March 10, 2005

OPTIMAL L_1 BANDWIDTH SELECTION FOR VARIABLE KERNEL DENSITY ESTIMATES

Alain BERLINET, Gérard BIAU and Laurent ROUVIÈRE *

Institut de Mathématiques et de Modélisation de Montpellier, UMR CNRS 5149, Equipe de Probabilités et Statistique, Université Montpellier II, Cc 051, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

Abstract

It is well established that one can improve performance of kernel density estimates by varying the bandwidth with the location and/or the sample data at hand. Our interest in this paper is in the data-based selection of a variable bandwidth within an appropriate parameterized class of functions. We present an automatic selection procedure inspired by the combinatorial tools developed in Devroye and Lugosi (2001). It is shown that the expected L_1 error of the corresponding selected estimate is up to a given constant multiple of the best possible error plus an additive term which tends to zero under mild assumptions.

Index Terms **— Variable kernel estimate, nonparametric estimation, partition, shatter coefficient.**

***AMS 2000 Classification*: 62G05.**

1 Introduction

Assume we are given an *i.i.d.* sample $X_1, \ldots, X_n$ drawn from an unknown probability density f on $\mathbb{R}^d$. One of the most popular estimates of f is the *fixed bandwidth kernel estimate* defined by

$$f_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big), \quad x \in \mathbb{R}^d, \tag{1.1}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel with $\int K = 1$ and $h > 0$ is the bandwidth (or smoothing parameter), see Rosenblatt (1956) or Parzen (1962). The terminology *fixed bandwidth* means that the parameter h is held constant across

*Corresponding author. Email: rouviere@ensam.inra.fr .

x and the X_i's (but it can depend on n). While the estimate (1.1) performs well for most regular densities, its capabilities are known to decrease when estimating more complex functions such as multimodal densities (Sain and Scott, 1996). Moreover, as the dimensionality increases, the so-called curse of dimensionality affects the quality of the estimation. Due to the sparseness of data in higher dimensions, multivariate neighborhoods are often empty, particularly in the tails of the density. Therefore, larger and larger bandwidths are necessary in the tails. However, this also has adverse effect of oversmoothing the main features (such as bumps and modes, see Sain, 2002). These drawbacks can be overcome, to some extent, by varying the bandwidth in order to better capture the local behavior of the underlying density. For that purpose, two big families of *variable* (*bandwidth*) *kernel estimates* have been considered in the literature.

The variable estimates of the first family have a bandwidth which is allowed to vary with the location x. Its members are often referred to as *balloon estimates* and take the form

$$f_n(x) = \frac{1}{nh(x)} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h(x)}\Big) .$$

Such estimates lead to substantial gains over the fixed bandwidth in higher dimensional spaces and, to some extent, circumvent the curse of dimensionality (Terrell and Scott, 1992).
The second family of variable kernel estimates was originally considered by Breiman, Meisel and Purcell (1977), who suggested varying the bandwidth at each sample point, leading to the so-called *sample point estimates*

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(X_i)} K\Big(\frac{x - X_i}{h(X_i)}\Big) . \tag{1.2}$$

An appealing property of the above estimate is that a good choice of $h(X_i)$ allows to reduce the bias. As a matter of fact, Abramson (1982) shows that the bias-rate usually reserved for fixed kernel estimates using negative fourth order kernels is actually achievable by estimates of the form (1.2). For a complete and comprehensive description of variable kernel estimates and their properties, we refer the reader to Jones (1990) who also discusses a variable bandwidth depending on both the location and the sample points.

To exploit the advantages offered by variable kernel estimates, one has to design a good data-dependent way of determining the bandwidth function. As an important (but negative) result towards this direction, Devroye and Lugosi (2000) show that it is impossible to find an optimal way of selecting the smoothing function if this latter is allowed to depend on the location x only. More precisely, consider the class of univariate variable estimates

$$f_{n,h(x)}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h(x)}K\Big(\frac{x-X_i}{h(x)}\Big),$$

where the bandwidth $h(x)$ is allowed to be any measurable function $h:\mathbb{R}\to(0,\infty)$ of x. Then a data-based variable kernel estimate has the form

$$f_{n,H(x)}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{H(x)}K\Big(\frac{x-X_i}{H(x)}\Big),$$

where it is understood that $H(x) = H(x; X_1, \ldots, X_n)$. Ideally, one would like to find $H(x)$ so that the expected error $\mathbf{E}\{\int |f_{n,H(x)}(x) - f(x)|\,\mathrm{d}x\}$ is close to the ideal value $\inf_{h:\mathbb{R}\to(0,\infty)}\mathbf{E}\{\int |f_{n,h(x)}(x) - f(x)|\,\mathrm{d}x\}$ for all densities. Unfortunately, Devroye and Lugosi (2000) prove that if K is a symmetric nonnegative square-integrable kernel with compact support, then

$$\inf_{H:\mathbb{R}^{n+1}\to(0,\infty)}\sup_{f\in\mathcal{F}_B}\frac{\mathbf{E}\{\int |f_{n,H(x)}(x) - f(x)|\,\mathrm{d}x\}}{\inf_{h:\mathbb{R}\to(0,\infty)}\mathbf{E}\{\int |f_{n,h(x)}(x) - f(x)|\,\mathrm{d}x\}} \geq Cn^{\frac{1}{10}},$$

where $\mathcal{F}_B$ denotes the class of nondecreasing, convex-shaped densities f on $[0,1]$ with $\sup_{(0,1)} f(x) \leq B$ and C is a positive universal constant. This inequality shows that even with the knowledge that $f \in \mathcal{F}_B$, one cannot efficiently design a variable bandwidth. In other words this class of variable bandwidth kernel estimates is too large to be optimized.

