Statistica Sinica **12**(2002), 965-990

UNIMODAL DENSITY ESTIMATION USING KERNEL METHODS

Peter Hall[1] and Li-Shan Huang[1,2]

[1] *Australian National University and* [2] *University of Rochester*

Abstract: We suggest a method for rendering a standard kernel density estimator unimodal: tilting the empirical distribution. It is proposed that the amount of tilting be chosen in order to minimise, subject to unimodality, the integrated squared distance between a conventional density estimator and its tilted version. This approach has an interesting data-compression aspect, in that the algorithm often implicitly summarises the dataset in a relatively small subsample as part of the process of enforcing unimodality. Another feature is that, no matter what the chosen bandwidth, the algorithm produces (with probability 1, for each sample size) a proper density estimate. Thus, it may be employed as an adjunct to any of the many popular bandwidth selection rules for density estimation. We show theoretically that in classes of densities that are of practical interest, the method enhances performance without suffering any deleterious first-order impact on asymptotic accuracy, for example as reflected in integrated squared error. In such cases, and except in places where the true density is virtually flat, the constrained density estimate is first-order equivalent to its unconstrained counterpart. The case where the number of modes is constrained to equal a number greater than 1 is also considered.

Key words and phrases: Bandwidth, biased bootstrap, integrated squared error, mode, nonparametric density estimation, order restricted inference, power divergence, tilting, weighted bootstrap.

1. Introduction

Spurious modes in a nonparametric density estimate can convey a particularly misleading impression of the underlying distribution. The person who views the estimate will not necessarily appreciate that they are the result of eccentricities of the estimator, rather than intrinsic features of the population. In this paper we suggest a method for constructing kernel density estimators that are constrained to be unimodal, or more generally to have k modes where $k \geq 1$ is given. Put simply, our method involves reweighting, or tilting, the empirical distribution in such a way that it differs as little as possible from its standard form, subject to the constraint of unimodality (or multimodality, if that is our goal) being satisfied. The resulting estimator is of the biased-bootstrap type, suggested in a general context by Hall and Presnell (1999).

There is of course a great deal of latitude in defining what is meant by "as little as possible", and in this respect our methods depart significantly from the proposal of Hall and Presnell (1999). We give most emphasis to an L_2 distance function that is tailored specifically to the problem of density estimation. It depends on the data, and is strongly influenced by choice of the bandwidth, h. One of the advantages of our proposal over alternative methods is that the amount of smoothing may be chosen without prejudice. The bandwidth can be selected prior to imposing the unimodality constraint, and our technique produces a uniquely defined, smooth unimodal density estimator based on that value of h. For several other techniques the amount of smoothing is implicit in the method and cannot be chosen separately. In particular, the method of Grenander (1956) imposes its own level of smoothing, which is generally regarded as being too little, especially in the neighbourhood of the mode of the density estimator. (This method usually produces a density estimator with a very "spikey" mode. In other senses too the estimator is quite rough.) Another approach, based on Silverman's (1981) test for unimodality, is to steadily increase the bandwidth until the density estimator first becomes unimodal; here, the constrained estimator is very closely linked to choice of bandwidth. In contrast to these methods, and to newer techniques derived from them, our approach produces an estimator which is intrinsically smooth and for which a virtually arbitrary bandwidth can be chosen.

We demonstrate theoretically that our estimator is well-defined in very general circumstances, and we derive L_2 convergence rates. It is shown that in many instances the L_2 accuracy of our estimator is asymptotically equivalent to that of its unconstrained counterpart, not just in terms of the order of magnitude of the rate but the constant multiplier as well. Therefore, the main effect of the constraint is to remove extraneous "wiggles", the number of which may be unboundedly large as sample size increases, without generally penalising overall performance. Our theory focuses on the case of constrained unimodality, but we note that virtually identical results are valid (with similar proofs) when the number of modes is constrained to be equal to k, provided the true density also has that number of modes. Our numerical work treats the cases $k = 1$ and 2.

The method of Grenander (1956) has been studied in depth by Bickel and Fan (1996), Wang (1996) and Birgé (1997). An alternative approach to unimodal density estimation has been suggested by Cheng, Gasser and Hall (1999). Tilting methods for ensuring monotonicity have been suggested by Hall and Huang (2001a, b), but the results and techniques discussed there are of a very different nature from those developed here. In particular, an integrated squared error measure of distance is not attractive in those applications, and there is no analogue of the influence of tail behaviour on unimodality. Moreover, the methodologies of Hall and Huang (2001a, b) relate to monotone regression and monotone hazard rate estimation, respectively, not unimodal density estimation as in the present paper.

Our method has a data-compression effect, in that the effective size of the sample is reduced by assigning zero weight to many data values. Of course, information in the zero-weighted data is not lost; it is incorporated into weights for the existing data, and into choice of the indices of those data. A similar effect is observed for the method of Grenander (1956), where the distribution function estimator corresponding to the unimodal density estimator is piecewise continuous with knots at only a portion of the sample values.

It should be noted that in most settings, imposing a shape constraint such as unimodality, based on slope, cannot be expected to asymptotically improve performance in terms of mean integrated squared error (MISE), over and above the level enjoyed by an unconstrained estimator. To appreciate why, note that if n denotes sample size and we use a bandwidth of larger order than $n^{\epsilon-1/3}$, for some $\epsilon > 0$, then the derivative of the estimator uniformly and consistently estimates that of the true density. Indeed, Bernstein-type bounds show that the probability that the gradient of the estimator is uniformly within δ of the gradient of the true density, equals $1 - O(n^{-\lambda})$ for each $\delta,\ \lambda > 0$. This means that, on intervals where the true density is bounded away from zero, the gradient of the density estimator has the same sign as that of the true density.

Thus, in the context of unimodal density estimation, and in large samples, the conventional estimator requires modification only in the neighbourhood of the mode and far out in the tails. The neighbourhood shrinks to zero, and the distance out in the tails moves further out, as sample size increase. Improving performance in these places, even reducing error to zero, will have only an asymptotically negligible effect on MISE. Our method admittedly alters the estimator at other places too, but this is only an artifact of its construction; the other changes result principally from the constraint $\sum_i p_i = 1$.

Assuming that the bandwidth is of size $n^{\epsilon-1/3}$ is of course a very mild condition. In the context of our work, a bandwidth of size $n^{-1/5}$ is optimal. More generally, for other "second order" estimator types a construction which optimises MISE performance ensures uniformly consistent estimation of gradients in the sense discussed above, and so again the estimators enjoy no first-order MISE gains in performance.

2. Methodology

Let $p_{\text{unif}} \equiv (1/n, \ldots, 1/n)$ denote the conventional uniform-empirical weight vector, and let $D(p)$, possibly depending on the data, denote a measure of the distance between p_{unif} and an element $p = (p_1, \ldots, p_n)$ of the class of n-variate probability vectors p. Suppose we have determined a bandwidth h for a standard kernel estimator $\hat{f}$ of the density f of the sampled distribution. For example, h might be defined using a plug-in rule, or cross-validation. Choose $p = \hat{p} = (\hat{p}_1, \ldots, \hat{p}_n)$ to minimise $D(p)$ subject to the modality constraint being satisfied, and take the final estimator of f to be $\hat{f}$ but with weight $\hat{p}_i$, instead of $1/n$, given to datum X_i in the sample $\mathcal{X} = \{X_1, \ldots, X_n\}$, for $1 \leq i \leq n$. If p is to be a probability vector then it is also necessary to impose the constraint that $\sum_i p_i = 1$.

Minimising $D(p)$ subject to the constraint amounts to enforcing the modality property subject to maximum fidelity to the data, where fidelity is measured by D. Obvious candidates for D include those based on measures of distance between distribution functions, for example the Kolmogorov-Smirnov distance

$$D_{\text{KS}}(p) \equiv \sup_x |\hat{F}(x|p) - \hat{F}(x|p_{\text{unif}})| \, ,$$

and the Cramér-Von Mises distance

$$D_{\text{CM}}(p) \equiv \int \{\hat{F}(x|p) - \hat{F}(x|p_{\text{unif}})\}^2 \, dx \, ,$$

where $\hat{F}(x|p)$ denotes the "weighted bootstrap" distribution function that places mass p_i on X_i.

Throughout our theoretical and numerical work we focus on the case where the data $\mathcal{X}$ are univariate, although at least formally there is no difficulty extending our methods and results to multivariate settings. In the univariate case $\hat{F}(x|p) = \sum_{i=1}^n p_i \, I(X_i \leq x)$, and so, for example,

$$D_{\text{CM}}(p) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} (np_i - 1)\,(np_j - 1)\, \min(X_i, X_j). \tag{2.1}$$

In a univariate setting the weighted density estimator has the form

$$\hat{f}(x|p) = \frac{1}{h} \sum_{i=1}^{n} p_i \, K\Big(\frac{x - X_i}{h}\Big) \, ,$$

where K is the kernel function and h denotes the bandwidth.

The approach above has a conceptual drawback, however, in that the distance measures D_{KS} and D_{CM} are not directly connected to the problem of density estimation. They measure "fidelity to the data" in a sense where fidelity relates to estimating a distribution function. It is arguably more appropriate to use smoothed-density versions of D_{KS} and D_{CM}:

$$d_{\mathrm{KS}}(p) \equiv \sup_x |\hat{f}(x|p) - \hat{f}(x|p_{\mathrm{unif}})|\,, \quad d_{\mathrm{CM}}(p) \equiv \int \{\hat{f}(x|p) - \hat{f}(x|p_{\mathrm{unif}})\}^2\, dx\,.$$

There are, however, numerical difficulties in performing constrained optimisation using d_{KS}, and moreover it measures fidelity in a way that is relatively uncommon in nonparametric density estimation. In particular, optimisation of the level of smoothing with respect to the uniform metric requires a bandwidth that is generally larger, particularly for large samples, than that given by common bandwidth-choice methods.

