ON THE NON-CONSISTENCY OF AN ESTIMATE OF CHIU

Luc Devroye
School of Computer Science
McGill University

ABSTRACT. We show that for some densities, a bandwidth selection method of Chiu (1991) for kernel density estimates is not consistent. While the method shows promise for some densities, it should be used with caution.

KEYWORDS AND PHRASES. Density estimation, kernel estimate, convergence, smoothing factor, bandwidth selection, empirical characteristic function, nonparametric estimation.

1991 MATHEMATICS SUBJECT CLASSIFICATIONS: Primary 62G05.

Authors' address: School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2K6. The authors' research was sponsored by NSERC Grant A3456 and FCAR Grant 90-ER-0291.

Introduction.

The purpose of this note is to give examples of densities for which one of Chiu's (1991) bandwidth selectors is not consistent. We consider an i.i.d. sample $X_1, \ldots, X_n$ drawn from a univariate density f, and estimate f by

$$f_{nh}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)$$

where K is the kernel (a function integrating to one), $K_h(x) = (1/h)K(x/h)$, and $h > 0$ is the smoothing factor (Akaike, 1954; Parzen, 1962; Rosenblatt, 1956). The fundamental problem in kernel density estimation is that of the joint choice of h and K in the absence of a priori information regarding f. Theoretical studies going back to Watson and Leadbetter (1963) show that the choice of h and K should not be split into two independent subproblems. Also, the choice of K largely depends upon the smoothness of f.

Watson and Leadbetter started from Parseval's identity

$$\mathbb{E} \int (f_{nh}(x) - f(x))^2 \, dx = \frac{1}{2\pi} \mathbb{E} \int |\varphi_{nh}(t) - \varphi(t)|^2 \, dt \, ,$$

where φ and φ_{nh} are the Fourier transforms of f and f_{nh}. Let ψ be the Fourier transform of the function K_h (note that h is absorbed in this definition). Then the expected L_2 error given above is minimal for the choice

$$\psi(t) = \frac{|\varphi|^2}{\frac{1}{n} + \frac{n-1}{n}|\varphi|^2} \, .$$

With this choice, the minimal expected L_2 error reduces to

$$\int \frac{|\varphi|^2(1 - |\varphi|^2)}{1 + (n-1)|\varphi|^2} \, .$$

These fundamental results were at the basis of a number of fine results:

A. Bullock Davis (1975, 1977) looked at the rate of decrease of the expected L_2 error for various rates of decrease of φ, when $\psi(t) = \xi(th)$ for a fixed form ξ. Here one is not allowed to vary the form of the kernel with n. She looked in particular at the Fourier kernel $K(x) = \sin(x)/(\pi x)$, and showed it to be nearly optimal for many densities.

B. Bullock Davis (1977) proposed letting $h = 1/t$, where t is the smallest t for which the estimate of the optimal $|\varphi|^2$ equals $1/(n+1)$. For the Fourier kernel $\sin(x)/(\pi x)$, this estimate is

$$|\widehat{\varphi}(t)|^2 = \frac{n}{n-1} \left(\left(\frac{1}{n} \sum_{i=1}^{n} \cos(tX_i) \right)^2 + \left(\frac{1}{n} \sum_{i=1}^{n} \sin(tX_i) \right)^2 \right) - \frac{1}{n-1} \, .$$

Bloxom (1979) provides some encouraging experimental results with this estimate.

C. Wahba (1981) and Hall and Marron (1988) adaptively estimate parameters of the optimal Fourier transform of the kernel under certain tail conditions on the characteristic function of f. This would require knowledge of the underlying class of densities.

D. Cline (1988) uses the Watson-Leadbetter result to point out that the optimal Fourier transform always is symmetric and positive. As the Fourier transform (say, ξ) of the Epanechnikov kernel takes negative values, it can always be replaced by ξ_+, its positive truncation, for a strict improvement in the expected L_2 error.

