On the Risk of Estimates for Block Decreasing Densities

Gérard Biau
Laboratoire de Probabilités et Statistique
Université Montpellier II
Cc 051, Place Eugène Bataillon
34095 Montpellier Cedex 5, France

and

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

Abstract. A density $f = f(x_1, \ldots, x_d)$ on $[0, \infty)^d$ is block decreasing if for each $j \in \{1, \ldots, d\}$, it is a decreasing function of x_j, when all other components are held fixed. Let us consider the class of all block decreasing densities on $[0, 1]^d$ bounded by B. We shall study the minimax risk over this class using n i.i.d. observations, the loss being measured by the L_1 distance between the estimate and the true density. We prove that if $S = \log(1 + B)$, lower bounds for the risk are of the form $C(S^d/n)^{1/(d+2)}$, where C is a function of d only. We also prove that a suitable histogram with unequal bin widths as well as a variable kernel estimate achieve the optimal multivariate rate. We present a procedure for choosing all parameters in the kernel estimate automatically without loosing the minimax optimality, even if B and the support of f are unknown.

Keywords and phrases. Multivariate density estimation, block decreasing density, minimax risk, nonparametric estimation, variable kernel estimate, bandwith selection.

2000 Mathematics Subject Classifications: Primary 62G05.

Research of the authors was supported by NSERC grant A3450 and by Université Montpellier II. Corresponding author: Luc Devroye, luc@cs.mcgill.ca.

§1. Introduction

A density f on $[0,\infty)^d$ is *block decreasing* if it is a non-increasing function of each of its arguments on $[0,\infty)$ when all other arguments are held fixed. Other definitions of multivariate monotonicity and multivariate unimodality have been proposed (see Dharmadhikari and Joag-Dev (1988) for a survey), but the block decreasing densities are rather natural as a family. For example, if $(U_1,\ldots,U_d)$ are independent uniform $[0,1]$ random variables and $(Y_1,\ldots,Y_d)$ is an arbitrary random vector in the positive quadrant, then $(U_1Y_1,\ldots,U_dY_d)$ has a block decreasing density. The same is true if U_1 is exponential or the absolute value of a normal variate. The class thus contains all arbitrary multivariate scale mixtures of basic random vectors with independent components. The purpose of this paper is to present a study of density estimates for this important multivariate class. Our results include several new multivariate density estimates that are tailored to the class of block decreasing densities and that are minimax optimal for certain subclasses. Actually, most of existing results on minimax optimality of estimates pertain to univariate data, and few generalizations to higher dimensions are available. The present work represents a first modest step towards the study of minimax optimal multivariate estimates. We will show below that the standard kernel estimate requires a major overhaul before it can be used for this class. We expect that for other classes, not treated here, other adjustments may be necessary, and the nature of these modifications will surely prove to be an interesting topic of future research.

Define the class $\mathcal{F}_B$ of all block decreasing densities on the unit hypercube $I_d=[0,1]^d$ of $\mathbb{R}^d$ bounded by B. More precisely, any element $f=f(x_1,\ldots,x_d)$ of $\mathcal{F}_B$ should satisfy $\int f(x)\,dx=1$, $f(x)=0$ if $x\notin I_d$, $0\le f(x)\le B$ if $x\in I_d$, and, for each $j\in\{1,\ldots,d\}$, f is a decreasing function of x_j as $x_j\uparrow 1$, when all other components are held fixed. Examples of block decreasing densities are given in Arnold, Castillo, and Sarabia (1999, pp. 58, 63, 68, 72, 88), including particular cases of beta conditional densities, Burr conditional densities and Pareto conditional densities. The reader can easily construct other examples by the multivariate scaling property mentioned in the first paragraph of this paper.

Assume now that we are given an i.i.d. sample of size n, $X_1,\ldots,X_n$, drawn from an unknown density $f\in\mathcal{F}_B$. Then the worst-case risk for a density estimate f_n of f $\big($a measurable mapping from $\mathbb{R}^d\times(\mathbb{R}^d)^n$ to $\mathbb{R}\big)$ is

$$\mathcal{R}(f_n,\mathcal{F}_B)=\sup_{f\in\mathcal{F}_B}\mathbb{E}\left\{\int|f_n(x;X_1,\ldots,X_n)-f(x)|\,dx\right\}$$

(as a sake of clarity, we will drop in the sequel the "dx" notation when no confusion is possible). The *minimax risk* $\big($Devroye (1987), Devroye and Györfi (1985), Devroye and Lugosi (2001)$\big)$ on $\mathcal{F}_B$ is given by

$$\mathcal{R}_n(\mathcal{F}_B)=\inf_{f_n}\mathcal{R}(f_n,\mathcal{F}_B), \tag{1}$$

and a density estimate f_n is *minimax optimal* if, for n large enough,

$$\mathcal{R}(f_n,\mathcal{F}_B)\le C\mathcal{R}_n(\mathcal{F}_B)$$

for some constant C. In the univariate setting ($d=1$), Birgé (1987a) proved that, if $S=\log(1+B)$, upper and lower bounds for the minimax risk (1) are of the form $C(S/n)^{1/3}$, where C is a universal constant not depending upon B or n. Observing that classical estimates like histograms or kernel estimates would not lead to the right bound, Birgé (1987b) also exhibited a minimax optimal estimate for the class $\mathcal{F}_B$ merely by considering a suitable histogram with geometrically increasing interval widths, and fine-tuning the geometrical rate of increase. Birgé's modified histogram provides us with a nice example of how estimates can be tailor made for certain classes of densities. We refer to this author for discussion and references on the problem of estimating unimodal and decreasing densities.

Our main purpose in this paper will be to generalize Birgé's results to block decreasing multivariate densities. In section 2, we show, using information-theoretic methods, that lower bounds for the d-dimensional minimax risk are proportional to $(S^d/n)^{1/(d+2)}$. In section 3, we prove that a multivariate generalization of Birgé's modified histogram is minimax optimal. In section 4, we propose a suitable variable kernel estimate with linear varying smoothing parameter that achieves the optimal rate. In section 5, we introduce a method for choosing all parameters in the kernel estimate automatically without loosing the minimax optimality, even if B and the support of f are unknown.

§2. Minimax lower bounds

We develop information-theoretic methods based on the work of Assouad (1982), Birgé (1986), and Bretagnolle and Huber (1979).

Theorem 1. *There exist positive constants C_1, C_2 and C_3, functions of d, such that*

$$\mathcal{R}_n(\mathcal{F}_B) \geq \frac{1}{4\Big[1+\big[1+\big(C_1 S^d/n\big)^{1/(d+2)}\big]^{1/d}\Big]^d}\left(\frac{C_1 S^d}{n}\right)^{1/(d+2)}$$

for $C_2 \leq S \leq C_3\, n^{1/d}$.

Proof. The first step of the proof is a discretization argument. Let ϵ be a positive real number and let $r \geq 1$ be an integer, both to be determined later. We partition the unit hypercube I_d into r^d cells

$$\mathcal{C}_{\mathbf{i}} = \prod_{j=1}^{d} [x_{i_j - 1}, x_{i_j}), \quad \mathbf{i} = (i_1, \ldots, i_d),$$

where $x_0 = 0$ and, for j in $\{1, \ldots, d\}$,

$$x_{i_j} = \frac{(1+\epsilon)^{i_j} - 1}{(1+\epsilon)^r - 1}, \quad i_j = 1, \ldots, r.$$

Observe that $x_r = 1$ and that $\sum_{i_j=1}^{r}(x_{i_j} - x_{i_j - 1}) = 1$. For every index $\mathbf{i} = (i_1, \ldots, i_d)$, set $\underline{\mathbf{i}} = i_1 + \ldots + i_d$ and define two functions $h_{\mathbf{i}}$ and $g_{\mathbf{i}}$ supported on $\mathcal{C}_{\mathbf{i}}$ as follows:

$$h_{\mathbf{i}} = \frac{\lambda^d\big[1+(1+\epsilon)^{1/d}\big]^d}{2^d(1+\epsilon)^{\underline{\mathbf{i}}}}\ , \quad g_{\mathbf{i}} = \frac{\lambda^d}{(1+\epsilon)^{\underline{\mathbf{i}}}}\prod_{j=1}^{d} g_{i_j},$$

where every g_{i_j} is a piecewise constant function on $[x_{i_j-1}, x_{i_j})$ defined by

$$g_{i_j} = \begin{cases} (1+\epsilon)^{1/d} & \text{on } \left[x_{i_j-1}, \frac{x_{i_j-1}+x_{i_j}}{2}\right) \\ 1 & \text{on } \left[\frac{x_{i_j-1}+x_{i_j}}{2}, x_{i_j}\right), \end{cases}$$

and

$$\lambda = \frac{2(1+\epsilon)}{r\epsilon\big[1+(1+\epsilon)^{1/d}\big]}\big[(1+\epsilon)^r - 1\big].$$