For these reasons we suggest using d_{CM} in preference to d_{KS}; we represent it by I below, denoting integral distance-measure. Provided K is symmetric, $I(p)$ admits a formula analogous to (2.1):

$$\begin{aligned} I(p) &= \int \{\hat{f}(x|p) - \hat{f}(x|p_{\mathrm{unif}})\}^2\, dx \\ &= (n^2 h)^{-1} \sum_{i=1}^n \sum_{j=1}^n (np_i - 1)\,(np_j - 1)\, L\Big(\frac{X_i - X_j}{h}\Big), \end{aligned} \tag{2.2}$$

where the function L denotes the convolution of K with itself. In particular, if K is the standard normal kernel, which is frequently used in practice, then $L(x)$ is proportional to $e^{-x^2/4}$ and so may be taken equal to this function.

In a numerical algorithm, unimodality may be enforced by insisting that $\hat{f}(\cdot|p)$ be monotone nondecreasing on the interval $(-\infty, \theta)$ and monotone nonincreasing on (θ, ∞), where θ is a candidate for the mode of $\hat{f}(\cdot|p)$. Tilting the empirical distribution so as to minimise $D(p)$ subject to achieving this constraint produces a probability vector $p = \hat{p}(\theta)$ that depends on the suggested mode. We would then select θ so as to minimise $D\{\hat{p}(\theta)\}$, although an alternative approach would be to take θ equal to a pre-determined quantity such as the location of the largest local maximum of $\hat{f}(\cdot|p_{\mathrm{unif}})$, or to an alternative estimator such as that suggested by Eddy (1980). The first of these options, at least, is a useful starting point for algorithms that find the value of θ that minimises $D\{\hat{p}(\theta)\}$, and will be used for that purpose in Section 4. The theory and numerical work in this paper will concentrate on choosing $\theta = \hat{\theta}$ to minimise $D\{\hat{p}(\theta)\}$, and so our weighted kernel estimator will use the probability weights $\hat{p}(\hat{\theta})$.

A minor drawback suffered by $I(p)$, and shared by the distance measures $D_{\mathrm{KS}}(p)$, $D_{\mathrm{CM}}(p)$ and $d_{\mathrm{KS}}(p)$, is that they are well-defined even when one or more components of p are negative. This means that simply minimising $I(p)$ with respect to p, subject to the modality constraint and $\sum_i p_i = 1$, may not produce a true probability distribution. We need to add the restriction that each $p_i \geq 0$, which requires an extra step in the numerical algorithm. Many of the general distance measures suggested by Cressie and Read (1984) for probability vectors are not well-defined unless all elements of p are nonnegative. Additionally, the Cressie-Read "divergences" have ready interpretation in terms of measures of information, or well-known statistical distances such as Hellinger distance. For these reasons they merit consideration in this setting. They have the form

$$D_\rho(p) = \frac{1}{\rho\,(1-\rho)} \left\{ n - \sum_{i=1}^{n} (np_i)^\rho \right\}$$

for $-\infty < \rho < \infty$ and $\rho \neq 0, 1$, and

$$D_0(p) = -\sum_{i=1}^{n} \log(np_i)\,, \quad D_1(p) = \sum_{i=1}^{n} p_i \log(np_i)\,.$$

The latter are both Kullback-Leibler divergences.

A major difference between $I(p)$ and many of the distance measures $D_\rho(p)$ is that $I(p)$ suffers very little penalty for reducing some of the p_i's to zero. By way of contrast, the conventional "likelihood-based" measure of distance $D_0(p)$, used for example in Owen's (1988, 1990) method of empirical likelihood, is infinite if some p_i vanishes. As a result, $D_0(p)$ is not suitable for enforcing unimodality. The measure $D_1(\rho)$ suffers less, but still provides substantially greater resistance to data compression (that is, to many of the p_i's being rendered equal to zero) than $I(p)$. Indeed, it is not unusual that between half and three-quarters of the values of p_i be set equal to zero when enforcing unimodality by minimising $I(p)$ for a sample of size about 50; see Section 4 for details. Of course, the information in these data is not lost. It is incorporated into the selection of data that are kept in the sample, and also into the values of nonzero weights p_i.

The data compression property can be useful, since it greatly reduces the amount of information that has to be retained in order to store a unimodal approximation to the original density estimator. Moreover, numerical results show that constraining by minimising $I(p)$ generally produces better performance than constraining by minimising $D_1(p)$, since the former distance measure focuses explicitly on minimising mean squared error.

In subsequent sections we take $\hat{p}$ and $\bar{p} = \bar{p}(\rho)$ to be the values of p that minimise $I(p)$ and $D_\rho(p)$, respectively, subject to the constraint of unimodality (or bimodality, in some numerical examples) and the condition $\sum_i p_i = 1$.

3. Theoretical Properties

Our first result shows that under very weak conditions, in particular only minimal assumptions about the bandwidth h, the probability vectors $\hat{p}$ and $\bar{p} =$

$\bar{p}(\rho)$ that confer unimodality are well defined and unique. By way of notation, a continuously differentiable density will be said to be strictly unimodal if there exists a unique point in the interior of its support at which f' vanishes. It follows that that point gives a global maximum of f; it is the mode of f. More generally, the mode of a continuously differentiable density f is a point x where f' vanishes, and where for some interval $\mathcal{L}$ containing x as an interior point, $f(x) \geq f(y)$ for all $y \in \mathcal{L}$, and $f(x) > \max\{f(y_1), f(y_2)\}$ for some $y_1,\ y_2 \in \mathcal{L}$ with $y_1 < x < y_2$. In the theorem below, a "continuous distribution" is a distribution that has a density.

Theorem 3.1. *Assume K is a symmetric, continuously differentiable, compactly supported, strictly unimodal density, that the data $\mathcal{X}$ are independent with a common continuous distribution, and that in the distance measure D_ρ we take $0 < \rho \leq 1$. Then for any value of h, with probability 1 there exists $p = \hat{p}$ that minimises $I(p)$ subject to $\hat{f}(\cdot|p)$ being unimodal, and also there exists $p = \bar{p} = \bar{p}(\rho)$ that minimises $D_\rho(p)$ subject to $\hat{f}(\cdot|p)$ being unimodal. If the support of K equals the interval $[-c, c]$, and if $X_{(1)} < \cdots < X_{(n)}$ denote the ordered sample values, then conditional on $\sup_{1\leq i\leq n-2}(X_{(i+2)} - X_{(i)}) \leq 4ch$, the probability that $\hat{p}$ and $\bar{p}$ are uniquely defined equals* 1.

The last part of the theorem is tailored to the case where f is compactly supported, on the interval $[A, B]$ say. To appreciate its implications in this setting, assume $F(A+x)$ and $F(B-x)$ decrease to zero like x^a and x^b, respectively, as $x \downarrow 0$; here $a, b > 0$. Then if $h \asymp n^{-1/5}$ and $d \equiv \max(a, b) < 5$, the probability that $\sup_{1\leq i\leq n-2}(X_{(i+2)} - X_{(i)}) \leq 4ch$ converges to 1 as $n \to \infty$. This result may be proved from Rényi's representation for order statistics, noting that $X_{(i+2)} - X_{(i)}$ is approximately equal to the sum of two independent exponential random variables divided by $n\,f(X_{(i)})$, and that the infimum of the latter diverges to infinity like $n^{1/d}$.

Theorem 3.1 continues to hold if we stipulate that $\hat{f}(\cdot|\hat{p})$ have k modes, for any fixed $k \geq 1$, provided (a) $k \leq n$, (b) the support of K equals $[-c, c]$ for some $c > 0$, and (c) $2ch$ is no larger than the supremum of the values of the smallest interpoint distance in subsets of size k of the full dataset. Theorem 3.1 does not hold if we measure distance using D_ρ with $\rho = 0$. Only slightly altered versions of Theorem 3.1 and the results below hold if we take K to be a Gaussian density.

The next theorem shows that under more restrictive but still quite weak assumptions about h, and a very mild smoothness condition on f, the constrained estimators $\hat{f}(\cdot|\hat{p})$ and $\hat{f}(\cdot|\bar{p})$ are consistent. We do not require f to be strictly unimodal.

Theorem 3.2. *Assume the conditions of Theorem* 3.1, *and in addition that K has two Hölder-continuous derivatives, $h = h(n) \to 0$ and $n^{(1/3)-\epsilon}h \to \infty$ for some $\epsilon > 0$, and f is continuous and unimodal (not necessarily with a unique mode). Then any probability vector $p = \hat{p}$ that minimises $I(p)$ subject to unimodality of $\hat{f}(\cdot|p)$, has the property that with probability* 1, *$\hat{f}(\cdot|\hat{p}) \to f$ in L_2. Furthermore, for any given $B > 1$, and with probability* 1, *for all sufficiently large n there exists a unique $p = \bar{p} = \bar{p}(\rho)$ that minimises $D_\rho(p)$ subject to $\hat{f}(\cdot|p)$ being unimodal and* $\sup_i \bar{p}_i \le Bn^{-1}$*; and $\hat{f}(\cdot|\bar{p}) \to f$ in L_2.*

Our next result describes convergence rates and related properties in the case of twice-differentiable densities with nonvanishing curvature at the mode, and for bandwidths of optimal size in the sense of the unconstrained estimator (i.e., $h \asymp n^{-1/5}$). The theorem accommodates a wide range of different tail behaviours of the density f. Sharper and more detailed results may be developed in specific cases.

Assume (i) K is a symmetric, compactly supported, strictly unimodal density with two Hölder-continuous derivatives, (ii) $h \asymp n^{-1/5}$, (iii) f is strictly unimodal with mode m and two bounded, square-integrable derivatives, and (iv) f'' is continuous in a neighbourhood of m and $f''(m) < 0$. Without loss of generality, the support of K is the interval $[-1, 1]$; we assume this below.

