Bin width selection in multivariate histograms by the combinatorial method

Luc Devroye
School of Computer Science
McGill University
Montreal, Canada H3A 2K6

and

Gábor Lugosi
Department of Economics
Universitat Pompeu Fabra
Ramon Trias Fargas 25–27
Barcelona, Spain

ABSTRACT. We present several multivariate histogram density estimates that are universally L_1-optimal to within a constant factor and an additive term $O(\sqrt{\log n/n})$. The bin widths are chosen by the combinatorial method developed by the authors in *Combinatorial Methods in Density Estimation* (Springer-Verlag, 2001). The present paper solves a problem left open in that book.

KEYWORDS AND PHRASES. Multivariate density estimation, nonparametric estimation, variable histogram estimate, bandwith selection.

2000 MATHEMATICS SUBJECT CLASSIFICATIONS: Primary 62G05.

Research of the authors was supported by NSERC grant A3450 and by DGI grant BMF2000-0807. Corresponding author: Luc Devroye, luc@cs.mcgill.ca.

§1. Introduction

We are asked to estimate a density f on $\mathbf{R}^d$ based on an i.i.d. sample $X_1, \ldots, X_n$ drawn from f. The estimators we are considering here generalize the standard d-dimensional regular histogram estimate, defined by an anchor point (in this case, the origin) and bin widths $h_1, \ldots, h_d$, one bin width per dimension. For integers $i_1, \ldots, i_d$, we define the rectangular cell

$$A(i_1, \ldots, i_d) = \prod_{j=1}^{d} (i_j h_j, (i_j + 1)h_j]$$

and denote by $N(i_1, \ldots, i_d)$ the number of data points in the cell $A(i_1, \ldots, i_d)$. Let $A(x)$ denote the unique cell to which $x \in \mathbf{R}^d$ belongs, let $|\cdot|$ denote Lebesgue measure of a set, and let $N(x)$ denote the number of data points in $A(x)$. Then the regular histogram estimate is

$$f_n(x) = \frac{N(x)}{n|A(x)|} = \frac{N(x)}{n \prod_{j=1}^{d} h_j}.$$

It is known that $\mathbf{E}\left\{\int |f_n - f|\right\} \to 0$ whenever the h_j's are functions of n only such that $\max_j h_j \to 0$ and $n \prod_{j=1}^{d} h_j \to \infty$ (Devroye and Györfi, 1985, p. 20, and Abou-Jaoude (1976a,b)). Of course, the real problem is to find the best h_j's. In particular, we are looking for data-based choices $H_1, \ldots, H_d$ such that, writing g_n for the histogram estimate with $H_1, \ldots, H_d$, and f_n for the histogram estimate with $h_1, \ldots, h_d$, we have

$$\mathbf{E}\left\{\int |g_n - f|\right\} \le C_n \inf_{h_1, \ldots, h_d} \mathbf{E}\left\{\int |f_n - f|\right\} + D_n$$

where C_n is small and D_n is of order smaller than most nonparametric rates, e.g., $D_n = O(\sqrt{\log n/n})$ would be a typical additive term. As the best error rate over all h_j's is often (but not always) larger than $\sqrt{\log n/n}$, an inequality of the type given above becomes useful, especially when C_n is near one or at least remains bounded. To save space, we say that a data-based bandwidth selection is L_1-optimal on a class of densities $\mathcal{F}$ if there are finite constants C and C' such that for each $f \in \mathcal{F}$, $\limsup_{n\to\infty} C_n \le C$, and $\limsup_{n\to\infty} D_n/\sqrt{\log n/n} \le C'$. We know of no bin width selection method that is L_1-optimal when $\mathcal{F}$ is the class of all densities. This is striking as for the multivariate kernel estimate, L_1-optimal bandwidths for all densities were developed by the authors (Devroye and Lugosi, 1996, 1997, 2001), based on a combinatorial method. In Devroye and Lugosi (2001), for $d = 1$, an attempt at an L_1-optimal bin width for histograms was developed, but it allowed only the selection of an optimal h_1 from the dyadic set $\{2^{-k}; k = 0, \pm 1, \pm 2, \ldots\}$. The purpose of this paper is to remove this condition, and to propose an L_1-optimal bin width in any dimension where $\mathcal{F}$ is the class of all densities with a finite p-th moment where p

is any positive number (the constant C' depends upon p and d, and C is universal).

We also extend the results for variable-binwidth histograms. We define large parametrized families of histogram estimates and show that L_1-optimality may be achieved even within such rich classes.

