorov-Smirnov distance

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 647–656.
© 1991 *Kluwer Academic Publishers.*

$$D_{mn}(\theta) = \sup_x \left| F_m(x) - H^{-1}(G_n(x);\theta) \right|$$

where $F_m(x)$ and $G_n(x)$ are the empirical d.f.'s of F and G, respectively. The minimum-distance estimator of θ is defined as the value $\hat{\theta}$ of θ, which minimizes $D_{mn}(\theta)$. Fukui and Miura give the asymptotic distribution of this $\hat{\theta}$.

Theorem (Fukui and Miura). If $m+n = N$ tends to infinity such that $m/N \to \lambda$, where $0< \lambda< 1$, then

$$P\{ N^{1/2}(\hat{\theta} - \theta_0) \leq y \}$$

$$\to P\{ \sup_t [\frac{B_1(t)}{\lambda^{1/2}} - \frac{B_2(H(t;\theta_0))}{(1-\lambda)^{1/2} h_t(t;\theta_0)} + \frac{h_\theta(t;\theta_0)}{h_t(t;\theta_0)} \times y]$$

$$+ \inf_t [\frac{B_1(t)}{\lambda^{1/2}} - \frac{B_2(H(t;\theta_0))}{(1-\lambda)^{1/2} h_t(t;\theta_0)} + \frac{h_\theta(t;\theta_0)}{h_t(t;\theta_0)} \times y] \geq 0 \}$$

where

$$m^{1/2}[F_m(x) - F(x)] \to B_1(F(x)), \qquad n^{1/2}[G_n(x) - G(x)] \to B_2(G(x)),$$

$B_1(t)$ and $B_2(t)$ are independent Brownian bridges, $h_t(t;\theta)$ and $h_\theta(t;\theta)$ are the partial derivatives of $H(t;\theta)$ with respect to t and θ, respectively. #

This limiting distribution is very complicated and does not seem to have a normal approximation. Although Fukui and Miura give an algorithm to compute this $\hat{\theta}$, it seems that a great deal of computation needs to be done before $\hat{\theta}$ can finally be found.

Our aim in this paper is to develop some methods which will provide estimators of θ involving less computational work while providing consistent estimation with limiting normal distributions.

These results can also be applied to test the hypothesis H_0: $\theta=\theta_0$. Notice that in many cases there exists a $\theta_0 \epsilon \Theta$, such that $H(x;\theta_0) = x$. Thus $G(x) = H(F(x);\theta_0) = F(x)$, so that it covers the usual two sample problem of testing H_0': F=G as a special case of testing H_0: $\theta=\theta_0$.

Although our results focus on a scalar θ, some results can be readily extended to the case when θ is a vector. The following examples illustrate some applications of the generalized Lehmann alternative models.

Example 1.1. The idea of proportional hazards was introduced by Cox(1972). If

$X_1,\ldots,X_n$ are independent r.v.'s with X_i having d.f. $G(x)=1-[1-F(x)]^{\theta_i}$, $i=1,\ldots,n$, then the ratio $r(x,\theta_i)/\ r(x,\theta_j)$ does not depend on x, where $r(x,\theta) = G'(x)/\ [1-G(x)]$ is the hazard function. This condition often appears to be at least approximately satisfied in many biological applications.

Example 1.2. Consider a mixture problem in which the d.f. of the observations is a linear combination of two d.f.'s $F(x)$ and $G(F(x))$, i.e.,

$$H(F(x);\theta) = (1-\theta)\ G(F(x)) + \theta\, F(x), \qquad 0\leq \theta\leq 1.$$

Assume that F is unknown but $G(\cdot)$ is known. It is often important to estimate the mixing parameter θ. If G also depends on an unknown parameter, then this is an example of a model with vector parameter.

This paper is arranged as follows. Section 2 contains results about an estimator $\hat{\theta}$ of θ obtained by matching the pth quantiles of the two samples. Section 3 deals with an estimator θ^* based on the Mann-Whitney statistic. In section 4, comparison is made between $\hat{\theta}$ and θ^* through examples. We will make the following assumptions throughout:

Assumption A. $F(x)$ is an absolutely continuous d.f. on R^1. $H(x;\theta)$ is a d.f. on $[0,1]$ for every $\theta=(\theta_1,\ldots,\theta_k)\epsilon\Theta \subset R^k$ such that

$$\frac{\partial H}{\partial x} > 0 \ \text{ and } \frac{\partial H}{\partial \theta_i} \neq 0.$$

In the case when θ is a real number, we may assume without loss of generality that $\frac{\partial H}{\partial \theta} < 0$ so that $h(x;\theta)$ is a strictly decreasing function of θ. Furthermore, assume that $\Theta = (a,b)$, where a and/or b may be infinity. Assume that $H(x;a+)=1$ and $H(x;b-)=0$ for any x, $0< x <1$. Assume that there exists a λ, $0< \lambda< 1$, such that $m/N \to\lambda$ as $m\to\infty$, $n\to\infty$, where $N=m+n$ is the total of the two sample sizes.

2. Estimation Based On Sample Quantiles

Let $X_1'<\cdots< X_m'$ and $Y_1'< \cdots < Y_n'$ be the order statistics of the X and Y samples, respectively. For a fixed positive number p, $0< p< 1$, the sample pth quantile of the Y sample is defined as $Y'_{[np]+1}$. Let ξ_p denote the population pth quantile of Y, then $G(\xi_p) = H(F(\xi_p);\theta) = p$. Substituting estimators $Y'_{[np]+1}$ for ξ_p, and the empirical d.f. F_m for F, we may write, approximately, $H(F_m(Y'_{[np]+1});\theta) = p$. Let X'_k be the largest observation in the X sample such

that $X'_k \le Y'_{[np]+1}$. When N=m+n is large, we would expect $F_m(Y'_{[np]+1})$ to be very close to $F_m(X'_k) = k/m$. Hence, our estimator $\hat{\theta}$ is the solution to the equation

$$H(k/m;\theta) = p \tag{2.1}$$

By Assumption A, $\frac{\partial H}{\partial \theta} < 0$ and $H(x;a+)=1$, $H(x;b-)=0$. This guarantees the existence and uniqueness of $\hat{\theta}$. For brevity we write Y_n^* for $Y'_{[np]+1}$. Let $\zeta_p = F(\xi_p)$ and let $H_2^{-1}(z,p)$ be the solution to the equation $H(z;\theta)=p$ with respect to θ for all z and p, such that $0< z< 1$, $0< p< 1$. Let $h_1(z;\theta)$ and $h_2(z;\theta)$ denote the partial derivatives of $H(z;\theta)$ with respect to z and θ, respectively. Thus $\hat{\theta} = H_2^{-1}(F_m(Y_n^*),p)$ and

$$N^{1/2}(\hat{\theta}-\theta_0) = N^{1/2}[H_2^{-1}(F_m(Y_n^*),p) - H_2^{-1}(\zeta_p,p)]$$

$$= N^{1/2}\frac{\partial}{\partial x}H_2^{-1}(x,p)\Big|_{x=\zeta_p} \cdot [F_m(Y_n^*) - \zeta_p] + o_p(1)$$

$$= -\left(\frac{N}{m}\right)^{1/2}\frac{h_1(\zeta_p;\theta_0)}{h_2(\zeta_p;\theta_0)} \cdot m^{-1/2}\sum_{i=1}^{m}[I(X_i \le Y_n^*) - \zeta_p] + o_p(1). \tag{2.2}$$

The last equality indicates that it suffices to find the limiting distribution of

$$A_N = m^{-1/2}\sum_{i=1}^{m}[I(X_i \le Y_n^*) - \zeta_p].$$

Applying the probability integral transformation F on both samples, $X_1,\ldots,X_m$ become $U_1,\ldots,U_m$, an i.i.d. sample from the uniform distribution on [0,1], and $Y_1,\ldots,Y_n$ become $Z_1,\ldots,Z_n$, an i.i.d. sample with d.f. $H(z;\theta)$. Although the values of $U_1,\ldots,U_m$ and $Z_1,\ldots,Z_n$ are unknown because F is unknown, the order relations in the combined samples remain the same since F is strictly increasing. Thus A_N can also be written as

$$A_N = m^{-1/2}\sum_{i=1}^{m}[I(U_i \le Z_n^*) - \zeta_p]$$

where $Z_n^* = Z'_{[np]+1}$.

Lemma 1. Let $\{X_i, i=1,\ldots,n\}$ be a sequence of continuous r.v.'s with continuous probability density function f(x). Let μ_p denote the population pth quantile and X_n^* the sample pth quantile. Suppose that μ_p is unique with $f(\mu_p)$ positive. Then the sequence of the random variables $n^{1/2}(X_n^*-\mu_p)$ has asymptotically a normal distribution with mean zero and variance τ^2, where

$$\tau^2 = \frac{p(1-p)}{[f(\mu_p)]^2} . \qquad \#$$

Lemma 2. Under Assumption A, if θ is the true value of the parameter, then A_N has a limiting normal distribution with mean zero and variance τ_0^2, where

$$\tau_0^2 = \zeta_p - \zeta_p^2 + \frac{\lambda\, p(1-p)}{(1-\lambda)[h_1(\zeta_p;\theta)]^2} .$$

Proof: Define

$$B_N = m^{-1/2} \sum_{i=1}^{m} [I(U_i \leq \zeta_p) - \zeta_p] + m^{1/2}(Z_n^* - \zeta_p).$$

Since the first term of B_N is the sums of i.i.d. r.v.'s, by the Lindeberg-Levy central limit theorem, it has an asymptotic normal distribution with mean zero and variance $\tau_1^2=\zeta_p(1-\zeta_p)$. By Lemma 1, the second term of B_N has a limiting normal distribution with mean zero and variance

$$\tau_2^2 = \frac{\lambda\, p(1-p)}{(1-\lambda)[h_1(\zeta_p;\theta)]^2} .$$

Since U_i and Z_n^* are independent, the two limiting normal distributions are also independent of each other. Hence B_N has a limiting normal distribution with mean zero and variance τ_0^2. It can be verified that $E(A_N-B_N)^2 \to 0$. So that A_N and B_N have the same asymptotic distribution. #

Now from Lemma 2 and (2.2), we obtain

Theorem 1. Assume θ is the true value of the parameter, then under Assumption A, the asymptotic distribution of the sequence of random variables $N^{1/2}(\hat{\theta}-\theta)$ has

a normal distribution with mean zero and variance σ_0^2, where

$$\sigma_0^2 = [h_2(\zeta_p;\theta)]^{-2} \left[\frac{h_1^2(\zeta_p;\theta)\,\zeta_p(1-\zeta_p)}{\lambda} + \frac{p(1-p)}{1-\lambda} \right]. \tag{2.3}$$

If $\hat{\zeta}_p$ is the pth quantile of H with this $\hat{\theta}$ substituted for θ, i.e., $\hat{\zeta}_p$ is the solution of

$$p = H(\hat{\zeta}_p;\hat{\theta})$$

and

$$\hat{\sigma}^2 = [\, h_2(\hat{\zeta}_p;\hat{\theta})\,]^{-2} \left[\frac{h_1^2(\hat{\zeta}_p;\hat{\theta})\,\hat{\zeta}_p(1-\hat{\zeta}_p)}{\lambda} + \frac{p(1-p)}{1-\lambda} \right].$$

then it is easily verified that $N^{1/2}(\hat{\theta}-\theta)/\hat{\sigma}$ is asymptotically N(0,1). One can also use this result to obtain asymptotic confidence limits for the unknown θ.

this result can also be used to test the hypothesis H: $\theta=\theta_0$. To improve the performance of the test statistic, one should minimize σ_0^2. (2.3) indicates that σ_0^2 depends on the quantities p, θ_0, λ and the function $H(x;\theta_0)$. All those quantities except p have been specified at the time of testing. However, we can choose an optimal value p_0 so that σ_0^2 is minimized for the given values θ_0, λ and the function $H(x;\theta_0)$. Thus the optimal estimator and the optimal test is to use this p_0 in the quantile matching method.

Example 2.1. Suppose that $H(x;\theta) = x^{\theta}$. Solving the equation

$$[F_m(Y_n^*)]^{\theta} = p$$

we get

$$\hat{\theta} = [\ln p]/\ln[F_m(Y_n^*)].$$

According to Theorem 1, $\mathrm{Var}(\hat{\theta}) = N^{-1}\sigma_0^2 + o(N^{-1})$. To compute σ_0^2, we find

$$h_1(x;\theta) = \theta\, x^{\theta-1}, \qquad h_2(x;\theta) = x^{\theta}\ln x, \qquad \mu_p = p^{1/\theta}.$$

Substituting these into (2.3), we have after some simplification

$$\sigma_0^2 = \frac{\theta^4}{\lambda(\ln p)^2}\left[\, p^{-1/\theta} - 1 + \frac{\lambda(1-p)}{(1-\lambda)p\theta^2} \right]. \tag{2.4}$$

3. Estimation Based on the Mann-Whitney Statistic

In section 2, the estimator $\hat{\theta}$ is developed based on the statistic involving $F_m(Y_n^*)$, the value of the empirical d.f. F_m at Y_n^*. This motivates us to consider the statistic

$$W_{xy} = \frac{1}{n} \sum_{i=1}^{n} F_m(Y_i),$$

the average of the empirical d.f. F_m at all the n Y observations. By the definition of the empirical d.f. F_m, W_{xy} can be written as

$$W_{xy} = \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} I(X_j \leq Y_i)$$

where $I(\cdot)$ is the indicator function. Notice that $W_{xy} = 1 - W_{yx}$, where

$$W_{yx} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} I(Y_i < X_j). \tag{3.1}$$

We will study the statistic W_{yx} which is known as the Mann-Whitney statistic, a special case of a U-statistic. If we apply the probability integral transformation F on both X and Y samples, W_{yx} can be expressed as

$$W_{yx} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} I(Z_i < U_j). \tag{3.2}$$

Let

$$M_1(\theta) = E_\theta[\, I(Z_1 < U_1)\,] = \int_0^1 H(u;\theta)\, du \tag{3.3}$$

be the expectation of W_{yx}. Under Assumption A, it is clear that $M_1(\theta)$ is a differentiable function. Since W_{yx} is a consistent and asymptotically normal(CAN) estimator for $M_1(\theta)$, we propose an estimator θ^* of θ as

$$\theta^* = M_1^{-1}(W_{yx}). \tag{3.4}$$

As in section 2, we will study the asymptotic behavior of the random variables $N^{1/2}(\theta^* - \theta)$ when the true value of the parameter is θ. Using the mean value

theorem, we have, from (3.4),

$$N^{1/2}(\theta^*-\theta) = [M_1'(\theta)]^{-1} N^{1/2}[W_{yx} - M_1(\theta)] + o_p(1). \tag{3.5}$$

Define $M_2(\theta) = P_\theta(Z_1 < U_1, Z_1 < U_2)$, $M_3(\theta) = P_\theta(Z_1 < U_1, Z_2 < U_1)$. Using conditional probabilities, $M_2(\theta)$ and $M_3(\theta)$ can be expressed as

$$M_2(\theta) = 2\int_0^1 (1-u)\, H(u;\theta)\, du\ , \tag{3.6}$$

$$M_3(\theta) = \int_0^1 H^2(u;\theta)\, du\ . \tag{3.7}$$

Theorem 2. Suppose that Assumption A holds. If θ is the true value of the parameter, then the random sequence $N^{1/2}(\theta^*-\theta)$ has an asymptotic normal distribution with mean zero and variance σ_1^2, where

$$\sigma_1^2 = [M_1'(\theta)]^{-2}\, \tau_1^2$$

$$\tau_1^2 = \frac{1}{\lambda} M_3(\theta) + \frac{1}{1-\lambda} M_2(\theta) - \frac{1}{\lambda(1-\lambda)} M_1^2(\theta). \tag{3.8}$$

Proof: From the theory of U-statistics (see for example, Lehmann(1975)), using expressions (3.3), (3.6) and (3.7) we have

$$\mathrm{Var}(W_{yx}) = \frac{1}{nm}\{\, M_1(\theta)[1-M_1(\theta)] + (n-1)[M_3(\theta)-M_1^2(\theta)]$$

$$+ (m-1)[M_2(\theta)-M_1^2(\theta)]\,\}.$$

Notice that

$$\frac{m}{N} \to \lambda,\ \frac{n}{N} \to 1-\lambda,\ \text{as } N\to\infty,$$

so

$$\mathrm{Var}\,[N^{1/2}(W_{yx} - M_1(\theta)] \to \tau_1^2\ .$$

It is well known that the limiting distribution of the statistic $N^{1/2}[W_{yx} - M_1(\theta)]$ is normal with mean zero and variance τ_1^2. Hence by (3.5), the limiting distribution of $N^{1/2}(\theta^*-\theta_0)$ is $N(0,\sigma_1^2)$. #

Example 3.1. Suppose that $H(x;\theta) = x^{\theta}$. Then

$$M_1(\theta)=\frac{1}{\theta+1}, \qquad M_2(\theta)=\frac{2}{(\theta+1)(\theta+2)}, \qquad M_3(\theta)=\frac{1}{2\theta+1}.$$

Hence

$$\theta^* = [W_{yx}]^{-1} - 1 \text{ and } \mathrm{Var}(\theta^*) = N^{-1}\sigma_1^2 + o(N^{-1})$$

with

$$\sigma_1^2 = (\theta+1)^4 \left[\frac{1}{\lambda(2\theta+1)} + \frac{2}{(1-\lambda)(\theta+1)(\theta+2)} - \frac{1}{\lambda(1-\lambda)(\theta+1)^2} \right]. \tag{3.9}$$

4. The Comparison between $\hat{\theta}$ and θ^*.

Between the two estimators $\hat{\theta}$ and θ^*, we should prefer the one with smaller asymptotic variance. In Theorems 1 and 2, we have given the asymptotic variances of $\hat{\theta}$ and θ^*, denoted by σ_0^2 and σ_1^2, respectively. Furthermore, in Examples 2.1 and 3.1, we derived estimators $\hat{\theta}$ and θ^* when $H(x;\theta) = x^{\theta}$. The specific expressions of σ_0^2 and σ_1^2 are given in (2.4) and (3.9). For a selected number of values of θ and λ, we compute σ_0^2 and σ_1^2. Here σ_0^2 is the minimum variance corresponding to the optimal value of p. The results are listed in Table 4.1.