Observe that $\displaystyle\int_{\mathcal{C}_{\mathbf{i}}} h_{\mathbf{i}} = 1/r^d$ since the Lebesgue measure of $\mathcal{C}_{\mathbf{i}}$ is

$$\lambda(\mathcal{C}_{\mathbf{i}}) = \prod_{j=1}^{d}(x_{i_j} - x_{i_j-1}) = \prod_{j=1}^{d} \frac{\epsilon(1+\epsilon)^{i_j-1}}{(1+\epsilon)^r - 1} = \frac{\epsilon^d(1+\epsilon)^{\underline{\mathbf{i}}-d}}{\big[(1+\epsilon)^r - 1\big]^d}\,.$$

Similarly,

$$\int_{\mathcal{C}_{\mathbf{i}}} g_{\mathbf{i}} = \frac{\epsilon^d(1+\epsilon)^{\underline{\mathbf{i}}-d}}{2^d\big[(1+\epsilon)^r - 1\big]^d}\,\frac{\lambda^d}{(1+\epsilon)^{\underline{\mathbf{i}}}}\sum_{k=0}^{d}\binom{d}{k}(1+\epsilon)^{k/d} = \frac{1}{r^d}\,.$$

Let us now compute the L_1 separation and the Hellinger closeness $\big($Devroye (1987) and Devroye and Lugosi (2001)$\big)$ of the hypercube of block decreasing densities f_θ, $\theta = (\theta_1, \ldots, \theta_{r^d}) \in \{0,1\}^{r^d}$ defined on each $\mathcal{C}_{\mathbf{i}}$ by $f_\theta = h_{\mathbf{i}}$ whenever the component of θ matching with $\mathcal{C}_{\mathbf{i}}$ is 1 and $f_\theta = g_{\mathbf{i}}$ otherwise. With respect to the L_1 separation, we have

$$\begin{aligned}
&\int_{\mathcal{C}_{\mathbf{i}}} |h_{\mathbf{i}} - g_{\mathbf{i}}| \\
&= \frac{\epsilon^d(1+\epsilon)^{\underline{\mathbf{i}}-d}}{2^d\big[(1+\epsilon)^r - 1\big]^d}\,\frac{\lambda^d}{(1+\epsilon)^{\underline{\mathbf{i}}}}\sum_{k=0}^{d}\binom{d}{k}\left|\left(\frac{1+(1+\epsilon)^{1/d}}{2}\right)^d - (1+\epsilon)^{k/d}\right| \\
&= \frac{1}{r^d\big[1+(1+\epsilon)^{1/d}\big]^d}\sum_{k=0}^{d}\binom{d}{k}\left|\left(\frac{1+(1+\epsilon)^{1/d}}{2}\right)^d - (1+\epsilon)^{k/d}\right| \\
&\geq \frac{1}{r^d\big[1+(1+\epsilon)^{1/d}\big]^d}\left[\left(\frac{1+(1+\epsilon)^{1/d}}{2}\right)^d - 1\right. \\
&\qquad\qquad\qquad\qquad \left. + (1+\epsilon) - \left(\frac{1+(1+\epsilon)^{1/d}}{2}\right)^d\right] \\
&= \frac{\epsilon}{r^d\big[1+(1+\epsilon)^{1/d}\big]^d} \\
&\overset{\text{def}}{=} \alpha.
\end{aligned}$$

As to the Hellinger closeness, write

$$\begin{aligned}
&\int_{\mathcal{C}_{\mathbf{i}}}\Big(\sqrt{h_{\mathbf{i}}}-\sqrt{g_{\mathbf{i}}}\Big)^2\\
&\quad=\frac{1}{r^d\big[1+(1+\epsilon)^{1/d}\big]^d}\sum_{k=0}^{d}\binom{d}{k}\Bigg[\bigg(\frac{1+(1+\epsilon)^{1/d}}{2}\bigg)^d+(1+\epsilon)^{k/d}\\
&\qquad\qquad -2\bigg(\frac{1+(1+\epsilon)^{1/d}}{2}\bigg)^{d/2}(1+\epsilon)^{k/(2d)}\Bigg]\\
&\quad=\frac{1}{r^d\big[1+(1+\epsilon)^{1/d}\big]^d}\Bigg[2\big[1+(1+\epsilon)^{1/d}\big]^d\\
&\qquad\qquad -2^{1-d/2}\big[1+(1+\epsilon)^{1/d}\big]^{d/2}\big[1+(1+\epsilon)^{1/(2d)}\big]^d.
\end{aligned}$$

By the Taylor's series expansion, we obtain

$$\big[1+(1+\epsilon)^{1/d}\big]^d\le 2^d\bigg[1+\frac{\epsilon}{2}+\frac{(3/2)^{d-2}(d-1)\epsilon^2}{8d}\bigg]\quad\text{for } 0<\epsilon\le 1,$$

$$\big[1+(1+\epsilon)^{1/d}\big]^{d/2}\ge 2^{d/2}\bigg[1+\frac{\epsilon}{4}+\frac{\epsilon^2}{8}\Big(\frac{1}{d}-1\Big)-\frac{\epsilon^2}{32d^2}-\frac{\epsilon^3}{32d^2}\Big(\frac{1}{d}-1\Big)-\frac{\epsilon^4}{128d^2}\Big(\frac{1}{d}-1\Big)^2\bigg]\quad\text{for } \epsilon>0,$$

and

$$\big[1+(1+\epsilon)^{1/(2d)}\big]^d\ge 2^d\bigg[1+\frac{\epsilon}{4}+\frac{\epsilon^2}{8}\Big(\frac{1}{2d}-1\Big)\bigg]\quad\text{for } 0<\epsilon\le 1.$$

Therefore, we are led, after tedious calculations, to

$$\int_{\mathcal{C}_{\mathbf{i}}}\Big(\sqrt{h_{\mathbf{i}}}-\sqrt{g_{\mathbf{i}}}\Big)^2\le\frac{2^d\epsilon^2(1+A\epsilon)}{Cr^d\big[1+(1+\epsilon)^{1/d}\big]^d}\,,$$

where

$$\begin{aligned}
A&=\frac{32d^4-24d^3-9d^2+11d+2}{64(3/2)^{d-2}d^3(d-1)+96d^4-96d^3+16d^2}>0,\\
C&=\frac{16d^2}{4(3/2)^{d-2}d(d-1)+6d^2-6d+1}>0.
\end{aligned}$$

Thus

$$\int_{\mathcal{C}_{\mathbf{i}}}\Big(\sqrt{h_{\mathbf{i}}}-\sqrt{g_{\mathbf{i}}}\Big)^2\le\frac{2^d\epsilon^2(1+A\epsilon)}{Cr^d\big[1+(1+\epsilon)^{1/d}\big]^d}\overset{\text{def}}{=}2-2\beta\,.$$

Plugging α and β into Assouad's lower bound theorem (Assouad (1982), Birgé (1986), Bretagnolle and Huber (1979)), we obtain

$$\mathcal{R}_n(\mathcal{F}_B)\ge\frac{r^d\alpha}{2}\Big(1-\sqrt{2-2\beta^n}\Big)\ge\frac{r^d\alpha}{2}\Big[1-\sqrt{2n(1-\beta)}\Big]$$

$$= \frac{\epsilon}{2\big[1+(1+\epsilon)^{1/d}\big]^d}\left[1-\sqrt{\frac{2^d n\epsilon^2(1+A\epsilon)}{Cr^d\big[1+(1+\epsilon)^{1/d}\big]^d}}\,\right].$$

We define $\lceil\cdot\rceil$ to be the nearest larger integer (or: ceiling) function. We can make the square root less than 1/2 if we take

$$r = \left\lceil\left(\frac{4.2^d n\epsilon^2(1+A\epsilon)}{C\big[1+(1+\epsilon)^{1/d}\big]^d}\right)^{1/d}\right\rceil .$$

Consequently,

$$\mathcal{R}_n(\mathcal{F}_B) \geq \frac{\epsilon}{4\big[1+(1+\epsilon)^{1/d}\big]^d}\,.$$

This last expression should be maximized with respect to ϵ, subject to the constraint that the set of densities $\mathcal{F}_\theta = \big\{f_\theta : \theta \in \{0,1\}^{r^d}\big\}$ forms a subclass of $\mathcal{F}_B$, i.e., that $\lambda^d \leq B$. Observe that λ^d is approximately equal to $\exp(dr\epsilon)$. Roughly speaking, this is at most B if $dr\epsilon \leq S$. Substituting r by its approximate value, $(4n\epsilon^2/C)^{1/d}$, we obtain that $d(4/C)^{1/d}n^{1/d}\epsilon^{(d+2)/d} \leq S$. This is why the value

$$\epsilon = \left(\frac{C_1 S^d}{n}\right)^{1/(d+2)},$$

with $C_1 = C/(4d^d) > 0$, is approximately optimal. We must insist that $C_1S^d \leq n$ to have $\epsilon \leq 1$. This value leads to the desired lower bound. However, we still have to verify that $\mathcal{F}_\theta \subset \mathcal{F}_{\mathcal{B}}$, i.e., that $\lambda^d \leq B$ with this choice of ϵ. This is done below. We begin with four small observations.

