24

Nonparametric Density Estimation

This chapter is an introduction to estimating densities if the underlying density of a sample of observations is considered completely unknown, up to existence of derivatives. We derive rates of convergence for the mean square error of kernel estimators and show that these cannot be improved. We also consider regularization by monotonicity.

24.1 Introduction

Statistical models are called *parametric models* if they are described by a Euclidean parameter (in a nice way). For instance, the binomial model is described by a single parameter p, and the normal model is given through two unknowns: the mean and the variance of the observations. In many situations there is insufficient motivation for using a particular parametric model, such as a normal model. An alternative at the other end of the scale is a *nonparametric model*, which leaves the underlying distribution of the observations essentially free. In this chapter we discuss one example of a problem of nonparametric estimation: estimating the density of a sample of observations if nothing is known a priori. From the many methods for this problem, we present two: kernel estimation and monotone estimation. Notwithstanding its simplicity, this method can be fully asymptotically efficient.

24.2 Kernel Estimators

The most popular nonparametric estimator of a distribution based on a sample of observations is the empirical distribution, whose properties are discussed at length in Chapter 19. This is a discrete probability distribution and possesses no density. The most popular method of nonparametric density estimation, the *kernel method*, can be viewed as a recipe to "smooth out" the pointmasses of sizes $1/n$ in order to turn the empirical distribution into a continuous distribution.

Let $X_1, \ldots, X_n$ be a random sample from a density f on the real line. If we would know that f belongs to the normal family of densities, then the natural estimate of f would be the normal density with mean $\overline{X}_n$ and variance S_n^2, or the function

$$x \mapsto \frac{1}{S_n\sqrt{2\pi}} e^{-\frac{1}{2}(x-\overline{X}_n)^2/S_n^2}.$$

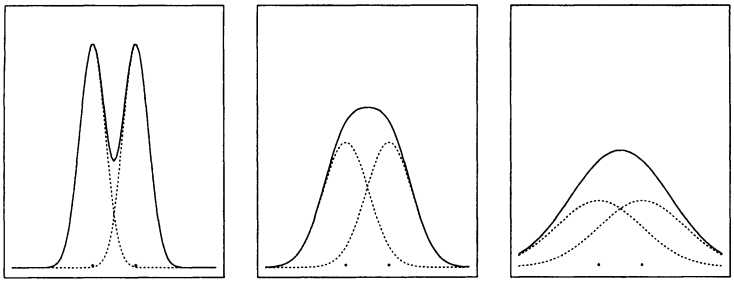

Figure 24.1. The kernel estimator with normal kernel and two observations for three bandwidths: small (*left*), intermediate (*center*) and large (*right*). The figures show both the contributions of the two observations separately (*dotted lines*) and the kernel estimator (*solid lines*), which is the sum of the two dotted lines.

In this section we suppose that we have no prior knowledge of the form of f and want to "let the data speak as much as possible for themselves."

Let K be a probability density with mean 0 and variance 1, for instance the standard normal density. A *kernel estimator* with *kernel* or *window* K is defined as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right).$$

Here h is a positive number, still to be chosen, called the *bandwidth* of the estimator. It turns out that the choice of the kernel K is far less crucial for the quality of $\hat{f}$ as an estimator of f than the choice of the bandwidth. To obtain the best convergence rate the requirement that $K \geq 0$ may have to be dropped.

A kernel estimator is an example of a *smoothing method*. The construction of a density estimator can be viewed as a recipe for smoothing out the total mass 1 over the real line. Given a random sample of n observations it is reasonable to start with allocating the total mass in packages of size $1/n$ to the observations. Next a kernel estimator distributes the mass that is allocated to X_i smoothly around X_i, not homogenously, but according to the kernel and bandwidth.

More formally, we can view a kernel estimator as the sum of n small "mountains" given by the functions

$$x \mapsto \frac{1}{nh}K\left(\frac{x - X_i}{h}\right).$$

Every small mountain is centred around an observation X_i and has area $1/n$ under it, for any bandwidth h. For a small bandwidth the mountain is very concentrated (a peak), while for a large bandwidth the mountain is low and flat. Figure 24.1 shows how the mountains add up to a single estimator. If the bandwidth is small, then the mountains remain separated and their sum is peaky. On the other hand, if the bandwidth is large, then the sum of the individual mountains is too flat. Intermediate values of the bandwidth should give the best results.

Figure 24.2 shows the kernel method in action on a sample from the normal distribution. The solid and dotted lines are the estimator and the true density, respectively. The three pictures give the kernel estimates using three different bandwidths – small, intermediate, and large – each time with the standard normal kernel.