Thus, one should constrain the class of possible bandwidth functions from which selection is made. This is precisely the problem that we address in the present paper, using a general multivariate data-based combinatorial methodology presented in Devroye and Lugosi (2001). More precisely, we will show how to select the smoothing function within an appropriate class so that the expected L_1 error of the corresponding selected estimate is up to a given constant multiple of the best possible error plus an additive term which tends to zero under mild assumptions. The paper is organized as follows. In Section 2, we present the multivariate selection procedure. We then specify the algorithm to the bandwidth function selection problem in Section 3. Examples are worked out in Section 4 for different models of variable kernel estimates, and Section 5 is devoted to the proofs.

2 Automatic parameter selection

Using ideas from Yatracos (1985), Devroye and Lugosi (2001) explore a new paradigm for the data-based or automatic selection of the free parameters of density estimates in general so that the expected L_1 error is within a given constant multiple of the best possible error. To summarize in the present context, assume we are given a class of density estimates parameterized by $\theta \in \Theta$ such that $f_{n,\theta}$ denotes the density estimate with parameter θ. Moreover, assume that each $f_{n,\theta}$ may be written in the form

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

where $K_\theta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a measurable function such that for each x, $\mathbf{E}\{|K_\theta(x, X)|\} < \infty$. Such estimates are called additive and regular (Devroye and Lugosi, 2001, Chapter 10). Examples of additive and regular estimates include the kernel, histogram, series and wavelet estimates. Now, let $m < n$ be an integer which splits the data $X_1, \ldots, X_n$ into

- a set $X_1, \ldots, X_{n-m}$ used for the construction of the density estimates;
- a validation set $X_{n-m+1}, \ldots, X_n$.

Introduce the class of random sets

$$\mathcal{A}_\Theta = \Big\{\big\{x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x)\big\} : (\theta, \theta') \in \Theta^2\Big\}$$

($\mathcal{A}_\Theta$ is the so-called *Yatracos class* associated with Θ) and define

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

where $\mu_m(A) = (1/m)\sum_{i=n-m+1}^{n} \mathbf{1}_{[X_i \in A]}$ is the empirical measure associated with the subsample $X_{n-m+1}, \ldots, X_n$. Then the *minimum distance estimate* f_n is defined as any density estimate selected among those $f_{n-m,\theta}$ with

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

Note that the $1/n$ term is added to ensure the existence of such a density estimate. According to Devroye and Lugosi (2001), Chapter 10, whenever $f_{n-m,\theta}$ is integrable (and not necessarily nonnegative), the selected f_n satisfies the following inequality, valid for all n and $m \leq n/2$:

$$\begin{aligned}\mathbf{E}\Big\{\int |f_n - f|\Big\} \leq 5\left(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\right) \inf_{\theta \in \Theta} \mathbf{E}\Big\{\int |f_{n,\theta} - f|\Big\} \\ + 8\mathbf{E}\left\{\sqrt{\frac{\log 2\,\mathbf{S}_{\mathcal{A}_\Theta}(m)}{m}}\right\} + \frac{5}{n}\end{aligned} \tag{2.1}$$

(for the sake of clarity, we drop the "dx" notation when no confusion is possible). Here, $\mathbf{S}_{\mathcal{A}_\Theta}(m)$ is the *Vapnik-Chervonenkis shatter coefficient* of the class of sets $\mathcal{A}_\Theta$ (Vapnik and Chervonenkis, 1971), defined by

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) = \max_{x_1,\ldots,x_m \in \mathbb{R}^d} \mathrm{Card}\big\{\{x_1,\ldots,x_m\} \cap A : A \in \mathcal{A}_\Theta\big\}.$$

This general methodology provides us with an automatic procedure to construct a density estimate f_n whose L_1 error is (almost) as small as that of the best estimate among the $f_{n,\theta}$, $\theta \in \Theta$. We emphasize that inequality (2.1) is nonasymptotic, that is, the bound is valid for all n. The rest of the analysis consists in obtaining upper bounds for the value of $\mathbf{S}_{\mathcal{A}_\Theta}(m)$.

