Statistica Sinica **6**(1996), 23-45

SOME PROBLEMS ON THE ESTIMATION OF UNIMODAL DENSITIES

Peter J. Bickel and Jianqing Fan

University of California and University of North Carolina

Abstract. In this paper, we study, in some new ways, the estimation of unimodal densities. Several methods for estimating unimodal densities are proposed: plug-in MLE, pregrouping techniques, linear spline MLE. Based on the maximum likelihood method, an automatic procedure for estimating a unimodal density as well as its mode is proposed. We also give asymptotic theory for the proposed estimators. An important consequence of this study is that having to estimate the location of a mode does not affect the limiting behavior of the proposed unimodal density estimate. Simulation studies illustrate the proposed methods.

Key words and phrases: Asymptotic distributions, MLE, modes, plug-in methods, pregrouping methods, unimodal densities.

1. Introduction

Nonparametric density estimation provides a useful technique of examining the overall structure of a set of data. A commonly used technique is the kernel method. The behavior of a kernel density estimate relies strongly on the choice of smoothing parameter (bandwidth). Data-driven bandwidth selection methods have been studied recently. One tries to minimize the Integrated Square Error (ISE) or the Mean ISE (MISE) or other related objects, and uses one of them as a measure of global effectiveness of a curve estimate. In practical density estimation, however, features such as shape and area under modes may be more interesting. ISE and MISE are not good criteria for these purposes. For example, the ISE of two curves can be very small, while the shapes of the two curves are quite different. When shape information is available, an alternative approach is to estimate a curve under shape restrictions. In this paper, we focus on a number of approaches to the estimation of a unimodal density with an unknown mode location. We describe our results and then point to some of the historical background of our approach.

To estimate a unimodal density, we first begin by introducing a plug-in maximum likelihood method. Let $\hat{f}_n(x;m)$ be the nonparametric maximum likelihood estimate under the restriction that the unknown density is unimodal with the mode location parameterized by m. Let $\hat{m}$ be a consistent estimate of the true location m_0 of the mode. Then, the plug-in version of the estimate is $\hat{f}_n(x;\hat{m})$. We show, in Section 2, that for *all* consistent estimates $\hat{m}$, $\hat{f}_n(x;\hat{m})$ converges at *the same rate* $n^{-1/3}$ *with the same asymptotic distribution.* The implication of this is that estimating an unknown density with unknown location of mode is not appreciably more difficult than estimating an unknown density with a known location of mode. This phenomenon was also observed by Birgé (1987c), who showed that among other properties, $n^{1/3}\|\hat{f}_n(\cdot,\hat{m}) - \hat{f}_n(\cdot;m_0)\|_1$ converges to zero for a particular choice of estimator $\hat{m}$. However, the current result holds for any consistent mode estimator although the result is local rather than global. This conclusion gives more support to Birgé's notion that the MLE is robust to mode estimation.

We then propose an automatic method for estimating the mode and the density based on the maximum likelihood method. A rate of convergence for the mode estimate is derived. The maximum likelihood estimate of the density is shown to have the same asymptotic properties as the case where the mode is known.

The graph of $\hat{f}_n(x;\hat{m})$ is quite spiky near the location $\hat{m}$. One way of reducing the spikiness problem is to use a pregrouping technique. The idea is to group the data into a number of groups first, and then to perform a form of MLE. We then prove that if the grouping is not too crude, the pregrouping version of the MLE does as well as the plug-in MLE in terms of pointwise weak convergence. This pregrouping technique also saves computing costs. Another way of reducing the peaking problem is the maximum penalized likelihood method. (See Woodroofe and Sun (1993)).

The discontinuity of the plug-in MLE is unsatisfactory. To deal with this problem, we introduce a maximum likelihood linear spline estimate. We give explicitly the form of the estimate. The asymptotic distribution of the estimate is derived when the mode is assumed known. Since not knowing the location of the mode is not a serious matter in estimating a unimodal density when the MLE is used, we expect but have not yet shown that such an estimate should also work well when we do not know the location of mode. A nice feature of such an estimate is that the location of the mode is determined *automatically* by the data. Again, the pregrouping technique can be used to guard against spiking problems and to reduce computation.

Various related issues are discussed in Section 3. No theory is available as yet, but we give some heuristics below.

An early work on estimating a density under shape restrictions is Grenander (1956), who estimated a decreasing density by using a maximum likelihood approach. The asymptotic distribution of the MLE at a point was found by Prakasa Rao (1969), and Groeneboom (1985). Recent developments in estimating a monotone density can be found in Birgé (1987a,b), who gives the behavior of nonparametric minimax risks. Wegman (1969, 1970a,b) proposed and studied the estimation of a unimodal density by finding the MLE for a modal interval of length ε. In particular, he found the pointwise asymptotic distribution of the MLE except for the modal interval, on which the MLE is not even consistent. We give a more natural MLE method, and derive the asymptotic distribution for all points except the mode itself. Mammen (1991a,b) made an interesting study of the shape restricted curve estimation in the context of the nonparametric regression setup. Various applications of the isotonic method can be found in Barlow and van Zwet (1970), Barlow el al. (1972), Robertson et al. (1988), Wang (1986), Ramsay (1988), among others.

2. Problems and Main Results

Let $f(x;m)$ be a unimodal density with mode location parameterized by m. Let $X_1 < \cdots < X_n$ be order statistics. Suppose that $X_1', \ldots, X_n'$ are i.i.d. from $f(x;m_0)$, where m_0 is the true location of the mode. If m_0 is known, the nonparametric maximum likelihood estimator $\hat{f}_n(x;m_0)$ is such that when $x > m_0$, $\hat{f}_n(x;m_0)$ is the left derivative of the least concave majorant of the empirical distribution function, and when $x < m_0$, $\hat{f}_n(x;m_0)$ is the right derivative of the greatest convex minorant of the empirical distribution. (See Grenander (1956).)

In applications, the true mode m_0 is typically unknown. Let $\hat{m}$ be a consistent estimator of m_0. Then, we use the estimator $\hat{f}_n(x;\hat{m})$ as an estimator of the unknown density $f(x;m_0)$. We call such an estimator the plug-in MLE.

Theorem 1. *Let $\hat{m}$ be a consistent estimate of the mode m_0 of the true underlying density, and $f'(x;m_0) \neq 0$ be the derivative of the density $f(x;m_0)$ with respective to x. Then,*

$$n^{1/3}\left|\frac{1}{2}f(x;m_0)f'(x;m_0)\right|^{-\frac{1}{3}}\left(\hat{f}_n(x;\hat{m}) - f(x;m_0)\right) \xrightarrow{\mathcal{L}} 2Z,$$

where the random variable Z is distributed as the location of the maximum of the process $(W(u) - u^2, u \in \Re)$, and $W(\cdot)$ is a standard two-sided Brownian motion on the real line $\Re$ originating from zero (i.e. $W(0) = 0$).

Remark 1. A striking feature of Theorem 1 is that for any consistent estimate $\hat{m}$, the plug-in MLE $\hat{f}_n(x;\hat{m})$ has the same asymptotic distribution as $\hat{f}_n(x;m_0)$. Birgé (1987c) showed that for a particular choice of $\hat{m}$, the L_1 norm of $\hat{f}_n(.;\hat{m}) - \hat{f}_n(\cdot;m_0)$ is also of order $o_P(n^{-1/2})$. Less satisfactory is the lack of real information about $\hat{f}(m_0;m_0)$.

We now propose a mode estimate based on the maximum likelihood method. Estimating the mode by the kernel method (Parzen (1962), Eddy (1980)) and greatest "clustering" method (Chernoff (1964), Venter (1967)) requires a choice of smoothing parameters. Unlike these traditional approaches, the maximum likelihood method is fully automatic. Let $\hat{f}_j(\cdot; X_j)$ be the maximum likelihood estimate for the data $\{X_i, i \neq j\}$ with mode location X_j. Let $\hat{j} = \arg\max_j \sum_{i\neq j} \log(\hat{f}_j(X_i; X_j))$. Then, the proposed estimate of the mode and of the density are

$$\hat{m}_{MLE} = X_{\hat{j}}, \qquad \hat{f}_{MLE}(\cdot) = \hat{f}_{\hat{j}}(\cdot; \hat{m}_{MLE}).$$

Herewith is a consistency result, which shows that $\hat{m}_{MLE}$ can be used in Theorem 1. Hence, the estimated density $\hat{f}_{MLE}$ has the same asymptotic property as in Theorem 1.