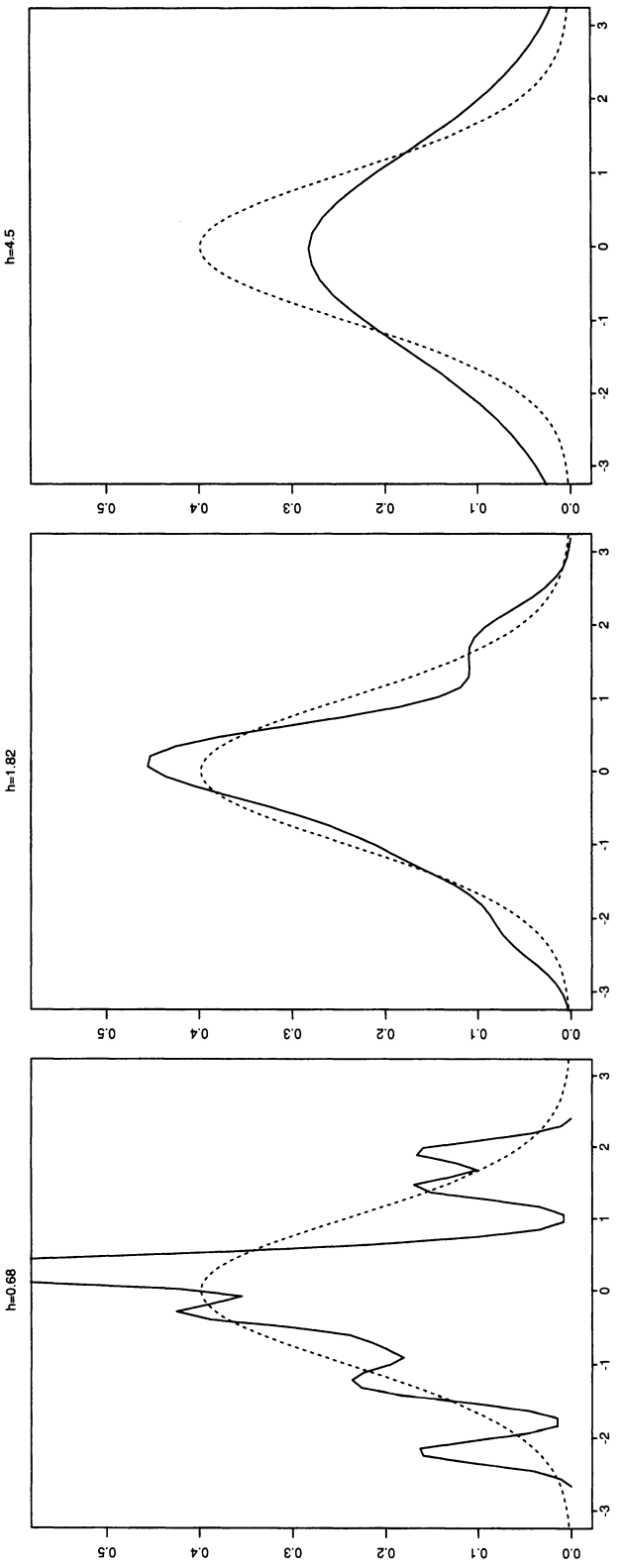

Figure 24.2. Kernel estimates for the density of a sample of size 15 from the standard normal density for three different bandwidths $h = 0{:}68$ (*left*), 1.82 (*center*), and 4.5 (*right*), using a normal kernel. The dotted line gives the true density.

A popular criterion to judge the quality of density estimators is the *mean integrated square error* (MISE), which is defined as

$$\begin{aligned}\mathrm{MISE}_f(\hat{f}) &= \int \mathrm{E}_f\big(\hat{f}(x) - f(x)\big)^2\,dx \\ &= \int \mathrm{var}_f\,\hat{f}(x)\,dx + \int \big(\mathrm{E}_f\,\hat{f}(x) - f(x)\big)^2\,dx.\end{aligned}$$

This is the mean square error $\mathrm{E}_f\big(\hat{f}(x) - f(x)\big)^2$ of $\hat{f}(x)$ as an estimator of $f(x)$ integrated over the argument x. If the mean integrated square error is small, then the function $\hat{f}$ is close to the function f. (We assume that $\hat{f}_n$ is jointly measurable to make the mean square error well defined.)

As can be seen from the second representation, the mean integrated square error is the sum of an integrated "variance term" and a "bias term." The mean integrated square error can be small only if both terms are small. We shall show that the two terms are of the orders

$$\frac{1}{nh}, \quad \text{and} \quad h^4,$$

respectively. Then it follows that the variance and the bias terms are balanced for $(nh)^{-1} \sim h^4$, which implies an optimal choice of bandwidth equal to $h \sim n^{-1/5}$ and yields a mean integrated square error of order $n^{-4/5}$.

Informally, these orders follow from simple Taylor expansions. For instance, the bias of $\hat{f}(x)$ can be written

$$\begin{aligned}\mathrm{E}_f\,\hat{f}(x) - f(x) &= \int \frac{1}{h}K\left(\frac{x-t}{h}\right) f(t)\,dt - f(x) \\ &= \int K(y)\big(f(x-hy) - f(x)\big)\,dy.\end{aligned}$$

Developing f in a Taylor series around x and using that $\int yK(y)\,dy = 0$, we see, informally, that this is equal to

$$\int y^2 K(y)\,dy\,\tfrac{1}{2}h^2 f''(x) + \cdots.$$

Thus, the squared bias is of the order h^4. The variance term can be handled similarly. A precise theorem is as follows.

24.1 ***Theorem.*** *Suppose that f is twice continuously differentiable with $\int \big|f''(x)\big|^2\,dx < \infty$. Furthermore, suppose that $\int yK(y)\,dy = 0$ and that both $\int y^2K(y)\,dy$ and $\int K^2(y)\,dy$ are finite. Then there exists a constant C_f such that for small $h > 0$*

$$\int \mathrm{E}_f\big(\hat{f}(x) - f(x)\big)^2\,dx \le C_f\left(\frac{1}{nh} + h^4\right).$$

Consequently, for $h_n \sim n^{-1/5}$, we have $\mathrm{MISE}_f(\hat{f}_n) = O(n^{-4/5})$.

Proof. Because a kernel estimator is an average of n independent random variables, the variance of $\hat{f}(x)$ is $(1/n)$ times the variance of one term. Hence

$$\begin{aligned}\operatorname{var}_f \hat{f}(x) &= \frac{1}{n}\operatorname{var}_f \frac{1}{h}K\left(\frac{x-X_1}{h}\right) \le \frac{1}{nh^2}\mathrm{E}_f K^2\left(\frac{x-X_1}{h}\right)\\ &= \frac{1}{nh}\int K^2(y) f(x-hy)\,dy.\end{aligned}$$

Take the integral with repect to x on both left and right sides. Because $\int f(x-hy)\,dx = 1$ is the same for every value of hy, the right side reduces to $(nh)^{-1}\int K^2(y)\,dy$, by Fubini's theorem. This concludes the proof for the variance term.

To upper bound the bias term we first write the bias $\mathrm{E}_f \hat{f}(x) - f(x)$ in the form as given preceding the statement of the theorem. Next we insert the formula

$$f(x+h) - f(x) = hf'(x) + h^2\int_0^1 f''(x+sh)(1-s)\,ds.$$

This is a Taylor expansion with the Laplacian representation of the remainder. We obtain

$$\mathrm{E}_f \hat{f}(x) - f(x) = \iint_0^1 K(y)\left[-hyf'(x) + (hy)^2 f''(x-shy)(1-s)\right]ds\,dy.$$

Because the kernel K has mean zero by assumption, the first term inside the square brackets can be deleted. Using the Cauchy-Schwarz inequality $(EUV)^2 \le EU^2EV^2$ on the variables $U = Y$ and $V = Yf''(x - ShY)(1-S)$ for Y distributed with density K and S uniformly distributed on $[0, 1]$ independent of Y, we see that the square of the bias is bounded above by

$$h^4\int K(y)y^2\,dy \iint_0^1 K(y)y^2 f''(x-shy)^2(1-s)^2\,ds\,dy.$$

The integral of this with respect to x is bounded above by

$$h^4\left(\int K(y)y^2\,dy\right)^2 \int f''(x)^2\,dx\,\frac{1}{3}.$$

This concludes the derivation for the bias term.

The last assertion of the theorem is trivial. ■

The rate $O(n^{-4/5})$ for the mean integrated square error is not impressive if we compare it to the rate that could be achieved if we knew a priori that f belonged to some parametric family of densities f_θ. Then, likely, we would be able to estimate θ by an estimator such that $\hat{\theta} = \theta + O_P(n^{-1/2})$, and we would expect

$$\mathrm{MISE}_\theta(f_{\hat{\theta}}) = \int \mathrm{E}_\theta\big(f_{\hat{\theta}}(x) - f_\theta(x)\big)^2\,dx \sim \mathrm{E}_\theta(\hat{\theta}-\theta)^2 = O\left(\frac{1}{n}\right).$$

This is a factor $n^{-1/5}$ smaller than the mean square error of a kernel estimator.