(1) $S \leq (11/12)^{(d+2)/d}(1/C_1)^{1/d}n^{1/d}$ is equivalent to $\epsilon \leq 11/12$;

(2) $r = 1 + (4n\epsilon^2/C)^{1/d} = 1 + SP(\epsilon)/(d\epsilon)$, where

$$P(\epsilon) = 1 + \frac{1}{d}\Big(A-\frac{1}{2}\Big)\epsilon + \frac{1}{2d}\Big(\frac{1}{2}-\frac{A}{d}\Big)\epsilon^2 + \frac{A}{4d^2}\epsilon^3 + \frac{1}{16d^2}\Big(\frac{1}{d}-1\Big)^2\epsilon^4 + \frac{A}{16d^3}\Big(\frac{1}{d}-1\Big)^2\epsilon^5;$$

(3) $\big((1+\epsilon)^x - 1\big)/x$ is nondecreasing in x for $x > 0$;

(4) $P(\epsilon) \geq 0.95$ and $P(\epsilon)\log(1+\epsilon) \leq \epsilon$ for $0 < \epsilon \leq 1$.

This is used in the following chain of inequalities:

$$\begin{aligned}\lambda &= \frac{2(1+\epsilon)}{r\epsilon\big[1+(1+\epsilon)^{1/d}\big]}\big[(1+\epsilon)^r - 1\big]\\ &\leq 2\frac{(1+\epsilon)e^{SP(\epsilon)/(d\epsilon)\log(1+\epsilon)}-1}{\epsilon+SP(\epsilon)/d}\,\frac{1+\epsilon}{1+(1+\epsilon)^{1/d}}\\ &\leq 2\frac{(1+\epsilon)e^{S/d}-1}{0.95S/d+\epsilon}\,\frac{1+\epsilon}{1+(1+\epsilon)^{1/d}}\end{aligned}$$

$$
\begin{aligned}
&\le 2\frac{(1+\epsilon)e^{S/d}-1}{0.95S/d+\epsilon} \\
&\le e^{S/d}-1 \\
&\le (e^{S}-1)^{1/d}.
\end{aligned}
$$

Only the fourth inequality requires explicit verification. This boils down to verifying

$$e^{S/d}\Big(2+\epsilon-0.95\frac{S}{d}\Big) \le 2-\epsilon-0.95\frac{S}{d}\,.$$

For $S \ge 60/19d$, we have $e^{S/d} \ge 23$ and $0.95S/d - 2 \ge 1 \ge \epsilon$. So, the left-hand side of the last inequality is at most equal to

$$23\Big(2+\epsilon-0.95\frac{S}{d}\Big).$$

Thus it suffices to check that

$$24\epsilon \le 22\Big(0.95\frac{S}{d}-2\Big).$$

This is immediate from the fact that $S \ge 60/19d$ and $\epsilon \le 11/12$. The proof of Theorem 1 is completed taking $C_2 = 60/19d$ and $C_3 = (11/12)^{(d+2)/d}(1/C_1)^{1/d}n^{1/d}$. □

Remark. On the constraints. Note that both constants C_2 and C_3 depend on the dimension. Calculations show that C_2 and C_3 are increasing slowly towards infinity. For $d = 1$, we have $C_2 = 3.16, C_3 = 0.19$. For $d = 2$, $C_2 = 6.31, C_3 = 1.92$. For $d = 3$, $C_2 = 9.47, C_3 = 3.28$. For $d = 4$, $C_2 = 12.63, C_3 = 4.55$. For $d = 5$, $C_2 = 15.79, C_3 = 5.81$. Therefore the constraints on S and n do not dramatically get more severe as d increases.

§3. Birgé's multivariate histogram estimate

Let ϵ be a positive real number and let $r \ge 1$ be an integer. As in section 2, we partition the unit hypercube I_d into r^d cells

$$\mathcal{C}_{\mathbf{i}} = \prod_{j=1}^{d}\Big[x_{i_j-1}, x_{i_j}\Big), \quad \mathbf{i} = (i_1,\ldots,i_d),$$

where $x_0 = 0$ and, for every j in $\{1,\ldots,d\}$,

$$x_{i_j} = \frac{(1+\epsilon)^{i_j}-1}{(1+\epsilon)^r-1}, \quad i_j = 1,\ldots,r.$$

We define

$$f_n = \sum_{\mathbf{i}} \frac{\mu_n(\mathcal{C}_{\mathbf{i}})}{\lambda(\mathcal{C}_{\mathbf{i}})} 1_{\mathcal{C}_{\mathbf{i}}}, \tag{4}$$

where μ_n denotes the empirical measure based on the sample $X_1,\ldots,X_n$, and where the summation extends over all $\mathbf{i}$ in $\{1,\ldots,r\}^d$. When $d = 1$, the estimate (4) reduces to the

1-dimensional histogram with unequal bin widths already defined in Birgé (1987b). For $d \geq 1$, we take the liberty of calling (4) *Birgé's multivariate histogram estimate.* From theorem 1, we recall that the lower bound for any estimate in the class $\mathcal{F}_B$ is proportional to $(S^d/n)^{1/(d+2)}$. Theorem 2 below implies that Birgé's multivariate histogram is minimax optimal.

THEOREM 2. *Birgé's multivariate histogram (4) on I_d with*

$$r = \left\lceil \left(R^2 n S^2\right)^{1/(d+2)} \right\rceil, \quad R = \frac{2 + 2^{d-3}(d-1)}{\sqrt{2^d - 1}},$$

and

$$\epsilon = e^{S/r} - 1$$

satisfies

$$\sup_{f \in \mathcal{F}_B} \mathbf{E}\left\{ \int |f_n - f| \right\} \leq C_1 \left(\frac{R^d S^d}{n} \right)^{1/(d+2)} + C_2 \left(\frac{R^d S^d}{n} \right)^{2/(d+2)},$$

for all $n \geq C_3 S^d$, where C_1, C_2 and C_3 are positive functions of d.

PROOF. Let f be any element of $\mathcal{F}_B$ and introduce the notation $g_{\mathbf{i}} = \int_{\mathcal{C}_{\mathbf{i}}} f/\lambda(\mathcal{C}_{\mathbf{i}})$. Then

$$\begin{aligned} \int |f_n - f| &= \sum_{\mathbf{i}} \int_{\mathcal{C}_{\mathbf{i}}} |f_n - f| \\ &\leq \sum_{\mathbf{i}} \int_{\mathcal{C}_{\mathbf{i}}} |f_n - g_{\mathbf{i}}| + \sum_{\mathbf{i}} \int_{\mathcal{C}_{\mathbf{i}}} |f - g_{\mathbf{i}}| \\ &\overset{\text{def}}{=} V_n + B_n, \end{aligned}$$

where V_n denotes "variation" and B_n denotes "bias". Define $\mathbf{i} = (i_1, \ldots, i_d)$, $\mathbf{1} = (1, \ldots, 1)$, $\mathbf{i} - \mathbf{1} = (i_1 - 1, \ldots, i_d - 1)$. We have

$$\begin{aligned} B_n &\leq \frac{1}{2} \sum_{\mathbf{i}} \lambda(\mathcal{C}_{\mathbf{i}}) \left[f(x_{\mathbf{i}-\mathbf{1}}) - f(x_{\mathbf{i}}) \right] \\ &\qquad \big(\text{since } f \text{ is block decreasing on } \mathcal{C}_{\mathbf{i}}, \text{ see Devroye (1987)}\big) \\ &\leq \frac{1}{2} \left[\sum_{\mathbf{i} \in \mathcal{I}_d} \lambda(\mathcal{C}_{\mathbf{i}}) B + \sum_{\mathbf{i} \in \mathcal{I}_d^c} \lambda(\mathcal{C}_{\mathbf{i}}) f(x_{\mathbf{i}-\mathbf{1}}) - \sum_{\mathbf{i}} \lambda(\mathcal{C}_{\mathbf{i}}) f(x_{\mathbf{i}}) \right], \end{aligned}$$

where $\mathcal{I}_d = \left\{ \mathbf{i} : \exists j \in \{1, \ldots, d\} \text{ with } i_j = 1 \right\}$, and $\mathcal{I}_d^c$ is the complement of $\mathcal{I}_d$ in $\{1, \ldots, r\}^d$. Observing that

$$\sum_{\mathbf{i} \in \mathcal{I}_d} \lambda(\mathcal{C}_{\mathbf{i}}) \leq \frac{d\epsilon}{(1+\epsilon)^r - 1},$$

we have

$$B_n \le \frac{1}{2}\left[\frac{Bd\epsilon}{(1+\epsilon)^r - 1} + \sum_{\mathbf{i}\in\mathcal{I}_d^c}(1+\epsilon)^d \lambda(\mathcal{C}_{\mathbf{i-1}})f(x_{\mathbf{i-1}}) - \sum_{\mathbf{i}}\lambda(\mathcal{C}_{\mathbf{i}})f(x_{\mathbf{i}})\right].$$

Using the fact that for $0 < \epsilon \le 1$, by Taylor's series with remainder,

$$(1+\epsilon)^d \le 1 + d\epsilon + 2^{d-3}d(d-1)\epsilon,$$

and, for a block decreasing density on I_d,

$$\sum_{\mathbf{i}\in\mathcal{I}_d^c}\lambda(\mathcal{C}_{\mathbf{i-1}})f(x_{\mathbf{i-1}}) \le 1,$$

we conclude that

$$B_n \le \frac{d\epsilon}{2}\left[\frac{B}{(1+\epsilon)^r - 1} + 1 + 2^{d-3}(d-1)\right].$$