3 Selecting a variable kernel estimate

In this section, we will be concerned with the selection of a bandwidth function in the variable kernel estimate, using the general combinatorial tools presented above. Moreover, to improve performance over ordinary kernel estimates for densities with varying behavior in different regions of the space, we shall use different parameterized smoothing functions in different regions of $\mathbb{R}^d$. For that purpose, let $\mathcal{P}_1$ (*resp.* $\mathcal{P}_2$) be a class of partitions of $\mathbb{R}^d$ such that each $P_1 = \{B_1^1,\ldots,B_{r_1}^1\} \in \mathcal{P}_1$ (*resp.* $P_2 = \{B_1^2,\ldots,B_{r_2}^2\} \in \mathcal{P}_2$) has at most r_1 (*resp.* r_2) cells. Denote by J the set $\{1,\ldots,r_1\} \times \{1,\ldots,r_2\}$ and by $\underline{\lambda} = (\lambda_{j_1j_2} : (j_1,j_2) \in J)$ a generic vector of $\mathbb{R}^{r_1r_2p}$.

To go straight to the point, we will assume that the variable bandwidth is a parameterized measurable function $h : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{r_1r_2p} \to (0,\infty)$ of the form

$$h(x,X_i,\theta) = \sum_{(j_1,j_2)\in J} \phi(x,X_i,\lambda_{j_1j_2}) \mathbf{1}_{B_{j_1}^1 \times B_{j_2}^2}(x,X_i),$$

where the parameter $\theta = (P_1,P_2,\underline{\lambda})$ and, for fixed x and X_i, each map $\lambda_{j_1j_2} \mapsto \phi(x,X_i,\lambda_{j_1j_2})$ is a polynomial function over $\mathbb{R}^p$ (the monomials are combinations of the components $\lambda_{j_1j_2}$) of degree no more than ℓ. To each partition $(P_1,P_2) \in \mathcal{P}_1 \times \mathcal{P}_2$ and parameter vector $\underline{\lambda} = (\lambda_{j_1j_2} : (j_1,j_2) \in J)$ we may now associate the corresponding variable bandwidth estimate $f_{n,\theta}(x)$ defined with $\theta = (P_1,P_2,\underline{\lambda})$. In other words, for $x \in \mathbb{R}^d$,

$$f_{n,\theta}(x) = \frac{1}{n}\sum_{i=1}^n \sum_{(j_1,j_2)\in J} \frac{\mathbf{1}_{[\phi(x,X_i,\lambda_{j_1j_2})>0]}}{\phi(x,X_i,\lambda_{j_1j_2})} K\Big(\frac{x-X_i}{\phi(x,X_i,\lambda_{j_1j_2})}\Big) \mathbf{1}_{B_{j_1}^1 \times B_{j_2}^2}(x,X_i), \tag{3.1}$$

with the usual convention $0 \times \infty = 0$. Observe that $f_{n,\theta}$ is not a bona fide density since it usually fails to integrate to one. Note also that we require the functions ϕ to be polynomial in their parameters $\lambda_{j_1j_2}$ *only*. This allows us to deal with a large choice of bandwidth models, see the examples in Section 4.

Now, we can use the combinatorial method described in Section 2 to select θ from the set

$$\Theta = \left\{(P_1, P_2, \underline{\lambda}) : P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2, \underline{\lambda} \in \mathbb{R}^{r_1 r_2 p}\right\},$$

and we let f_n be the resulting minimum distance estimate. To apply (2.1), we merely need to obtain upper bounds for the mth shatter coefficient $\mathbf{S}_{\mathcal{A}_\Theta}(m)$ of the Yatracos class associated with Θ. With a slight abuse of notation we denote by $\mathbf{S}_{\mathcal{P}}(j)$ the jth shatter coefficient of the class of sets $B^1 \times B^2$, where B^1 (*resp.* B^2) is any cell of any partition in $\mathcal{P}_1$ (*resp.* $\mathcal{P}_2$), and, for simplicity, we assume that $K(x) = c\mathbf{1}_{[||x|| \le 1]}$, where c is an appropriate normalizing factor.

Proposition 3.1 *If $\mathcal{A}_\Theta$ is the Yatracos class defined by*

$$\Theta = \left\{(P_1, P_2, \underline{\lambda}) : P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2, \underline{\lambda} \in \mathbb{R}^{r_1 r_2 p}\right\},$$

then

$$\begin{aligned}\mathbf{S}_{\mathcal{A}_\Theta}(m) \le & 2^{1+(2+4p)r_1 r_2}\, \ell^{4 r_1 r_2 p}\left[(m(n-m)+1)(2(n-m)-1)(m+1)\right]^{2 r_1 r_2 p} \\ & \times \mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)^{2 r_1 r_2}.\end{aligned}$$

Note that the above upper bound is non random. The proof of Proposition 3.1 relies on a lemma of Bartlett, Maiorov, and Meir (1998). This lemma bounds the number of distinct sign vectors that can be generated using polynomial functions. However, we emphasize that other types of functional dependencies are feasible. For example, Theorem 8.14, page 124 in Anthony and Bartlett (1999) provides a portmanteau result for more general function classes in terms of the number of arithmetic operations required for computing the functions. Combining the result of Proposition 3.1 with (2.1) leads to the following performance bound for the minimum distance estimates f_n.