This loss in efficiency is only a modest price. After all, the kernel estimator works for every density that is twice continuously differentiable whereas the parametric estimator presumably fails miserably if the true density does not belong to the postulated parametric model.

Moreover, the lost factor $n^{-1/5}$ can be (almost) covered if we assume that f has sufficiently many derivatives. Suppose that f is m times continuously differentiable. Drop the condition that the kernel K is a probability density, but use a kernel K such that

$$\int K(y)\,dy = 1, \qquad \int yK(y)\,dy = 0, \ldots, \qquad \int y^{m-1}K(y)\,dy = 0,$$
$$\int |y|^m K(y)\,dy < \infty, \qquad \int K^2(y)\,dy < \infty.$$

Then, by the same arguments as before, the bias term can be expanded in the form

$$\begin{aligned} \mathrm{E}_f \hat{f}(x) - f(x) &= \int K(y)\bigl(f(x-hy) - f(x)\bigr)\,dy \\ &= \int K(y)\frac{1}{m!}(-1)^m h^m y^m f^{(m)}(x)\,dy + \cdots \end{aligned}$$

Thus the squared bias is of the order h^{2m} and the bias-variance trade-off $(nh)^{-1} \sim h^{2m}$ is solved for $h \sim n^{1/(2m+1)}$. This leads to a mean square error of the order $n^{-2m/(2m+1)}$, which approaches the "parametric rate" n^{-1} as $m \to \infty$. This claim is made precise in the following theorem, whose proof proceeds as before.

24.2 *Theorem.* *Suppose that f is m times continuously differentiable with $\int \bigl|f^{(m)}(x)\bigr|^2 dx < \infty$. Then there exists a constant C_f such that for small $h > 0$*

$$\int \mathrm{E}_f\bigl(\hat{f}(x) - f(x)\bigr)^2\,dx \le C_f\left(\frac{1}{nh} + h^{2m}\right).$$

Consequently, for $h_n \sim n^{-1/(2m+1)}$, we have $\mathrm{MISE}_f(\hat{f}_n) = O(n^{-2m/(2m+1)})$.

In practice, the number of derivatives of f is usually unknown. In order to choose a proper bandwidth, we can use *cross-validation* procedures. These yield a data-dependent bandwidth and also solve the problem of choosing the constant preceding $h^{-1/(2m+1)}$. The combined procedure of density estimator and bandwidth selection is called *rate-adaptive* if the procedure attains the upper bound $n^{-2m/(2m+1)}$ for the mean integrated square error for every m.

24.3 Rate Optimality

In this section we show that the rate $n^{-2m/(2m+1)}$ of a kernel estimator, obtained in Theorem 24.2, is the best possible. More precisely, we prove the following. Inspection of the proof of Theorem 24.2 reveals that the constants C_f in the upper bound are uniformly bounded in f such that $\int \bigl|f^{(m)}(x)\bigr|^2 dx$ is uniformly bounded. Thus, letting $\mathcal{F}_{m,M}$ be the class of all probability densities such that this quantity is bounded by M, there is a constant $C_{m,M}$ such that the kernel estimator with bandwidth $h_n = n^{-1/(2m+1)}$ satisfies

$$\sup_{f \in \mathcal{F}_{m,M}} \mathrm{E}_f \int \bigl(\hat{f}_n(x) - f(x)\bigr)^2\,dx \le C_{m,M}\left(\frac{1}{n}\right)^{2m/(2m+1)}.$$

In this section we show that this upper bound is sharp, and the kernel estimator rate optimal, in that the maximum risk on the left side is bounded below by a similar expression for *every* density estimator $\hat{f}_n$, for every fixed m and M.

The proof is based on a construction of subsets $\mathcal{F}_n \subset \mathcal{F}_{m,M}$, consisting of 2^{r_n} functions, with $r_n = \lfloor n^{1/(2m+1)} \rfloor$, and on bounding the supremum over $\mathcal{F}_{m,M}$ by the average over $\mathcal{F}_n$. Thus the number of elements in the average grows fairly rapidly with n. An approach, such as in section 14.5, based on the comparison of $\hat{f}_n$ at only two elements of $\mathcal{F}_{m,M}$ does not seem to work for the integrated risk, although such an approach readily yields a lower bound for the maximum risk $\sup_f \mathrm{E}_f\big(\hat{f}_n(x) - f(x)\big)^2$ at a fixed x.

The subset $\mathcal{F}_n$ is indexed by the set of all vectors $\theta \in \{0,1\}^{r_n}$ consisting of sequences of r_n zeros or ones. For $h_n = n^{-1/(2m+1)}$, let $x_{n,1} < x_{n,2} < \cdots < x_{n,n}$ be a regular grid of meshwidth $2h_n$. For a fixed probability density f and a fixed function K with support $(-1,1)$, define, for every $\theta \in \{0,1\}^{r_n}$,

$$f_{n,\theta}(x) = f(x) + h_n^m \sum_{j=1}^{r_n} \theta_j K\left(\frac{x - x_{n,j}}{h_n}\right).$$

If f is bounded away from zero on an interval containing the grid, $|K|$ is bounded, and $\int K(x)\,dx = 0$, then $f_{n,\theta}$ is a probability density, at least for large n. Furthermore,

$$\int \big|f_{n,\theta}^{(m)}(x)\big|^2\,dx \le 2\int \big|f^{(m)}(x)\big|^2\,dx + 2h_n r_n \int \big|K^{(m)}(x)\big|^2\,dx.$$

It follows that there exist many choices of f and K such that $f_{n,\theta} \in \mathcal{F}_{m,M}$ for every θ.

The following lemma gives a lower bound for the maximum risk over the parameter set $\{0,1\}^r$, in an abstract form, applicable to the problem of estimating an arbitrary quantity $\psi(\theta)$ belonging to a metric space (with metric d). Let $H(\theta,\theta') = \sum_{i=1}^r |\theta_i - \theta_i'|$ be the *Hamming distance* on $\{0,1\}^r$, which counts the number of positions at which θ and θ' differ. For two probability measures P and Q with densities p and q, write $\|P \wedge Q\|$ for $\int p \wedge q\,d\mu$.