To bound V_n, we let $Z_{\mathbf{i}}$ be a binomial random variable with parameters n and $p_{\mathbf{i}} = \int_{\mathcal{C}_{\mathbf{i}}} f$, and we proceed as follows:

$$\begin{aligned}
\mathbb{E}\{V_n\} &= \sum_{\mathbf{i}} \frac{\lambda(\mathcal{C}_{\mathbf{i}})\mathbb{E}\{|Z_{\mathbf{i}} - \mathbb{E}\{Z_{\mathbf{i}}\}|\}}{n\lambda(\mathcal{C}_{\mathbf{i}})} \\
&\le \frac{1}{n}\sum_{\mathbf{i}}\sqrt{\mathbb{E}\Big\{\big(Z_{\mathbf{i}} - \mathbb{E}\{Z_{\mathbf{i}}\}\big)^2\Big\}} \quad \text{(Cauchy-Schwarz inequality)} \\
&= \frac{1}{n}\sum_{\mathbf{i}}\sqrt{np_{\mathbf{i}}(1-p_{\mathbf{i}})} = \frac{r^d}{\sqrt{n}}\frac{1}{r^d}\sum_{\mathbf{i}}\sqrt{p_{\mathbf{i}}(1-p_{\mathbf{i}})} \\
&\le \frac{r^d}{\sqrt{n}}\sqrt{\frac{1}{r^d}\sum_{\mathbf{i}}p_{\mathbf{i}}\Big(1 - \frac{1}{r^d}\sum_{\mathbf{i}}p_{\mathbf{i}}\Big)} \quad \text{(Jensen's inequality)} \\
&= \sqrt{\frac{r^d - 1}{n}}.
\end{aligned}$$

Combining the bounds on B_n and V_n, we obtain

$$\mathbb{E}\Big\{\int|f_n - f|\Big\} \le \frac{d\epsilon}{2}\left[\frac{B}{(1+\epsilon)^r - 1} + 1 + 2^{d-3}(d-1)\right] + \sqrt{\frac{r^d-1}{n}}.$$

We choose $\epsilon = e^{S/r} - 1$ and insist on $\epsilon \le 1$. Resubstitution yields

$$\begin{aligned}
\mathbb{E}\Big\{\int|f_n - f|\Big\} &\le \frac{d\epsilon}{2}\big[2 + 2^{d-3}(d-1)\big] + \sqrt{\frac{r^d-1}{n}} \\
&= A\epsilon + \sqrt{\frac{r^d-1}{n}} \quad \big(\text{with } A \overset{\text{def}}{=} d\big(1 + 2^{d-4}(d-1)\big)\big) \\
&\le A(e^{S/x} - 1) + \sqrt{\frac{(2^d-1)\,x^d}{n}} \quad \text{for any } r - 1 < x \le r.
\end{aligned}$$

In the last inequality, we used the bound $r^d - 1 < (2^d - 1)\, x^d$ for $r - 1 < x$, which follows from the binomial identity. The derivative of the function $x \mapsto A(e^{S/x} - 1) + \sqrt{(2^d - 1)\, x^d/n}$ is zero for the solution of

$$x^{(d+2)/2} = \Big[2A/\big(d\sqrt{2^d - 1}\big)\Big]\sqrt{n}Se^{S/x}.$$

Roughly speaking, this solution increases with n, so that $e^{S/x}$ approaches 1 as $n \to \infty$. Therefore, the choice

$$x = \big(R^2 n S^2\big)^{1/(d+2)}, \quad r = \lceil x \rceil \quad \text{and} \quad R = 2A/\Big(d\sqrt{2^d - 1}\Big)$$

will do. Plugging this back into the upper bound for the expected L_1 error, we obtain

$$\begin{aligned}
&\mathbf{E}\Big\{ \int |f_n - f| \Big\} \\
&\le \sqrt{\frac{(2^d - 1)\, x^d}{n}} + A\Big(\frac{S}{x} + \frac{S^2}{2x^2} e^{S/x}\Big) \quad \Big(\text{use } e^u - 1 \le u + \frac{u^2}{2} e^u \text{ for } u > 0\Big) \\
&\le \Big(\sqrt{2^d - 1} + \frac{A}{R}\Big)\Big(\frac{R^d S^d}{n}\Big)^{1/(d+2)} + \frac{A}{2R^2}\Big(\frac{R^d S^d}{n}\Big)^{2/(d+2)} \exp\big((R^d S^d/n)^{1/(d+2)}/R\big) \\
&\le \Big(\sqrt{2^d - 1} + \frac{d\sqrt{2^d - 1}}{2}\Big)\Big(\frac{R^d S^d}{n}\Big)^{1/(d+2)} + \frac{d\sqrt{2^d - 1} e^{1/R}}{4R}\Big(\frac{R^d S^d}{n}\Big)^{2/(d+2)} \\
&\qquad \text{for } R^d S^d \le n.
\end{aligned}$$

Observing that $\epsilon \le 1$ when $S^d/\big(R^2 \log\big(2(d+2)\big)\big) \le n$, the proof of theorem 2 is complete if we set

$$\begin{aligned}
C_1 &= \sqrt{2^d - 1} + d\sqrt{2^d - 1}/2\ , \\
C_2 &= d\sqrt{2^d - 1}\, e^{1/R}/(4R)\ , \\
C_3 &= \max\Big[R^d, 1/\big(R^2 \log\big(2(d+2)\big)\big)\Big]. \ \square
\end{aligned}$$

Remark. On the constraints. Note that both the constraints on S and n become more severe as the dimension increases. We have $C_3 = 2, 2.08, 3.45, 18.20, 353.15$ for $d = 1, 2, 3, 4, 5$, respectively.

§4. A minimax optimal variable kernel estimate

Let $K = 1_{[-1/2,1/2]^d}$ be the uniform kernel, let $r \ge 1$ be an integer and let $\epsilon = e^{S/r} - 1$. For $x = (x_1, \ldots, x_d) \in \mathbf{R}^d$, we define the function

$$\mathbf{h}(x) = \big(h(x_1), \ldots, h(x_d)\big),$$

where, for $u \in \mathbf{R}$,

$$h(u) = \frac{\epsilon}{1 + \frac{\epsilon}{2}}\Big[u + \frac{1}{(1+\epsilon)^r - 1}\Big].$$

We also set $\tilde{\mathbf{h}}(x) = h(x_1)\dots h(x_d)$. In this framework, the *variable kernel estimate* f_n of f with kernel K and smoothing parameter $\mathbf{h}$ $\big($Devroye and Lugosi (2000, 2001)$\big)$ reads, for $x \in \mathrm{R}^d$,

$$f_n(x) = \frac{1}{n\tilde{\mathbf{h}}(x)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\mathbf{h}(x)}\right),$$

where the argument of K is a vector with components $(x_j - X_{ij})/h(x_j)$, $1 \le j \le d$. With our choice of K,

$$f_n(x) = \frac{1}{n\tilde{\mathbf{h}}(x)} \sum_{i=1}^{n} 1_{\mathcal{W}_{d,x}}(X_i), \tag{5}$$

where $\mathcal{W}_{d,x}$ stands for the d-dimensional hypercube $[x_1 - h(x_1)/2, x_1 + h(x_1)/2] \times \ldots \times [x_d - h(x_d)/2, x_d + h(x_d)/2]$. The following theorem states that the kernel estimate (5) is minimax optimal.

THEOREM 3. *The variable kernel estimate* (5) *with*

$$r = \left\lceil \left(R^2 n S^2\right)^{1/(d+2)} \right\rceil \quad \text{and} \quad R = \frac{3 + 2^{d-3}(d-1)}{\sqrt{6^d}}$$

satisfies

$$\sup_{f \in \mathcal{F}_B} \mathbf{E}\left\{\int |f_n - f|\right\} \le C_1 \left(\frac{R^d S^d}{n}\right)^{1/(d+2)} + C_2 \left(\frac{R^d S^d}{n}\right)^{2/(d+2)},$$

for all $n \ge C_3 \max\left(S^d, S^{-2}\right)$, *where* C_1, C_2 *and* C_3 *are positive functions of* d.