Theorem 3.1 *Let $\mathcal{A}_\Theta$ be the Yatracos class defined by*

$$\Theta = \left\{(P_1, P_2, \underline{\lambda}) : P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2, \underline{\lambda} \in \mathbb{R}^{r_1 r_2 p}\right\}.$$

Assume that for every $(a, \lambda) \in \mathbb{R}^d \times \mathbb{R}^p$

$$\int \frac{1}{\phi(x,a,\lambda)}\, K\Big(\frac{x-a}{\phi(x,a,\lambda)}\Big)\, \mathrm{d}x < \infty .$$

Then, for $m \le n/2$, we have

$$\begin{aligned}\mathbf{E}\Big\{\int |f_n - f|\Big\} &\le 5\Big(1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}}\Big)\inf_{\theta\in\Theta}\mathbf{E}\Big\{\int |f_{n,\theta} - f|\Big\} \\ &\quad + 8\Bigg[\frac{\big(1+(2+4p)r_1r_2\big)\log 2 + 4r_1r_2p\log\ell + 2r_1r_2\log \mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)}{m} \\ &\quad + \frac{2r_1r_2p\log\big[(m(n-m)+1)(2(n-m)-1)(m+1)\big]}{m}\Bigg]^{1/2} + \frac{5}{n}\,.\end{aligned}$$

Theorem 3.1 generalizes to more complex bandwidths a result of Devroye, Lugosi, and Udina (2000), who partition the sample and use a different fixed bandwidth in each cell of the partition. The complexity of the class of possible partitions among which we select appears in the bounds. Larger families of partitions offer better flexibility, but it is more difficult to select the best among them. To make the above theorem useful, the classes of partitions $\mathcal{P}_k$ ($k = 1, 2$) have to be restricted in such a way that

$$\frac{\log \mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)}{m} \to 0 \quad \text{as } n \to \infty\,.$$

Recall that $\mathbf{S}_{\mathcal{P}}(m) \le \mathbf{S}_{\mathcal{P}_1}(m)\mathbf{S}_{\mathcal{P}_2}(m)$, where $\mathbf{S}_{\mathcal{P}_k}(m)$ stands for the mth shatter coefficient of the class of sets which are cells of any partition in $\mathcal{P}_k$ ($k = 1, 2$). As a simple but important example, consider $d = 1$, and let $\mathcal{P}_k$ be the class containing all partitions of the real line into at most r_k intervals. Then $\mathbf{S}_{\mathcal{P}_k}(m)$ is just the mth shatter coefficient of the class of all intervals, which equals $m(m+1)/2+1$. More generally, if $\mathcal{P}_k$ stands for the class of partitions of $\mathbb{R}^d$ into at most r_k rectangles, then the shatter coefficient $\mathbf{S}_{\mathcal{P}_k}(m)$ is known to be bounded by $(m+1)^{2d}$. Of course, other multivariate examples, such as tree or Voronoi partitions, are feasible (see Devroye, Györfi, and Lugosi, 1996, Chapter 13). In all those standard examples,

$$\log \mathbf{S}_{\mathcal{P}_k}\big(m(n-m)\big) = \mathrm{O}(\log n)\,.$$

Therefore, keeping r_k, p and ℓ fixed, and considering for example the choice $m = n/\log n$, we obtain

$$\mathbf{E}\Big\{\int |f_n - f|\Big\} \le 5\Big(1 + \mathrm{O}\Big(\frac{1}{\sqrt{\log n}}\Big)\Big)\inf_{\theta\in\Theta}\mathbf{E}\Big\{\int |f_{n,\theta} - f|\Big\} + \mathrm{O}\Big(\frac{\log n}{\sqrt{n}}\Big)\,.$$

Since in most cases of interest, the optimal L_1 error tends to zero much slower than $1/\sqrt{n}$, this inequality means that, asymptotically, the error of the minimum distance estimate stays within a constant factor multiple of the best possible error. Note also that r_k, p and ℓ are allowed to tend to infinity with n, but should not increase so fast that the second term in the upper bound in Theorem 3.1 starts dominating. We will not be concerned with the actual details of the minimization algorithm. We realize however that more work is needed to make the present method computationally feasible.

Of course, most kernel functions used in practice are not naive kernels. However, most kernels can be well approximated by Riemann kernels of the form

$$K(x) = \sum_{i=1}^{k} \alpha_i \mathbf{1}_{A_i}(x) \,,$$

$k < \infty$ and $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$. Thus, at the price of slightly worse constants in the bounds, the results presented in the present paper can be easily adapted to all important kernels. For a complete presentation, we refer the reader to Devroye and Lugosi (2001), Chapter 11.

We close this section by exhibiting a collection of densities for which the error commited by the selected variable kernel density estimate is less than the usual kernel rate $\mathrm{O}(n^{-2/5})$. We start with a result of Devroye and Lugosi (2000, Lemma 1). These authors use ideas of Sain and Scott (2002) to prove that

$$\sup_{f \in \mathcal{F}_B} \mathbf{E}\Big\{ \int |f_{n,h(x)} - f| \Big\} \leq \sqrt{\frac{4B}{n}} \,,$$

where $\mathcal{F}_B$ is the class of non decreasing, convex-shaped densities on $[0,1]$ with $\sup_{[0,1]} f(x) \leq B$, and $f_{n,h(x)}$ is the variable kernel density estimate corresponding to the variable bandwidth

$$h(x) = \sup \big\{ z > 0 : f * K_z(x) = f(x) \big\} \,. \tag{3.2}$$

Here

$$K = \mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]} \quad \text{and} \quad K_z(x) = \frac{1}{z} K\Big(\frac{x}{z}\Big) \,.$$