Let the interval (α, β) be contained within the support $\mathcal{S}$ of f, and suppose α and β, which we take to be functions of n, converge to the left- and right-hand extremities, respectively, of $\mathcal{S}$ as $n \to \infty$ (one or both of the extremities may be infinite), in such a way that (I) $f(\alpha + h)/f(\alpha)$ and $f(\beta)/f(\beta + h)$ are bounded, (II) for some $\epsilon > 0$ and all $\eta > 0$, $f(x) = O\{f'(x)^2 n^{(2/5)-\epsilon}\}$ uniformly in $x \in \mathcal{T}_\eta \equiv (\alpha, m - \eta) \cup (m + \eta, \beta)$, and (III) $nF(\alpha) \to \infty$ and $n\{1 - F(\beta)\} \to \infty$. (Under conditions (i)$-$(iv), sequences α, β with these properties always exist.) Let F denote the distribution function corresponding to f, and (IV) put $\lambda = F(\alpha) + 1 - F(\beta)$ and

$$\Lambda = \int_{-\infty}^{\alpha} f(x)\, f(x + 2h)\, dx + \int_{\beta}^{\infty} f(x)\, f(x - 2h)\, dx\,.$$

Let $\|\cdot\|$ denote the usual L_2 norm for functions.

Theorem 3.3. *Assume conditions* (*i*) $-$ (*iv*), *and that α, β, λ, Λ satisfy* (*I*) $-$ (*IV*). (*a*) *If $p = \hat{p}$ minimises $I(p)$ subject to unimodality of $\hat{f}(\cdot|p)$, then*

$$\|\hat{f}(\cdot|\hat{p}) - f\|^2 = O_p(\Lambda + \lambda^2 + n^{-4/5})\,. \tag{3.1}$$

(*b*) *If $p = \bar{p} = \bar{p}(\rho)$ minimises $D_\rho(p)$ subject to $\hat{f}(\cdot|p)$ being unimodal and* $\sup_i p_i \le Bn^{-1}$, *where $B > 1$ is arbitrary but fixed, then*

$$\|\hat{f}(\cdot|\bar{p}) - f\|^2 = O_p(h^{-1/2}\lambda + n^{-4/5})\,. \tag{3.2}$$

(*c*) *Provided* $\alpha,\ \beta,\ \lambda,\ \Lambda$ *may be constructed such that* $\Lambda+\lambda^2=o(n^{-4/5})$, *or* $h^{-1/2}\lambda=o(n^{-4/5})$, *we have, respectively,*

$$\frac{\|\hat f(\cdot|\hat p)-f\|}{\|\hat f(\cdot|p_{\text{unif}})-f\|}\to 1 \quad \textit{in probability, or} \tag{3.3}$$

$$\frac{\|\hat f(\cdot|\bar p)-f\|}{\|\hat f(\cdot|p_{\text{unif}})-f\|}\to 1 \quad \textit{in probability.} \tag{3.4}$$

In the cases of infinitely supported densities with standard exponential, standard normal, or regularly varying (with exponent $-r$) tails, appropriate values of $|\alpha|$ and β are respectively $(\frac{2}{5}-\epsilon)\log n$, $\{(\frac{4}{5}-\epsilon)\log n\}^{1/2}$ and $n^{(2-\epsilon)/\{5(r+2)\}}$ for some $\epsilon>0$. Corresponding values of λ and Λ are given by (IV) above. The case of compactly supported f will be treated in detail shortly.

Results (3.1) and (3.2) provide upper bounds to convergence rates, while (3.3) and (3.4) give greater detail – they show that, under suitable regularity conditions, integrated squared errors of the constrained estimator $\hat f(\cdot|p)$ (for $p=\hat p$ or $p=\bar p$) and its conventional, unconstrained counterpart $\hat f(\cdot|p_{\text{unif}})$ are asymptotically identical. Thus, nothing is either gained or lost (in an asymptotic sense), in terms of integrated squared error, by imposing the constraint of unimodality. Of course, an obvious gain is guaranteed unimodality. In addition to (3.3) and (3.4) it may be shown that, with p denoting either $\hat p$ or $\bar p$, and assuming the conditions of part (c), $\|\hat f(\cdot|p)-f(\cdot|p_{\text{unif}})\|=o_p(n^{-2/5})$.

It may be proved that if we define mean integrated squared error by MISE $=E\{\|\hat f(\cdot|p_{\text{unif}})-f\|^2\}$, then, under conditions (i)–(iv), $\|\hat f(\cdot|p_{\text{unif}})-f\|$ divided by the square root of MISE converges to 1 in probability. Therefore, we may use MISE$^{1/2}$ instead of $\|\hat f(\cdot|p_{\text{unif}})-f\|$ in the denominators on the left-hand sides of (3.3) and (3.4), without affecting their validity.

Theorems 3.2 and 3.3 have straightforward analogues in contexts where the number of modes is constrained to equal any fixed number $k\geq 1$, rather than simply $k=1$. For example, in the case of k modes, regularity conditions (ii) and (iii) imposed in Theorem 3.3 should be changed respectively to: (iii)$'$ f has k uniquely-defined modes and two bounded, square-integrable derivatives, and (iv)$'$ f'' is continuous in neighbourhoods of each of the k local maxima, where it is strictly negative, and each of the $k-1$ local minima, where both f and f'' are strictly positive. Then, for the same definitions of $a,\ \beta,\ \lambda,\ \Lambda$ as before, Theorem 3.3 holds without change. The proofs are also virtually identical.

For the sake of brevity we explore the consequences of Theorem 3.3 further only in the case of the distance function $I(p)$, and for a compactly supported density f that is polynomially decreasing at its boundaries. Specifically, assume that in addition to satisfying conditions (iii) and (iv), f is compactly supported with support interval $[A,B]$, and for $j=0,1,2$, the following condition (v) holds:

$$f^{(j)}(x) \asymp (x-A)^{a-j} \text{ as } x \downarrow A, \text{ and } f^{(j)}(x) \asymp (B-x)^{b-j} \text{ as } x \uparrow B\,,$$

where $a,b \geq 2$ and the relation $u(x) \asymp v(x)$ means that $u(x)/v(x)$ is bounded away from zero and infinity as $x \uparrow B$. (The assumption $a,b \geq 2$ is necessary in order for f to have two bounded derivatives, as required by assumption (iii).) Beta (s,t) densities with $s,t \geq 3$ are examples of densities that satisfy (iii), (iv) and (v). We claim that, under conditions (i)$-$(v), (3.3) is true. Equivalently, the integrated squared errors of the constrained estimator $\hat{f}(\cdot|\hat{p})$ and its conventional, unconstrained counterpart are asymptotically equal. Call this claim $\mathcal{C}_1$.

We also assert that if (i)$-$(v) hold then the distribution of the number of modes, $M=M(n)$ say, of the unconstrained estimator satisfies

$$\liminf_{n\to\infty} P(M \geq k) > 0 \quad \text{for all} \quad k \geq 1. \tag{3.5}$$

Moreover, if f decreases in one or both tails like a polynomial of degree greater than or equal to 4 (that is, if $\max(a,b) \geq 4$, where a and b are as in condition (v)), then

$$\lim_{n\to\infty} P(M \geq k) = 1 \quad \text{for all} \quad k \geq 1. \tag{3.6}$$

We shall refer to these two results as claim $\mathcal{C}_2$. It follows from $\mathcal{C}_2$ that constraining $\hat{f}$ to be unimodal can make a substantial difference to its "wiggliness", and so to its appearance, even if it does not appreciably alter the value of integrated squared error.

To verify $\mathcal{C}_1$, take $A=0$ without loss of generality, and note that condition (I), applied at the left-hand end of the support interval, is satisfied if

$$\alpha = \alpha(n) \to 0 \quad \text{in such a manner that} \quad h/\alpha \text{ is bounded.} \tag{3.7}$$

Condition (II), again applied at the left-hand end of the support, is satisfied if, for some $\epsilon \in (0,2)$,

$$\alpha \geq \alpha_\epsilon \equiv n^{-(2-\epsilon)/\{5(a-2)\}}, \tag{3.8}$$

where, since we are assuming in addition (3.7), we may interpret $1/(a-2)$ as $+\infty$ if $a=2$. We take $\alpha = n^{-a_\epsilon}$, where $5a_\epsilon \equiv \min\{1,(2-\epsilon)/(a-2)\}$. Then both (3.7) and (3.8) are satisfied. This choice of α also ensures that $nF(\alpha) \to \infty$, which is property (III) in the left-hand tail.

Put $\lambda_A \equiv F(\alpha) \asymp \alpha^{a+1}$ and $\Lambda_A = \int_0^\alpha f(x)\,f(x+2h)\,dx \asymp \alpha^{2a+1}$. Then $\lambda_A^2 = o(\Lambda_A)$, and for $\epsilon > 0$ sufficiently small, Λ_A is of smaller order than $n^{-4/5}$

if and only if either $a = 2$, or $a > 2$ and, for some $\epsilon' \in (0, 2)$, $(2a + 1)\min(a - 2, 2 - \epsilon') > 4\,(a - 2)$. These conditions hold for all $a \geq 2$. The analogous result in the right-hand tail is also true. Therefore, we may choose α, β such that $\Lambda + \lambda^2 = o(n^{-4/5})$, and so (3.3) holds, verifying claim $\mathcal{C}_1$.

To verify claim $\mathcal{C}_2$ we note that, using a slight modification of methods of Mammen, Marron and Fisher (1992), it may be proved that if we modify (3.5) by replacing M by the number, M_1 say, of modes of the *unconstrained* estimator in any sufficiently small neighbourhood of the origin, then (3.5) holds. (In fact, M_1 has a proper, nondegenerate asymptotic distribution.) Since $M \geq M_1$ then the unmodified form of (3.5) must be true. To appreciate why (3.6) is correct, observe that for each $k \geq 1$, the spacings between adjacent values of the $k + 1$ smallest order statistics (in the left-hand tail of the distribution with density f) may be represented as $n^{-1/(a+1)}\,\xi_i$ for $1 \leq i \leq k$, where the random variables $\xi_i = \xi_i(n)$ satisfy

$$\liminf_{n\to\infty} \inf_{1\leq i\leq k} P(\xi_i > x) > 0 \quad \text{for all} \quad x > 0 \quad \text{and each } k\,. \tag{3.9}$$

If $a \geq 4$ then, since they are of size $n^{-1/(a+1)}$, the spacings in the left-hand tail are of the same size as, or larger than, the bandwidth h, and so in view of (3.9), the number of modes in the lower tail must diverge to infinity in probability. A similar argument applies in the upper tail, and so (3.6) must hold if $\max(a, b) \geq 4$.