From Table 4.1, it is seen that for small values of θ, the estimator $\hat{\theta}$ is superior to θ^*, while θ^* is better for other values of θ. In other words, there is no overall winner. In particular, if we test the two sample problem H_0: F=G by using this model, i.e., testing H_0: $\theta=1$, the estimator θ^* is preferred.

TABLE 4.1 Asymptotic variances of $\hat{\theta}$ and θ^* for the Lehmann alternative model $H(x;\theta) = x^{\theta}$.

	$\lambda = 1/3$	$\lambda = 1/2$	$\lambda = 2/3$
θ=0.1	σ_0^2=0.121	σ_0^2=0.120	σ_0^2=0.141
	σ_1^2=0.117	σ_1^2=0.135	σ_1^2=0.188
θ=0.5	σ_0^2=1.840	σ_0^2=1.655	σ_0^2=1.860
	σ_1^2=1.519	σ_1^2=1.463	σ_1^2=1.772
θ=1	σ_0^2=6.949	σ_0^2=6.177	σ_0^2=6.949
	σ_1^2=6.000	σ_1^2=5.333	σ_1^2=6.000
θ=1.5	σ_0^2=16.01	σ_0^2=14.24	σ_0^2=15.96
	σ_1^2=14.57	σ_1^2=12.39	σ_1^2=13.31

5. REFERENCES

Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A.(1983). 'Information and asymptotic efficiency in parametric-nonparametric models', Ann. Statist. **11**, pp. 432 - 452.

Brooks, R. J.(1974). 'Bayesian analysis of the two-sample problem under the Lehmann alternatives', Biometrika, **61**, pp. 501 - 507.

Cox, D. R.(1972). 'Regression models and life tables', J. Roy. Statist. Soc. **B34**, pp. 187 - 220.

Fukui, M. and Miura, R.(1988). 'Kolmogrov-Smirnov estimation for the generalized Lehmann alternative models: two sample problem', in K. Matusita(ed.), Statistical Theory and Data Analysis II, Elsevier Science Publisher B. V. (North-Holland), pp. 101 - 110.

Lehmann, E. L.(1953). 'The power of rank tests', Ann. Math. Statist. **24**, pp. 23 - 43.

Lehmann, E. L.(1975). 'Nonparametrcs: statistical methods based on ranks', Holden-Day, San Francisco.

Savage, I. R.(1980). 'Lehmann alternatives', in B. V. Gnedenko, M. L. Puri and I. Vincze (eds.), Nonparametric Statistical inference II, North-Holland Publishing Company, pp. 795 - 821.

Young, D. H.(1973). 'Distributions of some censored rank statistics under Lehmann alternatives for the two sample case', Biometrika, **60**, pp. 543 - 549.

ON THE INTEGRABLE AND APPROXIMATELY INTEGRABLE LINEAR STATISTICAL MODELS

B. YA. LEVIT
University of Maryland
College Park Campus
Department of Mathematics
College Park, Maryland 20742 USA

ABSTRACT. The notion of integrable nonparametric models will be introduced. For integrable models the second order optimality theory is, to a large extent, similar to that relevant in parameter estimation theory. In particular the conditions for second order admissibility of the sample distribution function (in terms of order n^{-2} of its covariance operator) are found.

Typically, however the nonparametric models may well occur to be only approximately integrable (a.i.). In that case any rate of improvement of the classical nonparametric procedures should be anticipated ranging from $(\log n)^{\alpha} n^{-2}$ to $(\log n)^{-\alpha} n^{-1}$, $\alpha > 0$.

Examples of a.i. models include estimation of the unknown distribution function as well as of its density and higher order derivatives. Thus the a.i. models suggest an useful approach for a unified treatment of regular and irregular nonparametric problems.

1. Introduction

The aim of this presentation is to outline a portion of the "next" order optimality theory in nonparametric estimation which may be tentatively entitled as "Linear functionals". The theory appears to be sufficiently comprehensive to make it sensible to keep the presentation at the least technical level restricted so far to the (matrix or operator valued) quadratic losses.

Even within this restrictive setup the emphasis will be lodged mainly not with deriving the most complete results, but rather with exposing two opposite types of models. With the first ones called "Linear integrable models," the second order optimality theory turns out to be, to a large extent, similar to that available in the parameter estimation theory, Levit (1982, 1985). As an evidence of that in the setting of n i.i.d. observations the next term of the risk expansion for the asymptotically optimal estimators is typically of order n^{-2}, i.e., same as in the case of smooth parameter estimation models.

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 657–667.
© 1991 *Kluwer Academic Publishers.*

However, one may argue that in the nonparametric context this is far from being a common case. Thus a more general class of so-called "approximately integrable models" will be introduced restricted, for the sake of exposure, even further to real valued functionals.

A useful lower bound for a local minimal risk in estimating such functions will be stated. Based on this lower bound it can be demonstrated that with the approximately integrable models any rate of the "next" order improvement of the asymptotically optimal estimators may be anticipated ranging from $(\log n)^{\alpha} n^{-2}$ to $(\log\cdots\log n)^{-\alpha} n^{-1}$, for arbitrary $\alpha > 0$. Clearly when this is the case the next order improvement may well challenge the asymptotic optimality property of a given first order efficient estimator.

Another enlightening point is that approximately integrable models appear to be well tailored to incorporate both regular (such as estimating cdf) and irregular (such as estimation of the derivatives of cdf) nonparametric problems. With this approach one discovers then a close relation between the optimal rates of improvement for the standard estimators of the regular functionals and the rates of estimability for the irregular ones.

2. Some Preliminaries and Definitions

Let $X_1,\cdots,X_n$ be an independent sample in a measurable space $(\mathcal{X},\mathcal{H})$ with a common distribution F being an element of a topological space $(\mathcal{F},\mathcal{T})$, where $\mathcal{F}$ is a given set of distributions on $(\mathcal{X},\mathcal{H})$ and $\mathcal{T}$ the topology on $\mathcal{F}$ induced by the distance in variation. The goal is to estimate an unknown value of a given function $\Psi(F) = \Psi(F,z)$ being an element of the Hilbert space $H = H_{\mu}(Z)$ of square integrable functions w.r.t. given σ-finite measure μ on a set Z.

We assume $\Psi(F)$ to be linear in F: $\Psi(F,z) = \int\psi(x,z)dF(x)$, where $\psi(x,\cdot) \in H$ for almost all x with respect to F, $F \in \mathcal{F}$.

For an arbitrary estimator $\Psi_n(z)$ consider the risk function $R_n(\Psi_n,F)$ being the covariance operator of the random field $\Psi_n(z) - \Psi(F,z)$: $R_n(\Psi_n,F) = \mathrm{Cov}_F(\Psi_n(z) - \Psi(F,z))$. With R_n considered as an operator from $H^* = H$ to H denote by $\|R_n\|$ the H-space derived norm of R_n.

DEFINITION 2.1. a) The function $\Psi(F)$ is called $\rho(n)$-rate estimable on $\mathcal{F}$ if there is an estimator Ψ_n such that locally uniformly in F

$$\|R_n(\Psi_n,F)\| = \mathcal{O}(\rho(n)), \qquad (n \to \infty); \tag{2.1}$$

b) at best $\rho(n)$-rate estimable on $\mathcal{F}$ if for any sequence $\rho'(n) = o(\rho(n))$, $n \to \infty$, and any nonempty vicinity $V \subset \mathcal{F}$ it is not $\rho'(n)$-rate estimable on V;

c) exactly $\rho(n)$-rate estimable on $\mathcal{F}$ if it satisfies both a) and b).

Next definition refers to the rates at which the risk of a given estimator may be reduced.

DEFINITION 2.2. An estimator Ψ_n is called,

a) (locally) $\rho(n)$-rate unimprovable, or admissible if for any nonzero covariance operator R and any nonempty vicinity $V \subset \mathcal{F}$ there does not exist an estimator Ψ'_n such that locally uniformly in $F \in \mathcal{F}$ (in $F \in V$),

$$R_n(\Psi'_n,F) \le R_n(\Psi_n,F) - \rho(n)\, R\mathbf{1}(F \in V) + o(\rho(n)), \quad n \to \infty$$

in the usual sense of positively definite operators.

Otherwise Ψ_n is called (locally) $\rho(n)$-rate improvable or inadmissible;

b) at best $\rho(n)$-rate improvable if it is locally unimprovable for any sequence $\rho'(n)$ such that $\rho(n) = o(\rho'(n))$, $n \to \infty$;

c) exactly $\rho(n)$-rate improvable if it is both $\rho(n)$-rate improvable and at best $\rho(n)$-rate improvable;

d) first order asymptotically optimal if it is $\rho(n)$-rate consistent, i.e., satisfies (2.1), and $\rho(n)$-rate locally admissible.

Let F_n be empirical cdf and $\hat{\psi}_n = \int \psi(x,z)dF_n(x)$. The following result readily follows from Levit (1978).

THEOREM 2.1. Assume that the following conditions are met:

1) the integral, $\iint \psi^2(x,z)dF(x)d\mu(z)$ converges locally uniformly in F;

2) for any $\ell \in H^*$ and every F ranging in an everywhere dense subset of $\mathcal{F}$ there exist a sequence of real valued functions $\psi_k(x)$, $x \in \mathcal{X}$, and positive numbers a_k such that

$$\lim_{k\to\infty} \int |\psi_k(x) - \ell\circ(\psi(x,\cdot) - \Psi(F,\cdot))|^2\, dF = 0$$

and the family $G_{k,c}$ of distributions defined by the relations

$$\frac{dG_{k,c}}{dF}(x) = \exp\{c\, \psi_k(x) - b(c)\}, \quad |c| < a_k$$

$$b(c) = \ln \int \exp\{c\, \psi_k(x)\}dF(x),$$

belongs to $\mathcal{F}$.

Then $\hat{\psi}_n$ is n^{-1}-rate consistent first order asymptotically optimal estimator of $\Psi(F)$.

Our nearest goal is to analyze higher order asymptotically optimal properties of Ψ_n.

3. Completely Integrable Linear Models

DEFINITION 3.1. $_*$Let us call the model $(\mathcal{F},\Psi(F))$ completely integrable if for any $\ell \in H^*$ and every F ranging in an everywhere dense subset of $\mathcal{F}$ the distribution G defined by the relations

$$\frac{dG}{dF}(x) = \exp\{\ell\circ\psi(x,\cdot)-b(\ell)\},$$

$$b(\ell) = \log\int\exp\{\ell\circ\psi(x,\cdot)\}dF(x),$$

belongs to $\mathcal{F}$.

THEOREM 3.1. Let the model $(\mathcal{F},\Psi(F))$ be completely integrable. Then the estimator $\hat{\psi}_n$ is n^{-2}-rate admissible.

EXAMPLE 3.1. Let $\mathcal{X} = R^k$, $k \geq 1$, $\Psi(F,x) = F(x)$ be the unknown distribution function. Assume that the space $(\mathcal{F},\mathcal{T})$ of distributions contains an everywhere dense subset of all Hölder continuous distribution functions with an exponent α, $0 \leq \alpha \leq 1$ (of all continuous cdf for $\alpha = 0$). Let $H = H_\mu(R^k)$ for some finite measure μ on R^k. The model is easily seen to be completely integrable resulting in the following.

PROPOSITION 3.1. The empirical cdf $F_n(x)$ is n^{-2}-rate admissible.

4. Integrable Linear Models

DEFINITION 4.1. The model $_*(\mathcal{F},\Psi(F))$ is called integrable or boundedly integrable if for any $\ell \in H^*$ and every F ranging in an everywhere dense subset of $\mathcal{F}$ there exists $a > 0$ such that the family distributions G_c defined by the relations

$$\frac{dG_c}{dF}(x) = \exp\{c\ \ell\circ\psi(x,\cdot)-b(c)\}, \quad |c| < a,$$

$$b(c) = \ln\int\exp\{c\ \ell\circ\psi(x,\cdot)\}dF(x),$$

belongs to $\mathcal{F}$.

THEOREM 4.1. Let the model $(\mathcal{F},\Psi(F))$ be boundedly integrable. Then the estimation $\hat{\psi}_n$ is at best n^{-2}-rate improvable.

EXAMPLE 4.1. Let $\psi(x)$ be a scalar function, $\Psi(F) = \int\psi(x)dF(x)$, $\xi_F^0(\nu) = P_F\{|\psi(X_1)| > \nu\}$. Suppose that the space $(\mathcal{F},\mathcal{T})$ contains an everywhere dense set of distributions F such that $\xi_F^0(\nu) = \mathcal{O}(\exp\{-\nu^\alpha\})$, $\nu \to \infty$, for some $\alpha > 0$.

If it is assumed that $\alpha > 1$, then the model $(\mathcal{F},\Psi(F))$ is completely integrable. If one assumes that $\alpha \geq 1$ then the model $(\mathcal{F},\Psi(F))$ is boundedly integrable.

The models with $\alpha > 0$ (approximately integrable models) will be studied in the next Section.

Typically with boundedly integrable models the estimators $\hat{\psi}_n$ (sample mean) would be exactly n^{-2}-rate improvable. Henceforth we restrict ourselves to the case $\psi(x) \in R^k$, $\Psi(F) = \int \psi(x)dF \in R^k$.

Denote by c_F the set of vectors c in R^k such that

$$b_F(c) = \ln \int \exp\{c'\psi(x)\}\, dF < \infty \quad \text{and let}$$

$$\frac{dG_c}{dF}(x) = \exp\{c'\psi(x) - b_F(c)\}, \qquad c \in c_F.$$

ASSUMPTION 0. For any $F \in \mathcal{F}$ there exists $c_F \in C_F$ such that $\int \psi(x)dG_{c_F} = 0$ and the integral $\int |\psi(x)|^2\, dG_{c_F}$ converges locally uniformly in F.

Let $C = \bigcup_{F\in\mathcal{F}} C_F$. Notice that $C = R^k$ for completely integrable models. To consider the case of integrable but not completely integrable models the following assumption is relevant.

ASSUMPTION 1. There exists a positive function $\omega(c)$ defined in a vicinity of C with bounded third-order derivatives such that the matrix $\dfrac{\partial^2\omega}{\partial c^2}$ is nonpositively definite nonidentically zero for $c \in C$.

Assumption 1 is satisfied, for example, if the set C is contained in a half-space of R^k.

THEOREM 4.2. Let Assumptions 0-1 be satisfied. Define an M-type estimator $\hat{c}_n$ by the relation

$$\sum_{i=1}^{n} e^{\hat{c}_n'\psi(X_i)} = \min_{c\in R^k} \sum_{i=1}^{n} e^{c'\psi(X_i)}$$

if the minimum is attainable and $\hat{c}_n = 0$ otherwise. Let

$$\Psi_n = \hat{\psi}_n + \frac{2}{n}\frac{\nabla\omega}{\omega}(\hat{c}_n). \tag{4.1}$$

Then locally uniformly in F

$$R_n(\Psi_n,F) = R_n(\hat{\psi}_n,F) + \frac{4}{n^2}\frac{\frac{\partial^2\omega}{\partial c^2}(c_F)}{\omega} + o(n^{-2}).$$

Hence $\hat{\psi}_n$ is n^{-2}-rate improvable on $\mathcal{F}$.

EXAMPLE 4.2. Let $\mathcal{F}_0 = \{F_0\}$ be a set of symmetric distributions in R^1 with exponentially decreasing tails such that $\gamma_1 \le (1-F_0(x))e^x \le \gamma_2$, $x > 0$ for some $0 < \gamma_1 \le \gamma_2 < \infty$. Assume that $\mathcal{F}$ is the set of all distributions F obtained from the original set $\mathcal{F}_0$ by "exponential contamination": $\frac{dF}{dF_0} = \exp\{cx - b(c)\}$. Let $\Psi(F) = \int x dF(x)$.

It is easy to verify that the model $(\mathcal{F},\Psi(F))$ is integrable and satisfies Assumptions 0-1. Hence, by Theorems 4.1, 4.2 $\hat{\psi}_n$ is in that case exactly n^{-2}-rate improvable.

5. Approximately Integrable Models

For the rest of the paper we restrict ourselves further to real valued functionals $\Psi(F)$ (not necessary of the form

$$\Psi(F) = \int \psi(x) dF(x) \tag{5.1}$$

as before).

DEFINITION 5.1. The model $(\mathcal{F},\Psi(F))$ will be called approximately integrable if there exists a sequence of functions $\psi_n(x) \in R^1$, $x \in \mathcal{X}$, and of numbers $a_n > 0$ such that, a) locally uniformly in F

$$\lim_n (n^{-1}\sigma^2_{n,F} + B^2_{n,F}) = 0$$

where

$$\sigma^2_{n,F} = \mathrm{Var}_F\, \psi_n(X_1), \quad B_{n,F} = E_F\, \psi_n(X_1) - \Psi(F). \tag{5.2}$$

b) For every F_0 ranging in an everywhere dense subset of $\mathcal{F}$ and any of its vicinity V there exists $\delta > 0$ such that the family of distributions $G_{n,c}$, $|c| < \delta a_n$,

$$\frac{dG_{n,c}}{dF}(x) = \exp\{c\,\psi_n(x) - b_n(c)\},$$

$$b_n(c) = \log \int \exp\{c\,\psi_n(x)\}\, dF(x), \tag{5.3}$$

belongs to $\mathcal{F}$.

Denote

$$B_{n,F}(c) = B_{n,G_{n,c}}, \quad \sigma^2_{n,F}(c) = \sigma^2_{n,G_{n,c}} \tag{5.4}$$

and in the case (5.1) with $\sigma_F^2 = \mathrm{Var}_F\, \psi(X_1) < \infty$,

$$\Delta_{n,F}(c) = \sigma^2_{G_{n,c}} - \sigma^2_{n,G_{n,c}}. \tag{5.5}$$

Let $\phi(a)$ denote class of continuously differentiable probability densities λ vanishing outside interval $(-a,a)$, with $\int(\lambda'(c))^2/\lambda(c)dc < \infty$.