It should be noted that there is a wealth of material on the histogram density estimate. For L_1, we refer to Devroye and Györfi (1985), Abou-Jaoude (1976a, 1976b), Chen and Zhao (1987), Devroye (1987) and Lugosi and Nobel (1996). For L_2, see Freedman and Diaconis (1981), and Zhao, Krishnaiah and Chen (1990). For the Hellinger distance, see Kanazawa (1988, 1992, 1993), and Barron, Birgé and Massart (1999). For the Kullback-Leibler distance, we refer to Rodriguez and Van Ryzin (1985, 1986). For the sup norm convergence, see Kim and Van Ryzin (1975). For all criteria, there have been attempts at obtaining optimal bin widths based on various principles. Cross-validation was attempted in an L_2 setting by Rudemo (1982). Stone (1985) established its near-optimality for all bounded densities. Cross-validation is known to fail for densities that have large peaks (and that are not square integrable), leading even to non-consistency. For the Hellinger distance, drawing on work by Barron, Birgé and Massart (1999), Castellan (2000) obtained optimality in a sense close to our definition for L_1-optimality under certain conditions on the density, including a compact support. Her method is a form of penalized maximum likelihood criterion. Akaike's criterion has been used in the design of bin widths by Taylor (1987), Atilgan (1990), Hall (1990), and Kanazawa (1993). Complexity minimization was suggested by Hall and Hannan (1988) and Yu and Speed (1990, 1992). For bin widths based on asymptotic analysis, we refer to Scott (1979), Lecoutre (1985), Kogure (1987) and Wand (1997). Birgé and Rozenholc (2002) provide a survey and a comparative simulation. Leave one out maximum likelihood has not been explicitly attempted, but it is easy to see that it must yield bins with at least two elements per occupied bin, and thus, the bin width must be larger than the distance between the largest two order statistics, an observation that immediately points out the inconsistency of this method for all distributions with larger than exponential tails. In most of the work cited, L_1 optimality was not the goal. Furthermore, the d-dimensional choice of bandwidths was not considered, so, to fill this void in the literature, we develop the combinatorial method.

§2. The combinatorial method

Let our density estimates be parametrized by $\theta \in \Theta$, where θ represents the vector of bandwidths $(h_1, \ldots, h_d)$. Let $f_{n,\theta}$ denote the histogram density estimate with parameter θ. Let $m < n$ be an integer picked to split the data into a set $X_1, \ldots, X_{n-m}$ used for constructing a density estimate, and a validation set $X_{n-m+1}, \ldots, X_n$. To make the notation more transparent, in the sequel we sometimes write $Y_1, \ldots, Y_m$ for $X_{n-m+1}, \ldots, X_n$, according to which choice is more convenient. The variables in the validation set are used to construct an empirical measure $\mu_m(A) = (1/m)\sum_{i=m}^{n} 1_{[Y_i \in A]}$. Introduce the class of sets

$$\mathcal{A} = \{\{x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x)\} : \theta \in \Theta, \theta' \in \Theta, \theta \neq \theta'\}$$

(these are the so-called "Yatracos sets") and define

$$\Delta_\theta = \sup_{A \in \mathcal{A}} \left| \int_A f_{n-m,\theta} - \mu_m(A) \right|.$$

We define *the minimum distance estimate* ψ_n as any density estimate selected from among those density estimates $f_{n-m,\theta}$ with

$$\Delta_\theta < \inf_{\theta^* \in \Theta} \Delta_{\theta^*} + 1/n.$$

The $1/n$ here is added to ensure the existence of such a density estimate.

For the minimum distance estimate ψ_n as defined above, we have

$$\int |\psi_n - f| \leq 3 \inf_{\theta \in \Theta} \int |f_{n-m,\theta} - f| + 4\Delta + \frac{3}{n},$$

where $\Delta = \sup_{A \in \mathcal{A}} \left| \int_A f - \mu_m(A) \right|$ (Devroye and Lugosi, 2001, Theorem 6.4). Note that $\inf_{\theta\in\Theta} \int |f_{n-m,\theta} - f|$ is not much larger than $\inf_{\theta\in\Theta} \int |f_{n,\theta} - f|$, that is, holding out m samples does not hurt: indeed, by Theorem 10.2 of Devroye and Lugosi (2001), if $0 < m \leq n/2$, then

$$\frac{\inf_{\theta\in\Theta} \mathbf{E}\{\int |f_{n-m,\theta} - f|\}}{\inf_{\theta\in\Theta} \mathbf{E}\{\int |f_{n,\theta} - f|\}} \leq 1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}.$$

This means that by decreasing the sample size to $n-m$, the performance of the best estimate in the class cannot deteriorate by more than a constant factor. If m is small relative to n, the loss in the L_1 error is negligible.

Next, we recall a bound for Δ based upon a technique introduced by Vapnik and Chervonenkis (1971). Introduce

$$\mu(A) = \mathbf{P}\{Y_1 \in A\} \ (A \subset \mathbf{R}^d).$$

Consider a class $\mathcal{B}$ of subsets of $\mathbf{R}^d$ and set $\Delta = \sup_{A\in\mathcal{B}} |\mu_m(A) - \mu(A)|$. For any set of points $y_1^m = \{y_1, \ldots, y_m\} \subset \mathbf{R}^d$, introduce the *empirical Vapnik–Chervonenkis shatter coefficient*, defined by

$$\mathbb{S}_{\mathcal{B}}(y_1^m) = |\{\{y_1, \ldots, y_m\} \cap A; A \in \mathcal{B}\}|.$$

Since $Y_1, \ldots, Y_m$ are random, $\mathbb{S}_{\mathcal{B}}(Y_1^m)$ becomes a random variable whose expected value appears in the following form of the Vapnik-Chervonenkis inequality:

$$\mathbf{E}\left\{\sup_{A\in\mathcal{B}} |\mu_m(A) - \mu(A)|\right\} \leq 2\mathbf{E}\left\{\sqrt{\frac{\log 2\mathbb{S}_{\mathcal{B}}(Y_1^m)}{m}}\right\}.$$

This inequality is proved in Theorem 3.1 of Devroye and Lugosi (2001). (Note that the form of the inequality given there is slightly different, but this version is straightforward from that proof.)