4. Numerical Properties

4.1. Implementation

In each of the simulated-data examples the sample size was $n = 50$, the number of simulations was 500, and the bandwidth was chosen using the method suggested by Sheather and Jones (1991). For simulated- and real-data examples alike, the Gaussian kernel was used and unimodality was enforced on the interval $(\min_i X_i - 2h, \max_i X_i + 2h)$. (Choosing the range-overlap to equal $2h$ accommodated the decay of the Gaussian kernel.)

Calculation of a unimodal density estimate, given a candidate value m_0 for the mode, was accomplished by selecting ν equally-spaced grid points, $x_1, \ldots, x_\nu$, on $(\min_i X_i - 2h, \max_i X_i + 2h)$, and minimising distance subject to the constraint that the derivatives at grid points were nonnegative if $x_j < m_0$ and nonpositive if $x_j > m_0$, as well as enforcing the condition $\sum_i p_i = 1$. A grid search was then performed to locate the value $m_0 = \widehat{m}$ that yielded the smallest value of distance. We took $\nu = 100$.

To calculate a bimodal density estimate, denote the modes by m_0 and m_2 and the antimode by m_1, where $m_0 < m_1 < m_2$. We chose p to minimise distance subject to the constraints $\hat{f}'(x_j|p) \geq 0$ for $x_j \in (\min_i X_i - 2h, m_0) \cup (m_1, m_2)$ and $\hat{f}'(x_j|p) \leq 0$ for $x_j \in (m_0, m_1) \cup (m_2, \max_i X_i + 2h)$, as well as $\sum_i p_i = 1$; and then we found m_0, m_1 and m_2 by grid search.

The constraint of unimodality is the combination of two order conditions on the density, and so is inherently nonlinear. It requires quadratic programming or a similar procedure. We used the NAG routine E04, which was both stable and fast.

To assess performance we approximated pointwise mean squared error, $\mathrm{MSE}(x) = E\{\hat{f}(x|\hat{p}) - f(x)\}^2$, by the average over all simulations of the value of $\{\hat{f}(x_j|\hat{p}) - f(x_j)\}^2$, where (for that simulation) x_j was the gridpoint nearest to x. On this occasion we confined attention to those simulated datasets for which $\hat{p} \neq p_{\mathrm{unif}}$; this allowed us to more sharply delineate performance differences between constrained and unconstrained estimators. Therefore, the expectation in the definition of $\mathrm{MSE}(x)$ should be interpreted as being conditional on the event that manipulation is necessary in order to achieve the desired number of modes.

Our examples concentrate primarily on the distance measure $I(p)$, which, because it focuses explicitly on L_2 fidelity, produces superior L_2 performance relative to $D_\rho(p)$. However, in some cases we give results for $D_\rho(p)$ for $\rho = 1$, in order to illustrate the very different choices of weights that result.

4.2. Simulation study

In our first example the distribution was standard normal. There, just 147 out of the 500 simulated datasets required manipulation in order to produce unimodality. Panel (a) of Figure 1 depicts estimates computed from that sample among the 500 that corresponded to the 475th largest value of $I(\hat{p})$ (approximately the 95th percentile). The constrained estimate for the distance $D_1(p)$ is also shown; the mode estimate for both was the same to two significant figures, $\widehat{m} = -0.28$. Both constrained estimates correct for a second mode in the right-hand tail.

The weight vectors $\hat{p}$ that respectively minimise $I(p)$ and $D_1(p)$ are very different, as panels (c) and (d) of Figure 1 show. The majority of sample values (36 out of $n = 50$) have been given zero weight after minimising $I(p)$, effectively compressing the dataset to a substantially smaller one that nevertheless provides very effective estimation of the target density, as may be seen from panel (a). Panel (b) plots pointwise $\mathrm{MSE}(x)$ for the constrained estimate based on minimising $I(p)$ (solid line) or $D_1(p)$ (long-dashed line), and the unconstrained estimate (dotted line), conditional on manipulation being necessary to achieve unimodality.

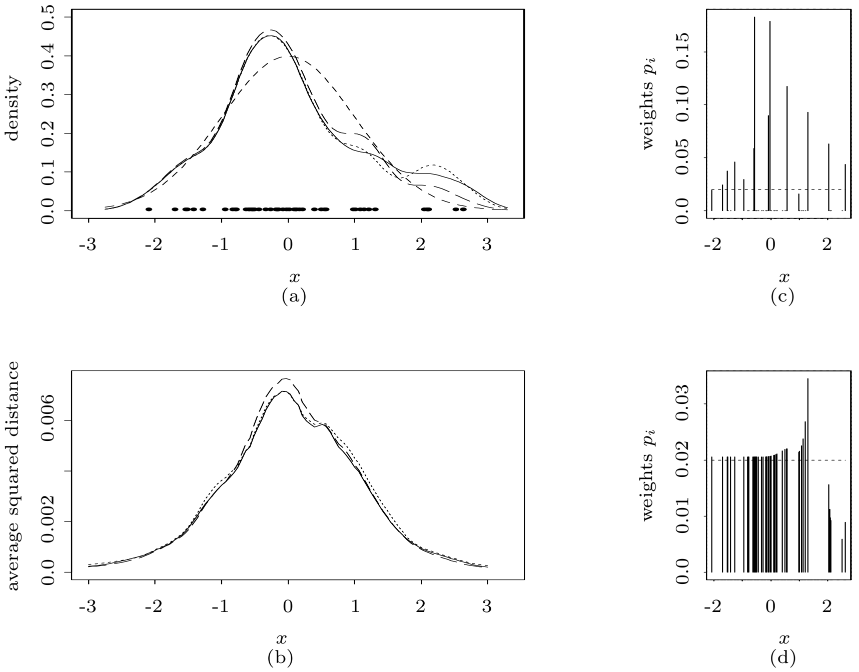

Figure 1. *Result of enforcing unimodality for data simulated from standard normal distribution.* For the simulated dataset whose value of $I(\hat{p}) = 5.04 \times 10^{-4}$ was at the 95th percentile, panel (a) shows the true density function (dashed line), the unconstrained kernel estimate with $h = 0.33$ (dotted line), the constrained estimate minimising $I(p)$ (solid line), and the version of this quantity for distance measure $D_1(p)$ (long-dashed line). Panel (b) plots the pointwise mean squared error for the constrained estimator minimising $I(p)$ (solid line) or $D_1(p)$ (long-dashed line), and the unconstrained estimator (dotted line). Panels (c) and (d) show the values of $\hat{p}_i$ and $\bar{p}_i$, as functions of X_i, after the constraint has been achieved for distance measures $I(p)$ and $D_1(p)$, respectively.

The slight but consistent reduction in mean squared error offered by the constrained estimator based on minimising $I(p)$ is clear from panel (b); it is still more obvious if we average left- and right-hand sides of the figure, exploiting the symmetry of the target distribution. The extent of improvement is least at the mode and in the tails. By way of comparison, MSE performance for the constrained estimator based on minimising $D_1(p)$ is inferior across the sample space, and in relative terms the greatest decrease in accuracy occurs at the mode.

A more detailed analysis shows this to be a consequence of the relatively high bias of the corresponding constrained estimator.

Our second example, the Beta $(2,2)$ distribution, is of methodological interest in that, unlike the normal, it is of the type addressed by regularity conditions imposed by Silverman (1983) and Mammen, Marron and Fisher (1991) in their studies of unimodality. Their asymptotic results show that, as sample size increases and for a bandwidth of size $n^{-1/5}$, such as that provided by the Sheather-Jones (1991) method used in our study, the probability of there being spurious modes in the tail of a density estimate converges to 0. Nevertheless we found that 152 of the 500 simulated samples required manipulation in order to achieve unimodality.

In this setting, the constrained estimator based on minimising $I(p)$ produced a marked reduction in pointwise MSE (conditional on manipulation being necessary). This was most pronounced at the mode, but occurred across the range of the sampled distribution. By way of comparison, the estimator based on minimising $D_1(p)$ had similar MSE performance to the unconstrained estimator, being slightly inferior at the mode and slightly superior elsewhere. In the case of the dataset that produced the 475th largest value of $I(\hat{p})$, the data-compression phenomenon noted earlier resulted in effective sample size being reduced from 50 to 13. For brevity we do not give graphs here.

We also considered bimodal densities f where the respective components were $N(-1.5, \sigma^2)$ and $N(1.5, \sigma^2)$; we took $\sigma = 0.5$ or 1. When $\sigma = 0.5$, f is the normal mixture density #7 of Marron and Wand (1992), while when $\sigma = 1$ it is a density treated by Minnotte (1997), with less clearly separated modes. When $\sigma = 0.5$, only 13 out of 500 datasets gave rise to a density estimate that was not bimodal, but the number rose to 180 when $\sigma = 1$. Enforcing bimodality slightly improved mean squared error performance in the first case, and provided substantial improvement in the second. For the latter results, to reduce computing time we took the modes and antimodes to be at their true values. Again, for brevity we do not give graphs here.

4.3. Buffalo snowfall data

These data represent the snowfalls, measured in inches, at Buffalo, New York, for each of the 63 winters from 1910/11 to 1972/73. Silverman ((1986), p.45), discussed properties of unconstrained estimates, computed using either of two bandwidths: $h = 12$, which gives a unimodal estimate, and $h = 6$, which produces a trimodal estimate. In the latter case the additional modes are small, and become “shoulders” of the unimodal estimate when h is increased to 12.

Using our methods there is no difficulty enforcing unimodality when $h = 6$; doing so has the effect of compressing sample size from 63 to 19. See Figure 2.

