Some Problems on the Estimation of Densities under Shape Restrictions

Peter J. Bickel
Department of Statistics
University of California
Berkeley, CA 94720

Jianqing Fan
Department of Statistics
University of North Carolina
Chapel Hill, N.C. 27514

Technical Report No. 258
June 1990

Supported by ONR N00014-89-J-1563
AMS 1980 subject classification. Primary 62G05. Secondary 62E20, 60G30.

Department of Statistics
University of California
Berkeley, California

Some Problems on the Estimation of Densities under Shape Restrictions

Peter J. Bickel[1]
Department of Statistics
University of California
Berkeley, CA 94720

Jianqing Fan
Department of Statistics
University of North Carolina
Chapel Hill, N.C. 27514

Abstract

The kernel and ordinary spline methods are not effective for capturing the shapes of unknown densities. To overcome this discrepancy, we study the problem of estimating densities under shape restrictions. We propose several methods for estimating unimodal densities: plug-in MLE, pregrouping techniques, linear spline MLE, and spline smoothing. Combining these techniques, an automatic spline procedure for estimating a unimodal density is proposed, which is justified to have a good performance. The asymptotic distributions of proposed estimators are also found. An important consequence of our study is that having to estimate the location of a mode does not affect the limiting behavior of our unimodal density estimate.

[1]Supported by ONR N00014-89-J-1563

Abbreviated title. Restricted density estimation.

AMS 1980 subject classification. Primary 62G05. Secondary 62E20, 60G30.

Key words and phrases. Asymptotic distributions, density estimation, minimum χ^2-estimates, MLE, modes, plug-in methods, pregrouping methods, splines, 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 (Parzen (1962)). The behavior of a kernel density estimate relies strongly on the choice of smoothing parameter (bandwidth). Data-driven bandwidth selection methods (Härdle *et al.* (1988), Rice (1984), among others) 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 shapes and areas 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 two curves are quite different. Thus, the kernel method can produce quite noisy a picture so that the overall structure of the data is hard to examine. To overcome this discrepancy, we propose estimating a global density under shape restrictions. As a start, we focus on the problem of estimating a unimodal density with an unknown mode location. Our ultimate goal is to develop a method to produce good pictures for estimating multimodal densities and to examine similarly other methods of estimation under order restrictions.

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 resulting estimate is the left slope of the least concave majorant of the empirical distribution function. The asymptotic distribution of the MLE was found by Prakasa Rao (1969), and Groeneboom (1985). Recent developments in estimating a monotone density can be found in Birgé (1987a,b), which give the behavior of nonparametric minimax risks. Wegman (1972 a, b) extends the problem to estimating a unimodal density by using an MLE method and proves the consistency of the MLE. Kiefer (1981) gave an illuminating survey of developments in the theory of estimating a monotonic function (e.g., density function, failure rate function). Of course, there is a long history of isotonic regression and its related problems. See Barlow

et al. (1972) and Robertson *et al.* (1988), Barlow and van Zwet (1970), Ramsay (1988), Kelly and Rice (1990), among others.

To estimate a unimodal density, we first begin by introducing a plug-in maximum likelihood method. We use this to understand how the location of the mode affects the estimate. An important result of section 2 is that slight misspecification of the location of a mode does not affect tremendously the behavior of our estimate; it misestimates the unknown density only on a tiny interval near the true mode. Practically, it means 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 fact makes it possible to estimate a multimodal density by using a plug-in method. The statement is further justified by considering a related problem in the population context: minimize the Kullback-Leibler discriminant information with the mode location misspecified.

Let's denote by $\hat{f}_n(x;m)$ 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 of mode m_0. Then, the plug-in version of the estimate is $\hat{f}_n(x;\hat{m})$. We show in section 2 that for *all* consistent estimate $\hat{m}$, $\hat{f}_n(x;\hat{m})$ converges at *the same rate* $n^{-1/3}$ *with the same asymptotic distribution.*

One can view the estimate $\hat{f}_n(x;m)$ as a histogram estimate with distinct bin width determined automatically by the data. We observe that the resulting picture of $\hat{f}_n(x;\hat{m})$ is quite spiky near the location $\hat{m}$. We introduce in section 3 a pregrouping technique to solve this peaking problem which reduces the computing costs as well, without affecting the behavior of the estimate $\hat{f}_n(x;\hat{m})$. The idea is to group the data into a number of groups first, and then to perform a form of minimum χ^2-estimate. We prove that if the pregrouping is not too crude, the pregrouping version of the MLE does as well as the plug-in MLE.

The discontinuity of the plug-in MLE is unsatisfactory. To deal with this problem, we introduce in section 4 a maximum likelihood linear spline estimate. We give explicitly the form of the estimate and its asymptotic distribution for the case where the mode location

is known. Since we demonstrate in section 2 that not knowing the location of mode is not a serious matter in estimating a unimodal density, we expect but have not yet shown that such an estimate should also work well when we don't know the location of mode. A nice feature of such an estimate is that the location of the mode as well as the number and locations of knots are determined *automatically* by data. Again, the pregrouping technique can be used to solve the peaking problem and to save the cost of computation.

Various procedures for practical applications are discussed in section 5. These include using spline methods to produce *smooth* pictures of the plug-in MLE, and determining the mode location *automatically* for the plug-in MLE.

Finally, in section 6, we further justify our theory and heuristics by statistical simulation. All estimates we propose here are fast to compute. Technical proofs are given in section 7.

2 Plug-in maximum likelihood estimate

In this section, we study problems of estimating a unimodal density by using a plug-in maximum likelihood estimate. We will derive the asymptotic distribution of the estimator, and demonstrate that for *any consistent* estimate of the location of the mode, the plug-in MLE behaves *the same* as when the location of mode is *known*. The result is explained by considering a problem of minimizing the Kullback-Leibler discriminant information for densities.

2.1 Plug-in MLE

Let $f(x; m)$ be a unimodal density with mode location parameterized by m. Suppose that $X_1', \cdots, X_n'$ are i.i.d. from $f(x; m_0)$, where m_0 is the true location of the mode. If m_0 is known, then an MLE of $f(x; m_0)$ is to find a density $\hat{f}_n(x; m_0)$, which maximizes the likelihood function among the class of unimodal densities with location of mode at m_0:

$$\max_{f \in \mathcal{F}_{m_0}} \prod_{i=1}^{n} f(X_i'), \tag{2.1}$$

where

$$\mathcal{F}_m = \{f : f \text{ is a unimodal density with mode location } m\}. \tag{2.2}$$

It has been proved (see Grander (1956), Robertson *et al* (1988)) that the solution to the problem (2.1) is 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.

Let $X_1, \cdots, X_n$ be the order statistics of the sample. Suppose a is such that $X_a < m_0 < X_{a+1}$. Denote $y_j = X_j$ for $j = 1, \cdots, a$, and $y_{a+1} = m_0$, $y_{j+1} = X_j$ for $j = a+1, \cdots, n$. Then, the explicit formula for the MLE is

$$\hat{f}_n(x; m_0) = \begin{cases} 0, & \text{if } x < y_1 \text{ or } x > y_{n+1} \\ f_j & \text{if } y_j \le x < y_{j+1}, j = 1, \cdots, a \\ f_j & \text{if } y_j < x \le y_{j+1}, j = a+1, \cdots, n \end{cases}, \tag{2.3}$$

where f_j is defined by

$$f_j = \begin{cases} \min_{a+1 \ge t > j} \max_{s \le j} \frac{t-s}{n(y_t - y_s)} & \text{if } j \le a \\ \min_{a+1 \le s \le j} \max_{t > j} \frac{t-s}{n(y_t - y_s)} & \text{if } j > a \end{cases}. \tag{2.4}$$

Now, the true mode is typically unknown. Let $\hat{m}$ be a consistent estimator of m_0 (e.g. by the kernel method (Parzen (1962), Eddy (1980)) or greatest "clustering" method (Chernoff (1964), Venter (1967)). Then, we use the estimator $\hat{f}_n(x; \hat{m})$ as an estimate of the unknown density $f(x, m_0)$. We call such an estimate the plug-in MLE.