To bound $\mathbb{E}\Delta$, we use this inequality with $\mathcal{B}$ replaced by our $\mathcal{A}$. Since the class $\mathcal{A}$ is random– its definition involves $X_1, \ldots, X_{n-m}$–, we use the inequality above conditionally, and the independence of $(X_1, \ldots, X_{n-m})$ and $(Y_1, \ldots, Y_m)$:

$$\begin{aligned}\mathbb{E}\Delta =& \mathbb{E}\left\{\mathbb{E}\left(\Delta | X_1, \ldots, X_{n-m}\right)\right\}\\ \leq& 2\mathbb{E}\left\{\mathbb{E}\left\{\sqrt{\frac{\log 2\mathbb{S}_{\mathcal{A}}(Y_1^m)}{m}}\Big| X_1, \ldots, X_{n-m}\right\}\right\}\\ =& 2\mathbb{E}\left\{\sqrt{\frac{\log 2\mathbb{S}_{\mathcal{A}}(Y_1^m)}{m}}\right\}\\ \leq& 2\sqrt{\frac{\mathbb{E}\left\{\log 2\mathbb{S}_{\mathcal{A}}(Y_1^m)\right\}}{m}}\end{aligned}$$

by Jensen's inequality, where the expected value is now taken with respect to all random variables X_i, Y_j, $i = 1, \ldots, n-m$, $j = 1, \ldots, m$. Thus, we readily obtain the following.

Theorem 1. *For all n, $m \leq n/2$, and f:*

$$\begin{aligned}\mathbb{E}\left\{\int |\psi_n - f|\right\} \leq& 3\left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right)\inf_{\theta\in\Theta}\mathbb{E}\left\{\int |f_{n,\theta} - f|\right\}\\ &+ 8\sqrt{\frac{\mathbb{E}\left\{\log 2\mathbb{S}_{\mathcal{A}}(Y_1^m)\right\}}{m}} + \frac{3}{n}.\end{aligned}$$

Devroye and Lugosi (2001, Lemma 10.5) showed that for $d = 1$, and with $h_1 \in \{2^k, k = \ldots, -2, -1, 0, 1, 2, 3, \ldots\}$,

$$\mathbb{S}_{\mathcal{A}}(Y_1^m) \leq (m+1)n^2,$$

uniformly over all $X_1, \ldots, X_{n-m}$ and $Y_1, \ldots, Y_m$. However, this estimate is not valid if we permit h_1 to take values in all of $(0, \infty)$. The contribution of this paper is to provide various bounds for the shatter coefficient.

§3. The shatter coefficient

The next lemma is the key combinatorial result needed to make Theorem 1 useful. Because of technical reasons, we restrict the class of histogram estimates such that the minimal bin width is not smaller than a parameter $a > 0$. Later we will see that the value of a may be chosen very small, say, of the order of n^{-2} and that this restriction becomes unimportant, since the optimal histogram estimates necessarily have bin widths exceeding this value.

Denote the components of the data vectors X_i by $X_{i,j}$ $(j = 1, \ldots, d)$ and the components of Y_i by $Y_{i,j}$ $(j = 1, \ldots, d)$.

Lemma 1. *Assume that* $\Theta = \{(h_1, \ldots, h_d) : a \le h_i, 1 \le i \le d\}$, *where* $0 < a < \infty$. *Then*

$$\mathbb{S}_{\mathcal{A}}(Y_1^m) \le (m+1)\left(\prod_{j=1}^{d}\left(n+1+\frac{1}{a}\sum_{i=1}^{n}|X_{i,j}|\right)\right)^2 .$$

Proof. It will help a lot to introduce for each X_i ($i = 1, \ldots, n$) its d-vector of bin numbers, $b_i = (b_{i,1}, \ldots, b_{i,d})$, where $b_{i,j}$ is the bin number for the j-th co-ordinate $X_{i,j}$ of X_i for a given value of $\theta \in \Theta$. That is, if $X_{i,j} \in (kh_j, (k+1)h_j]$, then $b_{i,j} = k$. Set $b = (b_1, \ldots, b_n)$, so that b is in fact a vector of nd bin numbers. As we vary $\theta \in \Theta$, we will first count the number of possible values for b. As we vary h_1 only, we note that the absolute value of the bin number of $X_{i,1}$ must lie between 0 and $(|X_{i,1}|/a)+1$. As h_1 increases from its minimal value to ∞, the bin numbers $b_{i,1}$, $1 \le i \le n$, can change at most

$$\sum_{i=1}^{n}\big((|X_{i,1}|/a)+1\big) = n + \frac{1}{a}\sum_{i=1}^{n}|X_{i,1}|$$

times. The number of possible values for $(b_{1,1}, \ldots, b_{n,1})$ is thus not more than one plus that number. Clearly then, the number of possible values for the vector b is at most