Notice that both integrable and completely integrable linear models are special cases of the approximately integrable ones with $a_n = 1$, $a_n = \infty$, correspondingly. All the results presented so far and those to follow are based on the following locally minimax lower bound.

THEOREM 5.1. Let F_0, V, $G_{n,c} \in V$, $|c| < \delta a_n$ be as in Definition 5.1, $\lambda_n(\cdot) \in \phi(\delta a_n)$. Then with the notations (5.2)-(5.5) the following inequalities obtain

$$\inf_{\Psi_n} \sup_{F\in V}(R_n(\Psi_n,F)-R_n(\hat{\psi}_n,F)) \geq -\int_{-\delta a_n}^{\delta a_n} \left[\frac{\lambda_n'(c)}{n\lambda_n(c)} - B_{n,F_0}(c)\right]^2 \lambda_n(c)dc$$

$$\inf_{\Psi_n} \sup_{F\in V}(R_n(\Psi_n,F)-n^{-1}\sigma^2_{n,F}) \geq$$

$$\int_{-\delta a_n}^{\delta a_n} \left[2n^{-1}\lambda_n'(c)B_{n,F_0}(c) - n^{-2}\frac{(\lambda_n'(c))^2}{\lambda_n(c)}\right]dc$$

and in the case of the functional (5.1) with locally bounded $\sigma_F^2 = \mathrm{Var}_F\psi(x)$

$$\inf_{\Psi_n} \sup_{F\in V}(R_n(\Psi_n,F)-n^{-1}\,\sigma_F^2) \geq$$

$$\int_{-\delta a_n}^{\delta a_n} \left[-n^{-1}\Delta_{n,F_0}(c)\lambda_n(c) + 2n^{-1}\,\lambda_n'(c)B_{n,F_0}(c) - n^{-2}\frac{(\lambda_n'(c))^2}{\lambda_n(c)}\right]dc.$$

We present next a few examples of approximately integrable models $(\mathcal{F},\Psi(F))$ where $\Psi(F) = \int x dF(x)$.

Denote $\hat{X}_n = n^{-1}\sum_{i=1}^{n} X_i$

$$\hat{\psi} = n^{-1} \sum_{i=1}^{n} X_i 1(|X_i| < \nu_n), \qquad \nu_n > 0$$

$$\bar{F}(x) = 1 - F(x) + F(-x), \qquad x > 0.$$

EXAMPLE 5.1. For given $\alpha, \beta, \gamma > 0$ and a real μ consider the class $\mathcal{F}$ of distribution functions $F(x)$, $(x \in R)$ such that for some $\gamma_1(F)$, $\gamma_2(F)$, $0 < \gamma_1(F) \le \gamma_2(F) < \gamma$ and locally bounded function $\gamma_3(F)$

$$\gamma_1(F) \le (x^{\mu} e^{-\beta x^{\alpha}})^{-1} \bar{F}(x) \le \gamma_2(F), \qquad x > \gamma_3(F) .$$

Let

$$\nu_n = \frac{1}{\beta} \left[\log 2\gamma n + \frac{\mu}{\alpha} \log \log 2\gamma n - \frac{\log \beta}{\alpha}\right]^{\frac{1}{\alpha}}$$

PROPOSITION 5.1. a) X_n is exactly $\frac{(\log n)^{\frac{2}{\alpha}}}{n^2}$ - rate improvable on $\mathcal{F}$ and b) locally uniformly in $F \in \mathcal{F}$

$$R_n(\psi_n, F) \le \frac{\sigma^2(F)}{n} - \frac{\gamma_1(F)}{4\gamma} \frac{(\log n)^{\frac{2}{\alpha}}}{n^2} (1 + o(1)), \qquad n \to \infty$$

EXAMPLE 5.2. Assume that for some $\alpha > 2$, $\gamma > 0$

$$\mathcal{F} = \{F | \gamma_1(F) \le \bar{F}(x) x^{\alpha} \le \gamma_2(F), \quad x > \gamma_3(F)\}$$

where $0 < \gamma_1(F) \le \gamma_2(F) < \gamma$, $\gamma_3(F)$ is locally bounded.

Let $\nu_n = \left[\frac{2\alpha\gamma}{\alpha-2} n\right]^{\frac{1}{\alpha}}$.

PROPOSITION 5.2. a) $\hat{X}_n$ is exactly $n^{-\frac{2(\alpha-1)}{\alpha}}$ - rate improvable on $\mathcal{F}$ and b) for $n \to \infty$ locally uniformly in F

$$R_n(\psi_n, F) \le \frac{\sigma^2(F)}{n} - \left[\frac{\alpha}{\alpha-2}\right]^{\frac{2}{\alpha}} \frac{\gamma_1(F)}{2^{\frac{2(\alpha-2)}{\alpha}} \gamma^{\frac{\alpha-2}{\alpha}}} n^{-\frac{2(\alpha-1)}{\alpha}} (1 + o(1)),$$

Denote $\log_k x = \underbrace{\log \log \ldots \log}_{k \text{ times}} x$

EXAMPLE 5.3. Assume that for some $\alpha > 1$ and an integer $k \ge 1$

$$\mathcal{F} = \{F | \ \gamma_1(F) \le \prod_{i=1}^{k-1} \log_i x \ (\log_k x)^{\alpha} x^2 \bar{F}(x) \le \gamma_2(F), \ x > \gamma_3(F)\}$$

for some $0 < \gamma_1(F) \le \gamma_2(F) < \infty$ and a locally bounded $\gamma_3(F)$.

Let

$$\nu_n = \begin{cases} \left(\dfrac{n}{\log n}\right)^{1/2} & , \text{ if } k > 1, \\ \left(\dfrac{n}{(\log n)^{\alpha}}\right)^{1/2} & , \text{ if } k = 1. \end{cases}$$

PROPOSITION 5.3. a) $\hat{X}_n$ is exactly $((\log_k n)^{\alpha-1} n)^{-1}$ - rate improvable on $\mathcal{F}$;

b) locally uniformly in F

$$R_n(\hat{\psi}_n, F) \le n^{-1}\sigma_F^2 - 2\gamma_1(F)((\log_k \sqrt{n})^{\alpha-1} n)^{-1} (1 + o(1)), \quad (n \to \infty)$$

and

c) $\hat{\psi}_n$ is second order admissible, or $((\log_k n)^{\alpha-1} n)^{-1}$-rate unimprovable on $\mathcal{F}$.

EXAMPLE 5.4. So far we have analized higher order asymptotic properties of $\hat{X}_n$ under progressively heavier tail behavior of the underlying distribution $F \in \mathcal{F}$. It is all but natural to inquire further what happens with this estimator while F ranges over the class,

$$\mathcal{F} = \{F : \gamma_1(F) \le x^2 \bar{F}(x) \le \gamma_2(F), \quad x > \gamma_3(F)\}$$

where $0 < \gamma_1(F) \le \gamma_2(F) < \infty$ and $\gamma_3(F)$ is locally bounded. Let $\nu_n = \sqrt{n}$.

PROPOSITION 5.4. a) The functional $\Psi(F) = E_F X$ is exactly $\log n/n$-rate estimable on $\mathcal{F}$;

b) $$R_n(\hat{\psi}_n, F) \le \frac{2\gamma_2(F) \log n}{n} (1+o(1)), \quad n \to \infty,$$

locally uniformly in F and

c) $\hat{\psi}_n$ is first-order asymptotically optimal and exactly n^{-1}-rate improvable on $\mathcal{F}$.

Proceeding further with heavy tailed distributions F, one is led to considering the nonregular linear functionals still covered by Theorem 5.1 which will allow optimal rates conclusions to be derived for such functionals.

EXAMPLE 5.5. Let for some α, γ, $1 < \alpha < 2$, $0 < \gamma < \infty$

$$\mathcal{F} = \{F \mid \gamma_1(F) \le x^{-\alpha}\,\bar{F}(x) \le \gamma_2(F),\quad x > \gamma_3(F)\}$$

where $0 < \gamma_1(F) \le \gamma_2(F) < \gamma$, $\gamma_3(F)$ being locally bounded and $\nu_n = \rho n^{1/\alpha}$, $\rho > 0$.

PROPOSITION 5.5. a) $\Psi(F)$ is exactly $n^{-\frac{2(\alpha-1)}{\alpha}}$-rate estimable on $\mathcal{F}$, and b) locally uniformly in F

$$R_n(\hat{\psi}_n, F) \le \left[\left(\frac{2\gamma_2(F)}{2-\alpha} - \gamma_1(F)\right)\rho^{2-\alpha} + \left(\frac{\alpha\gamma_1(F)}{\alpha-1}\right)^2 \rho^{2(1-\alpha)}\right]\times$$
$$n^{-\frac{2(\alpha-1)}{\alpha}}\,(1 + o(1)).$$

EXAMPLE 5.6. Let for some integer $k \ge 1$ and given $\alpha > 1$, $\gamma > 0$

$$\mathcal{F} = \{F \mid \gamma_1(F) \le x\left[\prod_{i=1}^{k-1} \log_i x\right](\log_k x)^{\alpha}\,\bar{F}(x) \le \gamma_2(F),\ x > \gamma_3(F)\}$$

with some $0 < \gamma_1(F) \le \gamma_2(F) < \gamma$, $\gamma_3(F)$ being locally bounded. Set $\nu_n = n$.

PROPOSITION 5.6. The functional $\psi(F) = E_F X$ is exactly $(\log_k n)^{-2(\alpha-1)}$-rate estimable on $\mathcal{F}$, and b) locally uniformly in F,

$$R_n(\hat{\psi}_n, F) \le \gamma_2^2(F)(\log_k n)^{-2(\alpha-1)}(1+o(1)), \qquad (n \to \infty).$$

We present next two more examples of approximately integrable models.

EXAMPLE 5.7. For $\gamma > 0$ let $\alpha + 1$ be the smallest integer not smaller than γ, and $\beta = \gamma - \alpha$. Denote by $\mathcal{F} = F(\gamma, \gamma_1, \gamma_2, \gamma_3)$, $\gamma > 1$, $\gamma_1, \gamma_2, \gamma_3 > 0$, class of distribution functions F in R^1 such that

1) $F(\cdot) \in C^{(\alpha)}(x-\gamma_3, x+\gamma_3)$
2) $F'(x) \ge \gamma_1$
3) $\sup\limits_{|y-x|<\gamma_3} |y-x|^{-\beta}\,|F^{(\alpha)}(y)-F^{(\alpha)}(x)| < \gamma_2$.

PROPOSITION 5.7. Let $\Psi(F) = F(x)$ and $\hat{F}_n(x)$ be the sample distribution function. The estimator $\hat{F}_n(x)$ is exactly $n^{-(1+\frac{1}{2\gamma-1})}$-rate improvable on $\mathcal{F}$.

It is instructive to compare the above result with that of Example 3.1. The estimator improving upon $\hat{F}_n(x)$ can be obtained as integrated kernel density estimator.

EXAMPLE 5.8. Another famous example of approximately integrable models is provided by $(\mathcal{F}, p(x))$ where $\mathcal{F}$ is the same as above and $p(x) = F'(x)$. The corresponding result is presented for the sake of its comparison with Proposition 5.7.

PROPOSITION 5.8. Let $\Psi(F) = P(x)$. Then $\Psi(F)$ is exactly $n^{-(1-\frac{1}{2\gamma-1})}$-rate estimable on $\mathcal{F}$.

References

Levit, B. Ya. (1978) 'Infinite-dimensional informational inequalities', Theor. Prob. Appl. 23, 371-377.

Levit, B. Ya. (1982) 'Minimax estimation and positive solutions of elliptic equations', Theor. Prob. Appl. 27, 563-586.

Levit, B. Ya. (1985) 'Second-order asymptotic optimality and positive solutions of Schrödinger's equation', Theor. Prob. Appl. 30, 333-363.

NONPARAMETRIC TECHNIQUES IN IMAGE ESTIMATION

A.B. TSYBAKOV
Institute for Problems of Information Transmission
USSR Academy of Sciences
101447 Moscow GSP-4
Ermolovoy str., 19
USSR

ABSTRACT. Image estimation is considered as a special kind of a nonparametric functional estimation problem. Classes of images and edges are introduced and estimates are proposed that have optimal rates of convergence on these classes. The influence of the choice of a design on the rates of convergeance is pointed out.

1. INTRODUCTION

In this paper, we consider the problems of image estimation and edge estimation, using the tools of nonparametric statistics. Images and edges are assumed to be nonrandom functions, belonging to some specific functional classes. In our model, the image f is a discontinuous function with two smooth parts: the informative part, where the values of f are large, and the background, where the values of f are small. On the boundary between these two parts, there is an abrupt change. This boundary is called edge. The edge should satisfy some regularity conditions. The function f is supposed to be "nice" everywhere except the edge. Assume that a sample of noisy observations of image values is given. Our goal is to estimate the edge and the image itself from such observations.

We propose some methods of image and edge estimation based on locally-polynomial approximations. The procedure of image estimation is a two-stage one. On the first stage, the background and the informative part are separated by classification techniques. This stage is equivalent to edge estimation. On the second stage, the locally-polynomial estimate is constructed separately on distinct parts of the image. We obtain the rates of convergence of the estimators. These rates depend on the edge smoothness and on the image smoothness outside the edge. It is proved that the convergence rates of the estimate are the best in an asymptotically minimax sense. Thus, we obtain the optimal rates of convergence for image and edge estimation problems. This can be used as a reference point in comparing various other estimators used in image analysis (see, e.g., Pratt (1978)).

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 669–677.
© 1991 *Kluwer Academic Publishers.*

An interesting consequence of the results given below concerns the choice of the design. It turns out that the standard equispaced grid design is not the best one for estimation of smooth edges, as far as the accuracy of estimators is measured by integrated error. It is shown that, in this case, the optimal rates can be achieved by using a random design.

2. IMAGE ESTIMATION

Let the image f be a function defined on the N-dimensional cube $X^N = [0,1]^N$. Assume that we have the sample $y_1,\ldots,y_n$, where $y_i = f(x_i) + \xi_i$, $i=1,\ldots,n$, ξ_i are independent random errors, and x_i are design points. We assume everywhere that $E(\xi_i) = 0$, $E(\xi_i^2) = \sigma^2 < \infty$. An important example of a design is the equispaced grid

$$\Gamma = \{x = (q_1,\ldots,q_N) \in X^N\colon n_1 q_i \in \{0,\ldots,n_1-1\}\},$$

where $n = n_1^N$, and n_1 is an integer.

Let β,L be some positive numbers, and let ℓ be the maximal integer such that $\ell<\beta$. Denote by $\Sigma_\beta = \Sigma_\beta(L,N)$ the class of all functions $g:X^N \to [0,1]$ having continuous partial derivatives up to order ℓ and such that the difference between g(z), $z \in X^N$, and its Taylor polynomial of order ℓ at a point $x \in X^N$ does not exceed $L|z-x|^\beta$ (this is the same functional class as that described in Stone (1980, 1982)). Here and in the sequel, $|\cdot|$ denotes the Euclidean norm.

Assume that for any image f there exists a partition of X^N into two disjoint sets $A_0 = A_0(f)$, $A_1 = A_1(f)$ (A_0 corresponds to the background and A_1 corresponds to the informative part). The edge $\mathcal{V}$,i.e., the boundary between A_0 and A_1, is assumed to be piecewise Lipschitz with a specified Lipschitz constant L_1 and a number of Lipschitz pieces that does not exceed a fixed m. We also assume that there exist $H>0$, $0<\gamma<1$ such that the sets A_0 and A_1 satisfy the (γ,H)-cone condition; i.e., for any $x \in A_0$ (respectively, A_1) there exist a vector $w \in R^N$ with $|w|=1$ such that the cone $\{z \in R^N\colon |z-x| \le H, (z-x)^T w \ge \gamma|z-x|\}$ is contained in A_0 (respectively A_1). Note that these conditions on $\mathcal{V}$ are rather weak and they seem to be satisfied for a large class of images.

Introduce the following classes of images Φ_β, $\beta>0$, on X^N. Let a,b be fixed numbers, and $0\le a<b\le 1$. Denote by Φ_0 the set of functions f such that

$$f(x) = aI\{x \in A_0\} + bI\{x \in A_1\},$$

where A_0 and A_1 satisfy the above conditions and I is the indicator function. For $\beta>0$, define Φ_β as the set of all functions f that can be represented as

$$f(x) = g_0(x)I\{x \in A_0\} + g_1(x)I\{x \in A_1\},\ g_0,g_1 \in \Sigma_\beta,$$

$$\inf_{x\in A_0} g_0(x) \ge b,\quad \sup_{x\in A_1} g_1(x) \le a,$$

$b>1$, $a>0$, and the sets A_0 and A_1 satisfy the above conditions.

Consider the following two-stage estimator of f. On the first stage, we perform a simple classification, and on the second stage, the image values are smoothed using the result of classification.

Classification. Introduce a partition of X^N into disjoint cubes Δ_k, $k=1,\ldots,K_n^N$, with edges of length $1/K_n$ where K_n is an integer such that $1/K_n \sim (C\log n/n)^{1/N}$, $n \to \infty$, $C>0$. For the cubes Δ_k, calculate the medians

$$\mathrm{med}_k = \mathrm{med}\{y_i : x_i \in \Delta_k\}.$$

Classify the cubes into "black" and "white" ones according to the following rule. For some threshold d such that $a<d<b$ decide that Δ_k is black if $\mathrm{med}_k > d$, and that Δ_k is white if $\mathrm{med}_k \le d$. A point $x \in X^N$ is called white (respectively, black) if it belongs to a white (respectively, black) cube.

Note that the number of points of the equispaced grid in each cube is roughly $C\log n$. Thus in view of the exponential error bound for deviations of the median, the classification error is decreasing faster than any prescribed power of n^{-1} for C large enough.