24.3 ***Lemma (Assouad).*** *For any estimator T based on an observation in the experiment $\big(P_\theta : \theta \in \{0,1\}^r\big)$, and any $p > 0$,*

$$\max_\theta 2^p \mathrm{E}_\theta d^p\big(T, \psi(\theta)\big) \ge \min_{H(\theta,\theta')\ge 1} \frac{d^p\big(\psi(\theta), \psi(\theta')\big)}{H(\theta,\theta')} \frac{r}{2} \min_{H(\theta,\theta')=1} \|P_\theta \wedge P_{\theta'}\|.$$

Proof. Define an estimator S, taking its values in $\Theta = \{0,1\}^r$, by letting $S = \theta$ if $\theta' \mapsto d\big(T, \psi(\theta')\big)$ is minimal over Θ at $\theta' = \theta$. (If the minimum is not unique, choose a point of minimum in any consistent way.) By the triangle inequality, for any θ, $d\big(\psi(S), \psi(\theta)\big) \le d\big(\psi(S), T\big) + d\big(\psi(\theta), T\big)$, which is bounded by $2d\big(\psi(\theta), T\big)$, by the definition of S. If $d^p\big(\psi(\theta), \psi(\theta')\big) \ge \alpha H(\theta,\theta')$ for all pairs $\theta \ne \theta'$, then

$$2^p \mathrm{E}_\theta d^p\big(T, \psi(\theta)\big) \ge \mathrm{E}_\theta d^p\big(\psi(S), \psi(\theta)\big) \ge \alpha \mathrm{E}_\theta H(S,\theta).$$

The maximum of this expression over Θ is bounded below by the average, which, apart

from the factor α, can be written

$$\frac{1}{2^r}\sum_{\theta}\sum_{j=1}^{r} \mathrm{E}_\theta |S_j - \theta_j| = \frac{1}{2}\sum_{j=1}^{r}\left(\frac{1}{2^{r-1}}\sum_{\theta:\theta_j=0}\int S_j\,dP_\theta + \frac{1}{2^{r-1}}\sum_{\theta:\theta_j=1}\int (1-S_j)\,dP_\theta\right).$$

This is minimized over S by choosing S_j for each j separately to minimize the jth term in the sum. The expression within brackets is the sum of the error probabilities of a test of

$$\overline{P}_{0,j} = \frac{1}{2^{r-1}}\sum_{\theta:\theta_j=0} P_\theta, \quad \text{versus} \quad \overline{P}_{1,j} = \frac{1}{2^{r-1}}\sum_{\theta:\theta_j=1} P_\theta.$$

Equivalently, it is equal to 1 minus the difference of power and level. In Lemma 14.30 this was seen to be at least $1 - \frac{1}{2}\|\overline{P}_{0,j} - \overline{P}_{1,j}\| = \|\overline{P}_{0,j} \wedge \overline{P}_{1,j}\|$. Hence the preceding display is bounded below by

$$\frac{1}{2}\sum_{j=1}^{r} \|\overline{P}_{0,j} \wedge \overline{P}_{1,j}\|.$$

Because the minimum $\overline{p}_m \wedge \overline{q}_m$ of two averages of numbers is bounded below by the average $m^{-1}\sum p_i \wedge q_i$ of the minima, the same is true for the total variation norm of a minimum: $\|\overline{P}_m \wedge \overline{Q}_m\| \geq m^{-1}\sum \|P_i \wedge Q_i\|$. The 2^{r-1} terms P_θ and $P_{\theta'}$ in the averages $\overline{P}_{0,j}$ and $\overline{P}_{1,j}$ can be ordered and matched such that each pair θ and θ' differ only in their jth coordinate. Conclude that the preceding display is bounded below by $\frac{1}{2}\sum_{j=1}^{r} \min \|P_\theta \wedge P_{\theta'}\|$, in which the minimum is taken over all pairs θ and θ' that differ by exactly one coordinate. ■

We wish to apply Assouad's lemma to the product measures resulting from the densities $f_{n,\theta}$. Then the following inequality, obtained in the proof of Lemma 14.31, is useful. It relates the total variation, affinity, and Hellinger distance of product measures:

$$\|P^n \wedge Q^n\| \geq \tfrac{1}{2}A^2(P^n, Q^n) = \tfrac{1}{2}\big(1 - \tfrac{1}{2}H^2(P, Q)\big)^{2n}.$$

24.4 ***Theorem.*** *There exists a constant $D_{m,M}$ such that for any density estimator $\hat{f}_n$*

$$\sup_{f\in\mathcal{F}_{m,M}} \mathrm{E}_f \int \big(\hat{f}_n(x) - f(x)\big)^2\,dx \geq D_{m,M}\left(\frac{1}{n}\right)^{2m/(2m+1)}.$$

Proof. Because the functions $f_{n,\theta}$ are bounded away from zero and infinity, uniformly in θ, the squared Hellinger distance

$$\int \big(f_{n,\theta}^{1/2} - f_{n,\theta'}^{1/2}\big)^2\,dx = \int \left(\frac{f_{n,\theta} - f_{n,\theta'}}{f_{n,\theta}^{1/2} + f_{n,\theta'}^{1/2}}\right)^2 dx$$

is up to constants equal to the squared L_2-distance between $f_{n,\theta}$ and $f_{n,\theta'}$. Because the

functions $K\big((x - x_{n,j})/h_n\big)$ have disjoint supports, the latter is equal to

$$h_n^{2m} \sum_{j=1}^{r_n} |\theta_j - \theta_j'|^2 \int K^2\left(\frac{x - x_{n,j}}{h_n}\right) dx = h_n^{2m+1} H(\theta, \theta') \int K^2(x)\, dx.$$

This is of the order $1/n$. Inserting this in the lower bound given by Assouad's lemma, with $\psi(\theta) = f_{n,\theta}$ and $d\big(\psi(\theta), \psi(\theta')\big)$ the L_2-distance, we find up to a constant the lower bound $h_n^{2m+1}\,(r_n/2)\,\big(1 - O(1/n)\big)^{2n}$. ■

24.4 Estimating a Unimodal Density

In the preceding sections the analysis of nonparametric density estimators is based on the assumption that the true density is smooth. This is appropriate for kernel-density estimation, because this is a smoothing method. It is also sensible to place some a priori restriction on the true density, because otherwise we cannot hope to achieve much beyond consistency. However, smoothness is not the only possible restriction. In this section we assume that the true density is monotone, or unimodal. We start with monotone densities and next view a unimodal density as a combination of two monotone pieces.

It is interesting that with monotone densities we can use maximum likelihood as the estimating principle. Suppose that $X_1, \ldots, X_n$ is a random sample from a Lebesgue density f on $[0, \infty)$ that is known to be nonincreasing. Then the maximum likelihood estimator $\hat{f}_n$ is defined as the nonincreasing probability density that maximizes the likelihood

$$f \mapsto \prod_{i=1}^{n} f(X_i).$$

This optimization problem would not have a solution if f were only restricted by possessing a certain number of derivatives, because very high peaks at the observations would yield an arbitrarily large likelihood. However, under monotonicity there is a unique solution.