$$\prod_{j=1}^{d}\left(n+1+\frac{1}{a}\sum_{i=1}^{n}|X_{i,j}|\right).$$

Consider regions R, R' of Θ on which the vector b is fixed (and takes two fixed values, possibly the same). For $\theta = (h_1, \ldots, h_d) \in R, \theta' = (h'_1, \ldots, h'_d) \in R'$, and Y_i, $i \le m$, note that $Y_i \in A(\theta, \theta')$ (i.e., $f_{n-m,\theta}(Y_i) > f_{n-m,\theta'}(Y_i)$) if and only if

$$\prod_{j=1}^{d} h_j > c(Y_i)\prod_{j=1}^{d} h'_j$$

where $c(y)$ is a fixed function of y only. That means that as $\theta \in R, \theta' \in R'$ are varied, the number of possible values for the vector

$$(z_1, \ldots, z_m) = \Big(1_{[Y_1 \in A(\theta,\theta')]}, \ldots, 1_{[Y_m \in A(\theta,\theta')]}\Big)$$

is at most $m+1$ (just let the ratio $\prod_{j=1}^{d} h_j / \prod_{j=1}^{d} h'_j$ vary from 0 to ∞, and consider passages through the values $c(Y_1), \ldots, c(Y_m)$). Thus, the shatter coefficient is not more than $m+1$ times the square of the number of possible values for the vector b:

$$\mathbb{S}_{\mathcal{A}}(Y_1^m) \le (m+1)\left(\prod_{j=1}^{d}\left(n+1+\frac{1}{a}\sum_{i=1}^{n}|X_{i,j}|\right)\right)^2 . \ \square$$

We have the following corollary of the previous lemma.

Lemma 2. *Assume that* $\Theta = \{(h_1,\ldots,h_d) : a \le h_i, 1 \le i \le d\}$, *where* $0 < a < \infty$. *Then, with* $0 < m < n$,

$$\mathbb{E}\left\{\log 2\mathbb{S}_{\mathcal{A}}(Y_1^m)\right\} \le (2d{+}1)\log(2n){+}2d\log\frac{1}{a}{+}2\sum_{j=1}^{d}\mathbb{E}\left\{\log\left(\max_{1\le i\le n}|X_{i,j}|+a\right)\right\}$$

Proof. By Lemma 1,

$$\begin{aligned}\log 2\mathbb{S}_{\mathcal{A}}(Y_1^m) &\le \log(2m+2)+2\sum_{j=1}^{d}\log\left(n+1+\frac{1}{a}\sum_{i=1}^{n}|X_{i,j}|\right)\\ &\le \log(2m+2)+2\sum_{j=1}^{d}\log\left((n+1)+\frac{1}{a}n\max_{1\le i\le n}|X_{i,j}|\right)\\ &\le (2d+1)\log(2n)+2d\log\frac{1}{a}+2\sum_{j=1}^{d}\log\left(\max_{1\le i\le n}|X_{i,j}|+a\right)\end{aligned}$$

as desired. $\square$

§4. Small-tailed distributions.

Combining Theorem 1 with Lemma 2 we obtain the following performance bound.

Theorem 2. *Assume that* $\Theta = \{(h_1,\ldots,h_d) : a \le h_i, 1 \le i \le d\}$, *where* $0 < a < \infty$. *Then, for all* n, $m \le n/2$, *and* f:

$$\begin{aligned}\mathbb{E}\left\{\int|\psi_n-f|\right\} \le 3\left(1+\frac{2m}{n-m}+8\sqrt{\frac{m}{n}}\right)&\inf_{\theta\in\Theta}\mathbb{E}\left\{\int|f_{n,\theta}-f|\right\}\\ &+8\sqrt{\frac{(2d+1)\log(2n)}{m}}\\ &+8\sqrt{\frac{2d\log\frac{1}{a}+2\sum_{j=1}^{d}\mathbb{E}\left\{\log\left(\max_{1\le i\le n}|X_{i,j}|+a\right)\right\}}{m}}+\frac{3}{n}.\end{aligned}$$

As a first example, let $\mathcal{G}$ be the class of all densities on $[-1,1]^d$. For these densities, by Theorem 2,

$$\mathbf{E}\left\{\int |\psi_n - f|\right\} \le 3\left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right) \inf_{\theta\in\Theta} \mathbf{E}\left\{\int |f_{n,\theta} - f|\right\}$$

$$+ 8\sqrt{\frac{(2d+1)\log(2n)}{m}} + 8\sqrt{\frac{(2d)\log(1+1/a)}{m}} + \frac{3}{n}.$$

Let us arbitrarily set $a = 1/n^2$ and $m = \lfloor \epsilon n \rfloor$ for $\epsilon \in (0,1)$ fixed. ($\lfloor\cdot\rfloor$ stands for integer part.) Then

$$\mathbf{E}\left\{\int |\psi_n - f|\right\}$$

$$\le \left(3 + 6\epsilon/(1-\epsilon) + 24\sqrt{\epsilon} + o(1)\right) \inf_{\theta\in\Theta} \mathbf{E}\left\{\int |f_{n,\theta} - f|\right\} + C\sqrt{\frac{\log n}{\epsilon n}}$$

where C is a constant depending only upon d.