PROOF. First observe that for $u \in \mathrm{R}$,

$$u - \frac{h(u)}{2} = \frac{1}{1 + \frac{\epsilon}{2}} u - \frac{\epsilon}{2 + \epsilon} \frac{1}{(1+\epsilon)^r - 1}$$

and

$$u + \frac{h(u)}{2} = \frac{1+\epsilon}{1 + \frac{\epsilon}{2}} u + \frac{\epsilon}{2 + \epsilon} \frac{1}{(1+\epsilon)^r - 1}.$$

Therefore, the variable kernel estimate (5) vanishes almost surely outside the d-dimensional hypercube

$$\mathcal{P}_d = \left[\frac{-\epsilon}{2(1+\epsilon)} \frac{1}{(1+\epsilon)^r - 1}, 1 + \frac{\epsilon}{2} + \frac{\epsilon}{2\left[(1+\epsilon)^r - 1\right]}\right]^d$$

(note that each component of $\mathbf{h}$ is positive on $\mathcal{P}_d$, so that f_n is well defined). Now, let $f \in \mathcal{F}_B$. As for Birgé's multivariate histogram, we split the expected L_1 error of the kernel estimate into a variation and a bias term, namely

$$\mathbb{E}\left\{\int |f_n - f|\right\} = \mathbb{E}\left\{\int_{\mathcal{P}_d} |f_n - f|\right\} \le \mathbb{E}\left\{\int_{\mathcal{P}_d} |f_n - \mathbb{E}\{f_n\}|\right\} + \int_{\mathcal{P}_d} |\mathbb{E}\{f_n\} - f|$$
$$\stackrel{\text{def}}{=} V_n + B_n.$$

We have for $x \in \mathcal{P}_d$,

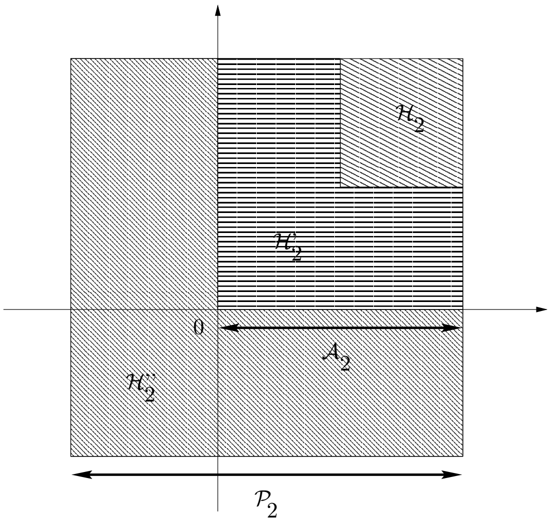
$$\Big|\mathbf{E}\{f_n(x)\} - f(x)\Big| \leq \frac{1}{\tilde{\mathbf{h}}(x)} \int_{\mathcal{W}_{d,x}} |f(t) - f(x)|\, dt.$$

Let us introduce the sets

$$\mathcal{A}_d = \left[0, 1 + \frac{\epsilon}{2} + \frac{\epsilon}{2\big[(1+\epsilon)^r - 1\big]}\right]^d,$$

$$\mathcal{H}_d = \left[\frac{\epsilon}{2\big[(1+\epsilon)^r - 1\big]}, 1 + \frac{\epsilon}{2} + \frac{\epsilon}{2\big[(1+\epsilon)^r - 1\big]}\right]^d,$$

$\mathcal{H}'_d = \mathcal{A}_d \setminus \mathcal{H}_d$ and $\mathcal{H}''_d = \mathcal{P}_d \setminus \mathcal{A}_d$. Figure 1 illustrates the respective positions of the sets $\mathcal{P}_d, \mathcal{A}_d, \mathcal{H}_d, \mathcal{H}'_d$ and $\mathcal{H}''_d$ in dimension 2.



Examples of sets $\mathcal{P}_2, \mathcal{A}_2, \mathcal{H}_2, \mathcal{H}'_2$ and $\mathcal{H}''_2$.

Since f is bounded and block decreasing on I_d, a moment's thought shows that, for $x \in \mathcal{P}_d$ and $t \in \mathcal{W}_{d,x}$,

$$|f(t) - f(x)| \leq \begin{cases} B & \text{if } x \in \mathcal{H}''_d \\ \max\Big\{f(x), B - f(x), f(x) - f\big(x + \frac{\mathbf{h}(x)}{2}\big)\Big\} & \text{if } x \in \mathcal{H}'_d \\ \max\Big\{f\big(x - \frac{\mathbf{h}(x)}{2}\big) - f(x), f(x) - f\big(x + \frac{\mathbf{h}(x)}{2}\big)\Big\} & \text{if } x \in \mathcal{H}_d. \end{cases}$$

Therefore,

$$B_n \leq \int_{\mathcal{H}''_d} B\, dx + \int_{\mathcal{H}'_d} B + f(x) - f\Big(x + \frac{\mathbf{h}(x)}{2}\Big)\, dx$$

$$+\int_{\mathcal{H}_d} f\Big(x-\frac{\mathbf{h}(x)}{2}\Big)-f\Big(x+\frac{\mathbf{h}(x)}{2}\Big)\,dx\,.$$

Moreover,

$$\max\big(\lambda(\mathcal{H}_d'),\lambda(\mathcal{H}_d'')\big)=\frac{d\epsilon}{2\big[(1+\epsilon)^r-1\big]}\,.$$

Thus

$$\begin{aligned}B_n\le\frac{3Bd\epsilon}{2\big[(1+\epsilon)^r-1\big]}&+\int_{\mathcal{H}_d} f\left(\frac{1}{1+\frac{\epsilon}{2}}x-\frac{\epsilon}{2+\epsilon}\frac{1}{(1+\epsilon)^r-1}\right)dx\\&-\int_{\mathcal{A}_d} f\left(\frac{1+\epsilon}{1+\frac{\epsilon}{2}}x+\frac{\epsilon}{2+\epsilon}\frac{1}{(1+\epsilon)^r-1}\right)dx\,.\end{aligned}$$

Using the (multivariate) change of variable

$$y_j=\frac{1}{1+\epsilon}x_j-\frac{\epsilon}{1+\epsilon}\frac{1}{(1+\epsilon)^r-1},\quad j=1,\ldots,d,$$

in the first integral leads to

$$\begin{aligned}&\int_{\mathcal{H}_d} f\left(\frac{1}{1+\frac{\epsilon}{2}}x-\frac{\epsilon}{2+\epsilon}\frac{1}{(1+\epsilon)^r-1}\right)dx\\&\quad=(1+\epsilon)^d\int_{\mathcal{H}_d^*} f\left(\frac{1+\epsilon}{1+\frac{\epsilon}{2}}x+\frac{\epsilon}{2+\epsilon}\frac{1}{(1+\epsilon)^r-1}\right)dx\,,\end{aligned}$$

where $\mathcal{H}_d^*$ stands for the d-dimensional hypercube

$$\mathcal{H}_d^*=\left[-\frac{\epsilon}{2(1+\epsilon)}\frac{1}{(1+\epsilon)^r-1},\frac{1+\frac{\epsilon}{2}}{1+\epsilon}-\frac{\epsilon}{2(1+\epsilon)}\frac{1}{(1+\epsilon)^r-1}\right]^d.$$

Consequently, using $(1+\epsilon)^d\le 1+d\epsilon+2^{d-3}d(d-1)\epsilon$ for $0<\epsilon\le 1$, and the fact that

$$\int_{\mathcal{H}_d^*} f\left(\frac{1+\epsilon}{1+\frac{\epsilon}{2}}x+\frac{\epsilon}{2+\epsilon}\frac{1}{(1+\epsilon)^r-1}\right)dx\le 1\,,$$

we deduce that

$$\begin{aligned}B_n\le d\epsilon&\left[\frac{3B}{2\big[(1+\epsilon)^r-1\big]}+1+2^{d-3}(d-1)\right]\\&+\int_{\mathcal{H}_d^*} f\left(\frac{1+\epsilon}{1+\frac{\epsilon}{2}}x+\frac{\epsilon}{2+\epsilon}\frac{1}{(1+\epsilon)^r-1}\right)dx\\&-\int_{\mathcal{A}_d} f\left(\frac{1+\epsilon}{1+\frac{\epsilon}{2}}x+\frac{\epsilon}{2+\epsilon}\frac{1}{(1+\epsilon)^r-1}\right)dx\,.\end{aligned}$$

Observe finally that

$$\begin{aligned}&\int_{\mathcal{H}_d^*} f\left(\frac{1+\epsilon}{1+\frac{\epsilon}{2}}x+\frac{\epsilon}{2+\epsilon}\frac{1}{(1+\epsilon)^r-1}\right)dx\\&\quad\le\frac{Bd\epsilon}{2\big[(1+\epsilon)^r-1\big]}+\int_{\mathcal{A}_d} f\left(\frac{1+\epsilon}{1+\frac{\epsilon}{2}}x+\frac{\epsilon}{2+\epsilon}\frac{1}{(1+\epsilon)^r-1}\right)dx\,.\end{aligned}$$

By definition of ϵ, we obtain

$$\begin{aligned}B_n &\le d\epsilon\Big[\frac{2B}{(1+\epsilon)^r-1}+1+2^{d-3}(d-1)\Big]\\ &= d\epsilon\big[3+2^{d-3}(d-1)\big]\\ &\overset{\text{def}}{=} A\epsilon.\end{aligned}$$