E. Cline (1990) gives precise asymptotic analysis of the expected L_2 error based upon the Fourier transform approach.

Finally, based upon recent developments in data-based bandwidth selection, several methods have been proposed for picking the bandwidth that have their origin in the expressions given by Watson and Leadbetter. Chiu (1991) has a plug-in method that is based upon the empirical characteristic function

$$\varphi_n(t) = \frac{1}{n} \sum_{j=1}^{n} e^{itX_j} .$$

Let Λ be the smallest positive t such that $|\varphi_n(t)|^2 \leq 3/n$ (where the constant 3 is a design parameter; Chiu recommends any constant > 1). Then use

$$h = \left(\frac{\int K^2}{Cn(\int x^2 K)^2} \right)^{1/5} ,$$

where

$$C = \frac{1}{\pi} \int_0^{\Lambda} t^4 (|\varphi_n(t)|^2 - 1/n)\, dt .$$

This is related to a plug-in method suggested by Park and Marron (1990), in which C is taken as

$$C = \frac{1}{\pi} \int_0^{\infty} t^2 (|\varphi_n(t)|^2 - 1/n) \psi^2(h't)\, dt ,$$

and $h' = C' s^{3/13} h^{10/13}$ is a pilot bandwidth, C' is a function of f (which in turn is estimated by the reference density approach) and s is a given measure of the scale of f. This yields an implicit equation in h, which must be solved. The non-consistency dealt with in this paper is due to a problem that is endemic in most L_2-based cross-validation methods including such "solve–an–equation" schemes.

An inequality for empirical characteristic functions.

The literature on empirical characteristic functions contains many strong results (Csörgö, 1981; Marcus, 1981, Keller, 1988), but none of these really fits our needs, as we require an inequality for

$$\mathbb{P}\left\{\sup_{|t|<\alpha} |\varphi(t) - \varphi_n(t)| > \beta\right\} ,$$

where β and α depend upon n in an arbitrary fashion. For β near $1/\sqrt{n}$, the results of Csörgö (1981) are useful. For β fixed, we are in large deviation territory (Keller, 1988). We believe that the following inequality is of independent general utility:

THEOREM 1. *Let X be a random variable with characteristic function φ and finite first moment, and let φ_n be the empirical characteristic function based upon an i.i.d. sample of size n drawn from X. Then, for $\alpha > 0$, $\beta > 0$ possibly dependent upon n,*

$$\mathbb{P}\left\{\sup_{|t|<\alpha} |\varphi(t) - \varphi_n(t)| > \beta\right\} \leq 4\left(1 + \frac{8\alpha\mathbb{E}|X|}{\beta}\right) e^{-n\beta^2/72} + o(1) ,$$

where the $o(1)$ term is uniform over all α and β.

PROOF. Define

$$\gamma = \frac{\beta}{4\mathbb{E}|X|} .$$

We find numbers $t_1 < t_2 < \cdots < t_k$ with the property that $t_1 = -\alpha$, $t_k = \alpha$, $|t_i - t_{i+1}| \leq \gamma$. Clearly, we can assure this with $k \leq 1 + 2\alpha/\gamma$. We begin with

$$\begin{aligned} \mathbb{P}\left\{\sup_{|t|<\alpha} |\varphi(t) - \varphi_n(t)| > \beta\right\} &\leq \mathbb{P}\left\{\sup_{|t-s|<\gamma} |\varphi(t) - \varphi(s)| > \beta/3\right\} \\ &\quad + \mathbb{P}\left\{\sup_{|t-s|<\gamma} |\varphi_n(t) - \varphi_n(s)| > \beta/3\right\} \\ &\quad + \sum_{i=1}^{k} \mathbb{P}\left\{|\varphi(t_i) - \varphi_n(t_i)| > \beta/3\right\} \\ &\stackrel{\text{def}}{=} I + II + III . \end{aligned}$$

Note that

$$|\varphi(t) - \varphi(s)| \leq \mathbb{E}|1 - e^{i(t-s)X}| \leq \mathbb{E}|(t-s)X| \leq \gamma\mathbb{E}|X| \leq \frac{\beta}{3}$$

when $|t-s| \le \gamma$. Therefore, $I \equiv 0$. Next, we let Y be the random variable that puts mass $1/n$ at each of the X_i's in the sample drawn from X. Then

$$\begin{aligned}|\varphi_n(t)-\varphi_n(s)| &\le \mathbb{E}|1-e^{i(t-s)Y}| \\ &\le \mathbb{E}|(t-s)Y| \\ &= |t-s|\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| .\end{aligned}$$