The solution must necessarily be a left-continuous step function, with steps only at the observations. Indeed, if for a given f the limit from the right at $X_{(i-1)}$ is bigger than the limit from the left at $X_{(i)}$, then we can redistribute the mass on the interval $(X_{(i-1)}, X_{(i)}]$ by raising the value $f(X_{(i)})$ and lowering $f(X_{(i-1)}+)$, for instance by setting f equal to the constant value $(X_{(i)} - X_{(i-1)})^{-1} \int_{X_{(i-1)}}^{X_{(i)}} f(t)\, dt$ on the whole interval, resulting in an increase of the likelihood. By the same reasoning we see that the maximum likelihood estimator must be zero on $(X_{(n)}, \infty)$ (and $(-\infty, 0)$). Thus, with $f_i = \hat{f}_n(X_{(i)})$, finding the maximum likelihood estimator reduces to maximizing $\prod_{i=1}^n f_i$ under the side conditions (with $X_{(0)} = 0$)

$$f_1 \geq f_2 \geq \cdots \geq f_n \geq 0,$$
$$\sum_{i=1}^{n} f_i\big(X_{(i)} - X_{(i-1)}\big) = 1.$$

This problem has a nice graphical solution. The *least concave majorant* of the empirical distribution function $\mathbb{F}_n$ is defined as the smallest concave function $\hat{F}_n$ with $\hat{F}_n(x) \geq \mathbb{F}_n(x)$ for every x. This can be found by attaching a rope at the origin $(0, 0)$ and winding this (from above) around the empirical distribution function $\mathbb{F}_n$ (Figure 24.3). Because $\hat{F}_n$ is

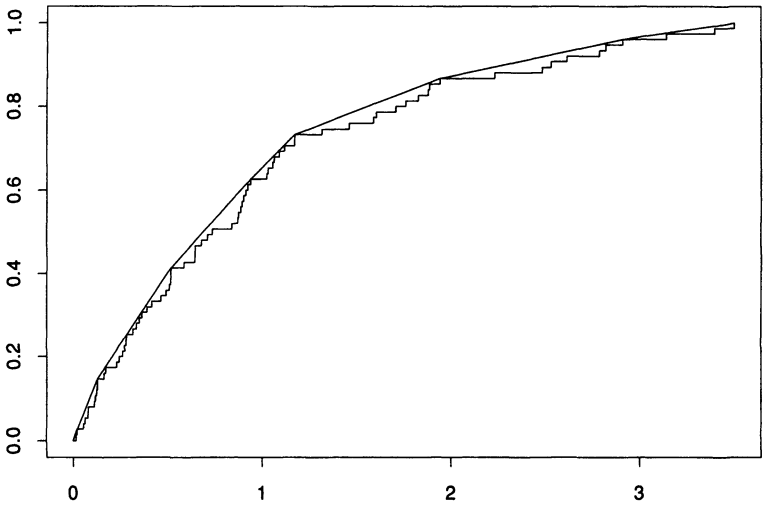

Figure 24.3. The empirical distribution and its concave majorant of a sample of size 75 from the exponential distribution.

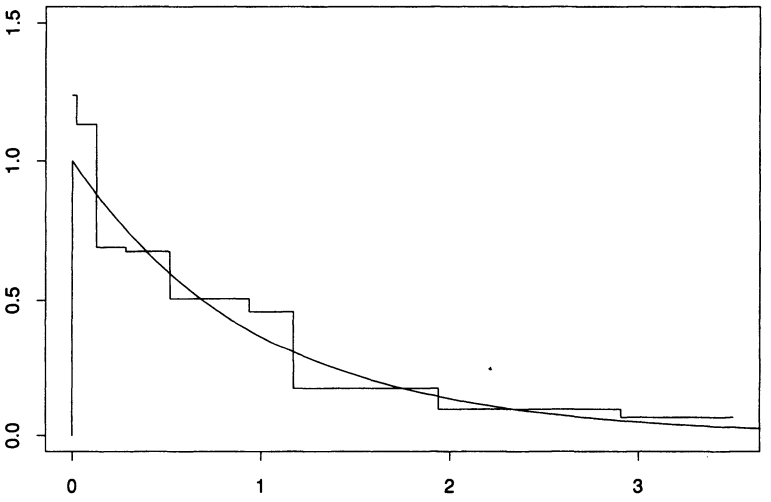

Figure 24.4. The derivative of the concave majorant of the empirical distribution and the true density of a sample of size 75 from the exponential distribution.

concave, its derivative is nonincreasing. Figure 24.4 shows the derivative of the concave majorant in Figure 24.3.

24.5 ***Lemma.*** *The maximum likelihood estimator $\hat{f}_n$ is the left derivative of the least concave majorant $\hat{F}_n$ of the empirical distribution $\mathbb{F}_n$, that is, on each of the intervals $(X_{(i-1)}, X_{(i)}]$ it is equal to the slope of $\hat{F}_n$ on this interval.*

Proof. In this proof, let $\hat{f}_n$ denote the left derivative of the least concave majorant. We shall show that this maximizes the likelihood. Because the maximum likelihood estimator

is necessarily constant on each interval $(X_{(i-1)}, X_{(i)}]$, we may restrict ourselves to densities f with this property. For such an f we can write $\log f = \sum a_i 1_{[0,X_{(i)}]}$ for the constants $a_i = \log f_i/f_{i+1}$ (with $f_{n+1} = 1$), and we obtain

$$\int \log f \, d\hat{F}_n = \sum_{i=1}^{n} a_i \, \hat{F}_n\big(X_{(i)}\big) \geq \sum_{i=1}^{n} a_i \, \mathbb{F}_n\big(X_{(i)}\big) = \int \log f \, d\mathbb{F}_n.$$

For $f = \hat{f}_n$ this becomes an equality. To see this, let $y_1 \leq y_2 \leq \cdots$ be the points where $\hat{F}_n$ touches $\mathbb{F}_n$. Then $\hat{f}_n$ is constant on each of the intervals $(y_{i-1}, y_i]$, so that we can write $\log \hat{f}_n = \sum b_i 1_{[0,y_i]}$, and obtain

$$\int \log \hat{f}_n \, d\hat{F}_n = \sum b_i \hat{F}_n(y_i) = \sum b_i \mathbb{F}_n(y_i) = \int \log \hat{f}_n \, d\mathbb{F}_n.$$

Third, by the identifiability property of the Kullback-Leibler divergence (see Lemma 5.35), for any probability density f,

$$\int \log \hat{f}_n \, d\hat{F}_n \geq \int \log f \, d\hat{F}_n,$$

with strict inequality unless $\hat{f}_n = f$. Combining the three displays, we see that $\hat{f}_n$ is the unique maximizer of $f \mapsto \int \log f \, d\mathbb{F}_n$. ■

Maximizing the likelihood is an important motivation for taking the derivative of the concave majorant, but this operation also has independent value. Taking the concave majorant (or convex minorant) of the primitive function of an estimator and next differentiating the result may be viewed as a "smoothing" device, which is useful if the target function is known to be monotone. The estimator $\hat{f}_n$ can be viewed as the result of this procedure applied to the "naive" density estimator

$$\tilde{f}_n(x) = \frac{1}{n\big(X_{(i)} - X_{(i-1)}\big)}, \qquad x \in \big(X_{(i-1)}, X_{(i)}\big].$$

This function is very rough and certainly not suitable as an estimator. Its primitive function is the polygon that linearly interpolates the extreme points of the empirical distribution function $\mathbb{F}_n$, and its smallest concave majorant coincides with the one of $\mathbb{F}_n$. Thus the derivative of the concave majorant of $\tilde{F}_n$ is exactly $\hat{f}_n$.