2.2 Asymptotic distribution of the plug-in MLE

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

$$n^{1/3} \left| \frac{1}{2} f(x; m_0) f'(x; m_0) \right|^{-\frac{1}{3}} (\hat{f}_n(x; \hat{m}) - f(x; m_0)) \xrightarrow{\mathcal{L}} 2Z, \tag{2.5}$$

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 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 $\hat{f}_n(x; \hat{m})$ behaves asymptotically the same as $\hat{f}_n(x; m_0)$. More precisely, in addition to (2.5), one also has

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

We conjecture that

$$\hat{f}_n(x; \hat{m}) - \hat{f}_n(x; m_0) = o_p\left(n^{-1/3}\right)$$

uniformly for $|x - m_0| \geq \varepsilon$.

We need Lemma 1 and Lemma 2 to prove Theorem 1.

Lemma 1. *If $x > m_1 \geq m_2$, then*

$$\hat{f}_n(x; m_1) \geq \hat{f}_n(x; m_2).$$

Proof. The intuition of the proof follows from the " pool-adjacent-violator" algorithm (see Robertson *et al.* (1988), page 8–10). Note that by (2.3) and (2.4), $\hat{f}_n(x; m_1)$ is computed by minimizing over a smaller set than $\hat{f}_n(x; m_2)$, and the spans $y_t - y_s$ for computing $\hat{f}_n(x; m_1)$ are no larger than those for computing $\hat{f}_n(x; m_2)$ over the range $t \geq a_{m_1}$, where a_{m_1} satisfies $X_{a_{m_1}} < m_1 < X_{a_{m_1}+1}$. Thus, the conclusion follows directly from (2.3) and (2.4).

2.3 A result on minimum Kullback-Leibler discriminant information

At a conceptual level, if we misspecify the location of the mode to be m, the MLE $\hat{f}_n(x; m)$ estimates the maximizer of the following problem

$$\max_{g \in \mathcal{F}_m} \int_{-\infty}^{+\infty} [\log g(x)] f(x; m_0) dx, \tag{2.6}$$

where $f(x; m_0)$ is the true underlying density, and $\mathcal{F}_m$ is defined by (2.2). The problem is equivalent to finding the solution of minimum Kullback-Leibler discriminant information:

$$\min_{g \in \mathcal{F}_m} \int_{-\infty}^{+\infty} \log[f(x; m_0)/g(x)] f(x; m_0) dx. \tag{2.7}$$

In this section, we find the explicit solution to the problem (2.6). From the solution (2.11), we see that misspecification of the mode location does not affect the MLE dramatically. It misestimates *only the density near the location of the mode.* This explains why the result of Theorem 1 holds for *any* consistent estimate of m_0.

To find the solution, let's assume that $m < m_0$. The other case can be treated similarly. The following lemma is also a crucial lemma for the proof of Theorem 1.

Lemma 2. *Suppose that $f(x)$ is a continuous unimodal density on $[m, \infty)$ with mode location m_0. Let $M \geq m_0$ be the smallest solution of the equation*

$$\int_m^M f(x)dx = f(M)(M - m). \tag{2.8}$$

Define

$$f^*(x) = \begin{cases} f(M) & \text{if } x \leq M \\ f(x) & \text{if } x > M \end{cases}. \tag{2.9}$$

Then $f^*(x)$ achieves

$$\max_{g \text{ is a decreasing density in } [m,\infty)} \int_m^\infty \log g(x) f(x) dx. \tag{2.10}$$

Theorem 2. *Suppose that $f(x; m_0)$ is a continuous unimodal density with mode location m_0, and $m < m_0$. Then the solution to the problem (2.7) is given by*

$$g(x; m) = \begin{cases} f(x; m_0) & \text{if } x \leq m \text{ or } x > M \\ f(M; m_0) & \text{if } m \leq x \leq M \end{cases}, \tag{2.11}$$

where $M \geq m_0$ is defined by

$$\int_m^M f(x; m_0)dx = f(M; m_0)(M - m). \tag{2.12}$$

Remark 2. From the solution (2.11), it is easy to show that

$$\sup_x |g(x;m) - f(x;m_0)| = O((m-m_0)^2), \tag{2.13}$$

provided the function $f(x;m_0)$ has a bounded second derivative near the mode m_0. Moreover, if $f(x;m_0)$ is strictly unimodal (strictly increasing on the one side of m_0, and strictly decreasing on the other side), then $M \to m_0$, as $m \to m_0$. Thus, if m is close to m_0, then $g(x;m) = f(x;m_0)$ except on the tiny interval $[m, M]$. Therefore, the MLE with the location misspecified to be m estimates the right value of the density $f(x;m_0)$ except on the small interval $[m, M]$.

3 Pregrouping Techniques

It is known (see Figures 1.1, 2.3, 2.5, 3.1, 3.3) that the MLE for estimating a unimodal density appears to be spiky near the mode. In this section, we introduce a pregrouping technique, which is used to solve the peaking problem of the MLE and to reduce the computing cost of the estimate.

The idea is to group the data first, and then apply the plug-in technique. To be more specific, let $\{I_j = (-t_j, t_{j+1}], j = 0, \pm 1, \pm 2, \cdots\}$ 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

$$F_n^*(x) = \frac{1}{n}(\#\text{ of } X_i's \le t_{j+1}), \text{ when } x \in (t_j, t_{j+1}]. \tag{3.1}$$

Let $\hat{f}_n^*(x;m)$ be the left derivative of the least concave majorant of $F_n^*(x)$, when $x > m$, and when $x < m$, the right derivative of the greatest convex minorant of $F_n^*(x)$. 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 $\hat{f}_n^*(x;m)$ 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}.

Note that $\hat{f}_n^*(x;m)$ is a solution of a form of minimum χ^2-discrepancy described as follows. Let $m \in (t_a, t_{a+1})$, and g be a unimodal density which is piecewise constant on intervals $[t_{j-1}, t_j)$, for $j \leq a$, $[t_a, m), (m, t_{a+1}]$, and $(t_j, t_{j+1}]$, for $j \geq a+1$. Write g_j as the height of the density g in these intervals. Let's denote n_j by the number of observations falling in $(t_{j-1}, t_j]$. Then, $\hat{f}_n^*(x;m)$ minimizes the χ^2-discrepancy:

$$\sum_j \left(g_j - \frac{n_j}{n w_j} \right)^2 w_j, \tag{3.2}$$

among all unimodal densities g just described, where w_j is the length of the above interval:

$$w_j = t_j - t_{j-1}; j < a, w_a = m - t_a, w_{a+1} = t_{a+1} - m, w_j = t_j - t_{j-1}; j > a+1.$$

Now, let's prove that by choosing the maximum span of the partitions of order $o(n^{-1/2})$, the estimate $\hat{f}_n^*(x;\hat{m})$ behaves the same as $\hat{f}_n(x,\hat{m})$.

We need the following two lemmas to prove Theorem 3. Lemma 3 tell us that the modified empirical distribution defined by (3.1) behaves almost the same as the empirical distribution, when the maximum span of the partitions is of order $o(n^{-1/2})$.

Lemma 3. *Let $X_1', \cdots, X_n'$ be i.i.d with a density $f(x)$. If f is bounded,*

$$\max_j |t_{j+1} - t_j| = o(n^{-1/2}), \tag{3.3}$$

then

$$\sup_x |\hat{F}_n(x) - F_n^*(x)| = o_p(n^{-1/2}), \tag{3.4}$$

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

Now, we prove for estimating a decreasing density that the pregrouping version of the MLE behaves asymptotically the same as the MLE.

Lemma 4. *Let $X_1', \cdots, X_n'$ be independent observations generated by a decreasing density f on $[0, \infty)$, which has a nonzero derivative $f'(t)$ at a point $t \in (0, \infty)$. If $\hat{f}_n^*(x)$ is the left derivative of the least concave majorant of $F_n^*(x)$, then*

$$n^{1/3} \left| \frac{1}{2} f(t) f'(t) \right|^{-1/3} (\hat{f}_n^*(t) - f(t)) \xrightarrow{\mathcal{L}} 2Z, \tag{3.5}$$

where the random variable Z was defined in Theorem 1.