Let us particularize formula (3.2) to the toy-class of linear densities on $[0,1]$, defined by

$$\mathcal{F} = \big\{ f(x) = ax + 1 - \frac{a}{2} : 0 \leq a \leq 2 \big\} \,.$$

Working out expression (3.2), one obtains that $h(x)$ falls into the general class of bandwidth functions of the form

$$h(x, \theta) = \sum_{j=1}^{3} \phi(x, \lambda_j) \mathbf{1}_{B_j}(x) \,,$$

where, for $j = 1, 2, 3$ and $\lambda_j = (\lambda_j^1, \lambda_j^2, \lambda_j^3) \in \mathbb{R}^3$,

$$\phi(x, \lambda_j) = \frac{1}{\lambda_j^1 x + 1 - \lambda_j^1/2} + \lambda_j^2 x + \lambda_j^3 \,,$$

and $P = \{B_1, B_2, B_3\}$ belongs to $\mathcal{P}$, the class of partitions of $[0,1]$ into at most three intervalls.

Using the notation $\underline{\lambda} = \left(\lambda_j^k : 1 \le j, k \le 3\right)$ for a generic vector of $\mathbb{R}^6$, we can run the combinatorial method to select θ from the set

$$\Theta = \left\{(P, \underline{\lambda}) : P \in \mathcal{P}, \underline{\lambda} \in \mathbb{R}^6\right\}.$$

Note that the class of considered bandwith functions is not polynomial in its parameters. Nevertheless our results are still valid in this extended framework (see just before Theorem 3.1). With the choice $m = n/\log n$, the selected minimum distance estimate f_n then satisfies the inequality

$$\sup_{f \in \mathcal{F}} \mathbf{E}\left\{\int |f_n - f|\right\} = \mathrm{O}\left(\frac{\log n}{\sqrt{n}}\right),$$

a much faster rate than the usual rate $\mathrm{O}(n^{-2/5})$.

4 Examples

Example 1 Biau and Devroye (2003) study the L_1 minimax risk over the multivariate class of bounded block decreasing densities. Extending former results of Birgé (1987a, 1987b), these authors show first that a suitable variable kernel estimate with linear varying bandwidth achieves the optimal minimax rate. Second, they exhibit by the present combinatorial method a data-dependent bandwidth of the form

$$h(x) = \sum_{i=0}^{q} a_i x^i,$$

and prove that the corresponding variable kernel estimate uniformly adapts over the class of bounded block decreasing densities. Clearly, this model falls into the general definition (3.1): just choose $P = \mathbb{R}^d$ and $\underline{\lambda} = (a_0, \ldots, a_q) \in \mathbb{R}^{q+1}$.

Example 2 Denote by V_d the volume of the unit sphere in $\mathbb{R}^d$ and let k be a positive real number. Terrell and Scott (1992) show that the following variable bandwidth

$$h(x,k) = \left(\frac{k}{nV_d f(x)}\right)^{1/d} \qquad (4.1)$$

is asymptotically equivalent to the k-nearest neighbor bandwidth presented by Loftsgaarden and Quesenberry (1965) (known to perform well as dimensionality increases). The problem is the selection of k. Since f is unknown, it has to be replaced by a pilot estimate computed from an independent sample. One can choose, for example, a fixed kernel estimate $\hat{f}$ designed with a Gaussian kernel and a (fixed) bandwidth selected by a data-driven method (Park and Marron, 1990). The resulting estimate is of the general form (3.1): just take $P = \mathbb{R}^d$, $\lambda = k^{1/d}$ and

$$\phi(x, \lambda) = \frac{\lambda}{\left(nV_d \hat{f}(x)\right)^{1/d}} .$$

Example 3 Abramson (1982) suggests using a variable bandwidth inversely proportional to the square root of the density at X_i, *i.e.*, $h(X_i) = \alpha f(X_i)^{-1/2}$. This adaptive choice performs well for small sample size and reduces the pointwise bias. We may still replace f by a pilot estimate computed from an independent sample of size q. For simplicity, assume that f is a density on $[0, 1]$ and, in place of f, plug the histogram estimate $\hat{f}$ anchored at 0 using at most r cells defined by the bin width vector $\tilde{h} = (\tilde{h}_1, \ldots, \tilde{h}_r)$, *i.e.*,

$$\hat{f}(x) = \sum_{j=1}^{r} \frac{\tilde{\mu}_q(B_j)}{\tilde{h}_j} \mathbf{1}_{B_j}(x) ,$$

where $B_j = (\sum_{k=1}^{j-1} \tilde{h}_k, \sum_{k=1}^{j} \tilde{h}_k]$ for $j = 1, \ldots, r$, and $\tilde{\mu}_q$ denotes the empirical measure computed from the independent sample. Here, we wish to select α and $\tilde{h}$ with the combinatorial method. In this context, the estimate suggested by Abramson (1982) reads

$$f_{n,\theta}(x) = \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{\sqrt{\tilde{\mu}_q(B_j)}}{\alpha\sqrt{\tilde{h}_j}} K\left(\frac{\sqrt{\tilde{\mu}_q(B_j)}(x - X_i)}{\alpha\sqrt{\tilde{h}_j}}\right) \mathbf{1}_{B_j}(X_i) , \qquad (4.2)$$

where $\theta = (\alpha, \lambda_1, \ldots, \lambda_r)$, and $\lambda_j = \sqrt{\tilde{h}_j}$. Note that the vector $\tilde{h}$ entirely determines the partition $(B_1, \ldots, B_r)$. Each estimate (4.2) is a member of the family (3.1).