Smoothing. Define the bandwidth $h_n \sim n^{-1/(2\beta^*+N)}$, where $\beta^* = \min\{\beta, N/2(N-1)\}$. For each $x \in X^N$, denote

$$I_n(x) = \{i : |x_i - x| \le h_n,\ x_i \text{ has the same colour as } x\},$$

$$m_n(x) = \mathrm{card}\, I_n(x).$$

The simplest smoother (for the case $0<\beta\le 1$) is defined as

$$f_n(x) = f_n^o(x) I\{0 \le f_n^o(x) \le 1\}$$

where

$$f_n^o(x) = \begin{cases} m_n(x)^{-1} \sum_{i \in I_n(x)} y_i & \text{if } m_n(x) \ne 0, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, f_n is the running mean over the appropriate neighborhoods. For $\beta > 1$, $f_n^o(x)$ is replaced by $p_{n,\ell}(x)$ where $p_{n,\ell}$ is the least squares polynomial of degree ℓ for regression with the data $\{(x_i, y_i),\ i \in I_n(x)\}$. For the special case that $\beta=0$, set

$$f_n^o(x) = \begin{cases} Y_{\mathrm{white}} & \text{if } x \text{ is white,} \\ Y_{\mathrm{black}} & \text{if } x \text{ is black,} \end{cases}$$

where

$Y_{white} = \text{ave}\{y_i: x_i \text{ is white}\}$, and Y_{black} is defined similarly.

Introduce two accuracy measures for image estimators, the discrete risk and the integrated risk, respectively:

$$R_D(f,T_n) = E\left[n^{-1}\Sigma_{z\in\Gamma}(f(z) - T_n(z))^2\right],\ R_I(f,T_n) = E\left[\int\left[f(z) - T_n(z)\right]^2 dz\right]$$

where T_n is an arbitrary estimate of f.

<u>Assumption 1</u>. $\{x_1,\ldots,x_n\} = \Gamma$.

<u>Theorem 1</u>. Let Assumption 1 be satisfied and let ξ_i, $i=1,\ldots,n$, be i.i.d. random variables with symmetric distribution function F such that $F(t) > 1/2$, $t>0$. Then, for C large enough,

$$\sup_{f\in\Phi_\beta} R(f,f_n) = O(\max\{n^{-2\beta/(2\beta+N)},\ (\log n/n)^{1/N}\}),\ \beta>0,$$

$$\sup_{f\in\Phi_\beta} R(f,f_n) = O((\log n/n)^{1/N}),$$

as n tends to ∞, where $R = R_D$ or $R = R_I$.

<u>Remark 1</u>. The appropriate choice of C depends only on F (see Tsybakov (1989)).

<u>Remark 2</u>. For $\beta>0$, the rate given in Theorem 1 is the maximum of two terms. The first term, of order $n^{-2\beta/(2\beta+N)}$, reflects the contribution of smooth parts of the image and it is the same as in standard nonparametric regression problems (Ibragimov, Hasminskii (1980), Stone (1982), Nemirovskii (1985), Nussbaum (1986)). The second term represents the contribution of the boundary. It reflects the fact that the edge is estimated uniformly with an accuracy of order $(\log n/n)^{1/N}$. For $\beta = 0$, only the edge term is present.

The proof of Theorem 1 for the discrete risk follows from Tsybakov (1989). The case of integrated risk is treated similarly.

3. EDGE ESTIMATION

In this section, we assume that the edge is a single-valued function of the first N-1 coordinates, $\vartheta: x^{N-1} \to [0,1]$, and that $\vartheta\in\Sigma_\delta(L_1,N-1)$, $N\geq 2$, $\delta>0$, $L_1<\infty$. We also assume that ϑ is bounded from below and from above: $1-H_1 \leq \vartheta(t) \leq H_1$, $t \in X^{N-1}$, for some $H_1 \in (1/2,1)$.

For $\delta \geq 1$ and $\beta \geq 0$, denote by $\Phi_{\beta,\delta}$ the subset of Φ_β for which ϑ satisfies the above conditions. Note that on $\Phi_{0,\delta}$, image estimation is equivalent to edge estimation. In this section, we consider only the classes $\Phi_{0,\delta}$ assuming, without loss of generality, that a=0, b=1, so that $f(x) = I\{x^N \leq \vartheta(t)\}$, where x^N is the Nth coordinate of x.

Thus, the image is assumed to be black and white (the set $\{x^N \le \vartheta(t)\}$ is black while its complement is white). This entails that for any edge estimator $\mathcal{W}_n : x^{N-1} \to [0,1]$

$$\sup_{\vartheta \in \Sigma_\delta} E\left(\int |\mathcal{W}_n(t) - \vartheta(t)|\,dt\right) = \sup_{f \in \Phi_{0,\delta}} R_I(f, T_n) \tag{1}$$

where $T_n(x) = I\{x^N \le \mathcal{W}_n(t)\}$ is the corresponding image estimator.

Note that if we do not require the edge estimator to be a single-valued function but rather a curve on X^N, then we could use an edge estimator based on the classification procedure of Section 2. In fact, define $\mathcal{W}_n$ as the boundary between black and white parts of the classified image. Instead of the L_1-error given by (1), take its extension for such general boundaries; i.e., the Lebesgue measure of symmetric difference between the black part of the image and the black part of the classified image. From Theorem 1, it follows that the expected value of such symmetric difference is of the order $(\log n/n)^{1/N}$. This is a rough estimate, and as we show now, it can be improved substantially if $\delta > 1$ (for $\delta = 1$ only the log-factor can be eliminated). The improvement is due to a random design and to a more involved construction of an edge estimator.

<u>Assumption 2</u>. The design points x_i, $i=1,\ldots,n$, are i.i.d. random vectors uniformly distributed on X^N.

Describe now the edge estimator. Introduce a partition of X^{N-1} into disjoint cubes Q_k, $k=1,\ldots,D_n^{N-1}$, with edges of length $1/D_n$, where D_n is an integer such that $D_n \sim n^{1/(N+\delta-1)}$, $n \to \infty$. Define the corresponding slices on X^N:

$$S_k = \{x \in X^N : x = (t, x^N),\ t \in Q_k,\ x^N \in [0,1]\}.$$

On each cube Q_k, define the estimator ϑ_n as a polynomial $p(t)$ of degree q (q is the maximal integer such that $q<\delta$), solving the following minimization problem

$$\sum_{x_i \in S_k} \left(y_i - I\{x_i^N \le p(t_i)\}\right)^2 \to \min_{p \in P_n} .$$

Here x_i^N is the Nth coordinate of x_i, t_i is the vector of first N-1 coordinates of x_i, and P_n is the finite set of polynomials defined as follows:

$$P_n = \{p(t) = p(t,\theta),\ \theta \in \mathcal{N}_n\}.$$

The quantity θ is the vector of coefficients of a polynomial $p(t) = p(t,\theta)$ of degree q and $\mathcal{N}_n$ is an ε-net ($\varepsilon \sim D_n^{-\delta}$) on the set $\{\theta : 1-H_1 \le p(t,\theta) \le H_1\}$.

Theorem 2. Let ξ_i, $i=1,\ldots,n$, be i.i.d. Gaussian random variables and let Assumption 2 be satisfied. Then, as $n \to \infty$,

$$\sup_{\vartheta\in\Sigma_\delta} E\left(\int|\vartheta_n(t) - \vartheta(t)|dt\right) = O\left(n^{-\delta/(N+\delta-1)}\right), \quad \delta>0.$$

Corollary. Under the assumption of Theorem 2 and as $n \to \infty$,

$$\sup_{f\in\Phi_{0,\delta}} R_I(f,f_n) = O\left(n^{-\delta/(N+\delta-1)}\right), \quad \delta>0,$$

where $f_n(x) = I\{x^N \le \vartheta_n(t)\}$.

Remark 3. Theorem 2 and the corollary remain valid for the multiplicative model $y_i = f(x_i)\xi_i$ where ξ_i, $i=1,\ldots,n$, are i.i.d. random variables such that $\xi_i = 1$ with probability $1-p_0$ and $\xi_i = 0$ with probability $p_0 (0<p_0<1/2)$.

Remark 4. Under the assumptions of Theorem 2, the same rate is valid if we replace the L_1-error by the $(1/p)$th power of the L_p-error, $p<\infty$.

4. LOWER BOUNDS

Let us give some minimax lower bounds, illustrating the sense in which the image and edge estimators described in Section 2 and Section 3 are the best estimators. This will lead to derivation of the optimal convergence rates for image and edge estimation problems. The optimal rates strongly depend on the choice of a risk function (discrete or integrated risk) and on the design. Two assumptions on the design will be used here: Assumption 1 and

Assumption 3. $x_1,\ldots,x_n$ are arbitrary X^N-valued random vectors such that $(x_1,\ldots,x_n)$ is independent of $(\xi_1,\ldots,\xi_n)$.

Some assumptions on the distribution function F of ξ_i are necessary in order to obtain nontrivial lower bounds. We assume throughout this section that the supports of $F(y-a)$ and $F(y-b)$ have a nonempty intersection (otherwise it is possible to distinguish the informative part from the background without error). This assumption can be written as $\mathcal{H}^2(b-a) < 2$, where

$$\mathcal{H}^2(t) = \int\left(\sqrt{dF(u)} - \sqrt{dF(u)}\right)^2$$

denotes the squared Hellinger distance.

Theorem 3. Let Assumption 3 be satisfied. Then

$$\liminf_n \sup_{f\in\Phi_{0,\delta}} n^{1/N} R_D(f,T_n) \ge c_1, \quad \delta \ge 1,$$

for an arbitrary estimate T_n. If, moreover, $\limsup_{t\to 0} \mathcal{H}^2(t)/t^2 < \infty$, then

$$\liminf_n \sup_{f\in\Phi_{0,\delta}} b_n R_D(f,T_n) \geq c_2, \quad \delta \geq 1,\ \beta > 0,$$

for an arbitrary estimate T_n. Here $b_n = \min\{n^{2\beta/(2\beta+N)}, n^{1/N}\}$, and c_1, c_2 are positive constants not depending on T_n or the design.

The proof of Theorem 3 is given in Tsybakov (1989).

Since $\Phi_\beta \supseteq \Phi_{\beta,\delta}$, $\delta \geq 1$, it follows from Theorem 3 that the rate of convergence of the image estimator proposed in Section 2 is the optimal one up to the log-factor. Moreover, the optimal rate does not depend on δ, the smoothness parameter of the edge. This is so because the risk R_D is calculated on the equispaced grid. Note that there is no design choice that can help in improving the rate with the discrete risk. For the integrated risk, however, such an improvement is possible as explained in Section 3. The next theorem establishes the optimality of the rate derived in Section 3.

<u>Theorem 4</u>. Let Assumption 3 be satisfied. Then

$$\liminf_n \sup_{f\in\Phi_{0,\delta}} n^{\delta/(N+\delta-1)} R_I(f,T_n) \geq c_1', \quad \delta \geq 1,$$

for an arbitrary estimate T_n. If, moreover, $\limsup_{t\to 0} \mathcal{H}^2(t)/t^2 < \infty$, then

$$\liminf_n \sup_{f\in\Phi_{0,\delta}} b_n' R_I(f,T_n) \geq c_2', \quad \delta \geq 1,\ \beta > 0,$$

for an arbitrary estimate T_n. Here $b_n' = \min\{n^{2\beta/(2\beta+N)}, n^{\delta/(N+\delta-1)}\}$, and c_1', c_2' are positive constants not depending on T_n or the design.

<u>Remark 5</u>. Using (1), we can rewrite the first inequality of Theorem 4 in terms of an edge estimation problem:

$$\liminf_n \sup_{\mathcal{V}\in\Sigma_\delta} n^{\delta/(N+\delta-1)} E\left(\int |\mathcal{W}_n(t)-\mathcal{V}(t)|dt\right) \geq c_1', \quad \delta \geq 1,$$

for any edge estimator $\mathcal{W}_n$.

The fact that the equispaced grid design does not obtain the rate $O\left(n^{-\delta/(N+\delta-1)}\right)$ has a simple explanation. Clearly, the distance between the consecutive gridpoints $n^{-1/N}$ is greater than $n^{-\delta/(N+\delta-1)}$ if $\delta > 1$. There exist two edge functions $\mathcal{V}(t)$ and $\mathcal{W}(t)$ for which the observed samples of y_i's are exactly the same, while the L_1-distance between $\mathcal{V}$ and $\mathcal{W}$ is of order $O(n^{-1/N})$. For example, we can take $\mathcal{V}(t) = n^{-1/N}/4$, $\mathcal{W}(t) = 3n^{-1/N}/4$. This immediately entails that for the equispaced grid design and for arbitrary $\mathcal{W}_n$,

$$\liminf_n \sup_{\mathcal{V}\in\Sigma_\delta} n^{1/N} E\left(\int |\mathcal{W}_n(t)-\mathcal{V}(t)|dt\right) \geq 0, \quad \delta \geq 1.$$

5. SUBOPTIMAL ESTIMATORS

In this section, we show that two other types of estimators used in image analysis do not achieve the optimal rates. These are the *linear* estimators and the *linewise processing* estimators.

Linear estimators are any statistics of the form

$$\Lambda_n(x) = \sum_{i=1}^{n} y_i \phi_{in}(x_1,\ldots,x_n,x), \quad x \in X^N,$$

where ϕ_{in} are arbitrary measurable functions.

Theorem 5. Let Assumption 3 be satisfied. Then for any linear estimate Λ_n,

$$\liminf_n \sup_{f\in\Phi_{0,1}} n^{1/(N+1)} R(f,\Lambda_n) \geq c_3,$$

for $R = R_I$, $R = R_D$, where c_3 does not depend on Λ_n.

The proof of this theorem follows from Tsybakov (1989).

Theorem 5 shows that nonlinearilty of the optimal image estimators is essential. Any linear estimator has a slower rate of convergence. For example, it follows from Theorem 1 that if N=2, then the convergence rate of f_n on the class $\Phi_{0,1}$ is of the order $(\log n/n)^{1/2}$, while for linear estimators the rate is not faster than $n^{-1/3}$, by Theorem 5.

Algorithms of linewise processing usually consist in smoothing the one-dimensional arrays of data along the lines of the equispaced grid. The relations between different lines are ignored. Since the linewise processing is applied for case that N=2, we also restrict the argument to this case. Assume that the image belongs to the class $\Phi_{0,\delta}$. Thus, the edge is a single-valued function $u=\vartheta(t)$, $u,t\in[0,1]$. The design is assumed to be on the equispaced grid, and the linewise processing is assumed to be performed along the lines parallel to the u-axis. Thus there are $n_1 = n^{1/2}$ lines, each line containing n_1 gridpoints. It follows from Theorem 5 that for linear smoothing along each line, the linewise discrete risk decreases not faster than $O(n_1^{-1/2}) = O(n^{-1/4})$. This entails that the overall discrete risk R_D does not tend to 0. Thus, the standard linewise processing fails. Nonetheless, we can consider an ideal procedure, taking into account all a priori information on the smoothness of ϑ and the fact that there is a jump of image values at the edge.

The ideal linewise procedure has two steps. On the first step, estimate the locations of jumps $a_1,\ldots,a_{n_1}$ for all n_1 lines in an optimal way. On the second step, smooth the estimates of $a_1,\ldots,a_{n_1}$ by nonparametric regression techniques, optimally for the given smoothness parameter δ. It can be shown that the optimal accuracy in estimating the location of a jump from a sample size n_1 is of the

order $n_1^{-1} = n^{-1/2}$ (see Ibragimov, Hasminskii (1979), Chapter 7, and Korostelev (1987) for the similar continuous-time problem). This can be interpreted as if the a_i's were observed with independent random errors whose standard deviation σ is of the order $n^{-1/2}$. On the second step, use a kernel edge estimator with a properly chosen kernel. This would lead to the IMSE proportional to $n_1^{-2\delta/(2\delta+1)} \cdot (\sigma^2)^{2/(2\delta+1)} = O(n^{-(\delta+2)/(2\delta+1)})$. From Theorem 2 and Remark 4, it follows that the optimal edge estimator proposed in Section 3 has an IMSE of the order $O(n^{-2\delta/(\delta+1)})$ which is better than the above rate if $\delta>1$.

REFERENCES

Ibragimov, I.A., and Hasminskii, R.Z., (1979). 'Asymptotic Estimation Theory', Nauka, Moscow.

Ibragimov, I.A., and Hasminskii, R.Z., (1980). 'On nonparametric estimation of regression'. Doklady Acad. Nauk SSSR 252, 780-784.

Korostelev, A.P. (1987). 'On minimax estimation of a discontinuous signal', Theory Probab. Appl., 32, 796-799.

Nemirovskii, A.S. (1985). 'On nonparametric estimation of smooth regression functions', Izv. Acad. Nauk SSSR, Tech. Kibern. No. 3, 50-60.

Nussbaum, M. (1986). 'On nonparametric estimation of a regression function that is smooth in a domain on R^k', Theory Probab. Appl. 31, 118-125.

Pratt, W.K. (1978). 'Digital Image Processing', J. Wiley, N.Y.

Stone, C.J. (1980). 'Optimal rates of convergence for nonparametric estimators', Ann. Statist. 8, 1348-1360.

Stone, C.J. (1982). 'Optimal global rates of convergence for nonparametric estimators', Ann. Statist. 10, 1040-1053.

Tsybakov, A.B. (1989). 'Optimal estimation accuracy of nonsmooth images', Problems of Inform. Transmission 25, 180-191.

REGULARIZED DECONVOLUTION ON THE CIRCLE AND THE SPHERE

ARNOUD C.M. VAN ROOIJ
Department of Mathematics
Katholieke Univ. Nijmegen
6525 ED Nijmegen
The Netherlands

FRITS H. RUYMGAART
Department of Mathematics
Texas Tech Univ.
Lubbock, TX 79409
U.S.A.