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 the condition (3.3) holds, then the conclusion of Theorem 1 follows with replacing $\hat{f}_n(x; \hat{m})$ by $\hat{f}_n^*(x; \hat{m})$.*

In practice, we can take the partition $\{t_j\}$ to be equally spaced grid points with span l_n. Theorem 3 shows that if $l_n = o(n^{-1/2})$ (i.e., the partition is not too crude to lose the detail of the data), $\hat{f}_n^*(x; \hat{m})$ has the same performance as the plug-in MLE $\hat{f}_n(x; \hat{m})$.

In simulations below, we always use partitions $\{t_j = \hat{m} + jl_n\}$ and group data away from the estimated mode: n_{j+1} repetitions at the grid point t_j when $j \leq -1$ and n_j repetitions at point t_j, when $j \geq 1$, where n_j is the number of observations falling in the interval $(t_{j-1}, t_j]$. Grouping in this way is more effective for solving peaking problems (compare Figures of odd numbers with those of even numbers).

Let's emphasize that the choice of l_n is driven to save computation and the spiking phenomenon of the plug-in MLE near the estimated mode. Usually, data is highly clustered near the mode so that the denominator of (2.4) is very small, which implies that the MLE near mode is unusually large. By using the pregrouping technique, the denominator of (2.4) would be more stable. Hence, the peaking problem can be solved. Note that the computational complexity of $\hat{f}_n(x; \hat{m})$ is $O(n \log n)$. If we take l_n close to $n^{-1/2}$, we would expect to group n data into $O(n^{1/2})$ groups. Thus, the computational complexity of the pregrouping version $\hat{f}_n^*(x; \hat{m})$ is only about $O(n^{1/2} \log n)$.

The strength of the pregrouping technique will become more clear in the next section, where we find a unimodal linear spline MLE estimate, which has computational complexity $O(n^2 \log n)$. The pregrouping technique will reduce the complexity to $O(n \log n)$.

4 Linear Spline MLE

The MLE does not produce nice pictures, because it produces *discontinuous* random bin width histograms. A reasonable next step is to find the MLE among linear spline density

estimates under the shape restrictions. An advantage over the usual spline method (see, e.g. Kelly and Rice (1990)) is that the number and the locations of knots of our approach are determined *automatically* by data.

In this section, we find the explicit form of the linear spline MLE, and its rate of convergence.

4.1 Solution to the problem

Let $X_1', \cdots, X_n'$ be i.i.d. with an unknown unimodal density f and denote $X_1, \cdots, X_n$ for the order statistics of the sample. Let

$$\begin{aligned} \mathcal{F}_L &= \{f : f \text{ is a continuous linear spline unimodal density} \\ &\qquad \text{on } [X_1, X_n] \text{ with knots at the data points}\}. \end{aligned} \tag{4.1}$$

We want to find a solution to the problem:

$$\arg \max_{f \in \mathcal{F}_L} \prod_{j=1}^{n} f(X_j'). \tag{4.2}$$

The solution to problem (4.2) can be computed explicitly by isotonic regression techniques (see Theorem 4). Let

$$\hat{f}_{aj} = \begin{cases} \min_{a+1 \le t > j} \max_{s \le j} \frac{t-s}{n[(X_t + X_{t-1})/2 - (X_s + X_{s-1})/2]}, & \text{when } j < a \\ \min_{a \le s \le j} \max_{t > j} \frac{t-s}{n[(X_t + X_{t-1})/2 - (X_s + X_{s-1})/2]}, & \text{when } j > a \\ \max \left\{ \max_{s \le a} \frac{a-s+1}{n[(X_{a+1} + X_a)/2 - (X_s + X_{s-1})/2]}, \right. & \\ \quad \left. \max_{t > a} \frac{t-a}{n[(X_t + X_{t-1})/2 - (X_a + X_{a-1})/2]} \right\}, & \text{when } j = a \end{cases}, \tag{4.3}$$

Here and hereafter we set $X_0 = X_1$, 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 $[X_1, X_n]$. Then, it will be shown in the proof of Theorem 4 that $\hat{f}_{nL}(x, a)$ is a density in $\mathcal{F}_L$ 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 $\hat{f}_{nL}(x, a)$, $a = 1, \cdots, n$. Then, we have the following result.

Theorem 4. *The solution to problem (4.2) is given by* $\hat{f}_{nL}(x;\hat{a})$.

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

$$F_n^*(x) = \frac{1}{n}\sum_{j=1}^{n+1} I_{\{z_j \le x\}}, \tag{4.4}$$

where I_A is the indicator of a set A, and $z_j = (X_j + X_{j-1})/2$. Let $\hat{f}_a^*(x)$ be the left derivative of the least concave majorant of $F_n^*(x)$ when $x > z_a$. Then by comparing (4.3) with (2.4), we have for $j > a$

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

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 to remedy the discontinuity of $\hat{f}_a^*(x)$. This identity gives a simple way of computing $\hat{f}_{aj}$ by using the "pool-adjacent-violators" algorithm. Consequently,

$$\sup_{x > z_{a+1}} \left|\hat{f}_{nL}(x;a) - \hat{f}_a^*(x)\right| \le \frac{1}{2}\max_{j \ge a}\left|\hat{f}_a^*(z_j) - \hat{f}_a^*(z_{j+1})\right|. \tag{4.5}$$

For the other case, we have a similar equation.

Remark 3. We could use the "pool-adjacent-violators" algorithm (see page 9 of Robertson *et al.* (1988)) to compute $\hat{f}_{aj}$ defined by (4.3), and then search for the best index $\hat{a}$ among n possible values of a. The computational complexity is about $O(n^2 \log n)$. Note that for such a MLE estimate, the location of the mode is automatically determined by the data. If it is too expensive to search among all possible a, one could adopt the plug-in idea given in section 2, or uses the pregrouping techniques. For estimating a decreasing density, we can simply take $\hat{a} = 1$.

4.2 Pregrouping version

Let's mention briefly pregrouping version of the linear spline MLE. The problem is exactly as in section 4, except using grouped data: casting observations in the interval $(t_j, t_{j+1}]$ to the point t_{j+1} (compare with section 3).

For simplicity, let's denote the grouped data by $(n_1, T_1), \ldots, (n_{n^*}, T_{n^*})$, which means, for example, n_1 repetitions of data at point T_1, where $T_1 < \cdots < T_{n^*}$.

Let's define (compare with (4.3))

$$\hat{f}^*_{aj} = \begin{cases} \min_{a+1 \geq t > j} \max_{s \leq j} \frac{\sum_{j=s}^{t-1} n_j}{n[(T_t + T_{t-1})/2 - (T_s + T_{s-1})/2]}, & \text{when } j < a \\ \min_{a \leq s \leq j} \max_{t > j} \frac{\sum_{j=s}^{t-1} n_j}{n[(T_t + T_{t-1})/2 - (T_s + T_{s-1})/2]}, & \text{when } j > a \\ \max \left\{ \max_{s \leq a} \frac{\sum_{j=s}^{a} n_j}{n[(T_{a+1} + T_a)/2 - (T_s + T_{s-1})/2]}, \right. & \\ \left. \quad \max_{t > a} \frac{\sum_{j=a}^{t-1} n_j}{n[(T_t + T_{t-1})/2 - (T_a + T_{a-1})/2]} \right\}, & \text{when } j = a \end{cases}, \tag{4.6}$$

where $T_0 = T_1$, and $T_{n^*+1} = T_{n^*}$. Let $\hat{f}^*_{nL}(x; a)$ be the function connecting points $(T_j, \hat{f}^*_{aj})$ by using lines, and 0 when x is out of the range $[T_1, T_{n^*}]$. Then, it is a continuous linear spline density on $[T_1, T_{n^*}]$, which is unimodal with mode location T_a. Let $\hat{f}^*_{nL}(x; \hat{a}^*)$ be the maximizer of the likelihood function among the n^* possible choices of densities $\hat{f}^*_{nL}(x, a)$, $a = 1, \cdots, n^*$. The density function $\hat{f}^*_{nL}(x; \hat{a}^*)$ is the desired pregrouping version linear spline MLE.

4.3 Asymptotic distributions

We only prove $\hat{f}_{nL}(x, 1)$ converges to the true density for estimating a decreasing density. For the unimodal case, we would expect that a similar result holds.