Example 4 Jones (1990) proposes to modify Abramson's estimate by considering α as a function of the estimation point. Keeping the same notation as in Example 3, with $\alpha = \alpha(x, \mu)$ polynomial in μ, the corresponding estimates may be written as

$$f_{n,\theta}(x) = \sum_{i=1}^{n} \sum_{j=1}^{r} \frac{\sqrt{\tilde{\mu}_q(B_j)}}{\alpha(x, \mu)\sqrt{\tilde{h}_j}} K\left(\frac{\sqrt{\tilde{\mu}_q(B_j)}(x - X_i)}{\alpha(x, \mu)\sqrt{\tilde{h}_j}}\right) \mathbf{1}_{B_j}(X_i) ,$$

with $\theta = (\mu, \lambda_1, \ldots, \lambda_r)$. The estimate of Jones still falls in the general class of variable bandwidth models (3.1).

5 Proofs

The proof of Proposition 3.1 will strongly rely on the following lemma. We make use of the function sgn(.) defined by

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if} \quad x \geq 0 \\ 0 & \text{if} \quad x < 0\,. \end{cases}$$

Lemma 5.1 (Bartlett, Maiorov, and Meire, 1998)
Suppose that $f_1(.), \ldots, f_m(.)$ are fixed polynomials of degree at most ℓ in p variables. Then the number of distinct sign vectors

$$\Big(\operatorname{sgn}\big(f_1(a)\big), \ldots, \operatorname{sgn}\big(f_m(a)\big)\Big)$$

that can be generated by varying $a \in \mathbb{R}^p$ is at most

$$2(2\ell)^p(m+1)^p\,.$$

Proof of Proposition 3.1 We introduce the notation $x_1, \ldots, x_{n-m}$ for the sample from $\mathbb{R}^{d(n-m)}$ used in the definition of $f_{n-m,\theta}$. 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 $\mathbb{R}^{dm}$ to be employed. We start as in Lemma 12.3, page 123 of Devroye and Lugosi (2001). For each $\theta \in \Theta$, consider the $m \times (n-m) \times r_1 \times r_2$ array z_θ with current element

$$z_\theta^{(t,i,j_1,j_2)} = \mathbf{1}_{[\phi(y_t,x_i,\lambda_{j_1j_2})>0]}\mathbf{1}_S\left(\frac{y_t - x_i}{\phi(y_t,x_i,\lambda_{j_1j_2})}\right)\mathbf{1}_{B^1_{j_1}\times B^2_{j_2}}(y_t,x_i)\,,$$

$t \leq m, i \leq n-m, j_1 \leq r_1$, $j_2 \leq r_2$, and $S = \{x : ||x|| \leq 1\}$. We shall first bound the number of different values the array z_θ can take as θ ranges through Θ. Fix temporarily $j_1 = J_1$ and $j_2 = J_2$, and consider the submatrix $u_\theta^{(J_1,J_2)}$ with current element $z_\theta^{(t,i,J_1,J_2)}$. Since there are $m(n-m)$ different pairs (y_t, x_i), the bit vector

$$(\mathbf{1}_{B^1_{J_1}\times B^2_{J_2}}(y_t,x_i) : t \leq m, i \leq n-m)$$

can take at most $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ values as θ ranges through Θ. Consequently, the number of different values the matrix $u_\theta^{(J_1,J_2)}$ can take is bounded by the product of $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ and the number of values the bit vector

$$(\mathbf{1}_{[\phi(y_t,x_i,\lambda_{J_1J_2})>0]}\mathbf{1}_{[\phi(y_t,x_i,\lambda_{J_1J_2})\geq ||y_t-x_i||]} : t \leq m, i \leq n-m)$$

can take as $\lambda_{J_1J_2}$ runs through $\mathbb{R}^p$. Since $\mathbf{1}_{[\phi(y_t,x_i,\lambda_{J_1J_2})>0]}\mathbf{1}_{[\phi(y_t,x_i,\lambda_{J_1J_2})\geq||y_t-x_i||]}$ equals

$$\begin{cases} 1 - \mathbf{1}_{[-\phi(y_t,x_i,\lambda_{J_1J_2})\geq 0]} & \text{if } y_t = x_i \\ \mathbf{1}_{[\phi(y_t,x_i,\lambda_{J_1J_2})-||y_t-x_i||\geq 0]} & \text{otherwise,} \end{cases}$$