ABSTRACT. In statistical errors in variables models one observes a noisy convolution of an unknown density with a known error distribution, and the problem is to recover the unknown density. This problem amounts to ill-posed inversion of an operator. Such inversion requires regularization. After a brief summary of regularization in general we turn to deconvolution of functions on Abelian groups and focus in particular on the circle. Finally we propose a kind of error in variables model on the sphere (see also Chang (1989)) and mention some ensuing problems.

1. Introduction

In statistics ill-posed problems may arise in situations where we observe or construct a noisy approximation $\hat{g}$ of

$$g = Kf, \tag{1.1}$$

where K is a known transformation of an unknown signal f. The problem is to recover the signal f from $\hat{g}$ which requires inversion of the operator K. This inversion, however, is often ill-posed; even when we restrict to the typical case where K is a one-to-one and continuous linear transformation of a Hilbert space $\mathbf{H}$ into itself, K^{-1} need not be defined on all of $\mathbf{H}$ (so that $\hat{g}$ may fall outside the range $K(\mathbf{H})$), and K^{-1} need not be continuous. Therefore so-called regularization of the inverse is needed.

The kind of statistical problem to which we will restrict ourselves here is the errors in variables model, where we observe $Y_i = X_i + E_i$, with $X_i \amalg E_i$ and the pairs $(X_1, E_1), ..., (X_n, E_n)$ and hence $Y_1, ..., Y_n$ i.i.d. In this model the unobservable errors E_i have a known density φ, say, the unobservable X_i have unknown density f and the observed $Y_1, ..., Y_n$ have a density g which can be estimated from the data; this density estimator will be denoted by $\hat{g}_n$. In this case the operator K is convolution with φ. Hence to recover f from $\hat{g}_n$ is a deconvolution problem; such problems are ill-posed indeed.

It is clear that in this kind of problem our Hilbert space $\mathbf{H}$ is a Hilbert space of functions $L_2(\mathbf{G}, \mathcal{A}, \mu_{\mathbf{G}})$, where $\mathbf{G}$ must at least have the structure of a topological group in order that convolution be defined. The symbol $\mu_{\mathbf{G}}$ denotes an invariant measure on

G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics, 679–690.
© 1991 *Kluwer Academic Publishers.*

$\mathbf{G}$, called Haar measure. For Abelian groups there is no distinction between left and right invariance. If $\oplus$ is the group operation, convolution is defined as

$$(\varphi \circledast f)(x) = \int_{\mathbf{G}} \varphi(x \ominus y) f(y) d\mu_{\mathbf{G}}(y), x \in \mathbf{G}, \tag{1.2}$$

where $x \ominus y = x \oplus y^{-1}$. A theory of regularized deconvolution is indeed possible for certain $\mathbf{G}$, in particular for locally compact Abelian groups.

In the simple examples below of such groups λ will denote Lebesgue measure, restricted or not to a part of $\mathbf{R}$:

(a) $\mathbf{G} = \mathbf{R}$ with $\oplus = +$ and $\mu_{\mathbf{G}} = \lambda$;

(b) $\mathbf{G} = (0, \infty)$ with $\oplus = \times$ and $\mu_{\mathbf{G}} = \mu_+$, the measure with density $(d\mu_+/d\lambda)(t) = t^{-1}, t \in (0, \infty)$;

(c) $\mathbf{G} = \mathbf{R}/\mathbf{Z} = \mathbf{S}$, the unit circle in the plane ($\mathbf{Z}$ the set of all integers); we will make the identification $\mathbf{S} = [0, 1)$, with $\oplus = +$ (modulo 1) and $\mu_{\mathbf{G}} = \mu_0$, Lebesgue measure on $[0, 1)$.

We will focus on $\mathbf{S}$ but briefly mention $\mathbf{R}$ and $(0, \infty)$. Regularized deconvolution on $\mathbf{S}$ will be considered in Section 4.

In higher dimensions the circle generalizes both to the unit sphere $\mathbf{S}^{d-1}$ in $\mathbf{R}^d$ or to $\mathbf{O}(d)$, the orthogonal group in $\mathbf{R}^d$. As we will see below, a modification of Chang's (1989) model leads to a kind of errors in variables model on the sphere, and one can easily imagine how errors in variables on $\mathbf{O}(d)$ arise. However, $\mathbf{S}^{d-1}$ is not even a group (although it is a homogeneous space), and $\mathbf{O}(d)$ is a group but not Abelian. In Section 5 we define an errors in variables model on the sphere and outline a possible solution. However, we cannot yet present the mathematical details but hope to do so in a forthcoming paper.

Before deconvolution can even be considered we have to face the problem of estimating g from the data $Y_1, ..., Y_n \in \mathbf{G}$. We consider this problem in Section 3 in the case where $\mathbf{G}$ is a Stiefel manifold, which is not in general a group (it is a homogeneous space) but which contains all the examples of present interest (the circle, the sphere, the orthogonal group) as special cases. In Section 2 we briefly discuss a general principle of regularization in Hilbert spaces.

Our references to the vast literature will be rather selective and restricted. An important but not specifically statistical reference in Tikhonov & Arsenin (1977) where the concept of regularization serves the same purpose but is somewhat differently defined. Their technique is applied in Vapnik (1982), often overlooked in the recent statistical literature, to some statistical problems. This author deals with compact operators but deconvolution is not included. For statistical deconvolution on the real line e.g. Carroll & Hall (1988), Fan (1988) and Zhang (1990). Many interesting practical examples can be found in Titterington (1985) and O'Sullivan (1986). In Carroll

et al. (1990) a general theory of regularized inversion in Hilbert spaces is reviewed, applied to deconvolution on abstract locally compact Abelian groups, and illustrated with a number of practical statistical problems. However, they don't consider errors in variables on the sphere.

2. Regularized inverses in Hilbert spaces

Let K be an injective operator of a Hilbert space $\mathbf{H}$ into itself; in other words let K be one-to-one and $K \in \mathcal{L}(\mathbf{H})$. By suitable preconditioning (replacing the equation $g = Kf$ by $Tg = TKf$ for suitable one-to-one $T \in \mathcal{L}(\mathbf{H})$) if necessary we may and will assume that K is *strictly positive Hermitian.* It is the contents of the spectral theorem, applied to such operators, that there exists a σ-finite measure space $(\mathcal{X}, \mathcal{A}, \mu)$, a unitary $U : \mathbf{H} \to L_2(\mathcal{X}, \mathcal{A}, \mu)$, a measurable $\gamma : \mathcal{X} \to \sigma(K)$, the spectrum of K, and a multiplication operator $M_\gamma : L_2(\mathcal{X}, \mathcal{A}, \mu) \to L_2(\mathcal{X}, \mathcal{A}, \mu)$, such that

$$K = U^{-1} M_\gamma U, \text{ where } M_\gamma \psi = \gamma \cdot \psi \text{ for } \psi \in L_2(\mathcal{X}, \mathcal{A}, \mu). \tag{2.1}$$

This is well-known for Euclidean spaces where an orthonormal matrix $O \in \mathbf{O}(d)$ exists such that $K = O^t \Lambda O$, with Λ a diagonal matrix. Loosely speaking (2.1) entails that, apart from unitary equivalence, K^{-1} is multiplication by γ^{-1}, an operator that indeed might turn out to be unbounded if $\mathbf{H}$ is infinite dimensional. (In the finite dimensional case $K^{-1} = O^t \Lambda^{-1} O$ which is still continuous but unstable in a sense when the smallest eigenvalue of K (and hence Λ) is close to 0.)

A regularized inverse of K is a sequence of operators $R_m \in \mathcal{L}(\mathbf{H}), m \in \mathbf{N}$, where m is a kind of smoothing parameter, with the property that R_m is close to K^{-1}:

$$\|R_m K f - f\| \to 0, \text{ it as } m \to \infty, \text{ for all } f \in \mathbf{H}. \tag{2.2}$$

If K happens to be invertible (and not just one-to-one) so that $K^{-1} \in \mathcal{L}(\mathbf{H})$, which is the case if $0 < \min\{\lambda : \lambda \in \sigma(K)\}$, we might take $R_m = K^{-1}$ for all $m \in \mathbf{N}$. But even in this case for stability considerations one might prefer to carry out a genuine regularization.

It is possible to construct a regularized inverse once the quantities in (2.1) are given. For $m \in \mathbf{N}$ let us define

$$\rho_m(x) = \begin{cases} \gamma^{-1}(x), & \text{for } x \in \mathcal{X} : \gamma(x) \geq 1/m \\ 0, & \text{elsewhere.} \end{cases} \tag{2.3}$$

It can be shown that (2.2) is satisfied with

$$R_m = U^{-1} M_{\rho_m} U, \text{ where } M_{\rho_m} \psi = \rho_m \cdot \psi \text{ for } \psi \in L_2(\mathcal{X}, \mathcal{A}, \mu). \tag{2.4}$$

This regularization is related to projection. Another interesting regularization is obtained by choosing $\rho_m = (\alpha(m) + \gamma)^{-1}$. In either case we have

$$\|R_m\| \leq m, \quad m \in \mathbf{N}. \tag{2.5}$$

We will use the regularization (2.4).

A proper choice of the regularization parameter m may depend on other parameters in the problem. Let $(\Omega, \mathcal{W}, \mathbf{P})$ be a probability space and the estimators $\widehat{g}_n : \Omega \to \mathbf{H}$ Borel measurable. We will assume that, given $f \in \mathbf{H}$, the element $g = Kf$ can be consistently estimated in the sense that $\mathbf{E}\|\widehat{g}_n - Kf\| = \delta_n(f) \to 0$, as $n \to \infty$, and we know that $\|R_m Kf - f\| = \varepsilon_m(f) \to 0$, as $m \to \infty$. Then it is possible to find for each $n \in \mathbf{N}$ an integer $m(n) \in \mathbf{N}$ (depending on f) such that

$$m(n) \to \infty \quad \& \quad \|R_{m(n)}\|\delta_n(f) \to 0, \ \textit{as } n \to \infty. \tag{2.6}$$

As an estimator for f we propose

$$\widehat{f}_n = R_{m(n)}\widehat{g}_n, \ \textit{with } m(n) \textit{ satisfying (2.6).} \tag{2.7}$$

Since also $\varepsilon_{m(n)}(f) \to 0$, as $n \to \infty$, for such $m(n)$ it follows that

$$\mathbf{E}\|\widehat{f}_n - f\| \leq \|R_{m(n)}\|\delta_n(f) + \varepsilon_{m(n)}(f) \to 0, \ \textit{as } n \to \infty. \tag{2.8}$$

If the rates of $\varepsilon_m(f)$ and $\delta_n(f)$ are known (see (2.5)) for the rate of $\|R_m\|$ it is possible to specify an approximately optimal rate for the r.h.s. in (2.8) by choosing the $m(n)$ in such a way that a balance between the rates of $\|R_{m(n)}\|\delta_n(f)$ and $\varepsilon_{m(n)}(f)$ is obtained.

For at least two important subclasses of operators further specification of the operator U and the function γ in (2.1) is possible: for compact and for convolution operators. By preconditioning Wicksell's problem (estimation of the true radius of a sphere when only radii of random two-dimensional cuts are observed) can be reduced to regularized inversion of a compact operator (Nychka & Cox (1989), Carroll et al. (1990); for recent results by a different approach see e.g., van Es & Hoogendoorn (1988)). Hence this provides a practical example in the first class and, as we have seen, errors in variables models belong to the second class. Here we will restrict ourselves to convolutions.

EXAMPLE 2.1. Let us first briefly consider convolution on $\mathbf{G} = \mathbf{R}$ with $\varphi \in L_1(\mathbf{R}, \mathcal{B}, \lambda)$, symmetric about 0, so that $\mathbf{H} = L_2(\mathbf{R}, \mathcal{B}, \lambda)$. In this case also $L_2(\mathcal{X}, \mathcal{A}, \mu) = L_2(\mathbf{R}, \mathcal{B}, \lambda)$, U is the usual Fourier transform $\mathcal{F}$, and γ the characteristic function of φ denoted by $\tilde{\varphi} = \sqrt{2\pi}\mathcal{F}\varphi$. With all these quantities made explicit the regularized inverse is easily

computed. The rate on the right in (2.8) can be improved by restricting f to a Sobolev space; it is still possible to determine U explicitly (Carroll et al.).

EXAMPLE 2.2. Let us next turn to convolution on $\mathbf{G} = (0,\infty)$ with a function $\varphi \in L_1\big((0,\infty),\mathcal{B},\mu_+\big)$, symmetric in the sense that $\varphi(x^{-1}) = \varphi(x)$ for $x \in (0,\infty)$. Here $\mathbf{H} = L_2\big((0,\infty),\mathcal{B},\mu_+\big)$ and it turns out that $L_2(\mathcal{X},\mathcal{A},\mu) = L_2(\mathbf{R},\mathcal{B},\lambda)$, with $Uf = \mathcal{F}_+ f = \mathcal{F}\ \big(f(\exp)\big)$ for $f \in L_2\big((0,\infty),\mathcal{B},\mu_+\big)$. Finally we have that $\gamma = \tilde{\varphi}_+ = \sqrt{2\pi}\mathcal{F}_+\varphi$. Surprisingly Laplace transforms can be related to this kind of convolution operator. In a statistical context the Laplace transform shows up in a mixture of exponential distributions. This connexion with the Laplace transform is considered in van Rooij & Ruymgaart (1990).

3. Density estimation on Stiefel manifolds

Downs gives an interesting example from vectorcardiography where the sample elements consist of pairs of orthonormal vectors in $\mathbf{R}^3$; such sample points lie in the Stiefel manifold
$\mathbf{S}$ (3,2). The general Stiefel manifold $\mathbf{S}(d,k)$ is defined as the submanifold of $\mathbf{M}(d,k)$, the set of $d \times k$ matrices ($d,k \in \mathbf{N}$ with $1 \le k \le d$) which could be identified with $\mathbf{R}^{kd}$, given by

$$\mathbf{S}(d,k) = \{O \in \mathbf{M}(d,k) : O^t O = I_k\}, \tag{3.1}$$

where I_k denotes the $k \times k$ identity matrix. The Euclidean inner product in $\mathbf{M}(d,k)$ satisfies $\langle M_1, M_2\rangle = \operatorname{tr} M_1^t M_2$. It is clear that $\mathbf{S}(d,k)$ is a compact subset of $\sqrt{k}\mathbf{S}^{kd-1}$. The Stiefel manifold is not only of interest in its own right but it generalizes both the sphere $\mathbf{S}^{d-1}(k=1)$ and the orthogonal group $\mathbf{O}(d)$ $(k=d)$.

Let $Y_1,\ldots,Y_n$ denote i.i.d. random elements in $\mathbf{S}(d,k)$ with common probability distribution Q. The Haar probability measure on $\mathbf{O}(d)$ induces an invariant probability measure $\overline{\mu}$ on $\mathbf{S}(d,k)$. Let us assume that Q has a continuous positive density g with respect to $\overline{\mu}$. Such a density can be approximated in the following way. Consider the subsets

$$C_O(t) = \{A \in \mathbf{S}(d,k) : \operatorname{tr} A^t O \ge t\}, \quad O \in \mathbf{S}(d,k), t \in [-k,k]. \tag{3.2}$$

We refer to these sets as caps (with center O and height t) because in the special case of the sphere they have such a form. Let us write

$$v(t) = \overline{\mu}\big(C_O(t)\big), \ \textit{independent of } O \in \mathbf{S}(d,k), \tag{3.3}$$

due to the invariance of $\overline{\mu}$. It can be shown that

$$P\big(C_O(t)\big)/v(t) \to g(O), \ \textit{as } t \uparrow k. \tag{3.4}$$

This suggests the estimators

$$\widehat{g}_n(O) = \widehat{P}_n\big(C_O(t_n)\big)/v(t_n), \textit{ for suitable } t_n \in [-k,k]. \tag{3.5}$$

where $\widehat{P}_n(D)$ is the empirical measure of a Borel set $D \subset \mathbf{S}(d,k)$.

It should be noted that a cap $C_O(t)$ can be produced by intersecting $\mathbf{S}(d,k)$ with a properly chosen closed halfspace in $\mathbf{M}(d,k)$. Let $\mathcal{H}$ denote the class of all such halfspaces. It is well-known that sharp exponential bounds can be obtained for the probabilities on the local fluctuations

$$\mathbf{P}\{\sup_{E\subset\mathcal{H}(D)} |\widehat{P}_n(E) - P(E)| \geq \lambda\}, \lambda \geq 0, \tag{3.6}$$

where $\mathcal{H}(D)$ is the class of subsets obtained by intersecting the measurable $D \subset \mathbf{S}(d,k)$ with all $H \in \mathcal{H}$. With the aid of these bounds the following result can be shown in an almost standard way.

THEOREM 3.1. *Let g be continuous on $\mathbf{S}(d,k)$ with strictly positive minimum, and let $c_n \to \infty$, as $n \to \infty$. Then we have*

$$\left(\frac{nv(t_n)}{c_n|\log v(t_n)|}\right)^{\frac{1}{2}} \sup_{O\in\mathbf{S}(d,k)} |\widehat{g}_n(O) - g_n(O)| \to 0, \textit{ as } n \to \infty, \textit{ a.s.}, \tag{3.7}$$

provided that

$$v(t_n) \to 0, \textit{ and } nv(t_n)/\{c_n|\log v(t_n)|\} \to \infty, \textit{ as } n \to \infty, \tag{3.8}$$

where $g_n(O) = \mathbf{E}\widehat{g}_n(O) = P\big(C_O(t_n)\big)/v(t_n)$.

Comparison of the nonrandom g_n with g follows the usual pattern and combination yields at least the MISE consistency

$$\mathbf{E}\|\widehat{g}_n - g\|^2 = \mathbf{E}\int_{\mathbf{S}(d,k)} \{\widehat{g}_n(O) - g(O)\}^2 d\overline{\mu}(O) \to 0, \textit{ as } n \to \infty, \tag{3.9}$$

as required in Section 2. Condition (3.8) is formulated in terms of the volumes $v(t_n)$. For practical implementation, however, it is the smoothing parameter t_n that we have to choose so that we have to "solve" condition (3.8) for t_n. With the help of Morse's lemma it is possible to convert a rate of $v(t_n)$ into a rate of t_n; see Hendriks et al. (1990) for the details.