Denote by Θ^* the set of all parameters (unrestricted by a, as in Θ). Ideally, we would like to replace the infimum over Θ in Theorem 2 by the infimum over Θ^*. The next lemma shows that with $a = n^{-2}$ this is indeed possible since, if n is sufficiently large, then deterministically, $\inf_{\theta\in\Theta^*-\Theta} \int |f_{n,\theta} - f| \ge 2/3$, that is,

$$3 \inf_{\theta\in\Theta^*-\Theta} \int |f_{n,\theta} - f| \ge 2$$

and thus the infimum over this range is unimportant.

LEMMA 3. *Let $\theta = (h_1, \ldots, h_d)$ be such that $\min_i h_i < 1/n^2$. Then there exists a constant $\gamma(f)$ such that for $n \ge \gamma(f)$,*

$$\int |f_{n,\theta} - f| \ge 2/3.$$

Therefore, for $n \ge \gamma(f)$,

$$\inf_{\theta\in\Theta} \mathbf{E}\left\{\int |f_{n,\theta} - f|\right\} = \inf_{\theta\in\Theta^*} \mathbf{E}\left\{\int |f_{n,\theta} - f|\right\}.$$

PROOF. Let M_j be a point with the property $\mathbf{P}\{|X_{1,j}| \ge M_j\} = 1/(3d)$, $j = 1, \ldots, d$. Set $M = \max_{j\le d} M_j$. Recall that by absolute continuity of f, there exists a function $R(u)$ such that if a set A has Lebesgue measure $\lambda(A) \le R(u)$, then $\int_A f \le u$. Note that

$$\int |f_{n,\theta} - f| \ge 2 \int_{f_{n,\theta}=0} f$$

$$= 2 - 2\int_{f_{n,\theta}>0} f$$

$$
\begin{aligned}
&= 2 - 2\int_{\exists j:|x_j|>M_j} f - 2\int_{\max_j(|x_j|/M_j)\le 1, f_{n,\theta}>0} f \\
&\ge 2 - 2d\frac{1}{3d} - 2\int_{\max_j(|x_j|/M_j)\le 1, f_{n,\theta}>0} f \\
&\ge 2 - 2/3 - 2/3
\end{aligned}
$$

if the Lebesgue measure of the set $\{\max_j(|x_j|/M_j) \le 1, f_{n,\theta} > 0\}$ is less than $R(1/3)$. But if i denotes the index of a coordinate for which $h_i < 1/n^2$, then the Lebesgue measure may be bounded by

$$
\prod_{k\ne i}(2M_k) \times (2nh_i) \le \frac{2\,(2M)^{d-1}}{n} \le R(1/3)
$$

for $n \ge 2(2M)^{d-1}/R(1/3)$. □

Summarizing, we obtain L_1 optimality for *all* densities in $\mathcal{G}$:

Theorem 3. *Let $\epsilon \in (0, 1/2]$ be fixed and let $m = \lfloor n\epsilon \rfloor$. Assume the density f has support in $[-1, 1]^d$, and consider the minimum distance estimate Ψ_n based on the restricted set of parameters $\Theta = \{(h_1, \ldots, h_d) : a \le h_i, 1 \le i \le d\}$, where $a = n^{-2}$. Then for n large enough,*

$$
\begin{aligned}
&\mathbf{E}\left\{\int |\psi_n - f|\right\} \\
&\quad\le \left(3 + 6\epsilon/(1-\epsilon) + 24\sqrt{\epsilon} + o(1)\right) \inf_{\theta\in\Theta^*} \mathbf{E}\left\{\int |f_{n,\theta} - f|\right\} + C\sqrt{\frac{\log n}{\epsilon n}}
\end{aligned}
$$

where $\Theta^* = \{(h_1, \ldots, h_d) : h_i > 0, 1 \le i \le d\}$. The $o(1)$ and C in this bound do not depend upon the individual density f, but the least n above which the inequality is true does depend upon f.

In the rest of this section we point out that the restriction to compactly supported densities is not necessary. In fact, L_1-optimality of the same estimate holds under the only assumption that each marginal of f has a finite p-th moment for some $p > 0$. Thus, the only densities excluded from the next L_1-optimality result are those with a truly heavy tail. Note that for such densities any regular histogram estimate is expected to perform very poorly. It remains an open question whether an analogue result remains true without any restriction on f. This is in contrast with the analogue problem for kernel density estimates for which L_1-optimality holds for all densities, see Devroye and Lugosi (1996).

THEOREM 4. *Consider the estimate of Theorem 3 and assume that* $M = \max_{j=1,\dots,d} \mathbf{E}|X_{1,j}|^p$ *is finite for some* $p > 0$. *Then for* n *large enough,*

$$\begin{aligned}
&\mathbf{E}\left\{\int |\psi_n - f|\right\} \\
&\quad\le \left(3 + 6\epsilon/(1-\epsilon) + 24\sqrt{\epsilon} + o(1)\right) \inf_{\theta\in\Theta^*} \mathbf{E}\left\{\int |f_{n,\theta} - f|\right\} \\
&\qquad + C\sqrt{\frac{\log n}{\epsilon n}} + C\sqrt{\frac{\log(Mn)}{p\epsilon n}}
\end{aligned}$$

where $o(1)$ *and* C *do not depend upon the individual density* f, *but the least* n *above which the inequality is true does depend upon* f.