Theorem 2. *Suppose that the tail of the underlying distribution satisfies $F(x) - F(-x) = 1 - o(x^{-1/\alpha})$ as $x \to +\infty$ for some $\alpha > 0$ and that the density $f(x; m_0)$ is bounded and unimodal with mode m_0. If the mode is uniquely defined, then $\hat{m}_{MLE}$ is a consistent estimate of the mode m_0.*

Remark 2. We show, in fact, that in addition to the conditions given in Theorem 2, if there exists a positive constant $k \geq 1$ and $c > 0$ such that in a neighborhood of m_0,

$$|f(y; m_0) - f(z; m_0)| \geq c|y - z|^k, \quad \text{for } y, z < m_0 \text{ and } y, z > m_0 \tag{2.1}$$

and the density is Lipschitz continuous at m_0, then

$$\hat{m}_{MLE} - m_0 = o_P\left(\left(n^{-1/2}\log^2(n)\right)^{1/(2k+1)}\right). \tag{2.2}$$

We conjecture that the estimates leading to (2.2) are too crude and that the rate is $n^{-1/(2k+1)}$. The heuristic basis of the conjecture is given in Section 5. The truth of this conjecture would imply that this estimate has convergence rate $O(n^{-1/5})$, the same rate as kernel based density estimate (Eddy (1980)), if $f''(m_0; m_0) < 0$ and has rate $O(n^{-1/3})$ if the density has a wedge (e.g. the triangular density). Wang (1994) showed that Birge's (1987c) result can be extended to any $\hat{m}$ which converges to m_0 no slower than $O_P(n^{-1/(2k+1)})$. Thus, if the conjecture is correct $\hat{m}_{MLE}$ give the appropriate rate for the L_1-norm: $\|\hat{f}_{MLE}(\cdot) - f(\cdot; m_0)\|_1 = O_P(n^{-1/3})$. Numerical support for the conjecture is given in Figures 5.1 and 5.2 and some heuristics for the conjecture are given following the proof of (2.2) in Section 5.

It has been observed empirically that the MLE for estimating a unimodal density appears to be spiky near the estimated mode. We suggest a pregrouping technique to reduce the spikiness and computation. The idea is to group the data first, and then apply the plug-in technique. Let $\{I_j = (-t_j, t_{j+1}], j = 0, \pm 1, \pm 2, \ldots\}$ be a partition of the real line, where $\{t_j\}$ is a sequence of increasing constants. Define a modified version of the empirical distribution function by

$$F_n^*(x) = \frac{1}{n}(\# \text{ of } X_i's \le t_{j+1}), \text{ when } x \in (t_j, t_{j+1}].$$

Let $\hat{f}_n^*(x; m)$ be the left derivative of the least concave majorant of $F_n^*(x)$ when $x > m$, and the right derivative of the greatest convex minorant of $F_n^*(x)$ when $x < m$. Let $\hat{m}$ be a consistent estimate of m_0. We call $\hat{f}_n^*(x; \hat{m})$ a "pregrouping" version of the plug-in MLE $\hat{f}_n(x; \hat{m})$. Note that the estimator $\hat{f}_n^*(x; \hat{m})$ is the plug-in MLE of the grouped data: taking all data in the interval $(t_j, t_{j+1}]$ to be t_{j+1}.

Intuitively, the coarser the partition of the interval, the less spiky the MLE. A natural question is how crude a partition can be so that the pregrouping MLE preserves the asymptotic properties of the usual MLE.

Theorem 3. *Let $\hat{m}$ be a consistent estimate of the mode m_0. Suppose that the function $f(\cdot; m_0)$ is bounded, and $f'(x; m_0)$ is nonzero at the point x. If $\max_j |t_{j+1} - t_j| = o(n^{-1/2})$, then the conclusion of Theorem* 1 *holds with $\hat{f}_n(x; \hat{m})$ replaced by $\hat{f}_n^*(x; \hat{m})$.*

Note that other smoothing methods should also yield the same behavior. For example, the kernel smoothing estimate for estimating a decreasing density would be the density of the least concave majorant of the smoothed empirical distribution (Mammen (1991a)).

The MLE, being a random bin width histogram, is not smooth. We can obtain a smoother estimate by finding the MLE satisfying the monotonicity restrictions among linear splines. The problem, of course, already appears in estimating a decreasing density. Let $X_1', \ldots, X_n'$ be a random sample from a decreasing density f and let $\mathcal{F}_L^D$ be the class of continuous linear spline decreasing densities on $[X_1, X_n]$ with knots at the data points. We wish to find:

$$\arg \max_{f \in \mathcal{F}_L^D} \prod_{j=1}^n f(X_j'). \tag{2.3}$$

The solution to problem (2.3) can be computed explicitly by isotonic regression techniques. Let

$$\hat{f}_{aj} = \begin{cases} \min_{a+1 \ge t > j} \max_{s \le j} \frac{t-s}{n(z_t - z_s)}, & \text{when } j < a, \\ \min_{a \le s \le j} \max_{t > j} \frac{t-s}{n(z_t - z_s)}, & \text{when } j > a, \\ \max\left\{\max_{s \le a} \frac{a-s+1}{n(z_{a+1} - z_s)}, \max_{t > a} \frac{t-a}{n(z_t - z_a)}\right\}, & \text{when } j = a, \end{cases} \tag{2.4}$$

where $z_j = (X_j + X_{j-1})/2$ with the convention that $X_0 = X_1$, and $X_{n+1} = X_n$. Let $\hat{f}_{nL}(x; a)$ be the function connecting the points $(X_j, \hat{f}_{aj})$ by using lines, and 0 when x is out of the data range $[X_1, X_n]$. The following two theorems describe the solution and the asymptotic behavior of the linear spline MLE.

Theorem 4. *The solution to problem* (2.3) *is given by* $\hat{f}_{nL}(x; 1)$.

Theorem 5. *Suppose that* $X_1', \ldots, X_n'$ *are independent observations from a decreasing density* f *on* $[0, \infty)$*, which has a nonzero derivative* $f'(x)$ *at a point* $x \in (0, \infty)$*. Then*

$$n^{1/3} \left| \frac{1}{2} f(x) f'(x) \right|^{-1/3} (\hat{f}_{nL}(x, 1) - f(x)) \xrightarrow{\mathcal{L}} 2Z,$$

where the random variable Z *was defined in Theorem* 1.

Linear splines can also be applied to the unimodal case. Let $\mathcal{F}_L^U$ be the class of linear spline unimodal densities on $[X_1, X_n]$ with knots at the data points. We wish to find

$$\arg \max_{f \in \mathcal{F}_L^U} \prod_{j=1}^{n} f(X_j'). \tag{2.5}$$

It will be shown in the proof of Theorem 6 that $\hat{f}_{nL}(x; a)$, defined above, is a density in $\mathcal{F}_L^U$ with mode location X_a. Let $\hat{f}_{nL}(x; \hat{a})$ be the maximizer of the likelihood function among the n possible choices of densities $\hat{f}_{nL}(x; a)$, $a = 1, \ldots, n$. Then, we have the following result.

Theorem 6. *The solution to problem* (2.5) *is given by* $\hat{f}_{nL}(x; \hat{a})$.