Theorem 5. *Suppose that $X'_1, \cdots, X'_n$ 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*

$$n^{1/3} \left| \frac{1}{2} f(x) f'(x) \right|^{-1/3} (\hat{f}_{nL}(x, 1) - f(x)) \xrightarrow{\mathcal{L}} 2Z, \tag{4.7}$$

where the random variable Z was defined in Theorem 1.

Thus, the linear spline shares the nice properties of MLE estimate defined in section 2.

5 Discussion

We have proposed the plug-in method to estimate unimodal densities, the pregrouping technique to solve peaking problems and to reduce computational cost, and the linear spline

approach to produce continuous pictures. Some important issues in applications include how to use higher order spline method to produce smooth curves, and how to estimate mode locations.

In the following discussion, we assume the pregrouping techniques have been applied so that the resulting data have a structure (n_1, T_1), ..., (n_{n^*}, T_{n^*}). (See section 4.2 for the exact meaning of our notations). Obviously, the discussion is also applicable to the original data with appropriate change of notations.

5.1 Higher order splines

To produce smooth pictures of the curve estimate, there are two possible approaches. The first approach is to smooth the plug-in MLE by using spline. More precisely, let's assume the plug-in MLE $\hat{f}_n^*(x; \hat{m})$ is of form: constants $\hat{f}_j$ ($\hat{f}_j \neq \hat{f}_{j+1}$) on interval (u_j, u_{j+1}) $j = 1, \ldots, n'$. Denote the midpoints of the intervals by

$$v_j = \frac{u_{j+1} + u_j}{2}, j = 1, \ldots, n' - 1.$$

Then, use a spline (e.g. cubic spline) curve to interpolate the $n' - 1$ points:

$$(v_j, \hat{f}_j), j = 1, \ldots, n' - 1.$$

Use the resulting curve as an estimate of the underlying density. The drawback of this approach is that the resulting curve is not necessary unimodal.

To overcome the drawback, we propose the following spline method. Let's use quadratic spline to interpolate a unimodal density to illustrate the idea. Let $m = \arg max_j \hat{f}_j$ be the index of the location of mode. First, for $j = 1, \cdots, m-1$, we use an increasing spline to interpolate the points $(v_j, \hat{f}_j)$: at the kth step, determine the quadratic spline which passes the points $(v_k, \hat{f}_k)$, and $(v_{k+1}, \hat{f}_{k+1})$, and which matches the first derivative at the point (v_k, f_k); if the quadratic spline is increasing at the interval $[v_j, v_{j+1}]$, use this quadratic spline; otherwise, use a linear spline that passes the two endpoints. Then, determine a quadratic spline to pass the three points near the mode: $(v_{m-1}, \hat{f}_{m-1})$, $(v_m, \hat{f}_m)$, and $(v_{m+1}, \hat{f}_{m+1})$. Finally,

use a similar criteria for $j = m+2, ..., n'-1$. In the simulations below, we always refer the spline interpolation to this approach with an initial by a linear spline connecting the points $(v_1, \hat{f}_1), (v_2, \hat{f}_2)$.

The second approach to smoothing the plugin MLE is using least square method to fit a unimodal density, by locating knots at points $v_1, \cdots, v_{n'-1}$. The discussion on this issue is beyond the intent of this paper.

5.2 Estimating the location of mode

To use our plug-in method, one has to estimate mode location. Various approaches have been proposed in the literature, which involve determining window size by users. Here, we propose a way of determining the mode location automatically.

Let's start with the grouped data: $(n_1, T_1), \ldots (n_{n^*}, T_{n^*})$. Using the data set

$$\{(n_j, T_j), j \neq a\},$$

one finds the MLE by with mode location T_a. Let's denote the resulting estimate by $\hat{f}_a(x)$. Choose $\hat{a}$ to maximize the "likelihood"

$$\prod_{j=1}^{n^*} \hat{f}_a^{n_j}(T_j).$$

Finally, use $\hat{f}_{\hat{a}}(x)$ to estimate the underlying density. Note that the mode location is estimated by $T_{\hat{a}}$. We will refer this technique to the "automatic selection of mode" in the next section.

6 Simulations

We generate data from 3 different distributions to demonstrate how our approaches work. The distributions are exponential, normal, and an asymmetric unimodal distribution (see (6.1)). The sample sizes are taken 200, 200 and 150, respectively, as we think they are rather small for nonparametric curve estimation.

Exponential Distribution

A random sample of size 200 is simulated from exponential distribution:

$$f(x) = \exp(-x) I_{\{x \geq 0\}}.$$

The simulated data is plotted at the bottom of Figure 1.1. The maximum likelihood estimate and its spline interpolation (see the first approach in section 5.1) are plotted in Figure 1.1. The peaking problem is visible. The problem is solved by using pregrouping technique (see Figure 1.2). We group the data into 25 groups with breaking points $t_j = 0.2 \times j$. The pregrouping version of MLE and its spline interpolation are plotted in Figure 1.2. It appears that the pregrouping version of spline interpolation estimate is very close to the truth. Figure 1.3 shows the linear spline MLE and its pregrouping version (see section 4). Figure 1.4 plots the default histogram method and kernel density estimate method in a statistical package SPLUS, for the purpose of comparison. It appears that our approach (spline interpolation in Figure 1.2) is much better.

Normal Distribution

A random sample of size 200 is simulated from Normal(0,1):

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

The simulated data is plotted at the bottom of Figure 2.1. The plugin MLE and its pregrouping version (partition at grid points $t_j = 0.25j$) with estimated mode 0 are plotted in Figure 2.1, and Figure 2.2 respectively. If mode location is estimated by the average of the data, which is 0.17, then the corresponding plugin MLE and its pregrouping version (with partition $t_j = 0.17 + j \times 0.25$) are plotted in Figure 2.3 and 2.4, respectively. Finally, we use the automatic procedure (see section 5.2) to determine the mode location. The mode location is estimated by 0.23, and the density estimated is plotted in Figure 2.5. Grouping data with partition at grid points $t_j = 0.2j$ first and then applying the automatic procedure in section 5.2, we estimate the mode location by 0.2, and the corresponding density estimate is plotted in Figure 2.6. Note that there are only about 100 observations used to estimate the monotonic pieces of the curve.

An Asymmetric Distribution

150 random samples are simulated from density

$$f(x) = \begin{cases} \frac{3}{8}\exp(-x) & x \geq 0 \\ \frac{3}{8}\exp(0.6x) & x < 0 \end{cases} . \tag{6.1}$$

The density and the simulated data are plotted in Figure 3.1. Roughly speaking, there are only about 56 observations in estimating the right piece of decreasing density and about 94 observations in estimating the left piece of increasing density. The estimated mode location -0.04 is used in the plugin MLE. To group the data, we use grid points $t_j = -0.04 + 0.15j$. The plugin MLE and its pregrouping version are plotted in Figure 3.1 and Figure 3.2. The estimates by using the automatic procedure suggested in section 5.2 are plotted in Figure 3.3 and Figure 3.4. The mode location estimated by the automatic procedure without pregrouping is 0.21, and with pregrouping is 0.11.

Conclusions

The above simulation suggests that pregrouping is a powerful technique for solving the peaking problem and reducing computational cost. Automatic procedures for selecting the location of the mode appear very good. For the normal model, they are almost as good as parametric estimation. Our simulation also suggests that if the partition is not too crude, the estimates are not very sensitive to the partitions.

7 Proofs

7.1 Proof of Theorem 1

We give the proof for $x > m_0$; the other case can be treated similarly.