Let us now turn to the analysis of V_n. From a straightforward adaptation of theorem 7.3, pp. 113 of Devroye (1987),

$$\mathbb{E}\Big\{\int_{\mathcal{P}_d}|f_n-\mathbb{E}\{f_n\}|\Big\}\le\int_{\mathcal{P}_d}\sqrt{\frac{\mathbb{E}\{f_n\}}{n\tilde{\mathbf{h}}}}\,,$$

and therefore

$$\begin{aligned}\mathbb{E}\Big\{\int_{\mathcal{P}_d}|f_n-\mathbb{E}\{f_n\}|\Big\} &\le \int_{\mathcal{P}_d}\sqrt{\frac{f-f+\mathbb{E}\{f_n\}}{n\tilde{\mathbf{h}}}}\\ &\le \int_{\mathcal{P}_d}\sqrt{\frac{f}{n\tilde{\mathbf{h}}}}+\int_{\mathcal{P}_d}\sqrt{\frac{|f-\mathbb{E}\{f_n\}|}{n\tilde{\mathbf{h}}}}\\ &\le \frac{1}{\sqrt{n}}\sqrt{\int_{\mathcal{P}_d}\frac{1}{\tilde{\mathbf{h}}}}+\frac{1}{\sqrt{n}}\sqrt{\int_{\mathcal{P}_d}|f-\mathbb{E}\{f_n\}|}\sqrt{\int_{\mathcal{P}_d}\frac{1}{\tilde{\mathbf{h}}}}\\ &\le \frac{1+\sqrt{A\epsilon}}{\sqrt{n}}\sqrt{\int_{\mathcal{P}_d}\frac{1}{\tilde{\mathbf{h}}}}\quad(\text{as } B_n\le A\epsilon)\\ &\le \frac{2}{\sqrt{n}}\sqrt{\int_{\mathcal{P}_d}\frac{1}{\tilde{\mathbf{h}}}}\quad\text{because we will insist that } A\epsilon\le 1.\end{aligned}$$

Elementary computations show that, for $0<\epsilon\le 1$,

$$\begin{aligned}\int_{\mathcal{P}_d}\frac{1}{\tilde{\mathbf{h}}} &= \Big(\frac{2+\epsilon}{2\epsilon}\Big)^d\prod_{j=1}^{d}2r\log(1+\epsilon)=\Big(\frac{2+\epsilon}{2\epsilon}\Big)^d 2^d r^d\big(\log(1+\epsilon)\big)^d\\ &\le \Big(\frac{2+\epsilon}{\epsilon}\Big)^d r^d\epsilon^d\le 3^d r^d.\end{aligned}$$

Putting all pieces together, we obtain

$$\mathbb{E}\Big\{\int|f_n-f|\Big\}\le A\epsilon+A'\sqrt{\frac{r^d}{n}}\,,$$

with $A'=2\sqrt{3^d}$. Using the inequality $r^d<2^dx^d$ for $r-1<x$, valid for $r\ge 2$, the end of the proof is similar to the end of the proof for Birgé's multivariate histogram, with

$$\begin{aligned}C_1 &= (d+2)\sqrt{6^d},\\ C_2 &= \frac{d\sqrt{6^d}\,e^{1/R}}{2R},\end{aligned}$$

$$C_3 = \max\left[R^d, 1/\Big(R^2 \log\left((d+2)(1+A^{-1})\right)\Big), \frac{2^{d+2}}{R^2}\right]. \square$$

Remark. Grenander's estimate. Grenander's estimate for univariate monotone densities (Grenander (1956)) has been shown to be minimax optimal for $\mathcal{F}_B$ (Birgé (1989)). The multivariate generalization of it was shown by Polonik (1995a, 1995b, 1998) to be the maximum likelihood estimate. However, it was not shown there to be minimax optimal, so that question remains open. For additional results on Grenander's estimate, we refer to Wegman (1969, 1970a, 1970b) and Groeneboom (1983).

Remark. On the constraints. Note that both the constraints on S and n become more severe as the dimension increases. We have $C_3 = 5.33, 47.02, 276.48$ for $d = 1, 2, 3$, respectively.

§5. Adaptive polynomial bandwidth kernel estimates

We saw in the previous section why kernel estimates with bandwidths of the form $a+bx$, $x \geq 0$ are important for monotone densities on the real line—they are minimax optimal modulo a universal constant. However, the form of bandwidth suggested in the previous section cannot be computed, since it involves unknown quantities. In $\mathbb{R}^d$, a d-dimensional bandwidth vector with j-th component $a_j + b_j x_j$ is similarly optimal for d-dimensional block decreasing densities. If one uses the values for a_j and b_j suggested in section 4, then it is necessary to at least know (or have a good estimate of) $f(0)$ and the smallest box $[0, s_1] \times \cdots \times [0, s_d]$ of unit probability. By rescaling the minimax results appropriately, we note that each h_j is of the form $s_j(a_j + b_j x_j / s_j)$, where a_j and b_j are functions of n, d and $\gamma(f) \overset{\text{def}}{=} f(0)\prod_{j=1}^d s_j$ only. However, what happens if f is not bounded or not of compact support? Or what happens when f is far removed from the densities that occur as worst cases in the minimax bound? Thus, to be really useful, we need to be able to choose the a_j's and b_j's automatically, so that we get near-optimal expected error rates for all densities, and near-optimal bounds for block monotone densities that match the minimax lower bounds derived earlier. This sort of universal robustness is derived here, based on the combinatorial method of Devroye and Lugosi (2001).

Let

$$\Theta = \left\{\theta = (h_1, \ldots, h_d) : \min_j h_j > 0\right\}$$

be the parameter space of all bandwidths. Consider the product kernel estimate on $\mathbb{R}^d$ defined by

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

where

$$K_\theta(x) = \prod_{j=1}^{d} \frac{1}{h_j} K_j\left(\frac{x^{(j)}}{h_j}\right) \ , \quad x = \left(x^{(1)}, \ldots, x^{(d)}\right),$$

and $K_1, \ldots, K_d$ are fixed 1-dimensional kernels integrating to one. In this estimate, the smoothing factor varies in each direction. This, of course, is the multivariate extension of the Akaike-Parzen-Rosenblatt kernel estimate. The automatic selection of the bandwidths has been discussed in Devroye and Lugosi (2001), where additional references may be found. The most recent results on the consistency of multivariate multiparameter kernel estimates are in Devroye and Krzyżak (2000).

The variable kernel estimate of the Breiman-Meisel-Purcell kind is defined by

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}\frac{1}{h_{j,i}}K\left(\frac{x^{(j)} - X_i^{(j)}}{h_{j,i}}\right) \ , \quad x = \left(x^{(1)}, \ldots, x^{(d)}\right),$$

where K is a 1-dimensional kernel (typically a density), and each X_i has its own set of d bandwidths, $(h_{1,i}, \ldots, h_{d,i})$, which possibly depends upon the data. The original paper by Breiman, Meisel, and Purcell (1977) had the 1-dimensional version of this, which was subsequently studied by Habbema, Hermans, and Remme (1978), Abramson (1982), Devroye (1985), Devroye and Penrod (1986), Hall and Marron (1988), Mielniczuk, Sarda, and Vieu (1989), Hall (1992), Terrell and Scott (1992), Marron, Hall, and Hu (1995), Hazelton (1996), Sain and Scott (1996, 1997), and Devroye and Lugosi (2000).

If K is a density, then so is f_n. A second estimate that is not a density but may nevertheless have interesting properties is one in which the smoothing factor depends upon x, not X_i, as in

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}\frac{1}{h_j(x)}K\left(\frac{x^{(j)} - X_i^{(j)}}{h_j(x)}\right),$$

where $h(x) = \big(h_1(x), \ldots, h_d(x)\big)$ is a given vector of functions of x. Typically, f_n is no longer a density. An interesting approach to local bandwidth selection could consist in parameterizing h in an appropriate way. For example, we may have a polynomial choice

$$h_j(x) = \sum_{\ell=0}^{k} a(j,\ell)\left(x^{(j)}\right)^{\ell}$$

for coefficients $a(j,\ell)$. Clearly, care must be taken to insure that $h_j(x) > 0$ for all x, so not all choices of coefficients are feasible. It is easy to avoid that problem, but that will not concern us here. For example, one might set

$$h_j(x) = \exp\left(\sum_{\ell=0}^{k} a(j,\ell)\left(x^{(j)}\right)^{\ell}\right)$$

or something similar.

Define $\theta = \big(a(j,\ell)\big)_{0\le\ell\le k, 1\le j\le d}$ as a $(k+1)d$-dimensional vector of parameters. The space for θ is Ω, where we only restrict $a(j,0) > 0$ for all j. Denote the variable kernel estimate with polynomial choice of each h_j by $f_{n,\theta}$. Our interest in this section is in the data-based selection of θ when the kernel is the uniform kernel $K = 1_{[-1/2,1/2]}$. We call the estimate with these choices of h_j and K the *polynomial bandwidth kernel estimate*. We are only interested in densities on $Q = [0, \infty)^d$ with our particular choice of θ. The purpose of this section is to show that there exists an algorithm for picking θ depending upon the data (this data-based choice will be called Θ) such that

$$\mathbb{E}\left\{\int |f_{n,\Theta} - f|\right\} \le C_1 \inf_{\theta \in \Omega} \mathbb{E}\left\{\int |f_{n,\theta} - f|\right\} + C_2 \sqrt{\frac{\log n}{n}}$$

where C_2 depends upon d and k only and C_1 is a universal constant. As Ω contains the standard product kernel density estimate, we note that modulo a factor C_1, the data-based estimate is as good as the best product kernel estimate, even if f were given to us beforehand. In particular, we have universal consistency: for all f,

$$\lim_{n\to\infty} \mathbb{E}\left\{\int |f_{n,\Theta} - f|\right\} = 0 \ .$$

If we apply the above result to the estimation of block decreasing densities on Q, we note that the kernel estimate of the previous section had $k = 1$, with all $2d$ coefficients given explicitly in terms of the supports s_j and a shape parameter $\gamma(f) = f(0)\prod_{j=1}^d s_j$. This setting is of course contained in Ω. For that choice, the kernel estimate of that section had an expected L_1 error $O\Big(\big(\big(\log(\gamma(f)+1)\big)^d/n\big)^{1/(d+2)}\Big)$. Let $\mathcal{F}$ denote the class of all block decreasing densities on Q. As the inequality above is valid for all n, we thus conclude that $f_{n,\Theta}$ has the same order for its expected L_1 error for densities that have $\gamma(f) < \infty$, and this is adaptive in a very strong sense:

$$\sup_{t>0} \sup_{\{f\in\mathcal{F}:\gamma(f)\le t\}} \frac{\mathbb{E}\left\{\int |f_{n,\Theta} - f|\right\}}{\left[\big(\log(t+1)\big)^d/n\right]^{1/(d+2)}} = O(1) \ .$$