Let us give a geometric interpretation of this result. Define a modified empirical distribution (strictly speaking, it is not a cdf)

$$\hat{F}_n^*(x) = \frac{1}{n} \sum_{j=1}^{n+1} I_{\{z_j \le x\}}, \tag{2.6}$$

where I_A is the indicator of the set A. Let $\hat{f}_a^*(x)$ be the left derivative of the least concave majorant of $\hat{F}_n^*(x)$ when $x > z_a$. Then we have for $j > a$,

$$\hat{f}_{aj} = \hat{f}_a^*(z_{j+1}). \tag{2.7}$$

In other words, $\hat{f}_{nL}(x; a)$ is a continuous version of $\hat{f}_a^*(x)$: $\hat{f}_{nL}(x; a)$ is obtained by connecting points $(X_i, \hat{f}_a^*(X_i))$ by lines. This identity gives a simple way of computing $\hat{f}_{aj}$ by using the "pool-adjacent-violators" algorithm, and an indication that MLE linear spline should not be very different from the MLE itself.

3. Discussion

We have proposed the maximum likelihood method to estimate unimodal densities, a pregrouping technique to reduce peaking problems and to save computational cost, and a linear spline approach to produce continuous pictures. Here are some computational details.

Amount of Pregrouping. In practice, we typically take the partition $\{t_j\}$ to be equally spaced grid points with span l_n. Theorem 3 suggests that the choice of l_n be not too large. Practically, we recommend choosing l_n such that the data are grouped into $25 \sim 50$ groups (Recall $5 \sim 15$ bins are suggested for histograms in many textbooks; we need more detail than that), depending on the number of data points. Our experience in simulations shows that such a resolution is detailed enough for practical purposes.

Bayesian Estimation of Mode. Let $\hat{f}_j(\cdot; X_j)$ be the maximum likelihood estimate for the data $\{X_i, i \neq j\}$ with mode location X_j and $L(j)$ be the likelihood of this estimate:

$$L(j) = \prod_{i \neq j} \hat{f}_j(X_i; X_j). \tag{3.1}$$

Define the Bayesian estimate of the mode by

$$\hat{m}_B = \sum_i \frac{L(i)}{\sum_j L(j)} X_i. \tag{3.2}$$

Our empirical experience via simulation shows that this estimator has a more stable variance than $\hat{m}_{MLE}$.

Smoothed MLE. As indicated at the end of Section 2, higher order spline MLE such as linear spline MLE does not produce a qualitatively different curve from the MLE itself. One possible way to produce a smoothed unimodal density is to impose a smoothness penalty on the likelihood function and then to maximize the penalized likelihood subject to the unimodality constraints. We do not explore in this direction because we do not know a simple optimization algorithm. An alternative way is to find a smoothed curve that basically (in a least squares sense) passes through the midpoints of the MLE histogram estimate. Unfortunately, the resultant curve is not necessarily unimodal. Herewith is our smoothing procedure.

Let $(x_1, z_1), \ldots, (x_N, z_N)$ denote the midpoints of the MLE histogram estimate $\hat{f}_{MLE}$ (i.e., x_i is the midpoint of the ith histogram bin and z_i is the height).

Let us take $x_2, x_6, \ldots, x_{4m+2}$ $(m = [(N-2)/4])$, as initial knots that may be deleted. Let corresponding power bases be

$$\begin{cases} B_j(x) = & (x - x_{4j+2})_+^3, \ j = 0, \ldots, m, \\ B_{m+1}(x) = & 1, \ B_{m+2}(x) = x, \ B_{m+3}(x) = x^2, \ B_{m+4}(x) = x^3. \end{cases}$$

Let $\log(f_s(x)) = \sum_1^{m+4} \theta_k B_k(x)$. Use the usual least squares to find θ_k that minimizes

$$\sum_1^N [\log(z_i) - \sum_1^{m+4} \theta_k B_k(x_i)]^2 w_i, \tag{3.3}$$

where w_i is the area of the histogram estimate on the ith bin.

Denote the least square estimate of (3.3) by $\hat{\theta}_j$ with standard error $\mathrm{SE}(\hat{\theta}_j)$. Then, delete the j_0th knot ($1 \le j_0 \le m$) having the smallest absolute t-value: $|\hat{\theta}_j|/\mathrm{SE}(\hat{\theta}_j)$, ($1 \le j \le m$).

Repeat the above deleting process (at each step delete one knot) until the absolute t-value is no smaller than 3. Let $\hat{x}_1, \ldots, \hat{x}_{\hat{j}}$ be the remaining knots with bases $B_j^*(x) = (x - \hat{x}_j)_+^3$, $j = 1, \ldots, \hat{j}$, and $B_{\hat{j}+1}^*(x) = 1$, $B_{\hat{j}+2}^*(x) = x$, $B_{\hat{j}+3}^*(x) = x^2$, and $B_{\hat{j}+4}^*(x) = x^3$, and estimates $\hat{\theta}_j, j = 1, \ldots, \hat{j} + 4$. Now, form the function

$$\hat{f}^*(x) = \exp\left(\sum_1^{\hat{j}+4} \hat{\theta}_j B_j^*(x)\right).$$

Normalize $\hat{f}^*(x)$ to be a density and denote the resulting function by $\hat{f}^{**}(x)$. Then, $\hat{f}^{**}(x)$ is a smoothed version of MLE, which will be presented in the next section. This kind of knot deletion idea was used in CART by Breiman et al. (1983).

4. Simulations

In this section, we use 4 simulated examples to illustrate the proposed procedures and to compare them with the kernel density estimate. For each example, we use sample size $n = 200$ and number of simulations 500. For 500 simulations, it is not possible to plot here all of these estimated curves. Instead, we select a representative simulation — the simulation whose average L_1-loss of the MLE at data points is median among 500 replications. The four simulated examples are

Example 1. exponential distribution: $f(x) = \exp(-x) I_{\{x>0\}}$ (4.1)

Example 2. Gaussian distribution: $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ (4.2)

Example 3. Asymmetric distribution:

$$f(x) = \tfrac{2}{3}(\exp(2x) I_{\{x \le 0\}} + \exp(-x) I_{\{x>0\}}) \tag{4.3}$$

Example 4. Triangular distribution: $f(x) = (1 - |x|)_+$ (4.4)

In the MLE fitting, we only assume that the density is unimodal with unknown mode, although density (4.1) is indeed decreasing. These densities represent different degrees of skewness and different weights of tails.

The kernel density estimate is defined as

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

with the bandwidth determined by the normal reference rule (see Silverman (1986)):

$$h = 1.06 s n^{-1/5}, \tag{4.5}$$

where $K(\cdot)$ is the standard Gaussian density and s is the sample standard deviation. Note that this choice of bandwidth is asymptotically optimal if the true density is normal. Thus, the kernel density performs well under model (4.2). In general, the above choice of bandwidth tends to oversmooth. Hence, it often produces a unimodal density and gives a good estimation of the mode location for symmetric densities. For these reasons, we would expect that the kernel density estimate with bandwidth (4.5) performs well for symmetric distributions.

Figures 1-4 depict the simulation results: The pregrouped MLE estimate with mode estimated by $\hat{m}_{MLE}$, smoothed MLE proposed in Section 3 and the kernel density estimate. The kernel density estimate does not estimate the tail of densities well and mis-estimates the peak when the distribution is asymmetric (e.g densities (4.1) and (4.3)).

Finally, we compare mode estimation by the MLE and by kernel density estimation. As we anticipated, the kernel density estimate performs better for symmetric densities and worse for asymmetric densities. In an attempt to understand the convergence rates, we simulated 500 times from (4.3) and (4.4) for $n = 50 * 2^j, (j = 0, \dots, 5)$, and computed the MSE of the mode estimation for three estimators: $\hat{m}_{MLE}$, $\hat{m}_B$, and the kernel density estimate. Figure 5 plots the logarithm of MSE against $\log_2(n)$ (hence the slope indicates the rate of convergence). For the symmetric density (4.4), the MLE method seems to have a rate comparable to the kernel density estimate except that the constant factors are larger. For the asymmetric distribution (4.3), the mode estimation by kernel has a much slower rate of convergence. Overall, the Bayesian estimation of mode (3.2) seems to have a smaller constant factor than the $\hat{m}_{MLE}$ (the rates are the same because the curves are parallel). For the symmetric density (4.4) the bias in the mode estimation is negligible (about 10 to 100 times smaller than the variance), whereas for the asymmetric density (4.4), the bias is not negligible. Figure 5.3 shows the bias and variance contribution in the logarithmic scale.