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

$$\begin{aligned} & P\left\{n^{\frac{1}{3}}l(x)(\hat{f}_n(x;\hat{m}) - f(x;m_0)) \leq t\right\} \\ = \; & P\left\{n^{\frac{1}{3}}l(x)(\hat{f}_n(x;\hat{m}) - f(x;m_0)) \leq t, |\hat{m} - m_0| \leq \varepsilon\right\} + o(1). \end{aligned} \tag{7.1}$$

Let $\hat{f}_N^*(x)$ be the MLE for a sample of size N from the monotone decreasing density

$$\frac{f(x, m_0+\varepsilon)}{1-F(m_0+\varepsilon)}, \quad x \geq m_0+\varepsilon.$$

Then, we can represent

$$\hat{f}_n(x, m_0+\varepsilon) = \left(1-\hat{F}_n(m_0+\varepsilon)\right) \hat{f}_N^*(x), \forall x > m_0+\varepsilon,$$

where $N = n[1-\hat{F}_n(m_0+\varepsilon)]$, $\hat{F}_n$ is the empirical distribution. By Prakasa Rao (1969), and Groeneboom (1985),

$$P\left\{N^{\frac{1}{3}}\left|\frac{f(x;m_0)f'(x;m_0)}{2[1-F(m_0+\varepsilon)]^2}\right|^{-\frac{1}{3}}\left(\hat{f}_N^*(x)-\frac{f(x;m_0)}{1-F(m_0+\varepsilon)}\right) \leq t\right\} \longrightarrow P\{2Z \leq t\}.$$

Equivalently,

$$P\left\{n^{\frac{1}{3}}\left|\frac{f(x;m_0)f'(x;m_0)[1-\hat{F}_n(m_0+\varepsilon)]^2}{2[1-F(m_0+\varepsilon)]^2}\right|^{-\frac{1}{3}}\right.$$
$$\left.\times\left(\hat{f}_n(x;m_0+\varepsilon)-f(x;m_0)\frac{1-\hat{F}_n(x+\varepsilon)}{1-F(m_0+\varepsilon)}\right) \leq t\right\} \longrightarrow P\{2Z \leq t\}. \qquad (7.2)$$

It follows from (7.2) that

$$P\left\{n^{\frac{1}{3}}\left|\frac{1}{2}f(x;m_0)f'(x;m_0)\right|^{-\frac{1}{3}}(\hat{f}_n(x;m_0+\varepsilon)-f(x;m_0)) \leq t\right\}$$
$$\longrightarrow P\{2Z \leq t\}, \qquad \forall t \in (-\infty, +\infty). \qquad (7.3)$$

Thus, it follows from Lemma 1, (7.1) and (7.3) that

$$\begin{aligned} & \liminf P\left\{n^{\frac{1}{3}}l(x)(\hat{f}_n(x;\hat{m})-f(x;m_0)) \leq t\right\} \\ \geq\ & \liminf P\left\{n^{\frac{1}{3}}l(x)(\hat{f}_n(x;m_0+\varepsilon)-f(x;m_0)) \leq t\right\} \\ =\ & P\{2Z \leq t\}. \end{aligned} \qquad (7.4)$$

Similarly, by (7.1) and Lemma 1, we have

$$\begin{aligned} & \limsup P\left\{n^{\frac{1}{3}}l(x)(\hat{f}_n(x;\hat{m})-f(x;m_0)) \leq t\right\} \\ \leq\ & \limsup P\left\{n^{\frac{1}{3}}l(x)(\hat{f}_n(x;m_0-\varepsilon)-f(x;m_0)) \leq t\right\} \end{aligned} \qquad (7.5)$$

The proof is completed, if we show that (7.5) has a limit (7.4). Let $f_\varepsilon^*(\cdot)$ be the solution to the problem:

$$\max_{g(\cdot) \text{ is a decreasing density on } [m_0-\varepsilon,\infty)} \int_{m_0-\varepsilon}^{\infty} [\log g(y)] \frac{f(y;m_0)}{1-F(m_0-\varepsilon)} dy.$$

Then, the solution is given explicitly by (2.9) with an appropriate change of notations. Note that by (2.9) for each fixed x, there exists an ε_x (independent of n) such that

$$f_{\varepsilon_x}^*(x) = \frac{f(x;m_0)}{1-F(m_0-\varepsilon)}.$$

Now, by the argument of Groeneboom (1985), one can show that

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

where $\hat{f}_{N_1}^{**}(\cdot)$ is the MLE over the class of decreasing densities on $[m_0-\varepsilon,\infty)$ based on data $X_j \ge m_0-\varepsilon$, 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 $\varepsilon \equiv \varepsilon_x$,

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

The conclusion follows from (7.5).

7.2 Proof of Theorem 2

Proof of Lemma 2. The idea of the proof is first to establish

$$\int_m^\infty [\log g(x)] f(x) dx \le \int_m^\infty [\log g(x)] f^*(x) dx, \tag{7.6}$$

where $g(x)$ is a decreasing density function on $[m,\infty)$. If (7.6) holds, then by (2.8)

$$\begin{aligned}\int_m^\infty \log g(x) f(x) dx &\le \int_m^\infty \log f^*(x) f^*(x) dx \\ &= \int_m^\infty \log f^*(x) f(x) dx,\end{aligned}$$

and consequently $f^*(\cdot)$ is a solution to the problem (2.10). Now, let's turn to prove (7.6). The proof uses an idea of Fan (1986). Let

$$a = \inf\{x : f(x) = f(M)\} \leq m_0.$$

Then, for any decreasing function $g(x)$,

$$\log g(x)[f(x) - f^*(x)] \geq \log g(a)[f(x) - f^*(x)]. \tag{7.7}$$

To see this, note that when $x \geq a$, $f(x) - f^*(x) \geq 0$. Thus, by the monotonicity of $g(x)$, (7.7) holds. Similarly, when $x < a$, $f(x) - f^*(x) \leq 0$, and (7.7) follows from the monotonicity of $\log g$ again. Integrating over the both sides of (7.7) from m to ∞, we obtain (7.6).

Proof of Theorem 2. For a unimodal density $g(x)$ with the location of mode m, write

$$g(x) = \alpha_g g_1(x) + (1 - \alpha_g) g_2(x). \tag{7.8}$$

where

$$\alpha_g = \int_{-\infty}^{m} g(x)dx, \quad g_1(x) = g(x)I_{(-\infty,m)}(x)/\alpha_g,$$

and $g_2(x)$ is defined similarly. Note that $g_1(x)$ is a non-decreasing density and $g_2(x)$ is a non-increasing density. Let $\beta = \int_{-\infty}^{m} f(x; m_0)dx$. It follows that

$$\begin{aligned} & \int \log g(x) f(x; m_0)dx \\ = \; & \beta \log \alpha_g + (1 - \beta)\log(1 - \alpha_g) + \beta \int_{-\infty}^{m} \log g_1(x) f(x; m_0)/\beta dx \\ & + (1 - \beta) \int_{m}^{\infty} \log g_2(x) f(x; m_0)/(1 - \beta)dx. \end{aligned} \tag{7.9}$$

Then the maximizers of (7.9) are give by

$$\alpha_g = \beta, g_1(x) = f(x; m_0)/\beta I_{(-\infty,m)}(x), \tag{7.10}$$

and $g_2(x)$ is defined by (2.9) replacing $f(x)$ by $f(x; m_0)/(1 - \beta)I_{[m,\infty)}(x)$. This completes the proof.

7.3 Proof of Theorem 3

Proof of Lemma 3. Let n_j be the number of observations in $(t_j, t_{j+1}]$. Then,

$$0 \leq F_n^*(x) - \hat{F}_n(x) \leq \max n_j/n.$$

Thus, for any $\varepsilon > 0$,

$$P\left\{\sqrt{n}\sup_x |\hat{F}_n(x) - F_n^*(x)| > \varepsilon\right\} \leq \sum_j P\left\{n_j > \sqrt{n}\varepsilon\right\}. \tag{7.11}$$

Note that the random variable n_j is distributed as Binomial(n, p_{nj}) with $p_{nj} = F(t_{j+1}) - F(t_j)$, where F is the cdf of the random variable X_1. Denote $p_n = \max_j p_{nj}$. Then,

$$p_n \leq [\sup f(x)] \max_j (t_{j+1} - t_j) = o(n^{-1/2}). \tag{7.12}$$

Now, we are going to prove that when n is large enough,

$$P\left\{n_j > \sqrt{n}\varepsilon\right\} \leq B p_{nj}^2. \tag{7.13}$$

If (7.13) holds, then the conclusion follows from (7.11), by the fact that

$$P\left\{\sqrt{n}\sup_x |\hat{F}_n(x) - F_n^*(x)| > \varepsilon\right\} \leq B\sum_j p_{nj}^2 \leq B p_n \longrightarrow 0.$$

Note that for any $c > 0$,

$$\begin{aligned} P\{n_j > \sqrt{n}\varepsilon\} &= P\left\{e^{n_j c} > e^{\sqrt{n}\varepsilon c}\right\} \\ &\leq \exp(-\sqrt{n}\varepsilon c) E\exp(n_j c) \\ &= \exp(-\sqrt{n}\varepsilon c)(1 - p_{nj} + e^c p_{nj})^n \\ &\leq \exp(-\sqrt{n}\varepsilon c + n p_{nj} e^c). \end{aligned} \tag{7.14}$$