PROOF. The result directly follows from Theorem 2 and Lemma 3 just an appropriate bound for $\mathbf{E}\left\{\log\left(\max_{1\le i\le n}|X_{i,j}| + a\right)\right\}$ is needed. To this end, observe that

$$\begin{aligned}
e^{p\mathbf{E}\left\{\log\left(\max_{1\le i\le n}|X_{i,j}|+a\right)\right\}} &\le \mathbf{E}\left\{e^{p\log\left(\max_{1\le i\le n}|X_{i,j}|+a\right)}\right\} \\
&\quad \text{(by Jensen's inequality)} \\
&\le \mathbf{E}\left\{\sum_{i=1}^n e^{p\log\left(|X_{i,j}|+a\right)}\right\} \\
&= n\mathbf{E}\left\{e^{p\log\left(|X_{1,j}|+a\right)}\right\} \\
&= n\mathbf{E}\left\{\left(|X_{1,j}|+a\right)^p\right\} \\
&\le n2^p\left(M + a^p\right)
\end{aligned}$$

and therefore

$$\mathbf{E}\left\{\log\left(\max_{1\le i\le n}|X_{i,j}| + a\right)\right\} \le \frac{1}{p}\log\left(n2^p\left(M+a^p\right)\right) .$$

Putting the pieces together, we obtain the desired claim. □

§5. Transformed histogram estimate

To guarantee L_1-optimality, one may artificially avoid heavy-tailed distributions by transforming the data beforehand. For example, applying the transformation $y := x/(1+|x|)$ to each co-ordinate separately, we may transform the data $X_1, \dots, X_n$ to data $X'_1, \dots, X'_n$ that are supported on $[-1,1]^d$. On the transformed data, we apply the combinatorial method with $a = 1/n^2$. The density of X_1 is f and that of X'_1 will be denoted by g. If ψ_n is the chosen histogram estimate, then the inverse transformation yields a density estimate ξ_n of f. Recall that strictly monotone transformations leave the L_1 distance invariant (see Devroye and Györfi, 1985). Thus by

Theorem 2 and Lemma 3, for all densities and all n large enough, the above method picks an estimate ξ_n with the property

$$\begin{aligned}\mathbf{E}\left\{\int|\xi_n-f|\right\}&=\mathbf{E}\left\{\int|\psi_n-g|\right\}\\&\le 3\left(1+\frac{2m}{n-m}+8\sqrt{\frac{m}{n}}\right)\inf_{\theta\in\Theta^*}\mathbf{E}\left\{\int|f_{n,\theta}-f|\right\}\\&\quad+8\sqrt{\frac{(2d+1)\log(2n)}{m}}+8\sqrt{\frac{(2d)\log(1+n^2)}{m}}+\frac{3}{n}.\end{aligned}$$

Here Θ^* denotes the space of all histogram bin widths $(h_1,\ldots,h_d):h_i\ge 0, 1\le i\le d$, and $f_{n,\theta}$ is the density corresponding to the transformed histogram estimate (which is not a histogram estimate). Thus, within the class of estimates thus described, the combinatorial method is L_1 optimal.

§6. Variable bandwidths

Another way of dealing with heavy-tailed densities is to allow bins to become wider in the tails. Of course, one would like to optimize the variable bandwidth. The purpose of this section is to explore this direction. For simplicity, assume that each component of X is concentrated on $[0,\infty)$, and consider bandwidths $h_j(\rho)=\sum_{\ell=1}^k a_{j,\ell}\phi_\ell(\rho)$, where ρ is the bin number for the j-th co-ordinate $j=1,\ldots,d$, the ϕ_ℓ are fixed positive functions, and the $a_{j,\ell}$ are unknown positive parameters. Such a parametrization of the bandwidth has been useful in kernel estimates for unimodal densities (Biau and Devroye, 2002), and, as we will show, works equally well for histogram estimates. For the j-th co-ordinate, on positive data, if bins are numbered $0,1,2,\ldots$, the thresholds separating the bins occur at $0,h_j(1),h_j(1)+h_j(2),\ldots$.

Note that if $k=1$ and $\phi_1\equiv 1$ then we recover the case of regular histograms discussed in the previous sections. If ϕ_ℓ is an increasing function, then such a choice allows bin widths to grow. One may, for example, take $\phi_\ell(\rho)=\rho^{\ell-1}$, $\ell=1,\ldots,k$ but any other choice is possible.

Thus, each density estimate is now parametrized by a kd-dimensional vector of the positive components $a_{j,\ell}$. Denote such a vector by θ and let Θ be the collection of all θ's. Based on this set of parameters, the minimum-distance estimate may be defined the same way as in the case of regular histograms and Theorem 1 remains valid. Once again, the heart of the matter is the combinatorial argument bounding the shatter coefficient $\mathbb{S}_{\mathcal{A}}(y_1^m)$, which is summarized in the next lemma. Once again, we need to restrict the set Θ to those parameters whose components are not too small.