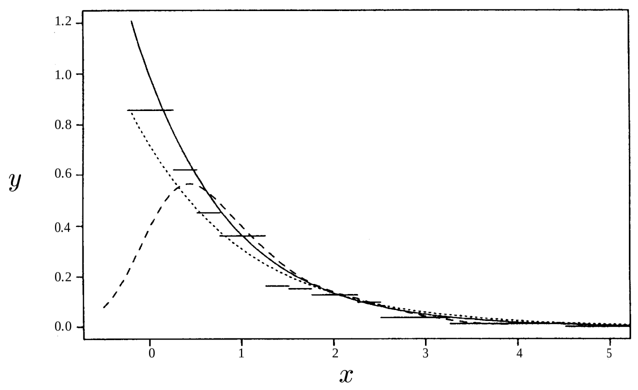

Figure 1.1. Example 1: Density Estimation

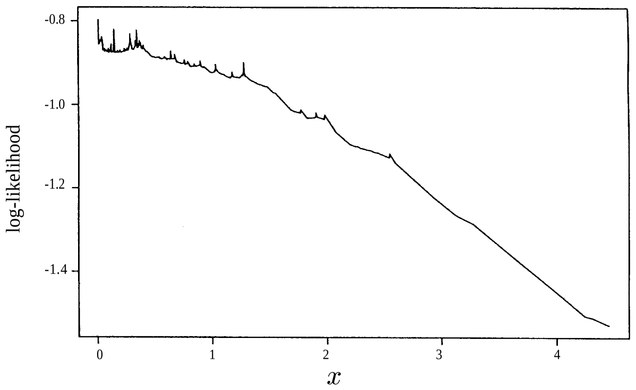

Figure 1.2. Example 1: Profile log-likelihood

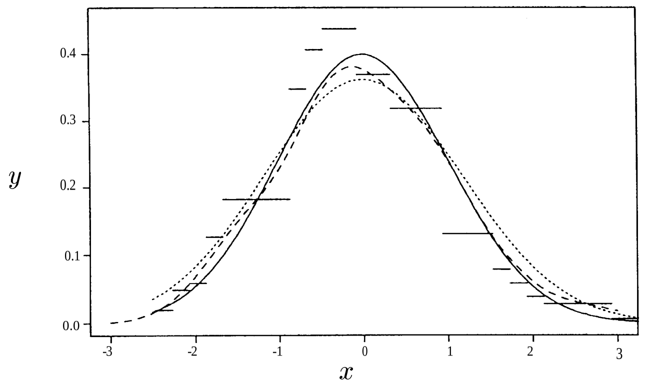

Figure 2.1. Example 2: Density Estimation

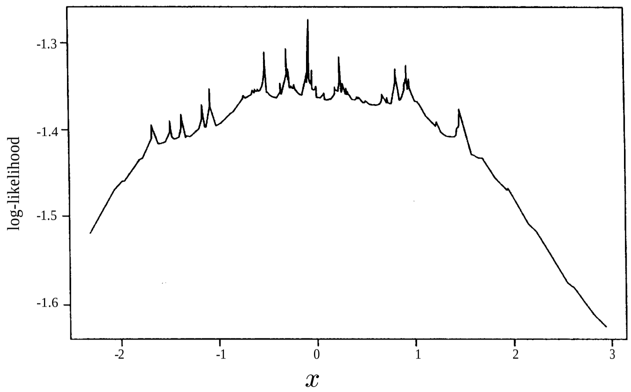

Figure 2.2. Example 2: Profile log-likelihood

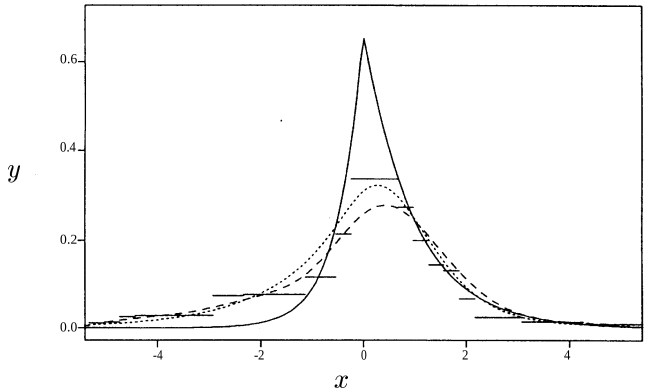

Figure 3.1. Example 3: Density Estimation

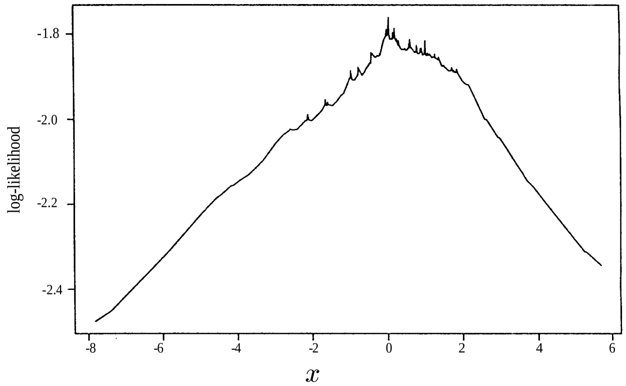

Figure 3.2. Example 3: Profile log-likelihood

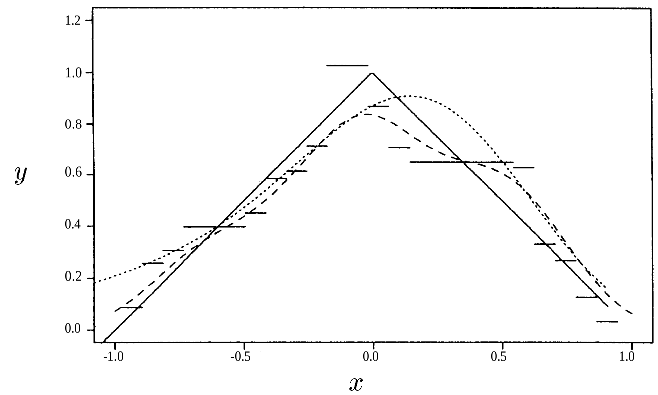

Figure 4.1. Example 4: Density Estimation

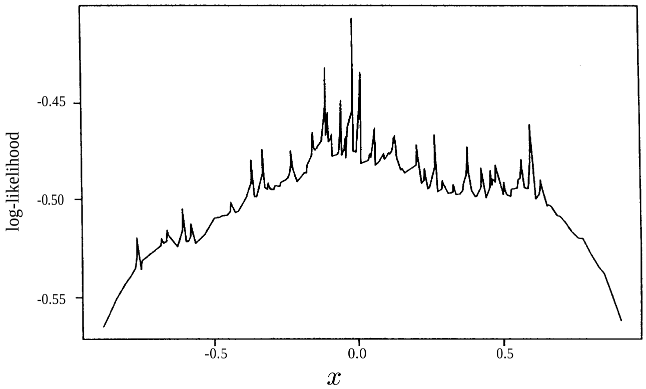

Figure 4.2. Example 4: Profile log-likelihood

Figure Captions

Figures 1-4. A representative estimated curve based on MLE and kernel density estimate with sample size $n = 200$. Figures 1.1-4.1 are the estimated curves. Solid curve — true density; solid step function — pregrouped MLE estimate; dashed line — kernel density estimate with bandwidth (4.5); dotted line — smoothed MLE. Figures 1.2 - 4.2 are plot of the logarithm of the profile likelihood: $\{X_j\}$ against $n^{-1} \log L(j)$ with $L(j)$ defined by (3.1).

Figure 5. Plot of the logarithm of mean square errors against the logarithm of the sample sizes for mode estimation. The slope in this log-log plot indicates the rate of convergence. Figure 5.1 is for asymmetric density (4.3) and Figure 5.2 is for symmetric density (4.4). Figure 5.3 gives the bias and variance decomposition for the asymmetric density (4.3) in the logarithmic scale. ■ — Bias; ▲ — variance. For the symmetric density (4.4), the bias is negligible.

