Statistica Sinica **15**(2005), 73-98

UNIMODAL KERNEL DENSITY ESTIMATION BY DATA SHARPENING

Peter Hall[1] and Kee-Hoon Kang[1,2]

[1]*Australian National University and* [2]*Hankuk University of Foreign Studies*

Abstract: We discuss a robust data sharpening method for rendering a standard kernel estimator, with a given bandwidth, unimodal. It has theoretical and numerical properties of the type that one would like such a technique to enjoy. In particular, we show theoretically that, with probability converging to 1 as sample size diverges, our technique alters the kernel estimator only in places where the latter has spurious bumps, and is identical to the kernel estimator in places where that estimator is monotone in the correct direction. Moreover, it automatically splices together, in a smooth and seamless way, those parts of the estimator that it leaves unchanged and those that it adjusts. Provided the true density is unimodal our estimator generally reduces mean integrated squared error of the standard kernel estimator.

Key words and phrases: Bandwidth, data sharpening, heavy tailed distributions, kernel methods, mean squared error, nonparametric density estimation, order constraint, tilting methods, unimodal density.

1. Introduction

Motivation for the assumption of unimodality usually has a Bayesian connection. It expresses a prior belief that the sampled distribution is homogeneous. Imposing it as a constraint, when constructing a nonparametric density estimator, is arguably the most effective way of incorporating knowledge of homogeneity into the final result, without sacrificing the particularly advantageous adaptivity conferred by nonparametric methods. The constraint of unimodality also makes the estimator particularly robust against undersmoothing, which, in the absence of a restriction on the number of modes, can seriously impair the qualitative appearance of the estimator.

There is an extensive literature on function estimation under shape constraints. Recently treated methods for unimodal density estimation include those considered by Wang (1995), Bickel and Fan (1996), Birgé (1997), Cheng, Gasser and Hall (1999) and Hall and Presnell (1999). A technique based on data sharpening was suggested by Braun and Hall (2001), but without any theoretical support and in the absence of clear guidance as to choice of distance function. In the

present paper we demonstrate, both theoretically and in numerical terms, that a version of data sharpening based on L_1 loss, or using closely related distance functions, performs particularly well. It produces smooth estimators with very good mean squared error performance, and with aesthetically attractive features which few other approaches enjoy.

When used for unimodal density estimation, data sharpening usually starts from a conventional kernel density estimator. It involves modifying the sample so as to ensure the standard estimator, applied to the altered rather than the original sample, is unimodal. In particular, for a given bandwidth and kernel the data are moved the least amount subject to the kernel estimator having the unimodal property. Provided the kernel is a probability density, altering the sample does not affect the basic properties that make standard kernel methods so attractive, for example the fact that they are nonnegative and integrate to 1.

If the sampled density truly is unimodal then, at least for moderately large samples, we expect the standard kernel estimator to depart from unimodality only in the tails and in the vicinity of the true mode. These are the places where the true density is relatively flat. There, a standard kernel density estimator tends to suffer from spurious wiggles that prevent it from reflecting qualitative features of the sampled density. Hence, in principle it is necessary only to modify the estimator in such places. We shall show that these are exactly the places where the data sharpening algorithm adjusts the data, and that with high probability it does not alter any data that are not very close to the mode or some distance out in one or other of the tails. Likewise, it alters the standard kernel estimator itself only very close to the mode or out in the tails. And it automatically splices together the adjusted and original forms of the kernel estimator at points where they join. The final estimator is very smooth; it enjoys as many derivatives as the kernel that was used in its construction.

Of course, these goals could be achieved in other ways, using more explicit techniques. One such approach would be to draw horizontal lines at appropriate heights across unwanted "valleys" in the conventional kernel estimator, so as to fill them in; to incorporate a degree of monotone smoothing at places where the lines met the graph of the conventional estimator, so as to ensure the final estimator was smooth; and to renormalise the result, so it integrated to 1. However, on account of the normalisation step this approach does not have the property that it equals the standard kernel estimator in places where the latter does not need adjustment. That can lead to significant performance difficulties, as we shall explain shortly.

Moreover, the fact that the explicitly "linearised" density estimator has perfectly flat sections means that it conveys the unwanted visual impression that there is something intrinsically special and interesting about those parts of the

true density. This can be a particular problem in the tails of the linearised density estimator, where a graph of the estimator will often decrease to zero in an eye-catching sequence of flat steps with curved edges. These and other aspects of the explicitly linearised estimator are of course no more than artifacts of the technique employed to construct it, although that is usually far from obvious to the casual observer. Such issues might be unimportant if the use to which the final estimate was put was a purely quantitative one, not influenced by the qualitative appearance of a graph of the estimator, but that is often not the case in practice.