Therefore,

$$\begin{aligned}II &\le \mathbb{P}\left\{\gamma\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \ge \frac{\beta}{3}\right\} \\ &\le \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \ge \frac{4\mathbb{E}|X|}{3}\right\} \\ &\to 0\end{aligned}$$

by the law of large numbers. Finally, we let ζ and η be the real and imaginary parts of the Fourier transform φ. Let ζ_n and η_n be the corresponding empirical functions. For example, $\zeta_n(t) = (1/n)\sum_{i=1}^{n}\cos(tX_i)$. Then, for fixed t_i,

$$\begin{aligned}\mathbb{P}\{|\varphi(t_i)-\varphi_n(t_i)| > \beta/3\} &\le \mathbb{P}\{|\zeta(t_i)-\zeta_n(t_i)| > \beta/6\} + \mathbb{P}\{|\eta(t_i)-\eta_n(t_i)| > \beta/6\} \\ &\le 4e^{-n\beta^2/72}\end{aligned}$$

by Hoeffding's inequality for bounded random variables (Hoeffding, 1963). This concludes the proof of Theorem 1. □

The main result.

In the Theorem below, we describe a simple class of densities for which Chiu's method is non-consistent. No attempt was made to obtain a general result.

THEOREM 2. *Let f be a density with finite first moment, and with real unimodal characteristic function φ satisfying $\varphi(t) \sim t^{-c}$ as $t \to \infty$, where $c < 1/2$ is a fixed constant. Then, if H is the bandwidth choice for Chiu's method, we have*

$$\liminf_{n\to\infty} \mathbb{E}\int |f_{nH} - f| > 0 .$$

AN EXAMPLE. The condition of the theorem is satisfied for the random variable $X = Z - Z'$, where Z and Z' are i.i.d. gamma random variables with parameter $a < 1/4$ (note that $\varphi(t) = 1/(1+t^2)^a$).

OTHER MODES OF CONVERGENCE. We can't give an L_2 version of the Theorem, as $\int f^2 = \infty$ for the densities under consideration. It should come as no surprise that it is precisely for these densities that problems occur, as the design of the method is L_2-based. This raises the interesting question of whether we should test for the finiteness of $\int f^2$ before applying an L_2-based bandwidth selector. Nevertheless, f_{nH} is undesirable by any standard as we will show that $nH \to 0$ in probability, so that we won't even have pointwise convergence at any point.

OTHER BANDWIDTH SELECTORS. Chiu's bandwidth selector shares its anomalous behavior with most of the L_2 cross-validation criteria. For example, a similar non-consistency was pointed out in Devroye (1989) for the original L_2 cross-validation method (Bowman, 1974; Rudemo, 1974). For a survey of other methods in this class, and for some fixes, see for example Jones and Kappenman (1992) or Marron (1988, 1989). Also, we have not considered Chiu's stabilized method or one of its modifications (Chiu, 1992).

PROOF. If the constant C in Chiu's method is such that $C/n^4 \to \infty$ in probability, then $nH \to 0$ in probability as well. By necessary conditions for consistency (Devroye and Györfi, 1985), this implies that for any kernel,

$$\liminf_{n\to\infty} \mathbb{E} \int |f_{nH} - f| > 0 \ .$$

Define a constant z with

$$\frac{1}{2} > z > \frac{4c}{5-2c} \ .$$

Define α as the solution of

$$\varphi(t) = n^{-z} \ .$$

Note in particular that as $n \to \infty$, $\alpha \sim n^{z/c}$. Define $\beta = (1/2)\inf_{|t|\leq\alpha} |\varphi(t)| = (1/2)\varphi(\alpha) = (1/2)n^{-z}$. Let A_n be the event that

$$\sup_{|t|<\alpha} |\varphi(t) - \varphi_n(t)| \leq \beta \ .$$

If A_n holds, then for $t \in [-\alpha, \alpha]$,

$$|\varphi_n(t)| \geq |\varphi(t)| - \beta \geq \inf_{|t|\leq\alpha} |\varphi(t)| - \beta \geq \frac{1}{2}n^{-z} \ .$$

For n large enough, and $|t| \le \alpha$, $\varphi_n(t) > \sqrt{3/n}$, and thus $\Lambda \ge \alpha$. As on $[-\alpha,\alpha]$, under A_n, $\varphi_n(t) \ge (1/2)\varphi(t)$ and $1/n \le (1/8)\varphi^2(t)$, we see that

$$\begin{aligned}
C &\ge \frac{1}{\pi}\int_0^\alpha t^4(|\varphi_n(\alpha)|^2 - 1/n)\,dt \\
&\ge \frac{1}{\pi}\int_0^\alpha t^4((1/4)|\varphi(\alpha)|^2 - 1/n)\,dt \\
&\ge \frac{1}{\pi}\int_0^\alpha t^4(1/8)|\varphi(\alpha)|^2\,dt \\
&\sim \frac{\alpha^{5-2c}}{8(5-2c)\pi} \\
&\sim \frac{n^{(z/c)(5-2c)}}{8(5-2c)\pi}\ .
\end{aligned}$$