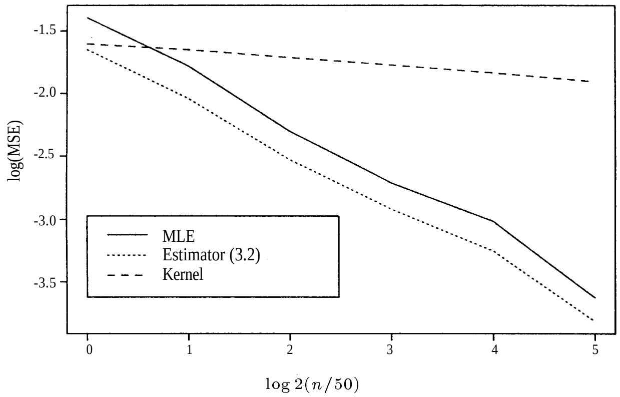

Figure 5.1. Example 3: MSE for mode estimation

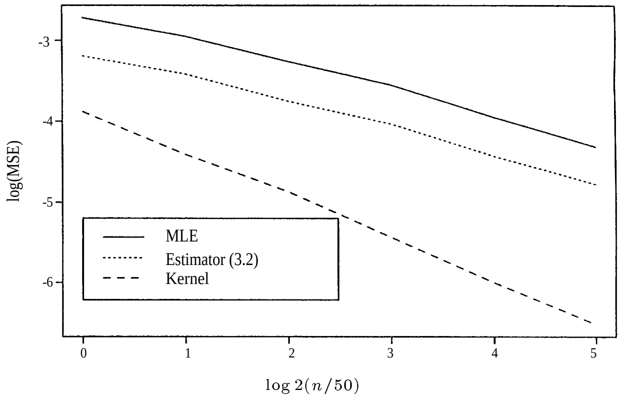

Figure 5.2. Example 5: MSE for mode estimation

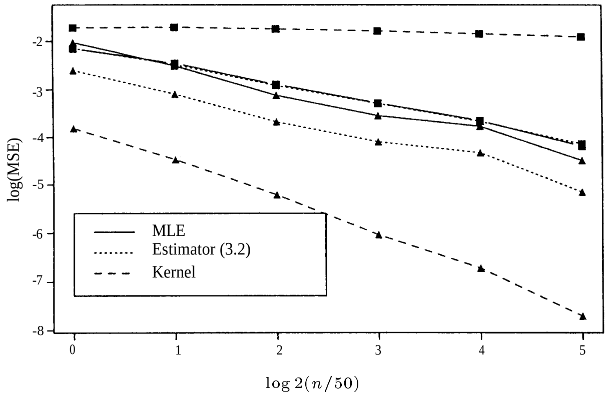

Figure 5.3. Example 3: Bias and Variance mode estimation

5. Proofs

Proof of Theorem 1. We give the proof for $x > m_0$; the other case can be treated similarly. First note that if $x > m_1 \geq m_2$, then

$$\hat{f}_n(x; m_1) \geq \hat{f}_n(x; m_2), \tag{5.1}$$

by the definition of the estimators.

Let $l(x) = |\frac{1}{2}f(x;m_0)f'(x;m_0)|^{-1/3}$. For any $\varepsilon > 0$, and $m_0 + \varepsilon < x$, by the consistency of $\hat{m}$,

$$\begin{aligned} &P\left\{n^{\frac{1}{3}}l(x)(\hat{f}_n(x;\hat{m}) - f(x;m_0)) \leq t\right\} \\ &= P\left\{n^{\frac{1}{3}}l(x)(\hat{f}_n(x;\hat{m}) - f(x;m_0)) \leq t, |\hat{m} - m_0| \leq \varepsilon\right\} + o(1). \end{aligned} \tag{5.2}$$

Note that $f(x; m_0)$ is decreasing when $x \geq m_0 + \varepsilon$. By a result of Prakasa Rao (1969) and Groeneboom (1985), we have

$$P\left\{n^{\frac{1}{3}}l(x)(\hat{f}_n(x;m_0+\varepsilon) - f(x;m_0)) \leq t\right\} \longrightarrow P\{2Z \leq t\}, \ \forall t \in (-\infty, +\infty). \tag{5.3}$$

Thus, the combination of (5.1), (5.2) and (5.3) leads to

$$\begin{aligned} &\liminf P\left\{n^{\frac{1}{3}}l(x)(\hat{f}_n(x;\hat{m}) - f(x;m_0)) \leq t\right\} \\ \geq\ &\liminf P\left\{n^{\frac{1}{3}}l(x)(\hat{f}_n(x;m_0+\varepsilon) - f(x;m_0)) \leq t\right\} = P\left\{2Z \leq t\right\}. \end{aligned} \tag{5.4}$$

Similarly, by (5.1) and (5.2), we have

$$\begin{aligned} &\limsup P\left\{n^{\frac{1}{3}}l(x)(\hat{f}_n(x;\hat{m}) - f(x;m_0)) \leq t\right\} \\ \leq\ &\limsup P\left\{n^{\frac{1}{3}}l(x)(\hat{f}_n(x;m_0-\varepsilon) - f(x;m_0)) \leq t\right\}. \end{aligned} \tag{5.5}$$

The proof is completed if we show that (5.5) has a limit (5.4). Let $f_\varepsilon = f(y;m_0)/(1 - F(m_0 - \varepsilon))$ and $f_\varepsilon^*(\cdot)$ be the solution to the problem:

$$\max_{g(\cdot) \text{ is a decreasing density on } [m_0-\varepsilon,\infty)} \int_{m_0-\varepsilon}^{\infty} [\log g(y)] f_\varepsilon(y) dy.$$

Then, $f_\varepsilon^*(y) = f_\varepsilon(a)1_{\{m_0-\varepsilon\leq y\leq a\}} + f_\varepsilon(y)1_{\{y>a\}}$, where a is chosen so that f_ε^* is a density function. See Bickel and Fan (1990) for a proof. Thus, for each fixed $x > m_0$, there exists ε_0 such that $f_\varepsilon^*(x) = f_\varepsilon(x)$ for $\varepsilon < \varepsilon_0$. By the argument of Groeneboom (1985), one can show that

$$P\left\{N_1^{1/3}|\frac{1}{2}f_\varepsilon^*(x)f_\varepsilon^{*\prime}(x)|^{-1/3}(\hat{f}_{N_1}^{**}(x) - f_\varepsilon^*(x)) \leq t\right\} \longrightarrow P\left\{2Z \leq t\right\},$$

where $\hat{f}^{**}_{N_1}(\cdot)$ is the MLE over the class of decreasing densities on $[m_0 - \varepsilon, \infty)$ based on data $X_j \geq m_0 - \varepsilon$, and $N_1 = n[1 - \hat{F}_n(m_0 - \varepsilon)]$. Using the fact that

$$\hat{f}_n(x; m_0 - \varepsilon) = \frac{N_1}{n} \hat{f}^{**}_{N_1}(x), \forall x > m_0 - \varepsilon$$

we have for $\varepsilon < \varepsilon_0$,

$$P\left\{ n^{1/3} l(x) [\hat{f}_n(x; m_0 - \varepsilon) - f(x; m_0)] \leq t \right\} \longrightarrow P\left\{2Z \leq t\right\}.$$

This, together with (5.4) and (5.5), leads to the desired conclusion.

We need the following two lemmas to prove Theorem 2.

Lemma 1. *Let $f(x; m_0)$ be a unimodal density with mode m_0 and*

$$G(m) = \sup_{g \in \mathcal{F}_m} \int \log g(x) f(x; m_0) dx, \tag{5.6}$$

where $\mathcal{F}_m$ is the class of unimodal densities with mode m. Then $G(m)$ is increasing when $m < m_0$ and is decreasing when $m > m_0$. If condition (2.1) *is satisfied, then for m in a neighborhood of m_0,*

$$G(m_0) - G(m) > c_1 |m_0 - m|^{2k+1} \tag{5.7}$$

for some $c_1 > 0$.