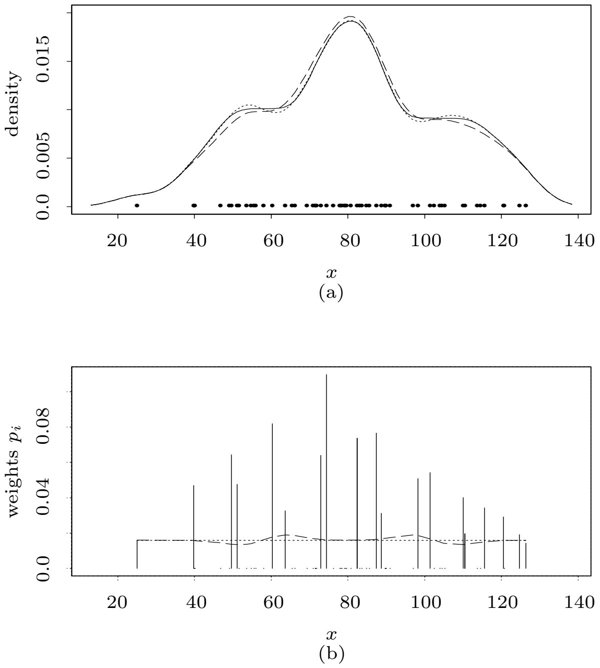

Figure 2. *Buffalo snowfall data.* The dotted line, solid line, and long-dashed line in panel (a) show the unconstrained kernel density estimate and its unimodal-constrained counterparts minimising $I(p)$ and $D_1(p)$, respectively, when $h = 6$. Panel (b) depicts the values of weights $\hat{p}_i$ on achieving the constraint, and the envelope curve for $\bar{p}$ (long-dashed line).

4.4. Chondrite data

The dataset consists of percentages of silica in 22 chondrite meteors. Minnotte (1997) constructed a kernel density estimate having three modes, at 22.76, 27.44, and 33.40, when $h = 0.7$. He also gave an estimate that equalled the first for $x \geq 27.44$, and was closest in L_1 distance to the first subject to the constraint that the mode at 22.76 failed to be statistically significant. However, while this estimate achieves the goal of removing the first mode, it has a flat section where the first mode used to be, and for this reason is not entirely satisfactory. An alternative approach is to remove the first mode by enforcing the constraint of bimodality. The result is the density estimate depicted in Figure 3. For the estimator based on minimising $I(p)$, grid search produces modes at 27.4 and 33.5 and an antimode at 31.05.

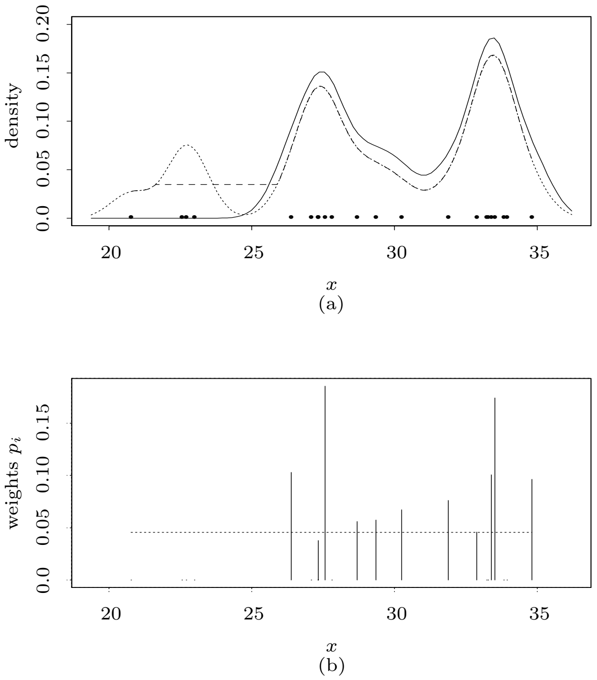

Figure 3. *Chondrite data.* Panel (a) shows the unconstrained kernel estimate with $h = 0.7$ (dotted line), the constrained estimate minimising $I(p)$ (solid line), and the constrained estimate of Minnotte (1997) (dashed line). Panel (b) shows the values of $\hat{p}_i$, plotted against X_i, after the constraint has been achieved.

5. Technical Arguments

Proof of Theorem 3.1. To prove the existence of $\hat{p}$ and $\bar{p}$ it suffices to show that for some p, $\hat{f}(\cdot|p)$ is unimodal. Now, by taking $p_i = 1$ for some $1 \le i \le n$, and of course each other p_i equal to 0, we obtain the unimodal estimator $\hat{f}(x|p) = h^{-1}K\{(x - X_i)/h\}$. In the case of $\bar{p} = \bar{p}(\rho)$ we should note in addition that, since $\rho \neq 0$, this choice of p is legitimate; the corresponding value of $D_\rho(p)$ is not infinite. Since the data have a continuous distribution then the configurations that would prevent $\hat{p}$ and $\bar{p}$ from being unique arise with probability 0.

Proof of Theorem 3.2. Put $\hat{f} = \hat{f}(\cdot|p_{\text{unif}})$. The assumptions in the theorem imply that $\sup|\hat{f} - f| \to 0$ with probability 1, which in turn implies that $\hat{f} \to f$ in L_2 with probability 1. If we show, in the case where distance is measured using $D_\rho(p)$, that $\bar{p}$ exists and satisfies

$$I(\bar{p}) \to 0 \quad \text{with probability } 1\,, \tag{5.1}$$

then the theorem will be proved in that setting. (Uniqueness of $\bar{p}$, given that it exists, again follows from the fact that the data have a continuous distribution.)

Moreover, since $I(\hat{p}) \le I(\bar{p})$, then we shall also have established the theorem in the case where distance is measured using $I(p)$.

Given $0 < \epsilon < 1$ and a probability vector p, let $\mathcal{A} = \mathcal{A}(\epsilon, p)$ denote the set of indices $1 \le i \le n$ such that $|np_i - 1| \le \epsilon$, and let $\widetilde{\mathcal{A}}$ equal the complement of $\mathcal{A}$ in $\{1, \ldots, n\}$. Note that, for each integer $j \ge 1$,

$$S_j \equiv \sup_{-\infty<x<\infty} (nh)^{-1} \sum_{i=1}^{n} L\Big(\frac{x - X_i}{h}\Big)^j = O(1) \tag{5.2}$$

with probability 1. In view of (2.2), and provided $p_i \le C_1 n^{-1}$ for some $C_1 > 1$,

$$I(p) \le \epsilon^2 S_1 + 2\epsilon S_1 n^{-1} \sum_{i\in\widetilde{\mathcal{A}}} |np_i - 1| + \max(C_1 - 1, 1)\, S_1 n^{-1} \sum_{i\in\widetilde{\mathcal{A}}} |np_i - 1|\,. \tag{5.3}$$

If $0 < \rho \le 1$ then for each $\epsilon > 0$ there exists a constant $C_2 = C_2(\epsilon, \rho) > 0$ such that

$$n^{-1} \sum_{i\in\widetilde{\mathcal{A}}} |np_i - 1| \le C_2\, D_\rho(p)\,. \tag{5.4}$$

For example, in the case $\rho = 1$, since $\sum_i p_i = 1$,

$$D_\rho(p) = \sum_{i=1}^{n} \Big\{p_i \log(np_i) - (p_i - n^{-1})\Big\} = n^{-1} \sum_{i=1}^{n} \psi(np_i)\,,$$

where $\psi(x) \equiv x \log(x/e) + 1$. The function ψ is nonnegative, vanishes only at $x = 1$, and satisfies $\psi(x) = O\{(x-1)^2\}$ as $x \to 1$, while $C_2(1, \epsilon)\, \psi(x) \ge |x - 1|$ for a constant $C_2 > 0$, uniformly in $|x - 1| \ge \epsilon$. These properties imply (5.4).

From (5.3) and (5.4) we deduce that $I(p) \le \epsilon^2 S_1 + 2\epsilon S_1 C_2 D_\rho(p) + \max(C_1 - 1, 1)\, S_1 C_2 D_\rho(p)$. Therefore, if we prove that for each fixed $C_1 > 1$ and that with probability 1 for all sufficiently large n,

$$\text{there exists a value } \tilde{p} \text{ of } p \text{ such that } \hat{f}(\cdot|\tilde{p}) \text{ is unimodal, } \sup_i \tilde{p}_i \le C_1 n^{-1}, \text{ and } D_\rho(\tilde{p}) \to 0 \text{ as } n \to \infty, \tag{5.5}$$

then (5.1) will be proved and Theorem 3.2 will follow.

To derive (5.5), note that in view of the conditions imposed on f, for each $0<\delta\le\frac{1}{2}$ there exists a strictly unimodal density f_δ that has support $\mathcal{I}=[a,b]$ for some $-\infty<a<b<\infty$, is continuous and piecewise linear on $\mathcal{I}$, is strictly monotone on each subinterval of $\mathcal{I}$ where it is linear, and satisfies $f_\delta(a+)\, f_\delta(b-)>0$,

$$\sup_{-\infty<x<\infty} |f(x) - f_\delta(x)| \le \delta \quad \text{and} \quad \sup_{-\infty<x<\infty} |f_\delta(x) f(x)^{-1} - 1| \le \tfrac{1}{2}\delta\,. \tag{5.6}$$

(Thus, f_δ has jump discontinuities at the ends of its support.) Let $m \in (a, b)$ be the mode of f_δ. Put $r_i = f_\delta(X_i)/\{nf(X_i)\}$ and $r = (r_1, \ldots, r_n)$. The latter is not necessarily a probability distribution, $\sum_i r_i$ may not equal 1, but nevertheless $\hat{f}(\cdot|r)$ is well-defined, and

$$E\{\hat{f}^{(j)}(x|r)\} = \int K(y)\, f_\delta^{(j)}(x-hy)\, dy \tag{5.7}$$

for $j = 0, 1$.