4. Regularized deconvolution on the circle

Continuing the examples in Section 2 let us now consider in a little more detail the convolution on $\mathbf{G} = \mathbf{S} = [0,1)$ with a function $\varphi \in L_1([0,1),\mathcal{B},\mu_0)$, symmetric in the sense that $\varphi(\ominus x) = \varphi(x)$ for $x \in [0,1)$ and with $\ominus x = \oplus x^{-1}$ where $\oplus$ is addition modulo 1. Clearly in this example we have $\mathbf{H} = L_2([0,1),\mathcal{B},\mu_0)$. A character on $\mathbf{S}$ is a mapping into the complex unit circle $\mathbf{T}$ which is a continuous group homomorphism. With pointwise multiplication and, in the case of the circle, uniform convergence these characters form a locally compact Abelian group Γ, called the character group of $\mathbf{G}$. Every $k \in \mathbf{Z}$, the set of all integers, determines a character ($i^2 = -1$)

$$\gamma_k(x) = \exp(2\pi i k x), x \in [0,1). \tag{4.1}$$

In this way a topological group isomorphism $\mathbf{Z} \to \Gamma$ is obtained. For simplicity we identify Γ with $\mathbf{Z}$.

We have already seen that in this case $\mu_{\mathbf{G}} = \mu_0$, Lebesgue measure on $[0,1)$, and as an invariant measure on Γ we take $\mu_\Gamma = \#$, the counting measure on $\mathbf{Z}$. Writing $d\mu_0(x) = dx$ we have the following formulas

$$(\varphi \circledast f)(x) = \int_0^1 \varphi(x \ominus y) f(y) dy, x \in [0,1), \tag{4.2}$$

where $\varphi \in L_1([0,1),\mathcal{B},\mu_0), f \in L_1([0,1),\mathcal{B},\mu_0) \cap L_2([0,1),\mathcal{B},\mu_0)$,

$$(\mathcal{F}_0\varphi)(k) = \int_0^1 \varphi(x)\exp(2\pi i k x)dx, \quad (\mathcal{F}_0^{-1}\psi)(x) = \sum\nolimits_{k\in\mathbf{Z}} \psi(k)\exp(-2\pi i k x), \tag{4.3}$$

where $k \in \mathbf{Z}, \varphi \in L_1([0,1),\mathcal{B},\mu_0), x \in [0,1), \psi \in \ell_1(\mathbf{Z})$. We also have

$$\mathcal{F}_0(\varphi \circledast f) = (\mathcal{F}_0\varphi)\cdot(\mathcal{F}_0 f), \textit{ for } \varphi, f \in L_1([0,1),\mathcal{B},\mu_0). \tag{4.4}$$

Because $L_2([0,1),\mathcal{B},\mu_0) \subset L_1([0,1),\mathcal{B},\mu_0)$ it is immediate that $\mathcal{F}_0$ is defined on $L_2([0,1),\mathcal{B},\mu_0)$. It now turns out that $L_2(\mathcal{X},\mathcal{A},\mu) = L_2(\mathbf{Z},\mathcal{P}(\mathbf{Z}),\#) = \ell_2(\mathbf{Z})$ and $U = \mathcal{F}_0$, since it can be shown that $\mathcal{F}_0 : L_2([0,1),\mathcal{B},\mu_0) \to \ell_2(\mathbf{Z})$ and that it is unitary indeed.

Hence, with $Kf = \varphi \circledast f$ for $f \in L_2([0,1),\mathcal{B},\mu_0)$ it follows from (2.1) and the remarks made above that $K = \mathcal{F}_0^{-1}M_\gamma\mathcal{F}_0$, where $\gamma = \mathcal{F}_0\varphi$ and $M_\gamma\psi = \gamma\cdot\psi$ for $\psi \in \ell_2(\mathbf{Z})$. Consequently, a regularized inverse is obtained by taking

$$R_m = \mathcal{F}_0^{-1}M_{\rho_m}\mathcal{F}_0, \textit{ where } M_{\rho_m}\psi = \rho_m\cdot\psi \textit{ for } \psi \in \ell_2(\mathbf{Z}), \tag{4.5}$$

and ρ_m as in (2.3) with $\gamma = \mathcal{F}_0\varphi$ (see also above) and $\mathcal{X} = \mathbf{Z}$. Given any $g \in L_2([0,1), \mathcal{B}, \mu_0)$ we can make this inversion more explicit by putting $R_m g = \mathcal{F}_0^{-1}(\rho_m \cdot \mathcal{F}_0 g) = \xi$. By formal manipulation (that can be justified) we obtain $\mathcal{F}_0\xi = \rho_m \cdot (\mathcal{F}_0 g) = (\mathcal{F}_0\mathcal{F}_0^{-1}\rho_m).\ (\mathcal{F}_0 g) = \mathcal{F}_0\big((\mathcal{F}_0^{-1}\rho_m) \circledast g\big)$ according to (4.4), from which we may deduce that $R_m g = (\mathcal{F}_0^{-1}\rho_m) \circledast g$. Because we know that both g and the original f are real functions we should take the real part in (4.3) and replace $\exp(2\pi ikx)$ by $\cos 2\pi kx, x \in [0,1)$. Thus we finally obtain

$$(\mathcal{F}_0^{-1}\rho_m)(x) = \sum\nolimits_{k:(\mathcal{F}_0\varphi)(k)\geq 1/m} \frac{\cos 2\pi kx}{(\mathcal{F}_0\varphi)(k)}, \quad x \in [0,1). \tag{4.6}$$

Summarizing we have proved the following result.

THEOREM 4.1. *Let $Kf = \varphi \circledast f, f \in L_2(\mathbf{S}, \mathcal{B}, \mu_0), \varphi \in L_1(\mathbf{S}, \mathcal{B}, \mu_0)$ with φ symmetric. Suppose that $\widehat{g}_n : \Omega \to L_2(\mathbf{S}, \mathcal{B}, \mu_0)$ are measurable estimators of $g = Kf$ with $\mathbf{E}\|\widehat{g}_n - g\| = \delta_n(f) \to 0$, as $n \to \infty$. Then the estimators*

$$\widehat{f}_n = R_{m(n)}\widehat{g}_n = \big(\mathcal{F}_0^{-1}\rho_{m(n)}\big) \circledast \widehat{g}_n, \quad n \in \mathbf{N}, \tag{4.7}$$

with $\mathcal{F}_0^{-1}\rho_{m(n)}$ given by (4.6), are MISE consistent provided that the $m(n), n \in \mathbf{N}$, satisfy (2.6).

EXAMPLE 4.1. For the practical importance of statistics with values in $\mathbf{S}$ we refer to Mardia (1975). Let us assume that we can observe $Y_1 = X_1 \oplus Z, \cdots, Y_n = X_n \oplus Z_n$, and that the known density of the Z_i is the so-called wrapped standard normal density

$$\varphi_0(x) = \sum\nolimits_{k\in\mathbf{Z}} \varphi(k+x), x \in [0,1), \tag{4.8}$$

where φ is the standard normal density on $\mathbf{R}$. For the computation of (4.5) we note that

$$(\mathcal{F}_0\varphi_0)(k) = \frac{1}{\sqrt{2\pi}}\int_0^1 \{\sum\nolimits_{k\in\mathbf{Z}} \exp(-\frac{1}{2}(x+k)^2 + 2\pi ikx)\}dx = \varphi(2\pi k), \quad k \in \mathbf{Z}. \tag{4.9}$$

This implies that $\{k \in \mathbf{Z} : (\mathcal{F}_0\varphi_0)(k) \geq 1/m\} = \{k \in \mathbf{Z} : |k| \leq a_m\}$, where $a_m = \{(1/2\pi^2)\log(m/\sqrt{2\pi})\}^{\frac{1}{2}}$. In this way we find the estimators

$$\widehat{f}_n = \{\sum\nolimits_{|k|\leq a_{m(n)}} \exp(2(\pi k)^2)\cos\big(2\pi k(\cdot)\big)\} \circledast \widehat{g}_n(\cdot), \tag{4.10}$$

for suitably chosen $m(n) \to \infty$, as $n \to \infty$.

Often φ will be a smooth density and if it is the density $g = \varphi \circledast f$ will also be smooth. If f is not a priori known to be smooth itself, the problem of recovering such

a possibly non-smooth f with the aid of a sample from the smooth g is hard. The difficulty of the inversion is reflected in the speed of convergence. For convolution on $\mathbf{R}$ the dependence of the rate on the number of derivatives is well-known (Carroll & Hall (1988), Fan (1988), Zhang (1990)); in Carroll et al. (1990) the convolution is restricted to a Sobolev space and the operator is again shown to be related with a multiplication but through unitary equivalence with a different Hilbert function space. For convolution on $(0,\infty)$ a similar kind of result is obtained in van Rooij & Ruymgaart (1990). The case of the circle $\mathbf{S}$ is very similar to the real line so that an outline will suffice.

Let us write $f' = f^{(1)}$ for the derivative of f on $\mathbf{S} = [0,1)$ in the sense that $\int_a^b f'(x)dx = f(b) - f(a)$ for all $0 \le a < b < 1$, and restrict ourselves to the class of smooth functions $\mathbf{H} = \{f \in L_2(\mathbf{S},\mathcal{B},\mu_0) : f^{(1)},\cdots,f^{(p)} \in L^2(\mathbf{S},\mathcal{B},\mu_0)\}$, for some $p \in \mathbf{N}$. This class is known to be a Hilbert space, called Sobolev space (Kantorovich & Akilov (1982)) when the inner product

$$\langle f_1, f_2\rangle_{\mathbf{H}} = \int_0^1 \{f_1(x)f_2(x) + \sum\nolimits_{j=1}^{p} f_1^{(j)}(x)f_2^{(j)}(x)\}dx, f_1, f_2 \in \mathbf{H}, \tag{4.11}$$

is used. The restriction of the convolution to $\mathbf{H}$ will be denoted by $K_{\mathbf{H}}f = \varphi \circledast f, f \in \mathbf{H}$. Because $(\varphi \circledast f)' = \varphi \circledast f'$ it is clear that $K_{\mathbf{H}} : \mathbf{H} \to \mathbf{H}$ and is strictly positive Hermitian. For the unitary equivalence we still use the Fourier transform $\mathcal{F}_0$ as defined in (4.3), but with different domain and range viz. $\mathcal{F}_0 : \mathbf{H} \to \ell_{2,\nu}(\mathbf{Z}) = L_2(\mathbf{Z},\mathcal{P}(\mathbf{Z}),\nu)$, where the measure ν has density

$$(d\nu/d\#)(k) = 1 + \sum\nolimits_{j=1}^{p}(2\pi k)^{2j}, k \in \mathbf{Z}, \tag{4.12}$$

with respect to the counting measure $\#$ on $\mathbf{Z}$.

For the isometry note in particular that, for any $\mathcal{F}_0 h = f$ with $h \in \mathbf{H}$ and $f \in \ell_{2,\nu}(\mathbf{Z})$, we have

$$\begin{aligned} \|h\|_{\mathbf{H}}^2 &= \|h\|^2 + \sum\nolimits_{j=1}^{p} \|h^{(j)}\|^2 = \|\mathcal{F}_0^{-1}f\|^2 + \sum\nolimits_{j=1}^{p} \|(\mathcal{F}_0^{-1}f)^{(j)}\|^2 = \\ &= \|f\|^2 + \sum\nolimits_{j=1}^{p}(2\pi k)^{2j}\|\mathcal{F}_0^{-1}((\cdot)^j f(\cdot))\|^2 = \\ &= \|f\|^2 + \sum\nolimits_{j=1}^{p}(2\pi k)^{2j}\|(\cdot)^j f(\cdot)\|^2 = \|\mathcal{F}_0 h\|_\nu, \end{aligned} \tag{4.13}$$

where $\|\cdot\|_\nu$ denotes the norm in $\ell_{2,\nu}(\mathbf{Z})$.

Furthermore, in the case of the circle it can be shown that the density estimators $\widehat{g}_n$ of the type discussed in Section 3 satisfy

$$\mathbf{E}\|\widehat{g}_n - g\| = \mathcal{O}(n^{-4/5}), \ as\ n \to \infty, \tag{4.14}$$

assuming the existence and continuity of g'' on $\mathbf{S}$ and provided that we choose $1-t_n = \mathcal{O}(n^{-1/5})$, as $n \to \infty$. This smoothness condition on g will be automatically satisfied since we will assume that φ is smooth.

THEOREM 4.2. *Suppose that* $\mathsf{E}\|\widehat{g}_n - g\| = \delta_n(f) = \mathcal{O}(n^{-\alpha})$, *as* $n \to \infty$, *for some* $\alpha > 0$ *and let* $\varphi_0 \in L_1(\mathbf{S}, \mathcal{B}, \mu_0)$ *be the wrapped normal density in (4.8). Moreover, let us assume that* $f \in \mathbf{H} \subset L_2(\mathbf{S}, \mathcal{B}, \mu_0)$, *implying that it is* p *times differentiable. Then there exist* $m(n)$ *satisfying (2.6) such that*

$$\mathsf{E}\|\widehat{f}_n - f\| = \mathcal{O}\big((\log n)^{\frac{1}{4}-p}\big), \text{ as } n \to \infty, \tag{4.15}$$

where $\widehat{f}_n$ *is given by (4.7).*

PROOF. Let us first fix an arbitrary $m \in \mathbf{N}$ and note that, with a_m as in Example 4.1 and exploiting the fact that $\mathcal{F}_0$ is an isometry,

$$\begin{aligned} \varepsilon_m(f) &= \|R_m\varphi \circledast f - f\| = \|\mathcal{F}_0^{-1}\big(\rho_m \cdot \mathcal{F}_0(\varphi \circledast f)\big) - f\| = \\ &= \|(\rho_m \cdot \mathcal{F}_0\varphi - 1)\mathcal{F}_0 f\| \leq \\ &\leq \|(\mathbf{1}_{-\infty,-a_m)}) \cdot \mathcal{F}_0 f\| + \|(\mathbf{1}_{(a_m,\infty)}) \cdot \mathcal{F}_0 f\|. \end{aligned} \tag{4.16}$$

The two terms in the upper bound can be dealt with in the same way so let us restrict ourselves to the last one.

Using $\|\cdot\|_\nu$ for the norm in $\ell_{2,\nu}(\mathbf{Z})$ as in (4.13), we find

$$\begin{aligned} \|(\mathbf{1}_{(a_m,\infty)}) \cdot \mathcal{F}_0 f\| &= \left\| \left(\frac{\mathbf{1}_{(a_m,\infty)}}{1+\sum_{j=1}^p (2\pi(\cdot))^{2j}} \right) \cdot (\mathcal{F}_0 f) \cdot \left(1 + \sum_{j=1}^p (2\pi(\cdot))^{2j}\right) \right\| = \\ &= \left\langle \frac{\mathbf{1}_{(a_m,\infty)}}{1+\sum_{j=1}^p (2\pi(\cdot))^{2j}}, (\mathcal{F}_0 f) \right\rangle_\nu \leq \|\mathcal{F}_0 f\|_\nu \cdot \left\| \frac{\mathbf{1}_{(a_m,\infty)}}{1+\sum_{j=1}^p (2\pi(\cdot))^{2j}} \right\|_\nu = \\ &= \|f\|_{\mathbf{H}} \cdot \left\{ \sum_{k>a_m} \frac{1}{1+\sum_{j=1}^p (2\pi k)^{2j}} \right\}^{\frac{1}{2}} = \mathcal{O}\big((\log m)^{\frac{1}{4}-\frac{1}{2}p}\big), \text{ as } m \to \infty. \end{aligned} \tag{4.17}$$

A proper balance between the values of the quantities on the right in (2.8) in the present case is obtained by choosing $m(n) \sim n^{\alpha/(2\alpha+2)}$, which results in the order on the right in (4.15). Q.E.D.

5. Errors in variables on the sphere

The sphere $\mathbf{S}^{d-1}$ is a homogeneous space under the orthogonal group $\mathbf{O}(d)$, and an element of $\mathbf{O}(d)$ acts on an element of $\mathbf{S}^{d-1}$, where this action will be multiplicatively written as in matrix multiplication. Throughout this section let μ be the invariant probability measure on $\mathbf{O}(d)$ and $\overline{\mu}$ the uniform probability measure on $\mathbf{S}^{d-1}$. All densities will be understood with respect to these measures. In Chang's (1989) spherical regression model an important role is played by random elements on $\mathbf{S}^{d-1}$ with density symmetric about a certain point. For various reasons (e.g. model tests) it might be interesting to describe such a density in terms of the symmetry point and a random element that represents the noise about that point, just like in the case of the real line. It turns out that, given $a \in \mathbf{S}^{d-1}$ and a random element $Z : \Omega \to \mathbf{O}(d)$, the random element $Za : \Omega \to \mathbf{S}^{d-1}$ always has a symmetric density about a. Moreover, all symmetric densities can be generated in this way. See Hendriks & Janssen (1990) where this problem is solved in the more general context of homogeneous spaces.

In a similar way distributions in the spherical regression model with errors in variables (Chang (1989)) can be generated by random elements. Given $X : \Omega \to \mathbf{S}^{d-1}$ and $Z : \Omega \to \mathbf{O}(d)$, one can consider the composition $Y = ZX : \Omega \to \mathbf{S}^{d-1}$. Under the usual assumptions let the density f of X be unknown, the density φ of Z be given, and suppose that we observe $Y_1 = Z_1\, X_1, ..., Y_n = Z_n X_n$ with density g. The density g can be estimated e.g. in the manner of Section 3 by $\widehat{g}_n$, say, using the sample elements Y_i. Let us here again focus on the nonparametric problem of recovering f from $\widehat{g}_n$; this is an extension of the errors in variables model on the circle considered in Section 4.