Since $\sqrt{n}p_n \to 0$, when n is large enough, we have

$$\varepsilon \log\left(n^{1/2} p_n\right)/2 \leq -1. \tag{7.15}$$

By taking $c = -\log(n^{1/2} p_{nj})$ in (7.14), we conclude from (7.15) that

$$\begin{aligned} P\left\{n_j > \sqrt{n}\varepsilon\right\} &\leq \exp\left(\sqrt{n}\varepsilon \log(n^{1/2} p_{nj}) + \sqrt{n}\right) \\ &\leq \left(n^{1/2} p_{nj}\right)^{\varepsilon\sqrt{n}/2}. \end{aligned} \tag{7.16}$$

Consequently, when $p_{nj} \leq 1/n$,

$$P\{n_j > \sqrt{n}\varepsilon\} \leq p_{nj}^{\varepsilon\sqrt{n}/4},$$

and when $p_{nj} > 1/n$, by (7.15) and (7.16),

$$P\{n_j > \sqrt{n}\varepsilon\} \leq \exp(\sqrt{n}\frac{\varepsilon}{2}\log(n^{1/2}p_{nj})) \leq \exp(-\sqrt{n}) \leq p_{nj}^2.$$

Thus, the condition (7.13) holds, as had to be shown.

Proof of Lemma 4. By Lemma 3, and the Hungarian embedding of Komlós *et al.* (1973), the process $F_n^*(t)$ has the following decomposition:

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

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 3. The conclusion follows from Lemma 4 and the proof of Theorem 1.

7.4 Proof of Theorem 4

We need only to prove that $\hat{f}_n(x;a)$ is the solution to the problem (4.2) with an additional constraint that the location of the mode is X_a. Let $f_j = f(X_j)$, for $f \in \mathcal{F}_L$. Then, the problem is equivalent to

$$\begin{aligned} &\max \sum_j \log f_j \\ \text{subject to} : \quad &(\textit{unimodality}) \quad f_1 \leq f_2 \leq \cdots \leq f_a \geq f_{a+1} \cdots \geq f_n, \qquad (7.17) \\ &(\textit{Area one}) \quad \sum_{j=1}^{n-1} \frac{f_{j+1} + f_j}{2}(X_{j+1} - X_j) = 1. \qquad (7.18) \end{aligned}$$

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 (7.18) can be rewritten as

$$\sum_{j=1}^{n} c_j f_j = 1. \tag{7.19}$$

Denote $g_j = 1/(nc_j)$ and $w_j = nc_j$. Then, the problem is equivalent to maximize $\sum_1^n \log f_j$ subject to (7.17) and $\sum_1^n (g_j - f_j)w_j = 0$. Consider the problem of isotonic regression

$$\min_f \sum_1^n (f_j - g_j)^2 w_j \tag{7.20}$$

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 (7.20) is given by (4.3) (see page 23 of Robertson *et al.* (1988)). The solution satisfies also (Theorem 1.3.6 of Robertson *et al.* (1988))

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

i.e. (7.18). Now, let's apply Lemma 4. Take a convex function $\Phi(u) = u \log u$. Then, $\hat{f}_a$ minimizes also

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

under the isotonic constraints, where $c = \sum \log g_j + n$. Since we are interested only in the class of isotonic regression satisfying (7.19), $\hat{f}_a$ maximizes

$$\sum_1^c \log f_j,$$

under the constraint (7.17) and (7.18). The desired conclusion follows.

Here, we quote Theorem 1.5.1 of Robertson *et al.* (1988):

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 (7.20), then f^* minimizes

$$\sum_j \Delta_\Phi(g_j, f_j)w_j$$

in the class of isotonic functions f.

7.5 Proof of Theorem 5

Let $l(x) = |f(x)f'(x)/2|^{-1/3}$. Note that 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 after (4.4). Thus, according to our geometric interpretation in section 4.1, with probability tending to one, we have

$$\hat{f}_1^*(x + \varepsilon_n) \leq \hat{f}_{nL}(x; 1) \leq \hat{f}_1^*(x - \varepsilon_n), \tag{7.21}$$

where $\hat{f}_1^*(x)$ was defined after (4.4).

Note that the modified empirical distribution defined by (4.4) satisfies

$$0 \leq F_n^*(x) - F_n(x) \leq \frac{1}{n},$$

where $F_n(\cdot)$ is the usual empirical cdf. Thus, by the same argument as Lemma 4, we have

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

Consequently, by (7.21),

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

The conclusion follows from a similar inequality:

$$\begin{aligned} & \liminf_n P\left\{n^{1/3}l(x)(\hat{f}_{nL}(x; 1) - f(x)) \leq t\right\} \\ \geq\ & P\{2Z \leq t\}, \forall t \in (-\infty, \infty) \end{aligned}$$

References

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

[2] 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 M. L. Puri (ed.) *Nonparametric Techniques in Statistical Inference*, Cambridge University Press, 159-173.

[3] Birgé, L. (1987a). Estimating a density under the order restrictions: Non-asymptotic minimax risk. *Ann. Statist.*, **15**, 995-1012.

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

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

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

[7] Fan, J. (1986). Shrinkage estimators and ridge regression estimators for elliptically contoured distributions. *Acta Math. Appl. Sinica*, **9**, 237-250. English translation in *Statistical Inference in Elliptically Contoured and Related Distributions*, (Fang, K.T. and Anderson, T.W. eds).

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

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

[10] Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen smoothing parameter selectors from their optimum? *J. Amer. Statist. Assoc.*, **83**, 86–101.

[11] Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose response curves and the assessment of synergism, *Manuscript.*

[12] Kiefer, J. (1981). Optimum rates for non-parametric density and regression estimates, under order restrictions. *Statistics and Probability: Essays in Honor of C. R. Rao* (G. Kallianpur, *et al.* eds), 419–427.

[13] 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.

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

[15] Prakasa Rao, B. L. S. (1969). Estimation of a unimodal density. *Sankhyā Ser. A*, 31, 23-26.

[16] Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statist. Sci.*, **3**, 425–441.

[17] Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12**, 1215-1230.

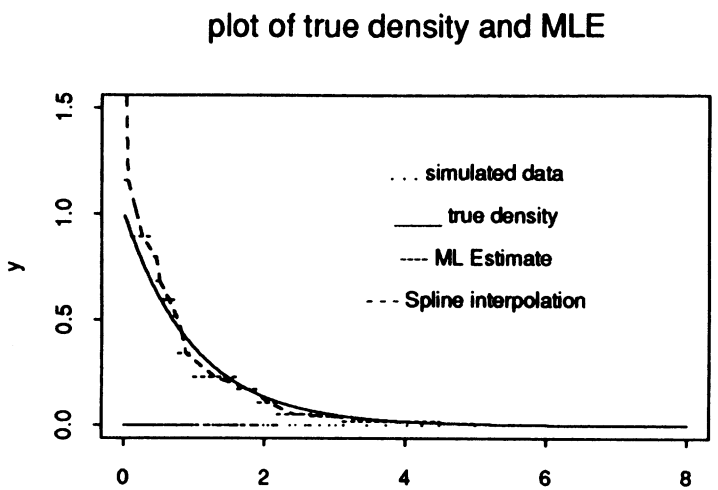
[18] Robertson, T., Wright, F.T., and Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. John Wiley & Sons, New York.

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

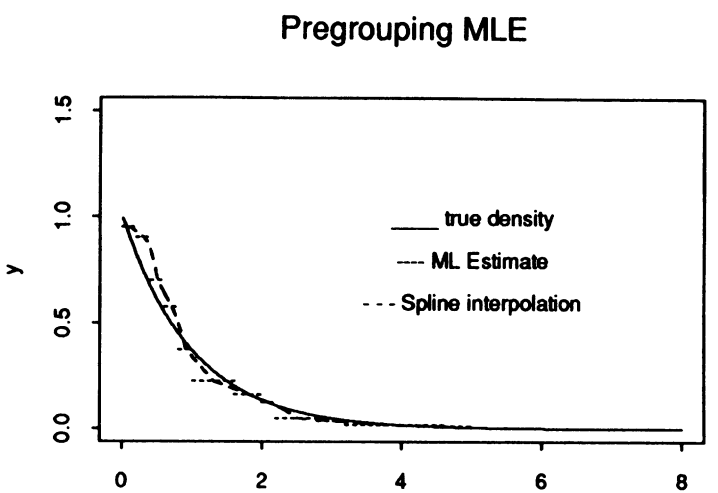
[20] Venter, J. H. (1967). On estimation of the mode. *Ann. Math. Statist.*, **38**, 1446-1455.