Without loss of generality, the support of K equals $[-1, 1]$. Using the properties (a) $n^{(1/3)-\epsilon}h \to \infty$ for some $\epsilon > 0$, (b) $\hat{f}^{(j)}(x|r)$, for $j = 1, 2$, equals a sum of independent random variables, and (c) K' and K'' are both Hölder-continuous, it may be proved that with probability 1,

$$\sup_{-\infty<x<\infty} |\hat{f}'(x|r) - E\{\hat{f}'(x|r)\}| = O(n^{-\epsilon_1})\,, \tag{5.8}$$

for some $\epsilon_1 > 0$, and

$$\sup_{-\infty<x<\infty} |\hat{f}''(x|r) - E\{\hat{f}''(x|r)\}| = O\{(n^{1-\epsilon_2}h^5)^{-1/2}\}\,, \tag{5.9}$$

for all $\epsilon_2 > 0$. Moreover, it may be shown from (5.7), and the fact that f_δ is strictly monotone on each subinterval of $\mathcal{I}$ where it is linear, that there exist constants $C_3 > 0$ and $0 < C_4 < 1$, depending only on K and f_δ, such that $E\{\hat{f}'(x|r)\} \geq C_3$ for $a + C_4h \leq x \leq m - C_4h$ and $E\{\hat{f}'(x|r)\} \leq -C_3$ for $m + C_4h \leq x \leq b - C_4h$. From these properties and (5.8) we deduce that, with probability 1 for all sufficiently large n and for this value of C_4,

$$\inf_{a+C_4h\leq x\leq m-C_4h} \hat{f}'(x|r) > 0\,, \qquad \sup_{m+C_4h\leq x\leq b-C_4h} \hat{f}'(x|r) < 0\,. \tag{5.10}$$

Next we investigate $\hat{f}(\cdot|r)$ in the neighbourhood of m, which without loss of generality we take to equal 0. Put $\mu(x) = E\{\hat{f}(x|r)\}$, and let γ_1 and γ_2 denote the gradients of f_δ immediately to the left and right, respectively, of 0. (Then, $\gamma_2 < 0 < \gamma_1$.) It may be proved that for x in a sufficiently small neighbourhood of 0, $\mu''(x) = h^{-1}(\gamma_2 - \gamma_1)\, K(x/h)$. From this property, the unimodality of K, the fact that the support of K equals $[-1, 1]$, the fact that the assumptions on h imposed in Theorem 3.2 imply that if $\epsilon_2 > 0$ is sufficiently small then $(n^{1-\epsilon_2}h^5)^{-1/2}$ is of smaller order than h^{-1}, and result (5.9), it may be proved that, for each $0 < C_4 < 1$, and with probability 1 for all sufficiently large n,

$$\sup_{-C_4h\leq x\leq C_4h} \hat{f}''(x|r) < 0\,. \tag{5.11}$$

Results (5.10) and (5.11), and the continuity of $\hat{f}(\cdot|r)$ imply that, with probability 1 for all sufficiently large n, we have for some $0 < C_4 < 1$:

$\hat{f}(\cdot|r)$ has a unique turning point, $\widehat{m}$ say, in $[a + C_4h, b - C_4h]$, and is strictly increasing on $[a+C_4h, \widehat{m})$ and strictly decreasing on $(\widehat{m}, b-C_4h]$. (5.12)

A similar argument may be used to prove that, if $[\hat{a}, \hat{b}]$ denotes the support of $\hat{f}(x|r)$, then with probability 1 for all sufficiently large n and for $0 < C_4 < 1$,

$$\begin{aligned}&\hat{a} < a < b < \hat{b}, \text{ and } \hat{f}(\cdot|r) \text{ is strictly increasing on}\\ &(\hat{a}, a + C_4 h] \text{ and strictly decreasing on } [b - C_4 h, \hat{b}).\end{aligned} \tag{5.13}$$

(To derive this result we make use of the fact that f_δ has jump discontinuities at the ends of its support.) Results (5.12) and (5.13) imply that

$$\hat{f}(\cdot|r) \text{ is strictly unimodal with mode } \widehat{m}. \tag{5.14}$$

Define $\bar{r} = n^{-1} \sum_i r_i$ and $q_i = (n\bar{r})^{-1} r_i$. In view of (5.6), $\bar{r} \geq 1 - \frac{1}{2}\delta \geq (1+\delta)^{-1}$ and $\bar{r} \leq 1 + \delta \leq (1-\delta)^{-1}$, and so for all i,

$$(1-\delta)\, r_i \leq q_i \leq (1+\delta)\, r_i\,. \tag{5.15}$$

(Hence, $\hat{f}(\cdot|q) = \bar{r}^{-1} \hat{f}(\cdot|r)$ and $(1-\delta)\, \hat{f}(\cdot|r) \leq \hat{f}(\cdot|q) \leq (1+\delta)\, \hat{f}(\cdot|r)$.) The vector $q = (q_1, \ldots, q_n)$ is a proper probability vector. In view of (5.14), with probability 1 for all sufficiently large n, $\hat{f}(\cdot|q)$ is unimodal. And by (5.6) and (5.15), given $\eta > 0$ and $C_1 > 1$, we may choose $\delta > 0$ so small that for all sufficiently large n, $\sup q_i \leq C_1 n^{-1}$ and $D_\rho(q) \leq \eta$. Letting $\eta = \eta(n)$ converge slowly to 0 as $n \to \infty$, and taking $\tilde{p} = q$, we obtain (5.5).

Proof of Theorem 3.3. Let M denote a symmetric, twice continuously differentiable, compactly supported, strictly unimodal probability density, with the property that $M'' < 0$ in a neighbourhood of the origin. Define M_K to be the density formed by convolving M with K, and let $c_1 > 0$. By increasing the scale of M sufficiently greatly we may ensure that $M_K''(u)$ is bounded below 0 uniformly in $|u| \leq c_1$. Equivalently, if we define

$$\psi_j(u) = \int K^{(j)}(v)\, M(u+v)\, dv, \tag{5.16}$$

then

$$\sup_{|u| \leq c_1} \psi_2(u) < 0\,. \tag{5.17}$$

Let m denote the true mode of f. Put $\check{p}_i = n^{-1}[1 + d + c_2 h^2 M\{(m - X_i)/h\}]$ for all i, and let $\tilde{p}_i = \check{p}_i$ if $X_i \in (\alpha, \beta)$, and $\tilde{p}_i = 0$ otherwise, where $c_2 > 0$ is a constant to be chosen later, and $d = d(\alpha, \beta, c_2)$ is a random variable chosen to ensure that $\sum_i \tilde{p}_i = 1$. Take $\check{p} = (\check{p}_1, \ldots, \check{p}_n)$ and $\tilde{p} = (\tilde{p}_1, \ldots, \tilde{p}_n)$; the former will generally not be a probability distribution. Let α, β be as in the statement of Theorem 3.3. Our first task is to prove that

$$\begin{aligned}&\text{for each } \epsilon > 0 \text{ we may select } c_1, c_2 \text{ such that}\\ &\liminf_{n\to\infty} P\Big\{\hat{f}(\cdot|\check{p}) \text{ has just one turning point, } \widehat{m} \text{ say, a local}\\ &\qquad\qquad \text{maximum in } (\alpha, \beta), \text{ and is strictly increasing}\\ &\qquad\qquad \text{in } (\alpha, \widehat{m}) \text{ and strictly decreasing in } (\widehat{m}, \beta)\Big\} \geq 1 - \epsilon\,.\end{aligned} \tag{5.18}$$

(Of course, choosing c_1, c_2 involves selection of the scale of M.)

Note that $\hat{f}(x|\check{p}) = (1+d)\,\hat{f}(x) + \delta(x)$, where

$$\delta(x) = \frac{c_2 h}{n} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big)\, M\Big(\frac{m - X_i}{h}\Big)\,.$$

Given a random variable Z, write $(1-E)Z$ for $Z - E(Z)$. Results of Komlós, Major and Tusnády (1975) may be used to prove that the stochastic process $\Delta_1(x) \equiv (1-E)\,\hat{f}''(x)$ may be approximated by a Gaussian process $\xi = \xi_n$, with zero mean and covariance structure equal to the asymptote of that of Δ_1, such that for each $C > 0$,

$$\sup_{|u| \le C} |\Delta_1(m + hu) - \xi(m + hu)| \to 0 \quad \text{in probability}\,. \tag{5.19}$$

In fact, the covariance of $\xi(m + hu_1)$ and $\xi(m + hu_2)$ equals $\gamma(u_1 - u_2)$, where

$$\gamma(u) = (nh^5)^{-1} f(m) \int K''(v)\, K''(u + v)\, dv\,.$$

Likewise it may be shown that $\Delta_2(x) \equiv (1-E)\,\delta''(x)$ satisfies $\Delta_2(m+hu) \to 0$ in probability, uniformly in $|u| \le C$. Furthermore, $E\{\hat{f}''(m + hu)\} = f''(m) + o(1)$ and $E\{\delta''(m + hu)\} = g(u) + o(1)$, both uniformly in $|u| \le C$, where $g(u) = c_2\, f(m)\, \psi_2(u)$ and ψ_2 is as at (5.16).