Thus, we need not to know the support, $f(0)$, or $\gamma(f)$ to get performance commensurate with the minimax lower bound, uniformly over all block decreasing f with $\gamma(f) \le t$. In other words, the estimate “adjusts” to the individual values $\gamma(f)$. The estimate of the previous section does not do that, as it only adjusts to the densities in the class with the highest value of $\gamma(f)$.

We now describe the manner in which we select the coefficients of the polynomial bandwidths. Split the sample in two parts, $X_1, \ldots, X_{n-m}$ and $X_{n-m+1}, \ldots, X_n$. Consider the subsets of $\mathbb{R}^d$ defined by $\{f_{n-m,\theta} > f_{n-m,\theta'}\}$, with $\theta, \theta' \in \Omega$. The collection of these sets is called a *Yatracos class* and is denoted by $\mathcal{A}$. Then select Θ so as to minimize

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

where $\mu_m(A) = (1/m)\sum_{j=n-m+1}^n 1_{[X_j \in A]}$ is the empirical measure for A based upon the second part of the sample. If the minimum does not exist, we select Θ such that

$$\sup_{A\in\mathcal{A}}\left|\int_A f_{n-m,\Theta} - \mu_m(A)\right| < \inf_{\theta^*\in\Omega}\sup_{A\in\mathcal{A}}\left|\int_A f_{n-m,\theta^*} - \mu_m(A)\right| + \frac{1}{n}.$$

For this method of choosing the polynomial bandwidths, we show the following non-asymptotic bound, valid for all n and all densities f on Q:

Theorem 4. *For the polynomial bandwidth kernel estimate with Θ as selected above, for any f on Q,*

$$\begin{aligned}&\mathbf{E}\left\{\int |f_{n,\Theta} - f|\right\}\\ &\quad\le 5\inf_{\theta\in\Omega}\mathbf{E}\left\{\int|f_{n,\theta}-f|\right\}\left(1+\frac{2m}{n-m}+8\sqrt{\frac{m}{n}}\right)+\frac{5}{n}\\ &\qquad+8\sqrt{\frac{2(k+1)d\log\big(d(n-m)\big)+\big(2(k+1)d+(k+1)^d\big)\log(2m)+3}{m}}.\end{aligned}$$

The choice $m = n/2$ yields

$$\begin{aligned}&\mathbf{E}\left\{\int |f_{n,\Theta} - f|\right\}\\ &\quad\le \inf_{\theta\in\Omega}\mathbf{E}\left\{\int|f_{n,\theta}-f|\right\}\left(15+40/\sqrt{2}\right)\\ &\qquad+8\sqrt{2}\sqrt{\frac{2(k+1)d\log(d/2)+\big(4(k+1)d+(k+1)^d\big)\log n+3}{n}}+\frac{5}{n}.\end{aligned}$$

The proof of theorem 4 follows from theorem 5 and lemma 1 below without further work. Many classical nonparametric density estimates may be written in the form

$$g_n(x) = \frac{1}{n}\sum_{i=1}^{n} K(x, X_i),$$

where $K : \mathbf{R}^d\times\mathbf{R}^d \to \mathbf{R}$ is a measurable function. Such estimates are called *additive* (Devroye and Lugosi, 2001, chapter 10). We say that the additive estimate g_n is *regular* if for each x, $\mathbf{E}\{|K(x,X)|\} < \infty$. The multivariate variable product kernel estimate introduced here is additive and regular.

Theorem 5 (Devroye and Lugosi, 2001, theorem 10.3). *Let the set Ω determine a class of regular additive density estimates (with possibly $\int f_{n-m,\theta} \neq 1$ for some or all θ). Then for all n, $m \le n/2$, and f:*

$$\begin{aligned}\mathbf{E}\left\{\int |f_{n,\Theta} - f|\right\} \le& 5\inf_{\theta\in\Omega}\mathbf{E}\left\{\int|f_{n,\theta}-f|\right\}\left(1+\frac{2m}{n-m}+8\sqrt{\frac{m}{n}}\right)\\ &+8\mathbf{E}\left\{\sqrt{\frac{\log 2\mathbb{S}_{\mathcal{A}}(m)}{m}}\right\}+\frac{5}{n}.\end{aligned}$$

Here $\mathbb{S}_{\mathcal{A}}(m)$ *is the* shatter coefficient *for* $\mathcal{A}$, *i.e., the maximal number of sets in* $\{(y_1,\ldots,y_m)\cap A : A\in\mathcal{A}\}$, *where the maximum is taken over all* $(y_1,\ldots,y_m)\in\mathrm{R}^{dm}$.

We will not be concerned with the actual details of the minimization algorithm. We realize that more work is needed to make the present method computationally feasible. Note that $\mathcal{A}$, and thus $\mathbb{S}_{\mathcal{A}}(m)$, depend upon $X_1,\ldots,X_{n-m}$. First we show that for any $\mathcal{A}$, there is a uniform bound for $\mathbb{S}_{\mathcal{A}}(m)$ over all values of $X_1,\ldots,X_{n-m}$.

Lemma 1. *Let* $f_{\ell,\theta}$ *be the polynomial bandwidth kernel estimate of order* k, *the degree of the polynomial on* R^d, *based on a fixed sample of size* ℓ. *Define the class of all sets* $\{x : f_{\ell,\theta} > f_{\ell,\theta'}\}$, *with* $\theta\neq\theta'\in\Omega$. *Then*

$$\mathbb{S}_{\mathcal{A}}(m) \le 2^{(k+1)^d+2+2(k+1)d} \times (d\ell)^{2(k+1)d} \times m^{2(k+1)d+(k+1)^d} \ .$$

Proof. We cut the proof in several parts. First we introduce the notation $x_1,\ldots,x_\ell$ for the sample from R^{dl} used in the definition of $f_{\ell,\theta}$. It is deterministic, and the bounds below will hold uniformly over all such samples. To compute the shatter coefficient, we will use $(y_1,\ldots,y_m)$ as the sample from R^{dm} to be employed. We begin by defining the vector

$$V(j,i,\theta) \stackrel{\text{def}}{=} \left(K\left(\frac{y_j^{(1)}-x_i^{(1)}}{h_1(y_j)}\right),\ldots,K\left(\frac{y_j^{(d)}-x_i^{(d)}}{h_d(y_j)}\right)\right) ,$$

where the dependence upon θ is through $h_1,\ldots,h_d$. Then define the $m\times\ell$ matrix $V(\theta)$ of vectors $V(j,i,\theta)$, $1\le j\le m$, $1\le i\le\ell$. We will first verify how many possible values this matrix can take as we vary θ. Fix $1\le j\le m$, $1\le i\le\ell$. Set $a=y_j^{(1)}-x_i^{(1)}$, $b=y_j^{(1)}$, and

$$P_1(b) = a_{10}+a_{11}b+\cdots+a_{1k}b^k \ .$$

Consider the values $K\big(a/P_1(b)\big)$ can take as $(a_{10},\ldots,a_{1k})$ varies. As $K=1_{[-1/2,1/2]}$, we note that $K\big(a/P_1(b)\big)=1_{[|P_1(b)|\ge 2|a|]}$. Note that $|P_1(b)|\ge 2|a|$ translates into

$$|a_{10}+a_{11}b+\cdots+a_{1k}b^k|\ge 2|a|$$

with the only variable items being $(a_{10},\ldots,a_{1k})$. Thus, we have two linear inequalities. Every (i,j) pair and every component index (of d possible components) yields a different pair of inequalities, for a total of $2dm\ell$ linear inequalities in the $(k+1)d$-dimensional space of the free parameters, $(a_{10},\ldots,a_{1k}),\ldots,(a_{d0},\ldots,a_{dk})$. It is well known from Cover's lemma (1965) that the number of regions in the partition defined by these inequalities does not exceed

$$2\sum_{s=0}^{(k+1)d-1}\binom{2dm\ell-1}{s} \le 2(2dm\ell)^{(k+1)d} \ .$$

Within any of these regions, the matrix $V(\theta)$ is fixed, with all kernels K taking precisely one value. We may do the same for $V(\theta')$, and thus note that there exists a partition of Ω^2 of size at most

$$\Big(2(2dm\ell)^{(k+1)d}\Big)^2$$

such that on any set of the partition, $\big(V(\theta), V(\theta')\big)$ is fixed.