Consider the rate of convergence of the maximum likelihood estimator. Is the assumption of monotonicity sufficient to obtain a reasonable performance? The answer is affirmative if a rate of convergence of $n^{1/3}$ is considered reasonable. This rate is slower than the rate $n^{m/(2m+1)}$ of a kernel estimator if $m > 1$ derivatives exist and is comparable to this rate given one bounded derivative (even though we have not established a rate under $m = 1$). The rate of convergence $n^{1/3}$ can be shown to be best possible if only monotonicity is assumed. It is achieved by the maximum likelihood estimator.

24.6 ***Theorem.*** *If the observations are sampled from a compactly supported, bounded, monotone density f, then*

$$\int \big(\hat{f}_n(x) - f(x)\big)^2 \, dx = O_P(n^{-2/3}).$$

Proof. This result is a consequence of a general result on maximum likelihood estimators of densities (e.g., Theorem 3.4.4 in [146].) We shall give a more direct proof using the convexity of the class of monotone densities.

The sequence $\|\hat{f}_n\|_\infty = \hat{f}_n(0)$ is bounded in probability. Indeed, by the characterization of $\hat{f}_n$ as the slope of the concave majorant of $\mathbb{F}_n$, we see that $\hat{f}_n(0) > M$ if and only if there exists $t > 0$ such that $\mathbb{F}_n(t) > Mt$. The claim follows, because, by concavity, $F(t) \le f(0)t$ for every t, and, by Daniel's theorem ([134, p. 642]),

$$\mathrm{P}\left(\sup_{t>0} \frac{\mathbb{F}_n(t)}{F(t)} > M\right) = \frac{1}{M}.$$

It follows that the rate of convergence of $\hat{f}_n$ is the same as the rate of the maximum likelihood estimator under the restriction that f is bounded by a (large) constant. In the remainder of the proof, we redefine $\hat{f}_n$ by the latter estimator.

Denote the true density by f_0. By the definition of $\hat{f}_n$ and the inequality $\log x \le 2(\sqrt{x} - 1)$,

$$0 \le \mathbb{F}_n \log \frac{\hat{f}_n}{\frac{1}{2}\hat{f}_n + \frac{1}{2}f_0} \le 2\mathbb{F}_n\left(\sqrt{\frac{2\hat{f}_n}{\hat{f}_n + f_0}} - 1\right).$$

Therefore, we can obtain the rate of convergence of $\hat{f}_n$ by an application of Theorem 5.52 or 5.55 with $m_f = \sqrt{2f/(f + f_0)}$.

Because $(m_f - m_{f_0})(f_0 - f) \le 0$ for every f and f_0 it follows that $F_0(m_f - m_{f_0}) \le F(m_f - m_{f_0})$ and hence

$$F_0(m_f - m_{f_0}) \le \tfrac{1}{2}(F_0 + F)(m_f - m_{f_0}) = -\tfrac{1}{2}h^2\big(f, \tfrac{1}{2}f + \tfrac{1}{2}f_0\big) \lesssim -h^2(f, f_0),$$

in which the last inequality is elementary calculus. Thus the first condition of Theorem 5.52 is satisfied relative to the Hellinger distance h, with $\alpha = 2$.

The map $f \mapsto m_f$ is increasing. Therefore, it turns brackets $[f_1, f_2]$ for the functions $x \mapsto f(x)$ into brackets $[m_{f_1}, m_{f_2}]$ for the functions $x \mapsto m_f(x)$. The squared $L_2(F_0)$-size of these brackets satisfies

$$F_0(m_{f_1} - m_{f_2})^2 \le 4h^2(f_1, f_2).$$

It follows that the $L_2(F_0)$-bracketing numbers of the class of functions m_f can be bounded by the h-bracketing numbers of the functions f. The latter are the $L_2(\lambda)$-bracketing numbers of the functions $\sqrt{f}$, which are monotone and bounded by assumption. In view of Example 19.11,

$$\log N_{[]}\big(2\varepsilon, \{m_f : f \in \mathcal{F}\}, L_2(F_0)\big) \le \log N_{[]}\big(\varepsilon, \sqrt{\mathcal{F}}, L_2(\lambda)\big) \lesssim \frac{1}{\varepsilon}.$$

Because the functions m_f are uniformly bounded, the maximal inequality Lemma 19.36 gives, with $J(\delta) = \int_0^\delta \sqrt{1/\varepsilon}\, d\varepsilon = 2\sqrt{\delta}$,

$$\mathrm{E}_{f_0} \sup_{h(f,f_0)<\delta} \big|\mathbb{G}_n(f - f_0)\big| \lesssim \sqrt{\delta}\left(1 + \frac{\sqrt{\delta}}{\delta^2\sqrt{n}}\right).$$

Therefore, Theorem 5.55 applies with $\phi_n(\delta)$ equal to the right side, and the Hellinger distance, and we conclude that $h(\hat{f}_n, f_0) = O_P(n^{-1/3})$.

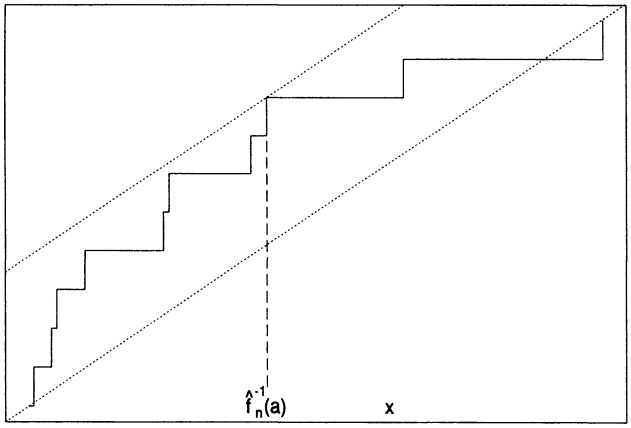

Figure 24.5. If $\hat{f}_n(x) \le a$, then a line of slope a moved down vertically from $+\infty$ first hits $\mathbb{F}_n$ to the left of x. The point where the line hits is the point at which $\mathbb{F}_n$ is farthest above the line of slope a through the origin.

The $L_2(\lambda)$-distance between uniformly bounded densities is bounded up to a constant by the Hellinger distance, and the theorem follows. ■

The most striking known results about estimating a monotone density concern limit distributions of the maximum likelihood estimator, for instance at a point.