[21] Wegman, E. J. (1970a). Maximum likelihood estimation of a unimodal function. *Ann. Math. Statist.*, **40**, 457–471.

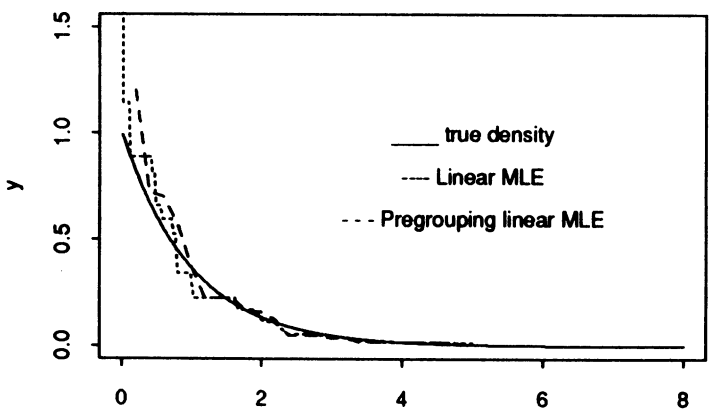
[22] Wegman, E. J. (1970b). Maximum likelihood estimation of a unimodal function, II. *Ann. Math. Statist.*, **40** 2169–2174.



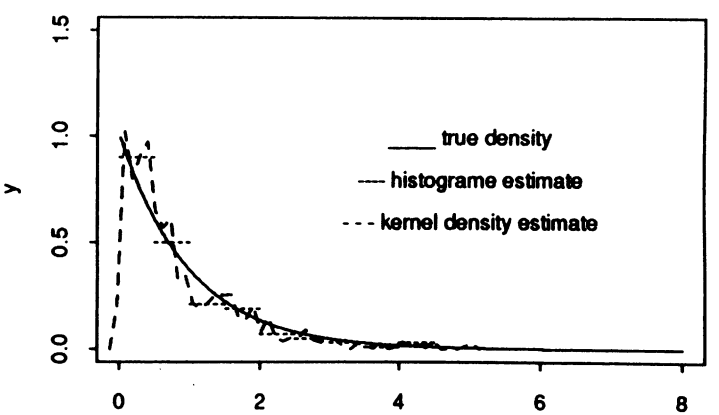
200 data generated from exponential(1)
Figure 1.1



200 data generated from exponential(1)
Figure 1.2

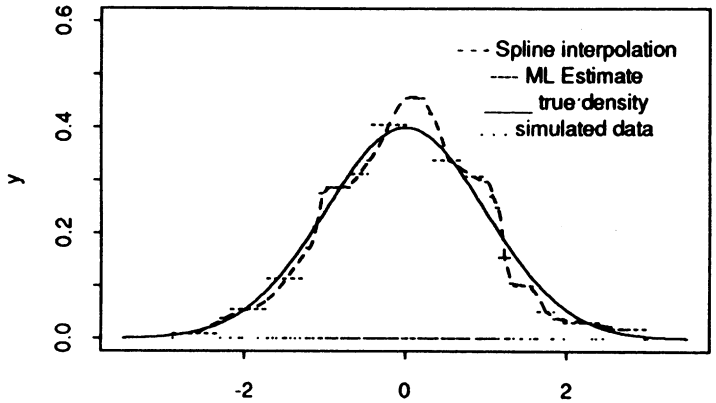


200 data generated from exponential(1)
Figure 1.3



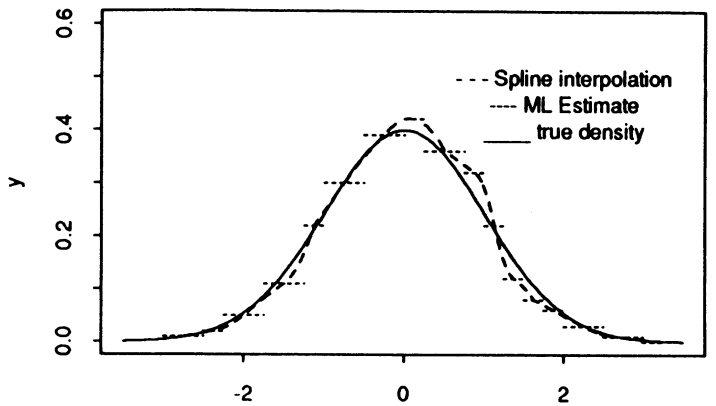
200 data generated from exponential(1)
Figure 1.4

plug-in method with mode = 0



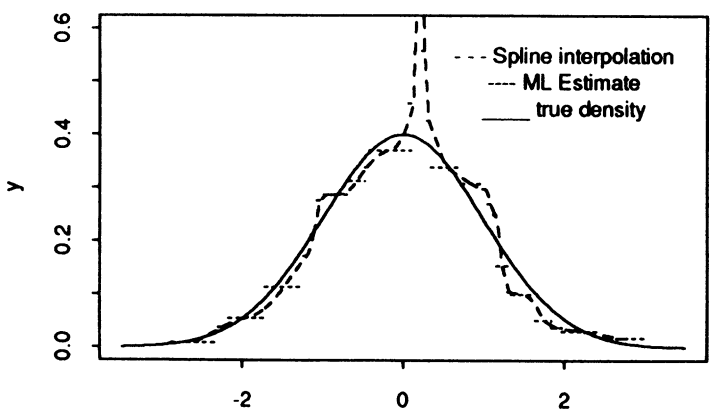
200 data generated from N(0,1)
Figure 2.1

pregrouping plug-in method with mode = 0



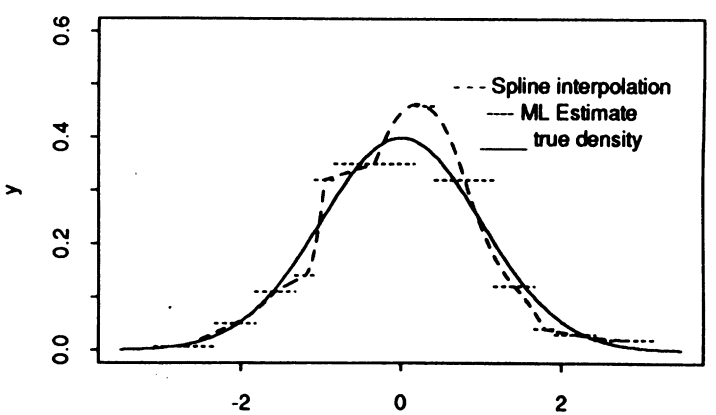
200 data generated from N(0,1)
Figure 2.2