Proof. Without loss of generality, we prove this lemma for the case $m < m_0$. First, the solution to the optimization problem (5.6) is given by

$$f_m(x) = h_m 1_{\{m \leq x \leq M_m\}} + f(x; m_0) 1_{\{x < m \text{ or } x > M_m\}}, \tag{5.8}$$

where $h_m = f(M_m; m_0)$ and M_m is a constant such that $f_m(x)$ is a density:

$$\int_m^{M_m} f(x; m_0) dx = h_m (M_m - m). \tag{5.9}$$

(See Bickel and Fan (1990) for a proof.) Given $m_2 < m_1 < m_0$, since $f_{m_2} \in \mathcal{F}_{m_1}$, we conclude that $G(m_1) \geq G(m_2)$. Therefore, $G(m)$ is increasing when $m \leq m_0$.

Next, we prove (5.7). First of all, by (5.8), we have

$$G(m_0) - G(m) = \int_m^{M_m} \log(f(x; m_0)/h_m) f(x; m_0) dx.$$

Evidently, as $m \to m_0$, $M_m \to m_0$ and

$$\sup_{m \leq x \leq M_m} |f(x; m_0)/h_m - 1| \to 0.$$

By Taylor's expansion, we obtain

$$\begin{aligned} G(m_0) - G(m) &= \int_m^{M_m} (f(x;m_0)/h_m - 1) f(x;m_0) dx \ (1 + o(m_0 - m)) \\ &= \int_m^{M_m} (f(x;m_0) - h_m)^2 dx/h_m \ (1 + o(m_0 - m)), \qquad (5.10) \end{aligned}$$

where the last equality follows from (5.9). Let $m^* \in (m, m_0)$ be the point such that $f(m^*;m_0) = f(M_m;m_0)$. When m is close to m_0, by (2.1) we have

$$\begin{aligned} G(m_0) - G(m) &\geq \frac{1}{2f(m_0;m_0)} \int_m^{M_m} (f(x;m_0) - f(m^*;m_0))^2 dx \\ &\geq \frac{1}{2f(m_0;m_0)} \int_m^{m_0} \left(c|x - m^*|^k\right)^2 dx \\ &= \frac{c^2}{2(2k+1)f(m_0;m_0)} \left(|m_0 - m^*|^{2k+1} + |m^* - m|^{2k+1}\right) \\ &\geq \frac{c^2}{2^{2k+2}(2k+1)f(m_0;m_0)} \Big| m_0 - m \Big|^{2k+1}. \end{aligned}$$

The conclusion follows from the last inequality.

Recall that $X_1 < \cdots < X_n$ denote the order statistics.

Lemma 2. *Suppose that the tail of the underlying distribution satisfies $F(x) - F(-x) = 1 - o(x^{-1/\alpha})$ as $x \to +\infty$ for some $\alpha > 0$ and that the density $f(x;m_0)$ is bounded. Then, the minimum and maximum spacing satisfy*

$$P\{\min_i (X_i - X_{i-1}) > n^{-2-\delta}\} \to 1, \quad P\{X_n - X_1 \leq n^{\alpha}\} \to 1,$$

for all $\delta > 0$.

Proof. According to Pyke (1965), the uniform spacing has the following representation:

$$(F(X_2) - F(X_1), \ldots, F(X_n) - F(X_{n-1})) \overset{d}{=} (\xi_1, \ldots, \xi_{n-1}) / \sum_{i=1}^{n+1} \xi_i,$$

where $\xi_1, \ldots, \xi_{n+1}$ are i.i.d. standard exponential random variables. Thus,

$$\begin{aligned} P\{n^{2+\delta/2} \min_i (F(X_{i+1}) - F(X_i)) > 1\} &= P\{\min_i \xi_i > n^{-1-\delta/2} \sum_{i=1}^{n+1} \xi_i / n\} \\ &\geq P\{\min_i \xi_i > 2n^{-1-\delta/2}\} + o(1) \to 1. \end{aligned}$$

Since $\sup_x f(x) \min_i (X_{i+1} - X_i) \geq \min_i (F(X_{i+1}) - F(X_i))$, we have

$$P\{\min_i (X_i - X_{i-1}) > n^{-2-\delta}\} \to 1.$$

It is easy to check, under our assumption on F, that

$$P\{X_n > n^\alpha/2\} \to 1 \quad \text{and} \quad P\{X_1 < -n^\alpha/2\} \to 1.$$

Thus, with probability tending to one, $X_n - X_1 \leq n^\alpha$. This completes the proof.

Proof of Theorem 2. Denote the log-likelihood by

$$G_n(X_j) = \sup_{g \in \mathcal{F}_{X_j}} \frac{1}{n} \sum_{i \neq j} \log g(X_i).$$

Since the maximum likelihood estimate $\hat{f}(x; X_j)$ is the right derivative of the greatest convex minorant of the empirical distribution when $x < X_j$, and the left derivative of the least concave majorant of the empirical distribution when $x > X_j$, then by Lemma 2, with probability tending to one, we have

$$\max_{i \neq j} \hat{f}(X_i; X_j) < n^3 \quad \text{and} \quad \min_{i \neq j} \hat{f}(X_i; X_j) > n^{-\alpha-1}.$$

Denote this set by Ω_n. The previous statement is equivalent to $P(\Omega_n) \to 1$. Thus, for $\omega \in \Omega_n$,

$$\begin{aligned} G_n(X_j) &= \sup_{\{\|\log g\|_\infty \leq d \log n;\ g \in \mathcal{F}_{X_j}\}} \frac{1}{n} \sum_{i \neq j} \log g(X_i) \\ &= \sup_{\{\|\log g\|_\infty \leq d \log n;\ g \in \mathcal{F}_{X_j}\}} \int \log g dP_n + O(\log n/n), \end{aligned}$$

where P_n is the empirical processes and $d = \max\{3, \alpha + 1\}$. Let $\mathcal{C}$ be the class of unimodal functions whose supnorm is bounded by 1. Then, for $\omega \in \Omega_n$

$$\max_j |G_n(X_j) - G(X_j)| \leq d \log n \sup_{g \in \mathcal{C}} \left| \int g(x)(dP_n - dP) \right| + O(\log n/n).$$

By empirical process theory (Theorem 37, Pollard (1984))

$$\sup_{g \in \mathcal{C}} \left| \int g(x)(dP_n - dP) \right| = o\left(a_n (\log n/n)^{1/2}\right) \text{ almost surely,}$$

for any sequence $a_n \to \infty$. Taking $a_n = \log^{0.25}(n)$, say, we have

$$\max_j |G_n(X_j) - G(X_j)| = o_P(\log^{1.75}(n)/\sqrt{n}). \tag{5.11}$$

If the mode m_0 is uniquely defined, then for small $\varepsilon > 0$, $f_{m_0-\varepsilon}(\cdot)$ and $f_{m_0-2\varepsilon}(\cdot)$ defined by (5.8) can not be identical. Thus, by the unimodality of $G(\cdot)$ in Lemma 1,

$$G(m_0-\varepsilon) > G(m_0-2\varepsilon) \quad \text{and} \quad G(m_0+\varepsilon) > G(m_0+2\varepsilon),$$

and hence

$$\inf_{|m-m_0|\le\varepsilon} G(m) > \sup_{|m-m_0|\ge 2\varepsilon} G(m).$$

Using this and (5.11), then $X_{\hat{j}} \in (m_0-2\varepsilon, m_0+2\varepsilon)$ with probability tending to one. That is, $\hat{m}_{MLE}$ is a consistent estimate of m_0.