Lemma 4. *Assume that $\phi_\ell(\rho)\ge 1$ for all $\ell=1,\ldots,k$ and positive integer ρ, and let $a>0$. Consider*

$$\Theta=\left\{\theta=(a_{j,\ell})_{j=1,\ldots,d;\ell=1,\ldots,k}:a<a_{j,\ell}<\infty\right\}.$$

Then

$$\mathbb{S}_{\mathcal{A}}(Y_1^m) \leq (m+1)(2n\rho_{\max}+1)^{2kd}$$

where

$$\rho_{\max} = \left\lceil \frac{1}{ka} \max_{j=1,\ldots,d} \max_{i=1,\ldots,n} X_{i,j} \right\rceil .$$

($\lceil\cdot\rceil$ denotes upper integer part.)

Proof. The proof is an extension of the argument of Lemma 1. Once again, we start by counting the possible different values of the nd-component vector b of bin numbers corresponding to the n data points $(X_1,\ldots,X_n) = (X_1,\ldots,X_{n-m},Y_1,\ldots,Y_m)$.

First of all observe that by the assumption $\phi_\ell(\rho) \geq 1$, the maximal bin number of any of the data points is at most $\rho_{\max}$. Consider any of these data points, say X_1, and concentrate on the first component only. Let ρ be a positive integer. Observe that the first bin number of X_i equals ρ ($\rho \leq \rho_{\max}$) if and only if

$$\sum_{t=1}^{\rho-1} h_1(t) \leq X_{i,1} < \sum_{t=1}^{\rho} h_1(t)$$

or equivalently, since $h_1(s) = \sum_{\ell=1}^k a_{1,\ell}\phi_\ell(s)$, if

$$\sum_{\ell=1}^k a_{1,\ell} z_\rho^- \leq X_{i,1} < \sum_{\ell=1}^k a_{1,\ell} z_\rho^+$$

where $z_\rho^- = \sum_{t=1}^{\rho-1} h_1(t)$ and $z_\rho^+ = \sum_{t=1}^{\rho} h_1(t)$. Thus, as we vary the parameters $a_{1,\ell}$, $\ell = 1,\ldots,k$ corresponding to the bin widths of the first component, the vector of n bin numbers for the n data points can take at most as many values as the number of different contiguous regions defined by the $2n\rho_{\max}$ hyperplanes of the form

$$\sum_{\ell=1}^k a_{1,\ell} z_\rho^- = X_{i,1} \quad \text{and} \quad \sum_{\ell=1}^k a_{1,\ell} z_\rho^+ = X_{i,1}, \quad i = 1,\ldots,n; \rho = 1,\ldots,\rho_{\max}$$

in the k-dimensional space of parameters. This number is well-known to be bounded by

$$\sum_{\ell=0}^k \binom{2n\rho_{\max}}{\ell} \leq (2n\rho_{\max}+1)^k$$

(see Schläffli, 1950). Clearly then, the number of possible values of the vector of all bin numbers, counting now all d components, is at most

$$(2n\rho_{\max}+1)^{kd}.$$

The rest of the proof is now identical to that of Lemma 1. □

Using Lemma 4, now it is easy to extend all arguments for regular histogram estimates. For example, it is immediate to obtain the analogue of Theorem 2 which states that if Θ is as in Lemma 4 then for all n, $m \le n/2$, and f,

$$\begin{aligned}\mathbf{E}\left\{\int |\psi_n - f|\right\} &\le 3\left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right) \inf_{\theta\in\Theta} \mathbf{E}\left\{\int |f_{n,\theta} - f|\right\} \\ &\quad + 8\sqrt{\frac{\log(2m+2) + 2kd\log\frac{1}{ka} + 2kd\mathbf{E}\left\{\log\left(\max_{i,j}|X_{i,j}| + ka\right)\right\}}{m}} + \frac{3}{n}.\end{aligned}$$

Now $\mathbf{E}\left\{\log\left(\max_{i,j}|X_{i,j}| + a\right)\right\}$ may be estimated the same way as in the proof of Theorem 4 to obtain L_1-optimality (with respect to the class Θ) for any density which has a finite p-th moment for some $p > 0$. It makes sense to take $a = 1/n^2$ since such a choice will not harm the L_1-optimality and includes all interesting choices of variable bandwidths. Note however that we cannot take $a = 0$.

§7. Acknowledgment

We thank the eagle-eyed referees. The first author gratefully acknowledges the hospitality of Universitat Pompeu Fabra during June 2002.

§8. References

S. Abou-Jaoude, “Conditions nécessaires et suffisantes de convergence L_1 en probabilité de l’histogramme pour une densité,” *Annales de l’Institut Henri Poincaré*, vol. 12, pp. 213–231, 1976a.

S. Abou-Jaoude, “La convergence L_1 et L_∞ de l’estimateur de la partition aleatoire pour une densité,” *Annales de l’Institut Henri Poincaré*, vol. 12, pp. 299–317, 1976b.

T. Atilgan, “On derivation and application of AIC as a data-based criterion for histograms,” *Communications in Statistics—Theory and Methods*, vol. 19, pp. 885–903, 1990.