plug-in method with estimated mode = 0.17

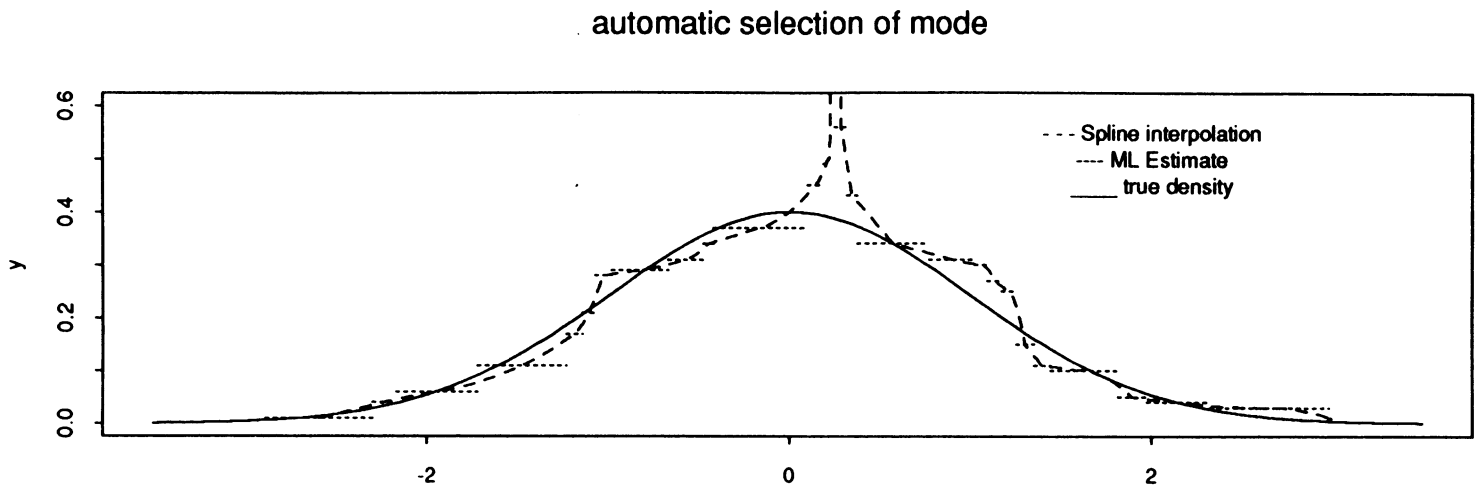


200 data generated from N(0,1)
Figure 2.3

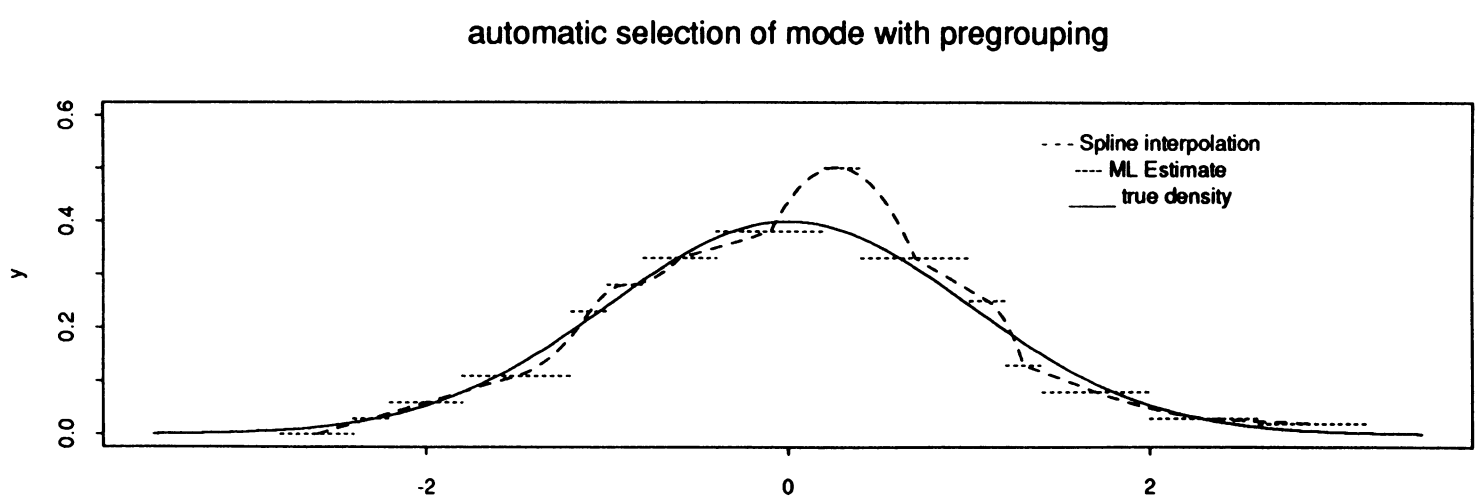
pregrouping plug-in method with mode = 0.17



200 data generated from N(0,1)
Figure 2.4

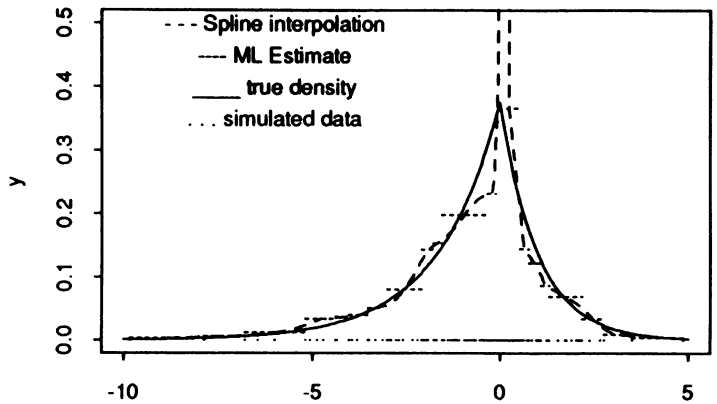


200 data generated from N(0,1)
Figure 2.5



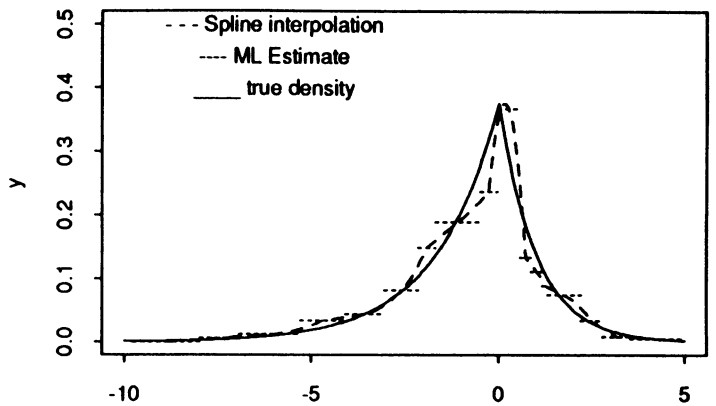
200 data generated from N(0,1)
Figure 2.6

plug-in method with estmated mode =-0.04



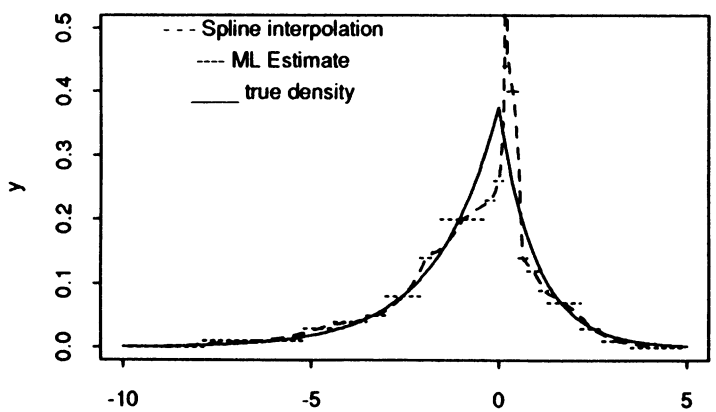
150 data generated from Eq. (6.1)
Figure 3.1

pregrouping plug-in method with mode = -0.04



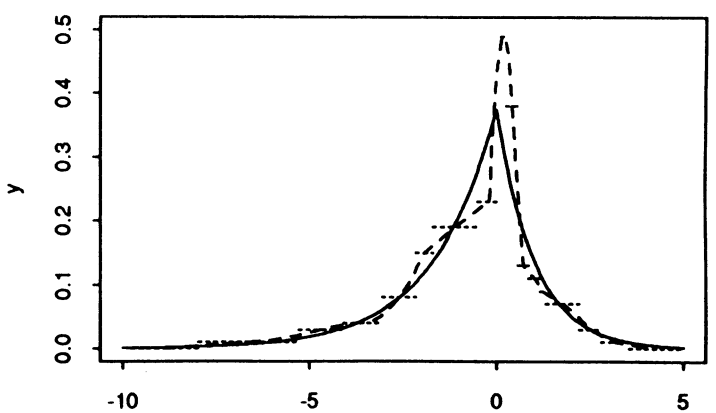
150 data generated from Eq. (6.1)
Figure 3.2

automatic selection of mode



150 data generated from Eq. (6.1)
Figure 3.3

automatic selection of mode with pregrouping



150 data generated from Eq. (6.1)
Figure 3.4