By our choice of z and c, the exponent of n on the right-hand side is greater than 4. We conclude that if $\mathbb{P}\{A_n\} \to 1$, then for any constant M,

$$\lim_{n\to\infty} \mathbb{P}\left\{C \ge Mn^4\right\} = 1\ ,$$

or, equivalently, for any constant $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left\{nH \le \epsilon\right\} = 1\ .$$

To prove that $\mathbb{P}\{A_n\} \to 1$, invoke Theorem 1. The constants α and β there are the same ones we introduced in the proof of Theorem 2. Note that $\beta = (1/2)n^{-z}$ with $z < 1/2$, and that α/β grows polynomially with n. Thus, $\mathbb{P}\{A_n\} \to 1$ as required. □

Acknowledgment.

I would like to thank the referee.

References

H. Akaike, "An approximation to the density function," *Annals of the Institute of Statistical Mathematics*, vol. 6, pp. 127–132, 1954.

B. Bloxom, "A Fourier integral density estimate: a Monte Carlo study," *Communications in Statistics B*, vol. 8, pp. 391–396, 1979.

A. W. Bowman, "An alternative method of cross-validation for the smoothing of density estimates," *Biometrika*, vol. 71, pp. 353–360, 1984.

S.-T. Chiu, "Bandwidth selection for kernel density estimation," *Annals of Statistics*, vol. 19, pp. 1883–1905, 1991.

S.-T. Chiu, "An automatic bandwidth selector for kernel density estimate," *Biometrika*, vol. 79, pp. 771–782, 1992.

D. B. H. Cline, "Admissible kernel estimators of a multivariate density," *Annals of Statistics*, vol. 16, pp. 1421–1427, 1988.

D. B. H. Cline, "Optimal kernel estimation of densities," *Annals of the Institute of Statistical Mathematics*, vol. 42, pp. 287–303, 1990.

S. Csörgö, "Limit behavior of the empirical characteristic function," *Annals of Probability*, vol. 9, pp. 130–144, 1981.

K. B. Davis, "Mean square error properties of density estimates," *Annals of Statistics*, vol. 5, pp. 1025–1030, 1975.

K. B. Davis, "Mean integrated square error properties of density estimates," *Annals of Statistics*, vol. 5, pp. 530–535, 1977.

L. Devroye, "On the non-consistency of the L2 cross-validated kernel density estimate," *Statistics and Probability Letters*, vol. 8, pp. 425–433, 1989.

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

P. Hall and J. S. Marron, "Choice of kernel order in density estimation," *Annals of Statistics*, vol. 16, pp. 161–173, 1988.

P. Hall and J. S. Marron, "Choice of kernel order in density estimation," *Annals of Statistics*, vol. 16, pp. 161–173, 1988.

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

M. C. Jones and R. F. Kappenman, "On a class of kernel density estimate bandwidth selectors," *Scandinavian Journal of Statistics*, vol. 19, pp. 337–349, 1992.

H.-D. Keller, "Large deviations of the empirical characteristic function," *Acta Scientiarium Mathematicarum Hungarica*, vol. 52, pp. 207–214, 1988.

M. B. Marcus, "Weak convergence of the empirical characteristic function," *Annals of Probability*, vol. 9, pp. 194–201, 1981.

J. S. Marron, "Automatic smoothing parameter selection: a survey," *Empirical Economics*, vol. 13, pp. 187–208, 1988.

J. S. Marron, "Automatic smoothing parameter selection: a survey," in: *Semiparametric and Nonparametric Economics*, ed. A. Ullah, pp. 65–86, Heidelberg, 1989.

B. U. Park and J. S. Marron, "Comparison of data-driven bandwidth selectors," *Journal of the American Statistical Association*, vol. 85, pp. 66–72, 1990.

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

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

M. Rudemo, "Empirical choice of histograms and kernel density estimators," *Scandinavian Journal of Statistics*, vol. 9, pp. 65–78, 1982.

G. Wahba, "Data-based optimal smoothing of orthogonal series density estimates," *Annals of Statistics*, vol. 9, pp. 146–156, 1981.

G. S. Watson and M. R. Leadbetter, "On the estimation of the probability density," *Annals of Mathematical Statistics*, vol. 34, pp. 480–491, 1963.