A. Barron, L. Birgé, and P. Massart, “Risk bounds for model selection via penalization,” *Probability Theory and Related Fields*, vol. 113, pp. 301–415, 1999.

G. Biau and L. Devroye, “On the risk of estimates for block decreasing densities,” *Journal of Multivariate Analysis*, to appear, 2002.

L. Birgé and Y. Rozenholc, "How many bins should be put in a regular histogram," Prépublication 721, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI et VII, France, 2002.

G. Castellan, "Sélection d'histogrammes ou de modèles exponentiels de polynomes par morceaux à l'aide d'un critère de type Akaike," Thèse, Mathématiques, Université de Paris-Sud, 2000.

X. R. Chen and L. C. Zhao, "Almost sure L_1-norm convergence for data-based histogram density estimates," *Journal of Multivariate Analysis*, vol. 21, pp. 179–188, 1987.

C. F. De Beer and J. W. H. Swanepoel, "Simple and effective number-of-bins circumference selectors for a histogram," *Statistics and Computing*, vol. 9, pp. 27–35, 1999.

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

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

L. Devroye and G. Lugosi, "A universally acceptable smoothing factor for kernel density estimates," *Annals of Statistics*, vol. 24, pp. 2499–2512, 1996.

L. Devroye and G. Lugosi, "Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes," *Annals of Statistics*, vol. 25, pp. 2626–2637, 1997.

L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*, Springer-Verlag, New York, 2001.

D. Freedman and P. Diaconis, "On the histogram as a density estimator: L_2 theory," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 57, pp. 453–476, 1981.

M. P. Gessaman, "A consistent nonparametric multivariate density estimator based on statistically equivalent blocks," *Annals of Mathematical Statistics*, vol. 41, pp. 1344–1346, 1970.

P. Hall and E. J. Hannan, "On stochastic complexity and nonparametric density estimation," *Biometrika*, vol. 75, pp. 705–714, 1988.

P. Hall, "Akaike's information criterion and Kullback-Leibler loss for histogram density estimation," *Probability Theory and Related Fields*, vol. 85, pp. 449–467, 1990.

Y. Kanazawa, "An optimal variable cell histogram," *Communications in Statistics, part A: Theory and Methods*, vol. 17, pp. 1401–1422, 1988.

Y. Kanazawa, “An optimal variable cell histogram based on the sample spacings,” *Annals of Statistics*, vol. 20, pp. 291–304, 1992.

Y. Kanazawa, “Hellinger distance and Kullback-Leibler loss for the kernel density estimator,” *Statistics and Probability Letters*, vol. 17, pp. 293–298, 1993.

Y. Kanazawa, “Hellinger distance and Akaike’s information criterion for the histogram,” *Statistics and Probability Letters*, vol. 17, pp. 293–298, 1993.

B. K. Kim and J. Van Ryzin, “Uniform consistency of a histogram density estimator and modal estimation,” *Communications in Statistics*, vol. 4, pp. 303–315, 1975.

A. Kogure, “Asymptotically optimal cells for a histogram,” *Annals of Statistics*, vol. 15, pp. 1023–1030, 1987.

J.-P. Lecoutre, “The L_2-optimal cell width for the histogram,” *Statistics and Probability Letters*, vol. 3, pp. 303–306, 1985.

G. Lugosi and A. Nobel, “Consistency of data-driven histogram methods for density estimation and classification,” *Annals of Statistics*, vol. 24, pp. 687–706, 1996.

C. C. Rodriguez and J. Van Ryzin, “Maximum entropy histograms,” *Statistics and Probability Letters*, vol. 3, pp. 117–120, 1985.

C. C. Rodriguez and J. Van Ryzin, “Large sample properties of maximum entropy histograms,” *IEEE Transactions on Information Theory*, vol. IT-32, pp. 751–759, 1986.

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

L. Schläffli, *Gesammelte Mathematische Abhandlungen*, Birkhäuser-Verlag, Basel, 1950.

D. W. Scott, “On optimal data-based histograms,” *Biometrika*, vol. 79, pp. 605–610, 1979.

C. J. Stone, “An asymptotically optimal histogram selection rule,” in: *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II*, (edited by L. Le Cam and R. A. Olshen), pp. 513–520, Wadsworth, Belmont, CA., 1985.

C. C. Taylor, “Akaike’s information criterion and the histogram,” *Biometrika*, vol. 74, pp. 636–639, 1987.

J. Van Ryzin, “A histogram method of density estimation,” *Communications in Statistics*, vol. 2, pp. 493–506, 1973.

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.

M. P. Wand, “Data-based choice of histogram bin width,” *The American Statistician*, vol. 51, pp. 59–64, 1997.

B. Yu and T. P. Speed, “Stochastic complexity and model selection II: histograms,” Technical Report, Department of Statistics, University of California, Berkeley, 1990.

B. Yu and T. Speed, “Data compression and histograms,” *Probability Theory and Related Fields*, pp. 195–229 , 1992.

L. C. Zhao, P. R. Krishnaiah, and X. R. Chen, “Almost sure L_r-norm convergence for data-based histogram estimates,” *Theory of Probability and its Applications*, vol. 35, pp. 396–403, 1990.