In spite of the fact that the composition is between functions defined on different spaces and $\mathbf{O}(d)$ is not even commutative, the convolution can still very similarly be defined. For $f \in L_2(\mathbf{S}^{d-1}, \overline{\mu})$ and given $\varphi \in L_1(\mathbf{O}(d), \mu)$, both spaces being endowed with their Borel σ-fields, we define the operator

$$g(x) = (Kf)(x) = (\varphi \circledast f)(x) = \int_{\mathbf{O}(d)} \varphi(O) f(O^t x) d\mu(O), \quad x \in \mathbf{S}^{d-1}, \tag{5.1}$$

where obviously $K : L - 2(\mathbf{S}^{d-1}, \overline{\mu}) \to L_2(\mathbf{S}^{d-1}, \overline{\mu})$.

This operator is easily seen to be bounded. The question for which φ the operator is injective needs to be answered. Given an injective operator K, inversion turns out to be ill-posed again. For regularization of the inverse irreducible representations will be an essential tool (Hewitt & Ross (1963)).

Acknowledgement

This work was done while the author was at the Department of Mathematics, Katholieke Universiteit Nijmegen, The Netherlands.

References

1. Carroll, R.J. and Hall, P. (1988) "Optimal rates of convergence for deconvolving a density", J. Amer. Statist. Assoc. 83, 1184-1186.
2. Carroll, R.J., van Rooij, A.C.M. and Ruymgaart, F.H. (1990) "Theoretical aspects of ill-posed problems in statistics", Tech. Report, Kath. Univ. Nijmegen.
3. Chang, T. (1989) "Spherical regression with errors in variables", Ann. Statist. 17, 293-306.
4. Es, A.J. van, and Hoogendoorn, A.W. (1988) "Kernel estimation in Wicksell's corpuscle problem", Biometrika, to appear.
5. Fan, J. (1988) "On the optimal rates of convergence for the nonparametric deconvolution problem", Tech. Report #157, Dept. Statist., Univ. of Cal., Berkeley.
6. Hendriks, H.W.M. and Janssen, J.H.M. (1990) "Densities on homogeneous spaces induced by random elements in the group", Tech. Report, Kath. Univ. Nijmegen, to appear.
7. Hendriks, H.W.M., Janssen, J.H.M., and Ruymgaart, F.H. (1989) "Strong uniform convergence of density estimators on Stiefel manifolds", Tech. Report, Kath. Univ. Nijmegen.
8. Hewitt, E. and Ross, K.A. (1963) Abstract Harmonic Analysis, I, Springer, New York.
9. Mardia, K.V. (1975) Statistics of Directional Data, Acad. Press, New York.
10. Nychka, D.W. and Cox, D.D. (1989) "Convergence rates for regularized solutions of integral equations from discrete noisy data", Ann. Statist. 17, 556-572.
11. O'Sullivan, F. (1986) "A statistical perspective on ill-posed inverse problems", Statist. Sci 4, 502-527.
12. Rooij, A.C.M. van and Ruymgaart, F.H. (1990) "Inversion of noisy Laplace transforms", Tech. Report, Kath. Univ. Nijmegen.
13. Tikhonov, A.N. and Arsenin, V.Y. (1977) Solutions of Ill-Posed Problems, Wiley, New York.
14. Titterington, D.M. (1985) "Common structure of smoothing techniques in statistics", Internat. Statist. Rev. 53, 141-170.
15. Vapnik, V. (1982) Estimation of Dependences Based on Empirical Data, Springer, New York.
16. Zhang, C.H. (1990) "Fourier methods for estimating mixing densities and distributions", Ann. Statist., to appear.

LIST OF ATTENDANTS

Dr. Rashid Ahmad
Department of Mathematics
Livingstone Tower
26 Richmond St.
University of Strathclyde
Glasgow G1 1XH, Scotland
UNITED KINGDOM

Professor Michael Akritas
20 G. Karaiskaki Street
19009 Rafina, Attica
GREECE

Professor Emin Anarim
Bogaziçi University
Department of Electrical
and Electronic Engineering
80815 Bebek - Istanbul
TURKEY

Professor Barry C. Arnold
Department of Statistics
University of California
Riverside, CA 92521

Mr. Bjφrn H. Auestad
Department of Mathematics
University of Bergen
5007 Bergen
NORWAY

Professor Adelchi Azzalini
Department of Statistical Sciences
University of Padua
Via S Francsco 33,
I-35121, Padua
ITALY

Dr. Nihat Badem
Dokuz Eylül University
Faculty of Engineering
& Architecture
Applied Mathematics Division
Bornova - IZMIR
TURKEY

Professor Andrew R. Barron
Department of Statistics
University of Illinois
725 S. Wright Street
Champaign, IL 61820

Ms. Gülay Basarir
Hacettepe University
Faculty of Science
Department of Statistics
06532 Beytepe - Ankara
TURKEY

Professor Rudolf Beran
Department of Statistics
367 Evans Hall
University of California
Berkeley, CA 94720

Professor Dr. Alain Berlinet
Université de Montpellier
Unité de Biométrie
ENSAM-INRA-USTL
9, place Pierre VIALA
F-34060 Montpellier Cédex
FRANCE

Professor P.K. Bhattacharya
Division of Statistics
469 Kerr Hall
University of Califonria
Davis, CA 95616

Professor Denis Bosq
Laboratoiré de Statistique
Université Paris VI
74 Rue Dunois
Paris, 75013
FRANCE

Ms. Maria Mori Brooks
Department of Statistics, CB#3260
Phillips Hall
University of North Carolina
Chapel Hill, NC 27599-3260

Professor Prabir Burman
Division of Statistics
469 Kerr Hall
University of California
Davis, CA 95616

Dr. Karen Chan
Department of Statistics
University of Glasgow
Mathematics Bldg., Rm 113
Glasgow, G12 800
UNITED KINGDOM

Professor Miklós Csörgő
Department of Mathematics
& Statistics
Carleton University
Ottawa, Ontario
CANADA K1S 5B6

Professor Antonio Cuevas
Departamento de Mathematicas
Facultad de Ciencias
Universidad Autonoma
Canto Blanco, 28049-Madrid
SPAIN

Professor Paul Deheuvels
7 Avenue du Château
92340 Bourg-la-Reine
FRANCE

Professor Miguel A. Delgado-Gonzalez
C/Linneo 32, 4a
Madrid 28005
SPAIN

Professor Luc Devroye
School of Computer Science
McGill University
Montreal PQ
CANADA H3A 2K6

Ms. Handan Dingiloğlu
Department of Statistics
Middle East Technical University
Ankara 06531
TURKEY

Dr. Joachim Engel
Sonderforschungsbereich 123
Universiät Heidelberg
Im Neuenheimer Feld 294
6900 Heidelberg
FEDERAL REPUBLIC OF GERMANY

Professor Ricardo Fraiman
Centro de Matemática
Eduardo Acevedo 1139 (11200)
Montevideo
URUGUAY

Mr. Mauro Gasparini
Department of Mathematics
University of Bologna
Bologna
ITALY

Professor Theodor Gasser
Zentralinst für Seelische
Gesundheit Abt Biostatistik
Postfach 5970
D-6800 Mannheim 1
FEDERAL REPUBLIC OF GERMANY

Professor Alexander Georgiev
Department of Biometry
Medical University of South Carolina
171 Ashley Avenue
Charleston, SC 29425-2503

Professor Wenceslao Gonzalez-Manteiga
Universidad de Santiago de Compostela
Departamento de Estadistica E.I.O.
Facultad de Mathematicas
Santiago de Compostela (La Coruña)
SPAIN

Dr. Pedro L. Gozalo
Instituto Universitario
Ortega Ygasset
Madrid
SPAIN

Dr. Gérard Grégoire
Laboratory of Modelling &
Computation, B.P. 53X,
38041 Grenoble Cédex
FRANCE

Dr. László Györfi
Technical University of Budapest
Stoczek u.2, Budapest XI
1521 HUNGARY

Professor Wolfgang Härdle
Universite Catholique de Louvain
34 Voie du Romas Pays
B-1348 Louvain-La-Neuve
BELGIUM

Professor Nancy Heckman
Department of Statistics
2021 West Mall
University of British Columbia
Vancouver, B.C.
CANADA V6T 1W5

Dr. Dimitrios Ioannides
Department of Mathematics
University of Patras
26110 Patras
GREECE

Professor S. Rao Jammalamadaka
Department of Statistics and
Applied Probability
University of California
Santa Barbara, CA 93106

Dr. Michael Chris Jones
Faculty of Mathematics
The Open University
Walton Hall, Milton Keynes
MK7 6AA UNITED KINGDOM

Professor Abram Kagan
Department of Mathematics
University of Maryland
College Park, MD 20742

Professor Dimitri Kazakos
Dept. of Electrical Engineering
Thorton Hall
University of Virginia
Charlottesville, VA 22901

Dr. C.P. Kitsos
23 Bithinias Street
Kalogreza (N. Ionia)
14234 Athens
GREECE

Professor Vassilios Klonias
Department of Mathematics
University of Crete
P.O.Box 1470
Iraklion, Crete
GREECE

Professor George Kokolakis
Department of Mathematics
National Technical University
of Athens
Zografou Campus
15773 Athens
GREECE

Professor Adam Krzyżak
Department of Computer Science
Concordia University
1455 De Maisonneuve Blvd. West
Montreal, Quebec
CANADA H3G 1M8

Dr. Boris Levit
Department of Mathematics & Informatics
University of Leiden
P.O. Box 9512
2300 RA Leiden
THE NETHERLANDS

Mr. Panagiotis Lorentziadis
25 Lelas Karagianni Street
Ptateia Americis
Athens 112 52
GREECE

Professor Y.P. Mack
Division of Statistics
469 Kerr Hall
University of California
Davis, CA 95616

Dr. Enno Mammen
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
6900 Heidelberg 1
FEDERAL REPUBLIC OF GERMANY

Professor James Stephen Marron
Department of Statistics
University of North Carolina
Chapel Hill, NC 27514

Professor Pascal Massart
U.A.-CNRS 743 Statistique Appliquee
Universite Paris Sud
Batiment 425, Orsay
91405 FRANCE

Professor Hans-Georg Müller
Division of Statistics
469 Kerr Hall
University of California
Davis, CA 95616

Mr. Andrew Neath
Division of Statistics
469 Kerr Hall
University of California
Davis, CA 95616

Mr. David J. Olive
Department of Statistics
101 Illini Hall
725 South Wright Street
University of Illinois
Champaign, IL 61820

Professor Takis Papaioannou
Department of Mathematics,
Statistics Unit
University of Ioannina
45110 Ioannina
GREECE

Professor Vassilios Papakonstantinou
Department of Mathematics
University of Patras
26110 Patras
GREECE

Professor Miroslaw Pawlak
Dept. of Electrical Engineering
The University of Manitoba
Winnipeg, Manitoba
CANADA R3T 2N2

Professor D.T.A. Pham
C.N.R.S. Laboratoire TIM3
Laboratoire TIM3, BP 53X
38041 Grenoble Cedex
FRANCE

Dr. Dimitris Politis
17 Ergotimou Street
11634 Athens
GREECE

Ms. Kathy Prewitt
Division of Statistics
469 Kerr Hall
University of California
Davis, CA 95616

Professor M.B. Rao
Department of Statistics
North Dakota State University
300 Minard
Fargo, ND 58105

Mr. David M. Reboussin
Department of Statistics
1210 West Dayton Street
University of Wisconsin
Madison, WI 53706

Professor Dr. R.-D. Reiss
Universität-GH-Siegen
Mathematics Department
Fachbereich 6 - Mathematik
Hölderlinstr. 3, 5900 Siegen
FEDERAL REPUBLIC OF GERMANY

Professor John L. Rice
Department of Mathematics, C-012
University of California, San Diego
La Jolla, CA 92093

Professor Murray Rosenblatt
Department of Mathematics C-012
University of California, San Diego
La Jolla, CA 92093

Professor George G. Roussas
Division of Statistics
469 Kerr Hall
University of California
Davis, CA 95616

Professor Frits H. Ruymgaart
Faculteit Wiskunde en Informatica
Katholieke Universiteit
Toernooiveld, 6525 ED Nijmegen
THE NETHERLANDS

Professor Alexander Samarov
Massachusetts Institute
of Technology
1 Amherst Street
Building E40-148
Cambridge, MA 02139

Mr. Antonis Sagris
7587 Champagnneor St. #3
Montreal, Quebec H3N 2K1
CANADA

Professor Pascal Sarda
Laboratoire de Statistque
Institut de Mathematiques - U.S.T.L.
Place Eugene Bataillon
34095 Montpellier
Cédex, FRANCE

Professor William R. Schucany
Department of Statistics
Southern Methodist University
Dallas, TX 75275

Dr. Getulio Borges da Silveira
Department of Economics
London School of Economics
Houghton St Aldwych
London WC2A 2AE
UNITED KINGDOM

Professor Ulrich Stadtmüller
Department of Mathematics
University of Ulm
Oberer Eselsberg
D-7900 Ulm
FEDERAL REPUBLIC OF GERMANY

Professor Winfried Stute
Mathematisches Institut
University of Giessen
Arndtstrasse 2, D-6300 Giessen
FEDERAL REPUBLIC OF GERMANY

Dr. Hüseyin Tatlidil
Department of Statistics
Faculty of Science
Hacettepe University
06532 Beytepe - Ankara
TURKEY

Professor Lanh Tat Tran
Department of Mathematics
Indiana University
Bloomington, IN 47405

Professor K. Truong
Department of Biostatistics
University of North Carolina
at Chapel Hill
CB#7400, Rosenau Hall
Chapel Hill, NC 27599-7400

Dr. A.B. Tsybakov
Institute for Problems of
Information Transmission
Academy of Sciences of the U.S.S.R.
Ermolovoy 19
U.S.S.R.

Ms. Ülkü Uzunogullari
6. Sokak 21/9
Bahcelievler/Ankara
TURKEY

Professor Edward van der Meulen
Department of Mathematics
Katholieke Universiteit Leuven
Celestijnenlaan 200 B
B-3030 Heverlee
BELGIUM

Professor Bert van Es
Department of Mathematics &
Computer Science
University of Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
THE NETHERLANDS

Professor Philippe Vieu
Laboratoire de Statistique
et Probabilities
University Paul Sabatier
118 Route de Narbonne
31062 Toulouse, Cedex
FRANCE

Mr. Bruce E. Wahlen
Department of Mathematics, C-012
University of California, San Diego
La Jolla, CA 92093

Dr. Matthew Wand
Universitat Heidelberg
Sonderforschungsbereich 123
Stochastische mathematische Modelle
Im Neuenheimer Feld 294
D-6900 Heidelberg 1
FEDERAL REPUBLIC OF GERMANY

Professor Jane-Ling Wang
Division of Statistics
469 Kerr Hall
University of California
Davis, CA 95616

Professor Sidney J. Yakowitz
Department of Systems Engineering
University of Arizona
Tucson, AZ 85721

Professor Yannis Yatracos
Department of Mathematics
& Statistics
University of Montreal
CP 6128, Suc. A
Montreal, Quebec
H3C 3J7 CANADA

Contributed Papers

ANTONIADIS, A. AND GRÉGORIE,* G.: Nonparametric estimation in change-point hazard rate models for censored data.

ARNOLD, B.C.: Nonparametric estimation of Lorenz curves.

BIRGÉ, L. AND MASSART,* P.: Lower bounds for rates of convergence of M.L. and related estimators.

BROOKS, M.M.: Bandwidth selection for intensity estimates.

DINGILOĞLU,* H. AND SUNGUR, E.: The use of copulas in extreme theory.

ENGEL, J.: Semiparametric model estimation in nonlinear regression.

FALK, M. AND REISS,* R.-D.: Estimating conditional curves: Nearest neighbor approach.

GOZALO, P.L.: A consistent model specification test for kernel estimation of regression on functions models.

KAGAN, A.: Quasi likelihood and the projection method for solving the maximum likelihood equation.

KAZAKOS, D.: Asymptotically tight error bounds for multi-hypothesis testing through large deviation theory.

KITSOS, C.P.: Nonparametric features on optimal experimental design.

KOKOLAKIS, G.E.: Nonparametric Bayesian prediction in multivariate binary discriminant analysis.

POLITIS, D.: Bootstrapping time series via autoregressive approximations.

RICE, J.: Bayesian restoration of single-channel patch champ recordings.

STADTMÜLLER, U.: Estimation in a multiplicative regression model.

UZUNOĞULLARI,* Ü. AND WANG,** J.-L.: Estimation of the hazard function and its derivatives under random truncation models.

WAND, M.P.: Exact mean integrated squared error for dependent data.

YAKOWITZ, S.: Nonparametric regression for Markov sequences without mixing assumptions.