Assume for the time being that

$$d \to 0 \quad \text{in probability}\,. \tag{5.20}$$

Combining the results from (5.19) down, and noting (5.17), we see that if $\epsilon > 0$ is given then by choosing c_1, c_2 sufficiently large, depending on C and the limit infimum of nh^5 as $n \to \infty$, we may ensure that

$$\liminf_{n\to\infty} P\Big\{(1 + d)\, \hat{f}''(x) + \delta''(x) < 0 \quad \text{for all} \quad |x - m| \le Ch\Big\} \ge 1 - \epsilon\,.$$

This implies that

$$\liminf_{n\to\infty} P\Big\{\hat{f}(\cdot|\check{p}) \text{ has at most one turning point on } [m - Ch, m + Ch]\Big\} \ge 1 - \epsilon\,. \tag{5.21}$$

Observe too that if K is supported on $[-1, 1]$, and if f is monotone on $[x - h, x + h]$, then

$$|E\hat{f}'(x)| = \int |f'(x - hu)|\, K(u)\, du \ge \tfrac{1}{2}\, |f'(x)|\,. \tag{5.22}$$

Property (I), stated prior to Theorem 3.3, enables it to be proved that for each ϵ, $\eta > 0$, $P\Big\{|(1-E)\,\hat{f}'(x)| \le f(x)^{1/2}\, n^{-(1/5)+\epsilon} \quad \text{for all} \quad x \in \mathcal{T}_\eta\Big\} \to 1$ as $n \to \infty$. (The set $\mathcal{T}_\eta$ was defined just prior to the statement of Theorem 3.3.) In conjunction with property (II) and (5.22) this implies that

$$P\Big[\operatorname{sgn}\{(m-x)\,\hat{f}'(x)\} > 0 \text{ for all } x \in \mathcal{T}_\eta\Big] \to 1 \tag{5.23}$$

as $n \to \infty$. Note too that

$$P\Big\{\hat{f}'(x|\check{p}) = (1+d)\,\hat{f}'(x) \text{ for all } x \in \mathcal{T}_\eta\Big\} \to 1\,. \tag{5.24}$$

Results of Komlós, Major and Tusnády (1975) may be used to show that for $\eta > 0$ fixed but sufficiently small, and assuming (5.20) holds and $C > 0$,

$$\begin{aligned}
&\liminf_{n\to\infty} P\Big[(1+d)\,\hat{f}'(x) + \delta'(x) > 0 \text{ for all } x \in (m-\eta, m-Ch)\,,\\
&\qquad\qquad \text{and } (1+d)\,\hat{f}'(x) + \delta'(x) < 0 \text{ for all } x \in (m+Ch, m+\eta)\Big]\\
&\geq \liminf_{n\to\infty} P\Big[\hat{f}'(x) > 0 \text{ for all } x \in (m-\eta, m-Ch)\,,\\
&\qquad\qquad \text{and } \hat{f}'(x) < 0 \text{ for all } x \in (m+Ch, m+\eta)\Big] \to 1\,,
\end{aligned} \tag{5.25}$$

where the last-stated convergence holds as $C \to \infty$. We give only an outline derivation. The last part of (5.25), i.e., the convergence result, was derived as part of the technical arguments of Mammen, Marron and Fisher (1992). To appreciate why the inequality between the two limit infima is valid, note that the pointwise standard deviation of $\delta'(x)$ is, at $O(h^3)$, an order of magnitude smaller than that of $\hat{f}'(x)$, which is $O(h)$. Furthermore, we may write $(1-E)\,\hat{f}'(x) = h\,\zeta(x) + o_p(h)$, where $\zeta = \zeta_n$ is a Gaussian process whose covariance equals that of $\hat{f}'$. Note too that, uniformly in x in a neighbourhood of the origin, $E\{\hat{f}'(x)\} = f'(x) + o(h)$; uniformly in $|u| \leq D$ for any $D > 0$, $f'(m+hu) = hu\,f''(m) + o(h)$; and uniformly in $-\infty < u < \infty$, $E\{\delta'(m+hu)\} = c_2\,f(m)\,h\,\psi_1(u) + o(h)$, where ψ_1 is defined at (5.16). Therefore, to derive the inequality between the limit infima at (5.25) it suffices to show that, with $u = u(x) = (x-m)/h$,

$$\begin{aligned}
&\liminf_{n\to\infty} P\Big[\zeta(x) + h^{-1}\,f'(x) + c_2\,f(m)\,\psi_1(u) > 0 \text{ for all } x \in (m-\eta, m-Ch)\,,\\
&\qquad\qquad \text{and } \zeta(x) + h^{-1}\,f'(x) + c_2\,f(m)\,\psi_1(u) < 0 \text{ for all } x \in (m+Ch, m+\eta)\Big]\\
&\geq \liminf_{n\to\infty} P\Big[\zeta(x) + h^{-1}\,f'(x) > 0 \text{ for all } x \in (m-\eta, m-Ch)\,,\\
&\qquad\qquad \text{and } \zeta(x) + h^{-1}\,f'(x) < 0 \text{ for all } x \in (m+Ch, m+\eta)\Big]\,.
\end{aligned} \tag{5.26}$$

Now, the function $\psi_1(u)$ is nonnegative for $u < 0$ and nonpositive for $u > 0$ (by virtue of the fact that the convolution of K and M is unimodal with mode 0), and so it shares the parity of the function $f'(x) = f'(m+hu)$ appearing in (5.26).

Therefore, adding the term $c_2\, f(m)\, \psi_1(u)$ to the argument of the probability on the right-hand side of (5.26), thereby obtaining the probability on the left-hand side, only increases the value of the probability in its limit.

Combining (5.20), (5.23), (5.24) and (5.25) we deduce that

$$\begin{aligned}
&\liminf_{n\to\infty} P\Big\{\hat{f}'(x|\check{p}) > 0 \text{ for all } x \in (\alpha, m - Ch)\,, \\
&\qquad\qquad\qquad \text{and } \hat{f}'(x|\check{p}) < 0 \text{ for all } x \in (m + Ch, \beta)\Big\} \\
&\geq \liminf_{n\to\infty} P\Big\{\hat{f}'(x) > 0 \text{ for all } x \in (\alpha, m - Ch)\,, \\
&\qquad\qquad\qquad \text{and } \hat{f}'(x) < 0 \text{ for all } x \in (m + Ch, \beta)\Big\} \to 1\,,
\end{aligned}$$

where the last-stated convergence holds as $C \to \infty$. Combining this result with (5.21) we deduce that (5.18) holds.

We still need to verify (5.20), however. To this end, let $\mathcal{I} = \mathcal{I}(\alpha, \beta)$ denote the set of indices i, $1 \leq i \leq n$, such that $X_i \notin (\alpha, \beta)$, and write $N = N(\alpha, \beta)$ for the number of elements of $\mathcal{I}$. For all sufficiently large n, $M\{(m - X_i)/h\} = 0$ for all $X_i \notin (\alpha, \beta)$, and so with probability converging to 1,

$$1 = \sum_{i=1}^{n} \tilde{p}_i = \frac{(1 + d)\,(n - N)}{n} + \frac{c_2 h^2}{n} \sum_{i=1}^{n} M\Big(\frac{m - X_i}{h}\Big)\,.$$

The last-written term is $O_p(h^3)$. Therefore, solving the displayed equation for d, and noting that $E(N/n) = \lambda \to 0$, we deduce that

$$d = n^{-1} N + O_p(h^3) = O_p(\lambda + h^3)\,, \tag{5.27}$$

which establishes (5.20).

It follows from the unimodality of K that if (5.18) holds then it remains true when $\hat{f}(\cdot|\check{p})$ is replaced by $\hat{f}(\cdot|\tilde{p})$ and α, β are replaced by the lower and upper extremities, $\hat{\alpha}$ and $\hat{\beta}$ say, of the support of $\hat{f}(\cdot|\tilde{p})$:

$$\begin{aligned}
&\text{for each } \epsilon > 0 \text{ we may choose } c_1, c_2 \text{ such that} \\
&\liminf_{n\to\infty} P\Big\{\hat{f}(\cdot|\tilde{p}) \text{ is strictly unimodal in } (\hat{\alpha}, \hat{\beta})\Big\} \geq 1 - \epsilon\,.
\end{aligned} \tag{5.28}$$

Indeed, if K is supported on $[-1, 1]$ then, with probability converging to 1 as $n \to \infty$, $\hat{\alpha} < \alpha < \beta < \hat{\beta}$ and $\hat{f}(\cdot|\check{p}) = \hat{f}(\cdot|\tilde{p})$ on $(\alpha + h, \beta - h)$. (The former result follows from (III).) Furthermore, with probability 1, $\hat{f}'(\cdot|\check{p}) \leq \hat{f}'(\cdot|\tilde{p})$ on $(\alpha, \alpha + h)$ (since $\hat{f}'(\cdot|\tilde{p})$ differs from $\hat{f}'(\cdot|\check{p})$ only in that, in this range, some of the kernel components that contribute negative gradients are missing from the former), and $\hat{f}'(\cdot|\tilde{p}) > 0$ on $(\hat{\alpha}, \alpha)$ (since all the kernel components that contribute to $\hat{f}(\cdot|\tilde{p})$ in this range have strictly positive gradients). These results have obvious analogues in the upper tail.

Let $\mathcal{J}$ be the set of indices i such that $(m - X_i)/h$ lies in the support of M. (Then, for all sufficiently large n, $\mathcal{J}$ is a subset of the complement of $\mathcal{I}$.) In view of (5.27) and the definition of $\tilde{p}_i$, $|n\tilde{p}_i - 1| = 1$ for $i \in \mathcal{I}$, $|n\tilde{p}_i - 1| = O_p(\lambda + h^2)$ uniformly in $i \in \mathcal{J}$, and $|n\tilde{p}_i - 1| = O_p(\lambda + h^3)$ uniformly in $i \notin \mathcal{I} \cup \mathcal{J}$. Since the numbers of elements of $\mathcal{I}$ and $\mathcal{J}$ equal $O_p(n\lambda)$ and $O_p(nh)$, respectively, then, for $0 < \rho \leq 1$,

$$\begin{aligned} D_\rho(\tilde{p}) &= O_p\biggl[n^{-1} \sum_{i=1}^{n} \min\left\{|n\tilde{p}_i - 1|, (n\tilde{p}_i - 1)^2\right\}\biggr] \\ &\leq O_p\{\lambda + h\,(\lambda + h^2)^2 + (\lambda + h^3)^2\} = O_p(\lambda + n^{-1})\,. \qquad (5.29) \end{aligned}$$