we conclude that

$$\begin{aligned}&\mathrm{Card}\{u_\theta^{(J_1,J_2)}:\theta\in\Theta\}\\&\le \mathrm{Card}\Big\{\big(\mathbf{1}_{[R_1(\lambda_{J_1J_2})\ge 0]},\ldots,\mathbf{1}_{[R_{m(n-m)}(\lambda_{J_1J_2})\ge 0]}\big):\lambda_{J_1J_2}\in\mathbb{R}^p\Big\}\,\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\,,\end{aligned}$$

where the $m(n-m)$ functions R_k's are defined by

$$R_k(\lambda_{J_1J_2})=\begin{cases}-\phi(y_t,x_i,\lambda_{J_1J_2}) & \text{if } y_t=x_i\\ \\ \phi(y_t,x_i,\lambda_{J_1J_2})-||y_t-x_i|| & \text{otherwise}\,.\end{cases}$$

Since the R_k's are polynomials over $\mathbb{R}^p$ of degree no more than ℓ, Lemma 5.1 shows that $u_\theta^{(J_1,J_2)}$ can take at most

$$2(2\ell)^p\,(m(n-m)+1)^p\,\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$$

values. If follows that the array z_θ can take at most

$$\Big[2(2\ell)^p\,(m(n-m)+1)^p\,\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\Big]^{r_1r_2},$$

and, similarly, that

$$\mathrm{Card}\big\{(z_\theta,z_{\theta'}):(\theta,\theta')\in\Theta^2\big\}\le\Big[2(2\ell)^p\,(m(n-m)+1)^p\,\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\Big]^{2r_1r_2}.$$

Write now $\mathcal{W}=\big\{(w,w'):(w,w')=(z_\theta,z_{\theta'})\text{ for some }(\theta,\theta')\in\Theta^2\big\}$. For fixed $(w,w')\in\mathcal{W}$, let $U_{(w,w')}$ denote the collection of all (θ,θ') such that $(z_\theta,z_{\theta'})=(w,w')$. For $(\theta,\theta')\in U_{(w,w')}$, we will use the following notation:

$$\theta=(P_1,P_2,\underline{\lambda})\quad\text{and}\quad\theta'=(P_1',P_2',\underline{\lambda}')\,,$$

with

$$P_1=\{B_1^1,\ldots,B_{r_1}^1\},\ P_2=\{B_1^2,\ldots,B_{r_2}^2\},\underline{\lambda}=(\lambda_{j_1j_2}:(j_1,j_2)\in J)$$

and

$$P_1'=\{B_1^{'1},\ldots,B_{r_1}^{'1}\},\ P_2'=\{B_1^{'2},\ldots,B_{r_2}^{'2}\},\ \underline{\lambda}'=(\lambda'_{j_1j_2}:(j_1,j_2)\in J).$$

For every $t\le m$, consider the sets

$$I_t=\big\{(i,j_1,j_2):z_\theta^{(t,i,j_1,j_2)}\neq 0,i\le n-m,j_1\le r_1,j_2\le r_2\big\}$$

and

$$I_t'=\big\{(i,j_1,j_2):z_{\theta'}^{(t,i,j_1,j_2)}\neq 0,i\le n-m,j_1\le r_1,j_2\le r_2\big\}\,.$$

Observe that y_t belongs to

$$A_{\theta,\theta'} = \{x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x)\}$$

if and only if

$$\sum_{(i,j_1,j_2)\in I_t} \frac{1}{\phi(y_t, x_i, \lambda_{j_1 j_2})} z_\theta^{(t,i,j_1,j_2)} > \sum_{(i,j_1,j_2)\in I'_t} \frac{1}{\phi(y_t, x_i, \lambda'_{j_1 j_2})} z_{\theta'}^{(t,i,j_1,j_2)} .$$

Within the set $U_{(w,w')}$, the values $z_\theta^{(t,i,j_1,j_2)}$ and $z_{\theta'}^{(t,i,j_1,j_2)}$ are fixed for all t, i, j_1 and j_2. Therefore, since the x_i's and y_t's are fixed,

$$\sum_{(i,j_1,j_2)\in I_t} \frac{1}{\phi(y_t, x_i, \lambda_{j_1 j_2})} z_\theta^{(t,i,j_1,j_2)}$$

may be written in the form

$$\frac{P_t(\underline{\lambda})}{Q_t(\underline{\lambda})},$$

where the functions $\underline{\lambda} \mapsto P_t(\underline{\lambda})$ (*resp.* $\underline{\lambda} \mapsto Q_t(\underline{\lambda})$) are polynomials over $\mathbb{R}^{r_1 r_2 p}$ of degree no more than $(n-m-1)\ell$ (*resp.* $(n-m)\ell$). Similarly, the quantity

$$\sum_{(i,j_1,j_2)\in I'_t} \frac{1}{\phi(y_t, x_i, \lambda'_{j_1 j_2})} z_{\theta'}^{(t,i,j_1,j_2)}$$

may be written in the form

$$\frac{P'_t(\underline{\lambda}')}{Q'_t(\underline{\lambda}')},$$

where P'_t (*resp.* Q'_t) are polynomials over $\mathbb{R}^{r_1 r_2 p}$ of degree no more than $(n-m-1)\ell$ (*resp.* $(n-m)\ell$). Set now $\tilde{\underline{\lambda}} = (\underline{\lambda}, \underline{\lambda}') \in \mathbb{R}^{2r_1 r_2 p}$ and, for $t = 1, \ldots, m$, define

$$\mathcal{R}_t(\tilde{\underline{\lambda}}) = P_t(\underline{\lambda})Q'_t(\underline{\lambda}') - P'_t(\underline{\lambda}')Q_t(\underline{\lambda}) .$$