Proof of (2.2). Let $\varepsilon_n = (\log^{1.75}(n)/n^{1/2})^{1/(2k+1)}$. In the sequel, we show that with probability tending to one, it is not possible to have $\hat{m}_{MLE}$ lies outside the interval $(m_0-\varepsilon_n, m_0+\varepsilon_n)$. By Lemma 1,

$$G(m_0) \ge \max\{G(m_0-\varepsilon_n), G(m_0+\varepsilon_n)\} + c_1\varepsilon_n^{2k+1} = \sup_{|m-m_0|\ge\varepsilon_n} G(m) + c_1\varepsilon_n^{2k+1}.$$

Since $f(\cdot)$ is Lipschitz continuous at m_0, it can easily deduced from (5.10) that

$$0 \le G(m_0) - \min\{G(m_0-\log n/n), G(m_0+\log n/n)\} \le O(\log n/n).$$

Consequently, when n is large,

$$\inf_{|m-m_0|\le\log n/n} G(m) > \sup_{|m-m_0|>\varepsilon_n} G(m) + c_1\varepsilon_n^{2k+1}/2. \tag{5.12}$$

It is easy to show that

$$P\{\text{at least one data point falls in } (m_0-\log n/n, m_0+\log n/n)\} \to 1.$$

Let X^* be a data point in $(m_0-\log n/n, m_0+\log n/n)$. By (5.11),

$$G_n(X^*) \ge G(X^*) + o_P\left(\log^{1.75}(n)n^{-1/2}\right).$$

For X_j such that $|X_j-m_0| \ge \varepsilon_n$, then by (5.11) and (5.12), when n is large, we have

$$\begin{aligned} G_n(X^*) &\ge G(X_j) + c_1\varepsilon_n^{2k+1}/2 + o_P\left(\log^{1.75}(n)n^{-1/2}\right) \\ &\ge G_n(X_j) + c_1\varepsilon_n^{2k+1}/2 + o_P\left(\log^{1.75}(n)n^{-1/2}\right) \\ &> G_n(X_j). \end{aligned}$$

Thus, with probability tending to one, the maximum of $G_n(\cdot)$ can not be achieved at the point X_j such that $|X_j-m_0| > \varepsilon_n$. Hence,

$$P\{|\hat{m}_{MLE}-m_0| \le \varepsilon_n\} \to 1,$$

and

$$\hat{m}_{MLE} - m_0 = O_p(\varepsilon_n) = o_P\left((\log^2 n/n)^{1/(2k+1)}\right).$$

The conclusion follows.

Heuristic basis of conjecture. We base our conjecture on the conjectured approximation,

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{\hat{f}_n(X_i; m)}{\hat{f}_n(X_i; m_0)} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f_m(X_i)}{f_{m_0}(X_i)} + O_P(n^{-1} + |m - m_0|^3) \tag{5.13}$$

uniformly for $m, m_0 \notin \{X_i : 1 \leq i \leq n\}$ where $f_m(\cdot)$ is defined by (5.8) and the conjecture that $\hat{m}$ behaves like the maximum of the left hand side of (5.13) for m as specified. If (5.13) holds it is easy to see from lemma 1 that, if $k \geq 1$,

$$\begin{aligned}&\sup\{\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_m(X_i)}{f_{m_0}(X_i)} : |m - m_0| \geq \epsilon\} \\ &= O_P(\epsilon^{(2k+1)/2} n^{-1/2}) + C\epsilon^{2k+1}, \qquad \text{where} \quad C > 0.\end{aligned} \tag{5.14}$$

Therefore, if $k \geq 1$ the sup in (5.14) must be achieved for $|m - m_0| = O(n^{-\frac{1}{2k+1}})$. Since the remainder in (5.13) is also $O_P(n^{-1})$ for $k \geq 1$ the conjecture follows. Our belief in (5.13) is based on the behavior in the corresponding parametric situation where $f = f(\cdot, \mu, \eta)$, the truth is $f(\cdot, \mu_0, \eta_0)$, $\hat{\eta}(\mu)$ is defined by

$$\hat{\eta}(\mu) = \max{}^{-1} \frac{1}{n}\sum_{i=1}^{n} \log f(X_i, \mu, \eta)$$

and $\eta(\mu)$ by $\max^{-1} \int \log f(x, \eta, \mu) f(x, \mu_0, \eta_0) dx$. If we Taylor expand

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i, \mu, \eta(\mu))}{f(X_i, \mu, \hat{\eta}(\mu))}$$

about $\hat{\eta}(\mu)$ and $\hat{\eta}(\mu)$ is assumed in the interior then, if $|\mu - \mu_0| = o(1)$, we expect

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} \left(\log \frac{f(X_i, \mu, \hat{\eta}(\mu))}{f(X_i, \mu_0, \hat{\eta}(\mu_0))} - \log \frac{f(X_i, \mu, \eta(\mu))}{f(X_i, \mu_0, \eta(\mu_0))}\right) \\ &= O_P(\{|\hat{\eta}(\mu) - \eta(\mu)|^2 - |\hat{\eta}(\mu_0) - \eta_0|^2\}).\end{aligned} \tag{5.15}$$

Finally, it seems plausible that

$$(\hat{\eta}(\mu) - \eta(\mu)) - (\hat{\eta}(\mu_0) - \eta(\mu_0)) = O_P(|\hat{\eta}(\mu_0) - \eta(\mu_0)||\mu - \mu_0|). \tag{5.16}$$

If we combine (5.15) and (5.16), identify η with the shape of f, and note that in our case we expect $\hat{\eta}(\mu) - \eta(\mu) = O_P(n^{-1/3})$ then (5.13) follows. Of course, there is much wrong with this argument. We do not have any assurance that bounds like (5.15) and (5.16) are valid since we know that $\hat{\eta}(\mu)$ is achieved on the boundary so that we cannot use Taylor expansions in function space. Nevertheless the conjecture looks promising to us.

Lemma 3. *Let $X_1', \ldots, X_n'$ be i.i.d with a density $f(x)$. If f is bounded and the maximum span of the partition satisfies the condition of Theorem 3, then*

$$\sup_x |\hat{F}_n(x) - F_n^*(x)| = o_p(n^{-1/2}),$$

where $\hat{F}_n$ is the empirical cdf of $X_1', \ldots, X_n'$.

We omit the proof of Lemma 3; (but see Bickel and Fan (1990)).

Proof of Theorem 3. We need only to prove the result for the decreasing density case; the unimodal case follows from the result of estimating a decreasing density and the proof of Theorem 1.

By Lemma 3, and the Hungarian embedding of Komlós et al. (1973), the process $F_n^*(t)$ has the following decomposition:

$$\begin{aligned} n^{1/2}(F_n^*(t) - F(t)) &= n^{1/2}\left(\hat{F}_n(t) - F(t)\right) + n^{1/2}\left(F_n^*(t) - \hat{F}_n(t)\right) \\ &= B_n(F(t)) + o_p(1), \end{aligned}$$

where $\{B_n, n \geq 1\}$ is a sequence of Brownian bridges, constructed on the same space as the $\hat{F}_n(t)$, the empirical process. The conclusion follows from the proof of Theorem 2.1 of Groeneboom (1985).

Proof of Theorem 4. The result follows from the proof of Theorem 6.

Proof of Theorem 5. Let $l(x) = |f(x)f'(x)/2|^{-1/3}$. By the proof of Lemma 2, with probability tending to one, the maximum spacing for the data set $\{X_i : X_i \in x \pm \varepsilon\}$ is of order $O(n^{-1}\log n)$, where ε is small enough so that $\inf_{y\in x\pm\varepsilon} f(y) > 0$. Thus, with probability tending to 1, the points $x - \varepsilon_n$, x, $x + \varepsilon_n$ are in different intervals of (z_j, z_{j+1}), where $\varepsilon_n = n^{-2/5}$, and z_j was defined in (2.6). Thus, by (2.7), we have with probability tending to one that

$$\hat{f}_1^*(x + \varepsilon_n) \leq \hat{f}_{nL}(x; 1) \leq \hat{f}_1^*(x - \varepsilon_n), \tag{5.17}$$

where $\hat{f}_1^*(x)$ was defined after (2.6).