24.7 ***Theorem.*** *If f is differentiable at $x > 0$ with derivative $f'(x) < 0$, then, with $\{\mathbb{Z}(h) : h \in \mathbb{R}\}$ a standard Brownian motion process (two-sided with $\mathbb{Z}(0) = 0$),*

$$n^{1/3}\big(\hat{f}_n(x) - f(x)\big) \rightsquigarrow \big|4f'(x)f(x)\big|^{1/3} \operatorname*{argmax}_{h\in\mathbb{R}}\big\{\mathbb{Z}(h) - h^2\big\}.$$

Proof. For simplicity we assume that f is continuously differentiable at x. Define a stochastic process $\{\hat{f}_n^{-1}(a) : a > 0\}$ by

$$\hat{f}_n^{-1}(a) = \operatorname*{argmax}_{s\ge 0}\big\{\mathbb{F}_n(s) - as\big\},$$

in which the largest value is chosen when multiple maximizers exist. The suggestive notation is justified, as the function $\hat{f}_n^{-1}$ is the inverse of the maximum likelihood estimator $\hat{f}_n$ in that $\hat{f}_n(x) \le a$ if and only if $\hat{f}_n^{-1}(a) \le x$, for every x and a. This is explained in Figure 24.5. We first derive the limit distribution of $\hat{f}_n^{-1}$. Let $\delta_n = n^{-1/3}$.

By the change of variable $s \mapsto x + h\delta_n$ in the definition of $\hat{f}_n^{-1}$, we have

$$n^{1/3}\big(\hat{f}_n^{-1} \circ f(x) - x\big) = \operatorname*{argmax}_{h\ge -n^{1/3}x}\big\{\mathbb{F}_n(x + h\delta_n) - f(x)(x + h\delta_n)\big\}.$$

Because the location of a maximum does not change by a vertical shift of the whole function, we can drop the term $f(x)x$ in the right side, and we may add a term $\mathbb{F}_n(x)$. For the same

reason we may also multiply the process in the right side by $n^{2/3}$. Thus the preceding display is equal to the point of maximum $\hat{h}_n$ of the process

$$n^{2/3}\big[(\mathbb{F}_n - F)(x + h\delta_n) - (\mathbb{F}_n - F)(x)\big] + n^{2/3}\big[F(x + h\delta_n) - F(x) - f(x)h\delta_n\big].$$

The first term is the local empirical process studied in Example 19.29, and converges in distribution to the process $h \mapsto \sqrt{f(x)}\,\mathbb{Z}(h)$, for $\mathbb{Z}$ a standard Brownian motion process, in $\ell^\infty(K)$, for every compact interval K. The second term is a deterministic "drift" process and converges on compacta to $h \mapsto \frac{1}{2}f'(x)h^2$. This suggests that

$$n^{1/3}\big(\hat{f}_n^{-1} \circ f(x) - x\big) \rightsquigarrow \operatorname*{argmax}_{h\in\mathbb{R}} \big\{\sqrt{f(x)}\,\mathbb{Z}(h) + \tfrac{1}{2}f'(x)h^2\big\}.$$

This argument remains valid if we replace x by $x_n = x - \delta_n b$ throughout, where the limit is the same for every $b \in \mathbb{R}$.

We can write the limit in a more attractive form by using the fact that the processes $h \mapsto \mathbb{Z}(\sigma h)$ and $h \mapsto \sqrt{\sigma}\mathbb{Z}(h)$ are equal in distribution for every $\sigma > 0$. First, apply the change of variables $h \mapsto \sigma h$, next pull σ out of $\mathbb{Z}(\sigma h)$, then divide the process by $\sqrt{f(x)\sigma}$, and finally choose σ such that the quadratic term reduces to $-h^2$, that is $\sqrt{f(x)\sigma} = -\frac{1}{2}f'(x)\sigma^2$. Then we obtain, for every $b \in \mathbb{R}$,

$$n^{1/3}\big(\hat{f}_n^{-1} \circ f(x - \delta_n b) - (x - \delta_n b)\big) \rightsquigarrow \left(\frac{\sqrt{f(x)}}{-\frac{1}{2}f'(x)}\right)^{2/3} \operatorname*{argmax}_{h\in\mathbb{R}} \big\{\mathbb{Z}(h) - h^2\big\}.$$

The connection with the limit distribution of $\hat{f}_n(x)$ is that

$$\begin{aligned}\mathrm{P}\Big(n^{1/3}\big(\hat{f}_n(x) - f(x)\big) \le -bf'(x)\Big) &= \mathrm{P}\big(\hat{f}_n(x) \le f(x - \delta_n b) + o(1)\big)\\ &= \mathrm{P}\Big(n^{1/3}\big(\hat{f}_n^{-1} \circ f(x - \delta_n b) - (x - \delta_n b)\big) \le b\Big) + o(1).\end{aligned}$$

Combined with the preceding display and simple algebra, this yields the theorem.

The preceding argument can be made rigorous by application of the argmax continuous-mapping theorem, Corollary 5.58. The limiting Brownian motion has continuous sample paths, and maxima of Gaussian processes are automatically unique (see, e.g., Lemma 2.6 in [87]). Therefore, we need only check that $\hat{h}_n = O_P(1)$, for which we apply Theorem 5.52 with

$$m_g = 1_{[0,x_n+g]} - 1_{[0,x_n]} - f(x_n)g.$$

(In Theorem 5.52 the function m_g can be allowed to depend on n, as is clear from its generalization, Theorem 5.55.) By its definition, $\hat{g}_n = \delta_n\hat{h}_n$ maximizes $g \mapsto \mathbb{F}_n m_g$, whence we wish to show that $\hat{g}_n = O_P(\delta_n)$. By Example 19.6 the bracketing numbers of the class of functions $\{1_{[0,x_n+g]} - 1_{[0,x_n]} : |g| < \delta\}$ are of the order δ/ε^2; the envelope function $|1_{[0,x_n+\delta]} - 1_{[0,x_n]}|$ has $L_2(F)$-norm of the order $\sqrt{f(x)\delta}$. By Corollary 19.35,

$$\mathrm{E}\sup_{|g|<\delta} |\mathbb{G}_n m_g| \lesssim \int_0^{\sqrt{\delta}} \sqrt{\log\frac{\delta}{\varepsilon^2}}\,d\varepsilon \lesssim \sqrt{\delta}.$$

By the concavity of F, the function $g \mapsto F(x_n + g) - F(x_n) - f(x_n)g$ is nonpositive and nonincreasing as g moves away from 0 in either direction (draw a picture.) Because

$f'(x_n) \to f'(x) < 0$, there exists a constant C such that, for sufficiently large n,

$$Fm_g = F(x_n + g) - F(x_n) - f(x_n)g \le -C\big(g^2 \wedge |g|\big).$$

If we would know already that $\hat{g}_n \xrightarrow{P} 0$, then Theorem 5.52, applied with $\alpha = 2$ and $\beta = \frac{1}{2}$, yields that $\hat{g}_n = O_P(\delta_n)$.