Observe that each $\mathcal{R}_t$ is a polynomial function over $\mathbb{R}^{2r_1 r_2 p}$ of degree no more than $(2(n-m)-1)\ell$. Therefore, applying again Lemma 5.1 we obtain

$$\begin{aligned}
&\operatorname{Card}\big\{\{\mathbf{1}_{[y_1\in A_{\theta,\theta'}]}, \ldots, \mathbf{1}_{[y_m\in A_{\theta,\theta'}]}\} : (\theta,\theta') \in U^2_{(w,w')}\big\} \\
&\quad\le \operatorname{Card}\big\{\{\mathbf{1}_{[\mathcal{R}_1(\tilde{\underline{\lambda}})>0]}, \ldots, \mathbf{1}_{[\mathcal{R}_m(\tilde{\underline{\lambda}})>0]}\} : \tilde{\underline{\lambda}} \in \mathbb{R}^{2r_1 r_2 p}\big\} \\
&\quad\le 2(2\ell)^{2r_1 r_2 p} \big((2(n-m)-1)(m+1)\big)^{2r_1 r_2 p} .
\end{aligned}$$

Putting all pieces together, we obtain

$$\begin{aligned}
\mathbf{S}_{\mathcal{A}_\Theta}(m) &\le 2(2\ell)^{2r_1 r_2 p}\big((2(n-m)-1)(m+1)\big)^{2r_1 r_2 p} \operatorname{Card}\mathcal{W} \\
&\le 2^{1+(2+4p)r_1 r_2}\, \ell^{4r_1 r_2 p} \big[(m(n-m)+1)(2(n-m)-1)(m+1)\big]^{2r_1 r_2 p} \\
&\qquad \times \mathbf{S}_{\mathcal{P}}(m(n-m))^{2r_1 r_2} .
\end{aligned}$$

■

Acknowledgments. The authors greatly thank an anonymous referee for his comments and suggestions. They are also indebted to the *UMR Biométrie et Analyse des Systèmes, ENSAM-INRA, Montpellier, France* for providing facilities.

References

[1] I. Abramson. On bandwidth variation in kernel estimates. *The Annals of Statistics*, 10:1217–1223, 1982.

[2] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

[3] P. L. Bartlett, V. Maiorov, and R. Meir. Almost linear VC-dimension bounds for piecewise polynomial networks. *Neural Computation*, 10:2159–2173, 1998.

[4] G. Biau and L. Devroye. On the risk of estimates for block decreasing densities. *Journal of Multivariate Analysis*, 86:143–165, 2003.

[5] L. Birgé. Estimating a density under order restrictions: nonasymptotic minimax risk. *The Annals of Mathematical Statistics*, 15:995–1012, 1987a.

[6] L. Birgé. On the risk of histograms for estimating decreasing densities. *The Annals of Mathematical Statistics*, 15:1013–1022, 1987b.

[7] L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19:135–144, 1977.

[8] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer–Verlag, New-York, 1996.

[9] L. Devroye and G. Lugosi. Variable kernel estimates: On the impossibility of tunning the parameters. In E. Giné and D. Mason, editors, *High-Dimensional Probability II*, pages 405–424. Springer–Verlag, New York, 2000.

[10] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer–Verlag, New York, 2001.

[11] L. Devroye, G. Lugosi, and F. Udina. Inequalities for a new data-based method for selecting nonparametric density estimates. In Madan L. Puri, editor, *Asymptotics in Statistics and Probability: Papers in Honor of George Gregory Roussas*, pages 133–154. VSP International Science Publishers, Zeist, The Netherlands, 2000.

[12] M. C. Jones. Variable kernel density estimates and variable kernel density estimates. *Australian Journal of Statistics*, 32:361–371, 1990.

[13] D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36:1049–1051, 1965.

[14] B. U. Park and J. S. Marron. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85(409):66–72, 1990.

[15] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[16] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27:832–837, 1956.

[17] S. R. Sain. Multivariate locally adaptive density estimators. *Computational Statistics and Data Analysis*, 39:165–186, 2002.

[18] S. R. Sain and D. W. Scott. On locally adaptive density estimation. *Journal of the American Statistical Association*, 436:1525–1534, 1996.

[19] S. R. Sain and D. W. Scott. Zero-bias locally adaptive density estimators. *Scandinavian Journal of Statistics*, 29:431–450, 2002.

[20] G. R. Terrell and D. W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20:1236–1265, 1992.

[21] 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*, 16:264–280, 1971.

[22] Y. G. Yatracos. Rates of convergence of minimun distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13:768–774, 1985.