Similarly, if $\mathcal{K}$ denotes the complement of $\mathcal{I} \cup \mathcal{J}$ in $\{1, \ldots, n\}$, and $S(\mathcal{H}_1, \mathcal{H}_2)$ represents the sum of $L_{ij} \equiv L\{(X_i - X_j)/h\}$ over $i \in \mathcal{H}_1$ and $j \in \mathcal{H}_2$, where $\mathcal{H}_1$ and $\mathcal{H}_2$ range over $\mathcal{I}$, $\mathcal{J}$, $\mathcal{K}$, then we may show from (2.2) and the definition of $\tilde{p}$ that

$$\begin{aligned} I(\tilde{p}) = O_p\Bigl[(n^2h)^{-1}\{&S(\mathcal{I},\mathcal{I}) + (\lambda + h^2)^2\, S(\mathcal{J},\mathcal{J}) + (\lambda + h^3)^2\, S(\mathcal{K},\mathcal{K}) \\ &+(\lambda + h^2)\, S(\mathcal{I},\mathcal{J}) + (\lambda + h^3)\, S(\mathcal{I},\mathcal{K}) + (\lambda + h^2)\,(\lambda + h^3)\, S(\mathcal{J},\mathcal{K})\}\Bigr]\,. \end{aligned}$$

The number of elements of $\mathcal{J}$ equals $O_p(nh)$, and so $S(\mathcal{J},\mathcal{J}) = O_p(n^2h^2)$. The expected value (conditional on X_i) of the sum of L_{ij} over $j \in \mathcal{K}$ with $j \neq i$ equals $O_p(nh)$, uniformly in i, and so $S(\mathcal{K},\mathcal{K}) = O_p(n^2h + n) = O_p(n^2h)$. Also, $S(\mathcal{I},\mathcal{J}) = 0$ for all sufficiently large n, since L is compactly supported. The expected value (conditional on X_i) of the sum of L_{ij} over $j \in \mathcal{K}$, with $j \neq i$, equals $O_p(nh)$, uniformly in i, and the expected number of elements of $\mathcal{I}$ equals $O(n\lambda)$, so $S(\mathcal{I},\mathcal{K}) = O_p(n^2\lambda h)$. Likewise, since the expected number of elements of $\mathcal{J}$ equals nh, $S(\mathcal{J},\mathcal{K}) = O_p(n^2h^2)$. Combining the results in this paragraph we deduce that

$$I(\tilde{p}) = O_p\{(n^2h)^{-1}\, S(\mathcal{I},\mathcal{I}) + \lambda^2 + n^{-1}\}\,. \qquad (5.30)$$

Let $\mathcal{I}_A$, $\mathcal{I}_B$ denote the sets of indices i such that $X_i \leq \alpha$, $X_i \geq \beta$, respectively, and note that L is supported on $[-2, 2]$. Then the sum of the off-diagonal terms in $E\{S(\mathcal{I}_A,\mathcal{I}_A)\}$ is bounded above by

$$n^2h \int_{-\infty}^{\alpha} f(x)\,dx \int_{-\infty}^{\infty} L(u)\,f(x - hu)\,du \leq n^2h \int_{-\infty}^{\alpha} f(x)\,f(x + 2h)\,dx \leq n^2h\Lambda\,.$$

The sum of the diagonal terms is of no more than this order. The expected value of the off-diagonal terms in $E\{S(\mathcal{I}_B,\mathcal{I}_B)\}$ may be bounded analogously, as n^2h times the integral from β to ∞ of $f(x)\,f(x - 2h)$, and the sum of the diagonal terms is not of larger order. Since K is compactly supported, $E\{S(\mathcal{I}_A,\mathcal{I}_B)\} = E\{S(\mathcal{I}_B,\mathcal{I}_A)\} = 0$ for all sufficiently large n. Combining these results we deduce that $E\{S(\mathcal{I},\mathcal{I})\} = O(n^2 h\Lambda)$, and hence, from (5.30), that

$$I(\tilde{p}) = O_p(\Lambda + \lambda^2 + n^{-1})\,. \tag{5.31}$$

Note too that, using (5.20) and the definition of $\tilde{p}$,

$$\sup_i n\tilde{p}_i \to 1 \quad \text{in probability}\,. \tag{5.32}$$

Define S_2 as at (5.2). We may prove from (2.2), using the Cauchy-Schwarz inequality, that for any probability distribution p,

$$\begin{aligned} I(p) &\le (n^2h)^{-1} \Big\{ \sum_{i=1}^n \sum_{j=1}^n (np_i - 1)^2\,(np_j - 1)^2 \Big\}^{1/2} \Big\{ \sum_{i=1}^n \sum_{j=1}^n L\Big(\frac{X_i - X_j}{h}\Big)^2 \Big\}^{1/2} \\ &\le S_2^{1/2}\, h^{-1/2}\, n^{-1} \sum_{i=1}^n (np_i - 1)^2\,. \end{aligned}$$

If $\sup_i np_i \le C_1$ for a fixed constant $C_1 > 1$, then $n^{-1} \sum_i (np_i - 1)^2$ is dominated by $C_2\, D_\rho(p)$, where $C_2 = C_2(C_1, \rho)$ does not depend on n or otherwise on p. Therefore,

$$I(p) \le C_2\, S_2^{1/2}\, h^{-1/2}\, D_\rho(p)\,. \tag{5.33}$$

From (5.28), (5.29) and (5.32) we deduce that a unimodal probability density is achievable by weighting using a probability vector $\tilde{p}$ for which both $D_\rho(\tilde{p}) = O_p(\lambda + n^{-1})$ and $\sup_i n\tilde{p}_i \to 1$ in probability. Therefore, the distribution $p = \bar{p}$ that minimises $D_\rho(p)$ subject to unimodality and $\sup_i n\bar{p}_i \le C_1$ must satisfy $D_\rho(\bar{p}) = O_p(\lambda + n^{-1})$. From this result, taking $p = \bar{p}$ in (5.33) and using (5.2), we deduce that

$$I(\bar{p}) = O_p\{h^{-1/2}(\lambda + n^{-1})\}\,. \tag{5.34}$$

More simply, if $p = \hat{p}$ minimises $I(p)$ subject to unimodality of $\hat{f}(\cdot|p)$ then, by (5.31),

$$I(\hat{p}) = O_p(\Lambda + \lambda^2 + n^{-1})\,. \tag{5.35}$$

Note that

$$\Big|\, \|\hat{f}(\cdot|p) - f\| - \|\hat{f}(\cdot|p_{\rm unif}) - f\| \,\Big| \le I(p)^{1/2}\,. \tag{5.36}$$

Since f'' is square-integrable then the mean integrated squared error of $\hat{f}(\cdot|p_{\rm unif})$, MISE say, equals $O(n^{-4/5})$. Indeed, writing κ and ϕ for the respective characteristic functions of K and f we have

$$\text{MISE} \le (nh)^{-1} \int K^2 + (2\pi)^{-1} \int \{\kappa(ht) - 1\}^2\, |\phi(t)|^2\, dt = O\{(nh)^{-1} + h^4\}\,.$$

Therefore,

$$\|\hat{f}(\cdot|p_{\text{unif}}) - f\| = O_p(n^{-2/5})\,. \tag{5.37}$$

A longer argument may be used to prove that

$$\text{MISE}^{-1}\,\|\hat{f}(\cdot|p_{\text{unif}}) - f\|^2 \to 1 \quad \text{in probability, and} \quad \text{MISE} \asymp n^{-4/5}\,; \tag{5.38}$$

the methods of Hall (1984) are employed to derive the first of these properties. By (5.35), (5.36) [with $p = \hat{p}$] and (5.37), we have $\|\hat{f}(\cdot|\hat{p}) - f\|^2 = O_p(\Lambda + \lambda^2 + n^{-4/5})$, which is result (3.1) in Theorem 3.3. Similarly, (3.2) follows from (5.34), (5.36) [with $p = \bar{p}$] and (5.37). Using (5.38) instead of (5.37) in these arguments we obtain (3.3) and (3.4).

Acknowledgement

The helpful comments of two reviewers are gratefully acknowledged.

References

Bickel, P. J. and Fan, J. (1996). Some problems on the estimation of unimodal densities. *Statist. Sinica* **6**, 23-45.

Birgé, L. (1997). Estimation of unimodal densities without smoothness assumptions. *Ann. Statist.* **25**, 970-981.

Cheng, M.-Y., Gasser, T. and Hall, P. (1999). Nonparametric density estimation under unimodality and monotonicity constraints. *J. Computat. Graph. Statist.* **8**, 1-21.

Cressie, N. A. C. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46**, 440-464.

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

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

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivariate Anal.* **14**, 1-16.

Hall, P. and Huang, L.-S. (2001a). Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.* **29**, 624-647.

Hall, P. and Huang, L.-S. (2001b). Nonparametric estimation of hazard rate under the constraint of monotonicity. *J. Comput. Graph. Statist.* **10**, 592-614.

Hall, P. and Presnell, B. (1999). Intentionally biased bootstrap methods. *J. Roy. Statist. Soc. Ser. B* **61**, 143-158.

Komlós, J., Major, P. and Tusnády, G. (1975). An approximation of partial sums of independent RV's, and the sample DF. I. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **32**, 111-131.

Mammen, E., Marron, J. S. and Fisher, N. I. (1992). Some asymptotics for multimodality tests based on kernel density estimates. *Probab. Theory Related Fields* **91**, 115-132.

Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20**, 712-736.

Minnotte, M. C. (1997). Nonparametric testing of the existence of modes. *Ann. Statist.* **25**, 1646-1660.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-249.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90-120.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53**, 683-690.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43**, 97-99.

Silverman, B. W. (1983). Some properties of a test for multimodality based on kernel density estimates. In *Probability, Statistics and Analysis* (Edited by J. F. C. Kingman and G. E. H. Reuter), 248-259. Cambridge University Press, Cambridge, UK.

Wang, Y. (1995). The L1 theory of estimation of monotone and unimodal densities. *J. Nonparametric Statist.* **4**, 249-261.

Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia.

E-mail: halpstat@pretty.anu.edu.au

Department of Biostatistics, University of Rochester Medical Center, Rochester, NY 14642, U.S.A.

E-mail: Lhuang@bst.rochester.edu

(Received May 2001; accepted February 2002)