Note that the modified empirical distribution defined by (2.6) satisfies $0 \leq \hat{F}_n^*(x) - \hat{F}_n(x) \leq 1/n$, where $\hat{F}_n(\cdot)$ is the usual empirical cdf. Thus, by the same argument as in the proof of Theorem 3, we have

$$P\left\{n^{1/3} l(x)(\hat{f}_1^*(x + \varepsilon_n) - f(x)) \leq t\right\} \longrightarrow P\{2Z \leq t\}, \quad \forall t \in (-\infty, \infty).$$

Consequently, by (5.17),

$$\begin{aligned} &\limsup_n P\left\{n^{1/3}l(x)(\hat{f}_{nL}(x;1)-f(x))\le t\right\} \\ \le\ &\limsup_n P\left\{n^{1/3}l(x)(\hat{f}_1^*(x+\varepsilon_n)-f(x))\le t\right\} \\ =\ &P\{2Z\le t\}, \quad \forall t\in(-\infty,\infty). \end{aligned}$$

The conclusion follows from a similar inequality:

$$\liminf_n P\left\{n^{1/3}l(x)(\hat{f}_{nL}(x;1)-f(x))\le t\right\} \ge\ P\{2Z\le t\}, \quad \forall t\in(-\infty,\infty).$$

We need the following lemma (Theorem 1.5.1 of Robertson et al. (1988)) to prove Theorem 6.

Lemma 4. *Suppose that $\Phi(\cdot)$ is differentiable, and convex on an interval I. Let $\Delta_\Phi(u,v)=\Phi(u)-\Phi(v)-(u-v)\Phi'(v)$. If f_j^* is a solution of problem* (5.21), *then f^* minimizes $\sum_j \Delta_\Phi(g_j,f_j)w_j$ in the class of isotonic functions f.*

Proof of Theorem 6. We need only prove that $\hat{f}_n(x;a)$ is the solution to the problem (2.5) with an additional constraint that the location of the mode is X_a. Let $f_j=f(X_j)$. Then the problem is equivalent to

$$\max \sum_j \log f_j$$

$$\text{subject to }: (\textit{unimodality}) \quad f_1\le f_2\le\cdots\le f_a\ge f_{a+1}\ge\cdots\ge f_n, \tag{5.18}$$

$$(\textit{Area one}) \quad \sum_{j=1}^{n-1}\frac{f_{j+1}+f_j}{2}(X_{j+1}-X_j)=1. \tag{5.19}$$

Write $c_j=(X_{j+1}-X_{j-1})/2$ with $X_0=X_1$, and $X_{n+1}=X_n$. Then the equality constraint (5.19) can be rewritten as

$$\sum_{j=1}^n c_jf_j=1. \tag{5.20}$$

Denote $g_j=1/(nc_j)$ and $w_j=nc_j$. Then, the optimization problem is equivalent to maximizing $\sum_1^n \log f_j$ subject to (5.18) and $\sum_1^n(g_j-f_j)w_j=0$. Consider the problem of isotonic regression

$$\min_f \sum_1^n (f_j-g_j)^2w_j \tag{5.21}$$

with a partial order $1 \preceq 2 \preceq \cdots \preceq a \succeq a+1 \succeq a+2 \succeq \cdots \succeq n$. Then, the solution to the problem (5.21) is given by (2.4) (see page 23 of Robertson et al. (1988)). The solution also satisfies (Theorem 1.3.6 of Robertson et al. (1988))

$$\sum_1^n (\hat{f}_{aj} - g_j) w_j = 0,$$

i.e. (5.19). Now, let us apply Lemma 4. Take a convex function $\Phi(u) = u \log u$. Then, $\hat{f}_a$ also minimizes

$$\sum_1^n (g_j \log g_j - g_j \log f_j - g_j + f_j) w_j = c - \sum_1^n \log f_j + n \sum_1^n c_j f_j,$$

under the isotonic constraints, where $c = \sum \log g_j - n$. Since we are interested only in the class of isotonic regression satisfying (5.20), $\hat{f}_a$ maximizes $\sum_1^n \log f_j$ under the constraints (5.18) and (5.19). The desired conclusion follows.

Acknowledgment

The research of the first author was supported by ONR N00014-89-J-1563. The research of the second author was supported in part by NSF Grant DMS-9005905 and NSF Grant DMS-9113527. We thank the associate editor and referees for the helpful suggestions and comments. In particular, we appreciate the profile likelihood heuristic raised in the review process.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). Statistical Inference under Order Restrictions. John Wiley, London.

Barlow, R. E. and van Zwet, W. R. (1970). Asymptotic properties of isotonic estimators for the generalized failure rate function, part I: strong consistency. In Nonparametric Techniques in Statistical Inference, (Edited by M. L. Puri), 159-173, Cambridge University Press.

Bickel, P. J. and Fan, J. (1990). Some problems on the estimation of densities under shape restrictions. Technical Report 258, Dept. of Statist., Univ. of California, Berkeley.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1983). CART: Classification and Regression Trees. Wadsworth, Belmont, CA.

Birgé, L. (1987a). Estimating a density under order restrictions: Nonasymptotic minimax risk. Ann. Statist. **15**, 995-1012.

Birgé, L. (1987b). On the risk of histograms for estimating decreasing densities. Ann. Statist. **15**, 1013-1022.

Birgé, L. (1987c). Robust estimation of unimodal densities. Unpublished manuscript.

Birgé, L. (1989). The Grenander estimator: A nonasymptotic approach. Ann. Statist. **17**, 1532-1549.

Chernoff, H. (1964). Estimation of the mode. Ann. Inst. Statist. Math. **16**, 31-41.

Eddy, W. F. (1980). Optimum kernel estimators of the mode. Ann. Statist. **8**, 870-882.

Grenander, U. (1956). On the theory of mortality measurement, Part II. Skand. Akt. **39**, 125–153.

Groeneboom, P. (1985). Estimating a monotone density. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer **Vol II** (Edited by L. M. Le Cam and R. A. Olshen), 539-555.

Komlós, J., Major, P. and Tusnády, G. (1975). An approximation of partial sums of independent r.v.'s and the sample d.f. Z. Wahrsch. verw. Gebiete **32**, 111-131.

Mammen, E. (1991a). Estimating a smooth monotone regression function. Ann. Statist. **19**, 724-740.

Mammen, E. (1991b). Nonparametric regression under qualitative smoothness assumptions. Ann. Statist. **19**, 741-759.

Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statist. **33**, 1065-1076.

Pollard, D. (1984). Convergence of Stochastic Processes. Springer-Verlag, New York.

Prakasa Rao, B. L. S. (1969). Estimation of a unimodal density. Sankhyā Ser.A, **31**, 23-36.

Pyke, R. (1965). Spacings. J. Roy. Statist. Soc. Ser.B **27**, 395-436.

Ramsay, J. O. (1988). Monotone regression splines in action. Statist. Sci. **3**, 425-461.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). Order Restricted Statistical Inference. John Wiley, New York.

Silverman, B. W. (1986), Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. Ann. Statist. **8**, 1348-1360.

Venter, J. H. (1967). On estimation of the mode. Ann. Math. Statist. **38**, 1446-1455.

Wang, J. L. (1986). Asymptotically minimax estimators for distributions with increasing failure rate. Ann. Statist. **14**, 1113-1131.

Wang, Y. (1994). The L_1 theory of estimation of monotone and unimodal densities. J. Nonparametr. Statist. to appear.

Wegman, E. J. (1969). A note on estimating a unimodal density. Ann. Math. Statist. **40**, 1661-1667.

Wegman, E. J. (1970a). Maximum likelihood estimation of a unimodal density function. Ann. Math. Statist. **41**, 457–471.

Wegman, E. J. (1970b). Maximum likelihood estimation of a unimodal density, II. Ann. Math. Statist. **41**, 2169–2174.

Woodroofe, M. and Sun, J. (1993). A penalized maximum likelihood estimate of $f(0+)$ when f is non-increasing. Statist. Sinica **3**, 501-515.

Department of Statistics, University of California, Berkeley, CA 94720, U.S.A.

Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, U.S.A.

(Received August 1992; accepted March 1995)