Fix such a region $\mathcal{R}$ of Ω^2. This fixes all values for $V(\theta)$ and $V(\theta')$. Next we are interested in the collection of indicators

$$Z \overset{\text{def}}{=} (Z_1, \ldots, Z_m) \overset{\text{def}}{=} \left(1_{\left[f_{\ell,\theta}(y_1) > f_{\ell,\theta'}(y_1)\right]}, \ldots, 1_{\left[f_{\ell,\theta}(y_m) > f_{\ell,\theta'}(y_m)\right]} \right) .$$

Indeed, the shatter coefficient is nothing but the number of different possible values of Z as (θ, θ') varies. We will calculate a bound W on the number of different possible values of Z when (θ, θ') varies within a region $\mathcal{R}$ of Ω^2, and show that W only depends upon d, ℓ, m, k. Then, by recalling the bound on the number of regions $\mathcal{R}$, we see that the shatter coefficient is bounded by

$$W \times \left(2(2dm\ell)^{(k+1)d}\right)^2 .$$

To compute W, we fix all values of

$$V(j,i,\theta) \overset{\text{def}}{=} \left(K\left(\frac{y_j^{(1)} - x_i^{(1)}}{h_1(y_j)}\right), \ldots, K\left(\frac{y_j^{(d)} - x_i^{(d)}}{h_d(y_j)}\right) \right)$$

for all j, i. For fixed j, let N_j be the number of x_i's, $1 \le i \le \ell$, for which

$$\prod_{s=1}^{d} K\left(\frac{y_j^{(s)} - x_i^{(s)}}{h_s(y_j)}\right) = 1 ,$$

where h_s depends upon θ through the coefficients in the polynomial

$$h_s(y_j) = a_{s0} + a_{s1} y_j^{(s)} + \cdots + a_{sk} \left(y_j^{(s)}\right)^k .$$

We will write h'_s if the coefficients of θ' are used instead. Let N'_j be the corresponding value of N_j. Clearly,

$$f_{\ell,\theta}(y_j) = \frac{N_j}{n \prod_{s=1}^{d} h_s(y_j)} , \quad f_{\ell,\theta'}(y_j) = \frac{N'_j}{n \prod_{s=1}^{d} h'_s(y_j)} .$$

Thus, $\left\{f_{\ell,\theta}(y) > f_{\ell,\theta'}(y)\right\}$ as a function of a generic $y \in \mathrm{R}^d$ is a set defined by an inequality of the form

$$\prod_{s=1}^{d} h'_s(y) > c \prod_{s=1}^{d} h_s(y) ,$$

where c is a fixed value. This is a polynomial inequality with each monomial being of the form

$$\left(y^{(1)}\right)^{p_1} \times \cdots \times \left(y^{(d)}\right)^{p_d}$$

and each $0 \le p_s \le k$ for all $1 \le s \le d$. The number of such monomials does not exceed $r \overset{\text{def}}{=} (k+1)^d$. By a mapping that makes each of the monomials a new variable, it is easy to see that considered as a set in $\mathbb{R}^r$,

$$\prod_{s=1}^{d} h'_s(y) > c \prod_{s=1}^{d} h_s(y)$$

is just a homogeneous linear inequality of the form $a_1 w_1 + \cdots + a_r w_r > 0$, with the coefficients a_i depending upon the pair (θ, θ') only. The shatter coefficient for a collection of m points in $\mathbb{R}^r$ for a collection of linear halfspaces is not more than $(m+1)^r$. Thus, in particular, the number of possible values of Z is not more than $(m+1)^r$ (see, e.g., Devroye and Lugosi, 2001), and therefore,

$$W \le (m+1)^{(k+1)^d} .$$

Putting everything together, we conclude that

$$\begin{aligned} \mathbb{S}_{\mathcal{A}}(m) &\le (m+1)^{(k+1)^d} \left(2(2dm\ell)^{(k+1)d}\right)^2 \\ &\le 2^{(k+1)^d + 2 + 2(k+1)d} \times (d\ell)^{2(k+1)d} \times m^{2(k+1)d + (k+1)^d} . \end{aligned}$$

This finishes the proof of lemma 2. □

§6. Acknowledgment.

The authors thank both referees for their comments and suggestions.

§7. References

I. Abramson, "On bandwidth variation in kernel estimates—a square root law," *Annals of Statistics*, vol. 10, pp. 1217–1223, 1982.

B. C. Arnold, E. Castillo, and J. M. Sarabia, *Conditional Specification of Statistical Models*, Springer-Verlag, New York, 1999.

P. Assouad, "Deux remarques sur l'estimation," *Comptes Rendus des Séances de l'Académie des Sciences. Série I*, vol. 10, pp. 1217–1223, 1982.

L. Birgé, "On estimating a density using Hellinger distance and some other strange facts," *Probability Theory and Related Fields*, vol. 71, pp. 271–291, 1986.

L. Birgé, "Estimating a density under order restrictions: nonasymptotic minimax risk," *Annals of Statistics*, vol. 15, pp. 995–1012, 1987a.

L. Birgé, "On the risk of histograms for estimating decreasing densities," *Annals of Statistics*, vol. 15, pp. 1013–1022, 1987b.

L. Birgé, "The Grenander estimator: a nonasymptotic approach," *Annals of Statistics*, vol. 17, pp. 1532–1549, 1989.

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

J. Bretagnolle and C. Huber, "Estimation des densités: Risque minimax," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 47, pp. 119–127, 1979.

T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Transactions on Electronic Computers*, vol. 14, pp. 326–334, 1965.

S. Datta, "On smoothing a Grenander estimator," Technical Report, Department of Statistics, University of Georgia, 1992.

L. Devroye, "A note on the L_1 consistency of variable kernel estimates," *Annals of Statistics*, vol. 13, pp. 1041–1049, 1985.

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

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

L. Devroye and G. Lugosi, "Variable kernel estimates: On the impossibility of tuning the parameters," in: *High-Dimensional Probability II*, edited by E. Giné, D. Mason, and J. Wellner, pp. 405–424, Springer-Verlag, New York, 2000.

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

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.

L. Devroye and A. Krzyżak, "New multivariate product density estimators," 2000. Unpublished.

S. Dharmadhikari and K. Joag-Dev, *Unimodality, Convexity and Applications*, Academic Press, New York, 1988.

U. Grenander, "On the theory of mortality measurement, part II," *Skandinavisk Aktuarietidskrift*, vol. 39, pp. 125–153, 1956.

P. Groeneboom, "Estimating a monotone density," 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. 539–555, Wadsworth, Belmont, C.A., 1983.

J. D. F. Habbema, J. Hermans, and J. Remme, "Variable kernel density estimation in discriminant analysis," in: *COMPSTAT 78*, edited by L. C. A. Corsten and J. Hermans, Physica-Verlag, Wien, 1978.

P. Hall, "On global properties of variable bandwidth density estimators," *Annals of Statistics*, vol. 20, pp. 762–778, 1992.

P. Hall and J. S. Marron, “Variable window width kernel estimates,” *Probability Theory and Related Fields*, vol. 80, pp. 37–49, 1988.

M. Hazelton, “Bandwidth selection for local density estimation,” *Scandinavian Journal of Statistics*, vol. 23, pp. 221–232, 1996.

J. S. Marron, P. Hall, and T. C. Hu, “Improved variable window estimators of probability densities,” *Annals of Statistics*, vol. 23, pp. 1–10, 1995.

J. Mielniczuk, P. Sarda, and P. Vieu, “Local data-driven bandwidth choice for density estimation,” *Journal of Statistical Planning and Inference*, vol. 23, pp. 53–69, 1989.

W. Polonik, “Density estimation under qualitative assumptions in higher dimensions,” *Journal of Multivariate Analysis*, vol. 55, pp. 61–81, 1995a.

W. Polonik, “Measuring mass concentrations and estimating density contour clusters—an excess mass approach,” *Annals of Statistics*, vol. 23, pp. 855–881, 1995b.

W. Polonik, “The silhouette, concentration functions and ML-density estimation under order restrictions,” *Annals of Statistics*, vol. 26, pp. 1857–1877, 1998.

S. R. Sain and D. W. Scott, “On locally adaptive density estimation,” *Journal of the American Statistical Association*, vol. 91, pp. 1525–1534, 1996.

S. R. Sain and D. W. Scott, “Zero-bias locally adaptive density estimators,” Technical Report, Rice University, Houston, 1997.

G. R. Terrell and D. W. Scott, “Variable kernel density estimation,” *Annals of Statistics*, vol. 20, pp. 1236–1265, 1992.

E. J. Wegman, “A note on estimating a unimodal density,” *Annals of Mathematical Statistics*, vol. 40, pp. 1661–1667, 1969.

E. J. Wegman, “Maximum likelihood estimation of a unimodal density function,” *Annals of Mathematical Statistics*, vol. 41, pp. 457–471, 1970a.

E. J. Wegman, “Maximum likelihood estimation of a unimodal density, II,” *Annals of Mathematical Statistics*, vol. 41, pp. 2160–2174, 1970b.