The majority of these difficulties can be removed by appropriately tuning the explicitly constructed estimator. However, the total number of subsidiary smooths and tapers that are required can be large, and developing a totally objective and effective procedure for implementing them produces an inelegant and tedious procedure, involving methodology that is aesthetically unattractive. Moreover, in the case of relatively heavy-tailed densities the linearisation step can add significantly to the probability mass of the density estimator. While this problem is eliminated by the normalisation step, the latter often creates difficulties of its own, by significantly increasing or reducing the height of the density estimator in the body of the distribution. That can impair performance, for example by seriously increasing mean squared error.

A related phenomenon occurs in the case of unimodal density estimation using data tilting methods (Hall and Huang (2002)), where the need to remove spurious wiggles in the tails of a conventional density estimator can result in a detrimental increase in the density estimator at other places, leading to poor mean squared error performance. Numerical results illustrating this point will be summarised in Section 4, and related phenomena can be observed for unimodal density estimators constructing using linearisation and related techniques.

In principle, data sharpening can be used to ensure unimodality for other estimator types. However, the extreme simplicity of standard kernel estimators, and the fact that they can be readily used with a fixed bandwidth, make them especially attractive on both aesthetic and computational grounds. By way of contrast, local likelihood methods for density estimation, which usually require spatially varying bandwidths in order to be implemented effectively, are unattractive.

The distance, $\widehat{D}$ say, through which the dataset has to be moved in order to ensure unimodality, can be used to test the null hypothesis that the density is unimodal. The hypothesis would be rejected if $\widehat{D}$ were too large. The null distribution could be estimated using bootstrap methods, by resampling from a unimodal distribution. One candidate for the latter would be the constrained distribution of the test statistic, although other options are available.

Multivariate applications of our method are also possible. Indeed, in the setting of two or more dimensions there are relatively few alternative approaches. Again, the distance through which the data have to be moved could be used as the basis for a test of unimodality.

Some early methods for density estimation under qualitative constraints, starting from Grenander's (1956) introduction of the technique that is now often referred to as nonparametric maximum likelihood, employed data tilting. See Prakasa Rao (1969) for an application to unimodal density estimation. Recent applications of related ideas include work of Bickel and Fan (1996) and Hall and Huang (2002), on unimodal density estimation. Fougères (1997) has shown that the method of monotone rearrangement, suggested by Hardy, Littlewood and Pólya (1952) and applied to a kernel density estimator, produces a consistent unimodal estimator when the true density is unimodal. Methods for estimating regression functions under qualitative constraints include those suggested by Friedman and Tibshirani (1984), Ramsay (1988), Kelly and Rice (1990), Qian (1994), Tantiyaswasdikul and Woodroofe (1994), Delecroix, Simioni and Thomas-Agnan (1995), Mammen and Thomas-Agnan (1999) and Hall and Huang (2001). Data sharpening methods were surveyed by Braun and Hall (2001).

The remainder of this paper is organised as follows. Section 2 introduces methodology. Main theoretical results are summarised in Section 3, and numerical properties are outlined in Section 4. Technical arguments for Section 3 are given in Section 5.

2. Methodology

A density f defined on the real line is said to be unimodal if there exists a point m_f, called a mode of f, such that f is increasing on $(-\infty, m_f)$ and decreasing on (m_f, ∞). (Here and below we use the terms "increasing" and "decreasing" to mean "nondecreasing" and "nonincreasing", respectively.) Assuming f has a continuous derivative on the real line we say that f is uniquely unimodal, or equivalent that it is unimodal with a unique mode, if there exists a unique point m_f in the interior of the support of f such that $f'(m_f) = 0$. Note that the definition of unique unimodality excludes densities with shoulders, as well as those with more than one turning point.

Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ denote the original dataset drawn from the population with density f. The conventional kernel estimator of f is

$$\hat{f}_{\mathcal{X}}(x) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.1}$$

where h is a bandwidth and K a kernel function. We wish to perturb the data as little as possible such that the constraint of unimodality is satisfied. To this end,

let $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$ and $\hat{f}_{\mathcal{Y}}$ be respectively a sharpened dataset derived from $\mathcal{X}$, and the corresponding kernel density estimator obtained on replacing $\mathcal{X}$ by $\mathcal{Y}$ at (2.1). We take $\mathcal{Y}$ to be the minimiser of $D(\mathcal{X}, \mathcal{Y}) = \sum_i \Psi(X_i - Y_i)$ subject to $\hat{f}_{\mathcal{Y}}$ being unimodal, where Ψ denotes a distance function. Examples of functions Ψ include $\Psi(x) = |x|^p$, to which we shall refer as L_p distance, and more general distance measures analogous to those used in the theory of M estimation, such as $\Psi(x) = \int_0^x \psi(y)\, dy$ where ψ denotes an antisymmetric function that is positive on the positive half line.

Forcing $\hat{f}_{\mathcal{Y}}$ to be unimodal means ensuring the existence of a quantity $\widehat{m}$ such that $\hat{f}'_{\mathcal{Y}}(x) \geq 0$ for $x \leq \widehat{m}$, and $\hat{f}'_{\mathcal{Y}}(x) \leq 0$ for $x \geq \widehat{m}$. In practical numerical terms we first choose a candidate m for $\widehat{m}$, and then choose $\mathcal{Y}(m)$ to minimise $\sum_i \Psi(X_i - Y_i)$ subject to $\hat{f}_{\mathcal{Y}}$ being unimodal. Finally, writing $\Delta(m)$ for the corresponding value of $\sum_i \Psi(X_i - Y_i)$, we select $\widehat{m}$ to minimise $\Delta(m)$. (In practice, we found that $\widehat{m}$ did not depend on the starting candidate.) Thus, our method produces an estimator of the mode as well as a unimodal density estimator. However, we shall not explore properties of $\widehat{m}$ here, except to note that, while its bias and error about the mean are of the same orders as those of the conventional mode estimator (defined as the point at which $\hat{f}_{\mathcal{X}}$ attains its global maximum), our $\widehat{m}$ is not asymptotically normally distributed.

As we in Section 4, excellent statistical performance, especially for heavy tailed distributions, is obtained very generally using distance measures such as L_1. In particular, for all the distributions with which we have experimented, this approach equals or outperformed methods based on L_p distance for $p > 1$. The excellent statistical properties of our methods based on L_1-type distances arise from the distance function being asymptotically linear, rather than increasing at a faster rate, for large values of its argument. This feature implies that only a relatively small penalty is imposed for moving an outlying data value a long distance to the main body of the sample, in order to ensure unimodality. In contrast, when using L_2 distance the algorithm tends to move data from the middle of the distribution to the tails, as well as moving them in the opposite direction; it shifts many data by small amounts, rather than a small number of data by a large amount, and this almost invariably impairs performance.

However, the fact that the distance function $|x|$ has a discontinuous derivative at the origin means that when using L_1 distance one tends to experience numerical difficulties in 10% to 20% of samples, depending on distribution type, when attempting to minimise $\sum_i |X_i - Y_i|$ subject to $\hat{f}_{\mathcal{Y}}$ being unimodal. This problem can be overcome by smoothing the loss function at its vertex.

One approach is to use a function such as $\Psi = \Psi_{\tan}$, constructed with $\psi(x) = \arctan(x)$. This Ψ is asymptotically linear, and perfectly convex, but has a smooth bowl shape at the origin, so its statistical performance is close to that for

L_1 distance, without the latter's numerical drawbacks. As in the setting of robust statistical methods, there are many opportunities for improving performance, for instance by using a non-convex loss function corresponding to a redescending weight. An example is the function Ψ based on $\psi(x) = \sin(x/a)\, I(|x| < \pi a)$. (Here, the parameter a would have to be chosen.) For discussion of this and many other opportunities for choosing ψ, see, for example, Andrews *et al.* (1972).

In numerical work in Section 4 we compare the above data sharpening methods with a data tilting approach suggested by Hall and Huang (2002), based on ideas discussed by Hall and Presnell (1999). To construct a unimodal density estimator by tilting the empirical distribution, we change the data weights instead of altering the data themselves. Accordingly, the kernel density estimator becomes

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

where $p_1, \ldots, p_n$ are chosen to minimise a measure $D(p)$ of the distance between the multinomial distribution $p = (p_1, \ldots, p_n)$ on n points, and the corresponding uniform distribution, subject to $\hat{f}(x|p)$ being unimodal. The resulting constrained density estimator will be denoted by $\hat{f}_{\rm tilt}$.