* Presenter of paper

** Also, a participant

INDEX

Abelian groups 680
Abramson 141
absolute error 566,569
accuracy of the bootstrap confidence intervals 114
actual average squared errors for estimation 203
adaptive 145,146,159
additive models 499
Ahmad 477
almost i.i.d. errors 115
almost stationary processes 389
almost sure convergence 509
AMISE 254
α-mixing 273,275,444,446
approximation error 561,567
AR(1) process 378
AR(2) model 383
asymptotic covariance 321
asymptotic distribution 321,648
asymptotic efficiency 167
asymptotic error exponent 634
asymptotic linearity 288
asymptotic normality 152,443,444,450,456
asymptotically efficient 321
asymptotically normal 443
autocorrelation 378
autocorrelation operator 409
autocovariance function 553
automatic estimate 16
automatic kernel estimates 15
automatic selection 201,548
autoregressive function(s) 271,278
autoregressive model 607
autoregressive processes 272
average curve 99
average derivative estimates 167
average derivative estimator 169
average squared error 274
Azuma inequality 352

bandwidth(s) 20,67,148,251,589,590
bandwidth choice 141
bandwidth factorization 255
bandwidth selection 241,274
Bayes estimates 339
Bayes optimal probability of error 567
Bayes risk consistencies 347
Bell 477
Bell-Pitman function 487
Beran 318
Bernstein('s) inequality 570,572,575
Bernstein-type inequality(-ies) 272,573
Bhattacharya and Mack 144
bias 379
biased cross-validation 256
Bickel 226
Bickel and Wichura 145
Bieren's estimate 16
Billingsley 145,225
bispectra 406
Blaschke's selection theorem 228
block-circulant 202
blur matrices 205
Boneva, Kendall and Stepanov 226
Bonferroni method 121
Boos 226,236
bootstrap 111,114,440,501,577,585,586
bootstrap confidence intervals 122
boundary effects 114
bounded influence function 297
bounded influence regression 297
bounded Lipschitz distance 232
bounds of the statistical risk 561
Bowman 238
Bradley 421
Bretagnolle-Hubert kernel 3
Brownian bridge process 410
Brownian bridges 215
Buckley and James 316
bumphunting 126
bump-hunting problem 226
bumps 226

cascade and parallel connections 541
Cauchy density 236,618
censored data 315
censoring variable 432
chaotic systems 531
Chebyshev's inequality 37
Chernoff 147,570,574
Chernoff's bound 31,34
Chernoff's bounding method 33
Choi 477
circle 680
classes of images 669,670

class-fit detectors 478
classification 671,673
classification error 671
classified image 673
clean U-statistic 117
codelength 563,569
compact 682
complexity(-ies) 561,563,566,567,568,570,571
complexity penalty(-ies) 561,562,568
complexity regularization 561-63,565-70,572-73
complexity regularization estimator 573
complexity regularized 569
composite alternative 631
conditonal density estimates 463
conditional distribution 616,617
conditional heteroskedasticity 297
conditional mean 614
conditional median estimator 616
conditional median function 613,614,616
conditional quantiles 617,621,625
conditional variance 618
conditionally independent 321
confidence band 215
confidence bands for the quantile function 215
confidence intervals 114,619
conservative confidence bands 45
consistency 129,130,555
consistent test 631
continuity of a Brownian Bridge process 221
continuous time process 509
continuous time stochastic process 553
control variables 531
convergence rates 249
converges at rate 570
converges to zero at rate 569
convex minorant 193,195
convolution 682,683
corrected v-fold cross-validation 606
correlated residuals 75
covariance operator 509
covariate 587,588,595,597
coverage probability(-ies) 577,578,579,583
Cristóbal, Faraldo and González-Manteiga 320
cross-validated 23,262
cross-validation11,67,241,261,274-77,279,366,603
cross-validation functional 322
Csörgő-Révész-Stute ([CRS]) conditions 19
Cuevas 236
cumulative hazard rate function 316
curve estimation 141

Dabrowska 318
DasGupta and Studden 339
data-based 249
data compression 561
data-driven bandwidth 20
data-driven estimator 26
deconvoluted kernel 615,616,623,624
deconvoluted kernel density estimators 614,615
deconvoluted kernel estimates 623,624
deconvoluted kernel estimator(s) 617,618,621
deconvoluted kernel function(s) 622,625
deconvoluted kernel median function estimators 616
deconvoluted kernel method 614
deconvoluted kernel regression function estimator(s) 616,619
deconvolution 203,614,618,625
degraded images 201
degrees of censorship 321
Deheuvels' estimate 16,40
DeMontricher 125,131
Demmler and Reinsch 344
Demmler-Reinsch basis 344
density 173
density estimates 19
density estimation 125,280,561
density estimation-free confidence bands 217
density estimation on Stiefel manifolds 683
density function 218
dependent competing risks 431
derivative(s) 69,145,160
design-adaptive 619,620,625
design-adaptive method 614
design-adaptive quantile estimators 626
design choice 675
deviation 148
Devroye 227
Devroye and Györfi 419
Dharmadhikari and Joag-dev 235
diagnostic capabilities 129
diagonal expansion 543
diagonal series expansion 543
differential entropy 81
diffusion process 553
dimensionality problem 280
discrete observations 553
distance between 119
distance measure 632

distortion function(s) 566,567,568,569,570
distribution-free bounds 358
distribution function 215
distribution of functionals 112
divided differences 339
Doeblin's hypothesis 443,444
Doob martingale 32,41
double exponential 618
dynamic systems 531

Eddy 148
edge estimation 669,672,675
edge estimator 673,675,677
efficient 145,146,159,164
Efron-Stein inequality 40
eigenvectors 509
elliptically contoured 173
empirical distribution function 19,35,261-62,264
empirical measure 91,633
empirical probability measure 38
empirical processes 19
empirical quantile function 213
energy transfer 406
entropy estimators 81
equispaced grid 670,671,675,676
error distribution 618
errors-in-variables 614,625,679,680
errors in variables model on the sphere 680,681,689
estimated mean squared error 367
estimating derivatives 241
estimation cross-validation 201
estimation of the derivatives 277
exponential bounds 347
exponential inequalities 31
extraneous statistical noise 488

FFTs 202
finite order kernels 4
fixed design 315
fixed design model 112
ϕ-mixing condition 273,445
Fourier integrals 407
Fourier matrix 202
Fourier representation 406
frame-grabbing 205
functional 243
functional laws of the iterated logarithm 19

Gamma 618
Gaussian mixture 477
Gaussian process(-es) 142,148,272
GCV 203
generalization of the MPLE 128
generalized additive models (GAM) 280
generalized kth divided difference 345
generalized Lehmann alternative model(s) 448,647
generalized maximum likelihood estimates 434
geometric ergodicity 444
geometrically ergodic 443
global rates of convergence 358
Good and Gaskins 125,131,226
goodness-of-fit 122
goodness-of-fit detectors 478
Gorodettski 420
Gram-Charlier kernels 10
Grenander 147
Györfi 419
Györfi, Härdle, Sarda and Vieu 419

Hall 231,234,238
halving approach 121
Hausdorff metric 228
Hart and Vieu 419
Hasminskii 672,677
hazard rate estimation 280
h-block cross-validation 603,604,608,609,610
Heckman and Woodroofe 341
Hellinger distance 567,570,573,574
heteroskedastic regression 287
heteroskedasticity 297
Hettmansperger 226
Hettmansperger and Klimko 226
high dimensional linear models 123
higher order kernel(s) 241,255
Hilbert space 679,687
Hilbertian autoregressive process 509
histogram 632
histogram estimate 36,42
Hoeffding 31,570
Hoeffding's inequality 31,33,571,572
homoskedastic 321
Huber 232
Huber's ψ-function 298
hypothesis testing 631

Ibragimov 420,672,677
identification problem 494

I-divergence 89
ill-posed problems 202,679
image analysis 669,676
image and edge estimation 669,674
image and edge estimators 674
image estimation 669,670,672
image estimator 673,675
image restoration 201
IMSL 129
IMSL-subroutines 238
increments 19
index of resolvability 561,563-65,567-69,571-74
indices of approximation 568
inflation rate 362
inflection points 226
influence function 126,127,129
integrable nonparametric models 657
integral equation 543
integrated mean squared errors (IMSE) 144
integrated squared error (ISE) 274
interaction 377
inverse estimation 191
inverse problem 191,195
inverse triangular density 622
inversion 195
IP-method 227
isotropic 407
Izenman and Tran 420

Johnson family of curves 379

Kaplan-Meier 316
Karlin 342
k-dimensional linear parametric model 120
kernel(s) 147,320,389,591
kernel classification rule 347
kernel density estimate(s) 94,463
kernel density estimation 141,241,419
kernel estimate(s) 20,42,126-27,331,443,588-89
kernel estimator(s) 67,112,146,261-62,264,616
kernel function 146,378
kernel nonparametric techniques 531
kernel of order k 146
kernel of order (m,p) 7
kernel-quantile-density estimator 216
kernel regression 361
kernel spectral estimates 122
kernel-type estimator 173
Klonias 126,130,131,251
k nearest neighbour 320
kn-NN estimate 331
Kolmogorov 480
Kolmogorov entropy 138
Kolmogorov-Smirnov distance 42,647
Kolmogorov-Smirnov statistics 35,481
Korostelev 677
Koul, Susarla and Van Ryzin 317,318
Krieger and Pickands 141
k-th nearest neighbor 22,39
Kullback-Leibler loss 234

Laguerre kernels 10
laws of the iterated logarithm 19
Lay 228
Leamer 339
least squares cross-validation 256
least squares estimate 136
least squares method 316
Lebesgue measure of symmetric difference 673
Legendre kernels 10
Legostaeva and Shiryayev 339
life time distributions 431
LIL 21
linear estimate 676
linear estimators 676
linear regression model 315
linear smoothing 676
linewise procedure 676
linewise processing 676
Lipschitz condition 253
local bandwidth 141
local bandwidth choice 367
local means 46
local medians 46
local polynomials 625
locally mixing (LM) 419
log-likelihood based distortions 569
log-odds ratio 567
logistic distortion 566
logistic distortion function 566
logistic regression 566
logistic transformation 621
Loève 130,131
L1-consistent density estimator 81
L1-discrepancy 620
L1-distance 89,632
L1-error 673,674
L1 norm 35,252
L1 strong consistency 463

Lp-error 674
L2-approximation error 563
L2-approximation to the empirical distribution function 6
L2-discrepancy 619
L2-distance 117
Lui 477

Mack and Müller 143
Mallow's CL 204
Mann-Whitney Statistic 647,649,652,653
Marcinkiewicz 130
Markov chain(s) 606
Markov process(-es) 272,443,444,445
Marron 231
martingale 32
martingale difference 635
martingale difference sequences 32
martingale-like 604,606
Masry 419
Massart's bound 35
maximal invariants 487
maximal statistical noise 478
maximum penalized likelihood 125
McDiarmid 32,34
mean absolute difference 206
mean integrated L1 and L2 errors 218
mean integrated square error (MISE) 21,241,252,274
mean squared error (MSE) 144
mean squared errors for estimation 203
measurement errors 614,616,617,620
measures of deviation 413
median(s) 671
M-estimators 297
Miller 316
minimal sufficient statistic 478
minimax Bayes estimate 339
minimax lower bounds 674
minimum complexity density estimation 561,575
minimum complexity estimator 574
minimum description-length criteria 561
minimum description-length paradigm 561
minimum description-length principle 562,567
minimum variance 655
mixing condition(s) 272,509
mixing processes 389
mixture 649
mode estimation 146
mode estimator(s) 159,160
model selection 561,562,565
mollifier 3
Monte Carlo 405
MPLE 126,127,128
MR-estimates 287
multidimensional mean squared error 322
multiparameter Gaussian process 164
multiple kernel estimation procedure for the density 11
multiple minima 209
multivariate density estimates 12
multivariate kernel estimate 247
multivariate nonparametric regression 163
Müller 144,161
Müller and Prewitt 163,164
Müller and Stadtmüller 142,162
Müller and Wang 161,162

Nadaraya 143,322
Nadaraya-Watson 69
Nadaraya-Watson estimate 378
Nadaraya-Watson estimator 143
Nadaraya-Watson weighting 322
Nash 130
Navier-Stokes equation 405
nearest neighbor density estimator 20
nearest neighbors 300
nearest neighbour estimate 16
Nemirovskii 672,677
neural net(s) 576
neural network(s) 561,564,565,575,576
noise estimate 209
nonlinear autoregressive time series 122
nonlinear functions 564,565
nonlinear model 531
nonlinear time series 493,541
nonparametric 553
nonparametric deconvolution feasible 626
nonparametric density 133
nonparametric density estimate 82
nonparametric density estimation 35
nonparametric estimate 443,444
nonparametric estimation 31,431
nonparametric FPE criterion 495
nonparametric maximum likelihood 192,193
nonparametric pilot estimator 226
nonparametric prediction 389
nonparametric regression 111,133,329,357,377,542,587-88, 590,595,597,672,676

nonparametric regression analysis 613
normal 618
not identically distributed errors 119
numerical optimization 130
Nussbaum 340,672,677

observational studies 587
optimal choice of h_n for mean integrated L1 and L2 errors 220
optimal convergence rates 686
optimal local bandwidth 162
optimal rates of convergence 669,674
optimisation 202
ordered data 553
ordinary cross-validation 603-06,608,610
ordinary kernel functions 623,624
ordinary kernel regression function 619
ordinary smooth 617,618
Orear and Cassel 226
orthogonal functions 543
orthogonal polynomials 11
orthogonal polynomials with weights 10
orthogonal series 389
orthogonal series estimate 544
oversmooth bandwidth 114
oversmoothed 207

pair-matching 587,588,595,597
pairwise resampling 116
Papantoni-Kazakos and Kazakos 419
parametric regression model 119
Pareto renewal process 478
Park 477
partition 632
partitioning estimate 333
Parzen 142,147,420
Parzen-Rosenblatt kernel estimate 3
peaks 141
penalized likelihood 340
penalty functional 125
Pham and Tran 419,420
Phillips Curve 363
piecewise concave/convex functions 134
piecewise convex-concave 226,229
piecewise monotone functions 134
pilot estimator(s) 162,254
plug-in estimator 26
plug-in method 256
plug-in rule 67
plug-in techniques 277
pointwise confidence intervals 121
Polasek 339
polynomial regression 561
polynomials 563,564,565,614
Pratt 677
prediction 203
prediction band(s) 577,579,581,582
prediction ellipsoid 581
prediction error 604
prediction interval(s) 577,578
prediction region(s) 578,581,586
predictor 509
Priestley and Chao 142
probablity centered 581,582
probability centering 577
probability density function 81
probability of error 567
project pursuit regression (PPR) 280
projection 499
pseudo-likelihood ratio test 379

quadratic splines 607
quantile(s) 443,444,446
quantile density function 218
quantile function(s) 213,620
quantile matching 652
quantile process 214
quantitative measures of smoothness 133

random delta sequence 14
random design 673
random right censorship 315
randomized procedures 489
rate(s) of convergence 251,569,572,577,579,585,617
rate of improvement 657
recursive kernel density estimates 39
regression 67,339
regression estimation 329
regression function 142,316,613,617,625
regression model 298
regression spline 133
regular and irregular nonparametric problems 657
regularization 201,562
regularized deconvolution 680
regularized deconvolution on the circle 685
regularized inverse 681
relative entropy 574
relative entropy distance 567,574

relative stability 37,39
repeated measurements 377
reproducing kernel 5
reproducing kernel Hilbert-subspace 5
residual resampling 113
resolvability 571,573
restricted minimax estimates 340
ρ-mixing 444,445,446,455
Robinson 419
robust regression 297
Romano 144,147,157,158
root-density 126
Rosenblatt 142,420
Roussas 273,419,420
running mean 671

Sacks and Ylvisaker 339,340
Sager 147
sample distribution function 657
sample of curves 99
Scheffe's theorem 225
score function 218
Scott, Tapia and Thompson 126,131
seasonal fluctuations 389
second order admissibiilty 657
second order optimality theory 657
second order spectra 413
selection bias 563
selection of lags 497
semiparametric 553
semiparametric estimation 647
semiparametric weighted least squares 299
semiparametric weighted M-estimator 300
Sethuraman 226
shape parameters 112,133
shift functions 99
shift operator 103
signal detection 477
Silverman 126,131,158,161,225-27,233-34,236
simple hypothesis 631
simulations 257
simultaneous coverage probability 121
simultaneous error bars 121
Slepian lemma 47
smoothed bootstrap 118
smoothed bootstrap estimate 255
smoothed cross-validation 255
smoothed quantile process 220
smoother 671
smoothing 671,676
smoothing factor 35
smoothing matrix 202
smoothing
parameter(s) 20,144,201,261,267,378,548
smoothing parameter selection 241,268
smoothing techniques 272
smoothness assumptions 133
smoothness classes 136
smoothness conditions 92
Sobolev space(s) 125,564,687
Speckman 340
spline(s) 563,564,604,605,607
squared deviance between the parametric
fit and the non parametric fit 121
squared error 566,569
squared error distortion 566,567
squared error distortion function 569
stable processes 480
stationary Gaussian process 610
stationary process 607
Steele's inequality 41
Stiefel manifold 680,683
Stone 231,234,670,672,677
Strassen 19
strictly positive Hermitian 681,687
strong consistency 329,443
strong mixing condition 419
strong uniform consistency of random
kernel-type density estimates 12
structural average 104
structural functional 102
structural points 99
Stute and González-Manteiga 320
sup-norm 19
super smooth 617,618,625
supremum distance 236
survival function 431
Sweeting 226,229,236
symmetric stable laws 486
system modeling 533

tail condition 93
tailor-designed methods 227
Tapia and Thompson 125,131
test statistic 632
testing statistical hypotheses 477
tightness 145,153,156
time series analysis 541
total variation 631
Tran 420

transformations 605
transformed kernel estimate 16
transition distribution function 443,444
treatment-covariate interaction 587-89,595-96
treatment main effect 588,589
trend curves 377
trigonometric series 563,564
Tsybakov 669,672,676,677
turbogenerator set 533
two-dimensional problem 74
two-sample detectors 478

unbiasedness 554
undersmoothed 207
unemployment rate 362
uniform deconvolution problem 191,192
uniform empirical quantile function 213
uniform mixing samples 463
uniform quantile process 214
uniform renewal processes 478
universal consistency 329
U-statistic 117,591

Vapnik-Chervonenkis bound 38
Vapnik-Chervonenkis dimension 38
variable kernel density estimates 242
variable kernel estimates 39
variance 410
v-fold cross-validation 603,604,605,609

Wahba 344
Watson 143,322
Watson-Nadaraya approach 277
Watson-Nadaraya estimate 273
Watson-Nadaraya kernel estimate 273
Watson-Nadaraya method 272
wave number vector 406
weak consistency 329
weak convergence 144,148,153,156,164
weakly stationary process 553
Weerahandi and Zidek 339
Wegman 419
weighted averaging of rounded points (WRPing) 113
weighted linear regression 620
weighted M-estimates 289
weighted R-estimators 287
Wiener process 149,411,488,554
wild bootstrap 111,119
wild bootstrapping 119
Wolf sunspot numbers 502
wreath product group 487

Yakowitz 419
Yoshihara 419

zero-one distortion 569
zero-one distortion function 566