The consistency of $\hat{g}_n$ can be shown by a direct argument. By the Glivenko-Cantelli theorem, for every $\varepsilon > 0$,

$$\sup_{|g| \ge \varepsilon} \mathbb{F}_n m_g \le \sup_{|g| \ge \varepsilon} Fm_g + o_P(1) \le -C \inf_{|g| \ge \varepsilon} \big(g^2 \wedge |g|\big) + o_P(1).$$

Because the right side is strictly smaller than $0 = \mathbb{F}_n m_0$, the maximizer $\hat{g}_n$ must be contained in $[-\varepsilon, \varepsilon]$ eventually. ■

Results on density estimators at a point are perhaps not of greatest interest, because it is the overall shape of a density that counts. Hence it is interesting that the preceding theorem is also true in an L_1-sense, in that

$$n^{1/3} \int \big|\hat{f}_n(x) - f(x)\big|\, dx \rightsquigarrow \int \big|4f'(x)f(x)\big|^{1/3}\, dx\, \mathrm{E} \operatorname*{argmax}_{h \in \mathbb{R}} \big\{\mathbb{Z}(h) - h^2\big\}.$$

This is true for every *strictly* decreasing, compactly supported, twice continuously differentiable true density f. For boundary cases, such as the uniform distribution, the behavior of $\hat{f}_n$ is very different. Note that the right side of the preceding display is degenerate. This is explained by the fact that the random variables $n^{1/3}\big(\hat{f}_n(x) - f(x)\big)$ for different values of x are asymptotically independent, because they are only dependent on the observations X_i very close to x, so that the integral aggregates a large number of approximately independent variables. It is also known that $n^{1/6}$ times the difference between the left side and the right sides converges in distribution to a normal distribution with mean zero and variance not depending on f. For uniformly distributed observations, the estimator $\hat{f}_n(x)$ remains dependent on all n observations, even asymptotically, and attains a $\sqrt{n}$-rate of convergence (see [62]).

We define a density f on the real line to be *unimodal* if there exists a number M_f, such that f is nondecreasing on the interval $(-\infty, M_f]$ and nondecreasing on $[M_f, \infty)$. The *mode* M_f need not be unique. Suppose that we observe a random sample from a unimodal density.

If the true mode M_f is known a priori, then a natural extension of the preceding discussion is to estimate the distribution function F of the observations by the distribution function $\hat{F}_n$ that is the least concave majorant of $\mathbb{F}_n$ on the interval $[M_f, \infty)$ and the greatest convex minorant on $(-\infty, M_f]$. Next we estimate f by the derivative $\hat{f}_n$ of $\hat{F}_n$. Provided that none of the observations takes the value M_f, this estimator maximizes the likelihood, as can be shown by arguments as before. The limit results on monotone densities can also be extended to the present case. In particular, because the key in the proof of Theorem 24.7 is the characterization of $\hat{f}_n$ as the derivative of the concave majorant of $\mathbb{F}_n$, this theorem remains true in the unimodal case, with the same limit distribution.

If the mode is not known a priori, then the maximum likelihood estimator does not exist: The likelihood can be maximized to infinity by placing an arbitrary large mode at some fixed observation. It has been proposed to remedy this problem by restricting the likelihood

to densities that have a modal interval of a given length (in which f must be constant and maximal). Alternatively, we could estimate the mode by an independent method and next apply the procedure for a known mode. Both of these possibilities break down unless f possesses some additional properties. A third possibility is to try every possible value M as a mode, calculate the estimator $\hat{f}_n^M$ for known mode M, and select the best fitting one. Here "best" could be operationalized as (nearly) minimizing the Kolmogorov-Smirnov distance $\|\hat{F}_n^M - \mathbb{F}_n\|_\infty$. It can be shown (see [13]) that this procedure renders the effect of the mode being unknown asymptotically negligible, in that

$$\int \left| \hat{f}_n^{\hat{M}}(x) - \hat{f}_n^{M_f}(x) \right| dx \leq 4\|\mathbb{F}_n - F\|_\infty = O_P\left(\frac{1}{\sqrt{n}}\right),$$

up to an arbitrarily small tolerance parameter if $\hat{M}$ only approximately achieves the minimum of $M \mapsto \|\hat{F}_n^M - \mathbb{F}_n\|_\infty$. This extra "error" is of lower order than the rate of convergence $n^{1/3}$ of the estimator with a known mode.

Notes

The literature on nonparametric density estimation, or "smoothing," is large, and there is an equally large literature concerning the parallel problem of nonparametric regression. Next to kernel estimation popular methods are based on classical series approximations, spline functions, and, most recently, wavelet approximation. Besides different methods, a good deal is known concerning other loss functions, for instance L_1-loss and automatic methods to choose a bandwidth. Most recently, there is a revived interest in obtaining exact constants in minimax bounds, rather than just rates of convergence. See, for instance, [14], [15], [36], [121], [135], [137], and [148] for introductions and further references. The kernel estimator is often named after its pioneers in the 1960s, Parzen and Rosenblatt, and was originally developed for smoothing the periodogram in spectral density estimation.

A lower bound for the maximum risk over Hölder classes for estimating a density at a single point was obtained in [46]. The lower bound for the L_2-risk is more recent. Birgé [12] gives a systematic study of upper and lower bounds and their relationship to the metric entropy of the model. An alternative for Assouad's lemma is Fano's lemma, which uses the Kullback-Leibler distance and can be found in, for example, [80].

The maximum likelihood estimator for a monotone density is often called the *Grenander estimator*, after the author who first characterized it in 1956. The very short proof of Lemma 24.5 is taken from [64]. The limit distribution of the Grenander estimator at a point was first obtained by Prakasa Rao in 1969 see [121]. Groeneboom [63] gives a characterization of the limit distribution and other interesting related results.

PROBLEMS

1. Show, informally, that under sufficient regularity conditions

$$\text{MISE}_f(\hat{f}) \sim \frac{1}{nh}\int K^2(y)\,dy + \frac{1}{4}h^4 \int f''(x)^2\,dx \left(\int y^2 K(y)\,dy\right)^2.$$

What does this imply for an optimal choice of the bandwidth?

2. Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with variance 1. Calculate the mean square error of the estimator $\phi(x - \overline{X}_n)$ of the common density.

3. Using the argument of section 14.5 and a submodel as in section 24.3, but with $r_n = 1$, show that the best rate for estimating a density at a fixed point is also $n^{-m/(2m+1)}$.

4. Using the argument of section 14.5, show that the rate of convergence $n^{1/3}$ of the maximum likelihood estimator for a monotone density is best possible.

5. **(Marshall's lemma.)** Suppose that F is concave on $[0, \infty)$ with $F(0) = 0$. Show that the least concave majorant $\hat{F}_n$ of $\mathbb{F}_n$ satisfies the inequality $\|\hat{F}_n - F\|_\infty \leq \|\mathbb{F}_n - F\|_\infty$. What does this imply about the limiting behavior of $\hat{F}_n$?