In Section 4 we take $D_1(p) = n \sum_i p_i \log(np_i)$, this being a particular form of power divergence (Cressie and Read (1984)). It gives relatively good performance in the context of unimodal density estimation, largely because it is robust against aberrations caused by reducing one or more weights p_i to 0. That operation is frequently necessary in order to eliminate outlying data that cause spurious bumps in the tails of unconstrained kernel density estimators. In comparison, the more conventional power divergence measure, $D_0(p) = -\sum_i \log(np_i)$, which is infinite when one or more values of p_i vanish, often leads to undefinable density estimators if the imposed constraint is unimodality. Nevertheless, even when data tilting uses D_1 it is outperformed by data sharpening based on $\Psi_{\rm tan}$.

3. Theoretical Properties

There is a variety of approaches to establishing theory for data sharpening unimodal density estimators, many of them tailored to specific distance measures D. In most instances the theory shows that in the case of compactly supported kernels, with high probability and away from the mode and the tails of the sampled distribution, the sharpened estimator $\hat{f}_{\mathcal{Y}}$ is identical to its conventional counterpart. For brevity and simplicity we shall describe only one approach here, developed for the case of L_1 distance; see (3.1) below. It has the advantage that, under a mild and straightforward condition, (3.3), the algorithm based on minimising L_1 distance guarantees not only that $\hat{f}_{\mathcal{Y}} = \hat{f}_{\mathcal{X}}$ along "most" of the support of f, but also that specific convergence rates are achieved uniformly on

the line. See Theorem 3.1. Alternative approaches, based on more general loss functions, do not appear to allow such a simple and elegant description of the properties of $\hat{f}_{\mathcal{Y}}$.

All our results extend easily to the case of estimating f under the constraint that it has $k \geq 1$ modes, assuming this condition is incorporated into the basic algorithm and that the modes of $\hat{f}_{\mathcal{Y}}$ are required to be at least a certain fixed, sufficiently small distance apart.

We initially assume the kernel K is a symmetric, compactly supported, uniquely unimodal probability density with two bounded derivatives. Call this condition $(\mathrm{C}_{K,1})$. Theorems 3.1–3.3 below hold without change if K is taken to be the Gaussian kernel, although the proofs are more elaborate in that case. However, if K is not compactly supported then it is not generally true that $\hat{f}_{\mathcal{Y}} = \hat{f}_{\mathcal{X}}$, with high probability, for most of the support of f.

Given a data sharpening algorithm $\mathcal{A}$ that takes the original dataset $\mathcal{X} = \{X_1, \ldots, X_n\}$ to $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$, where Y_i denotes the image of X_i, we define

$$D(\mathcal{X}, \mathcal{Y}) = \sum_{i=1}^{n} |X_i - Y_i| \,. \tag{3.1}$$

We apply subscripts to $\mathcal{A}$ and $\mathcal{Y}$ to denote specific algorithms and the sharpened datasets that they produce, respectively. In particular, let $\mathcal{A}_0$ denote the version of $\mathcal{A}$ that selects $\mathcal{Y} = \mathcal{Y}_0$ to minimise $D(\mathcal{X}, \mathcal{Y})$ subject to $\hat{f}_{\mathcal{Y}}$ being unimodal.

Let $\mathcal{F}_0$ be the class of unimodal densities f with the properties: the support of f is an interval with endpoints $a_f < b_f$ (not necessarily finite), the mode m_f of f is unique and lies in (a_f, b_f), f has a bounded, uniformly continuous derivative on $(-\infty, \infty)$, and $|f'|$ is bounded away from 0 on the set $\mathcal{S}(c) = [c_1, c_2] \cup [c_3, c_4]$ whenever $c = (c_1, \ldots, c_4)$ satisfies

$$a_f < c_1 < c_2 < m_f < c_3 < c_4 < b_f \,. \tag{3.2}$$

Put $h_0 = n^{-1/5}$. The notation $h \asymp h_0$ means that the ratio h/h_0 is bounded away from zero and infinity as $n \to \infty$.

Theorem 3.1. *Let $f \in \mathcal{F}_0$, and assume D is given by* (3.1)*, that* $(\mathrm{C}_{K,1})$ *holds and that $h \asymp h_0$ as $n \to \infty$. Suppose too that there exists a particular data sharpening algorithm $\mathcal{A}_1$, not necessarily $\mathcal{A}_0$, which with probability* 1 *produces a unimodal density estimator $\hat{f}_{\mathcal{Y}}$, and which moves $\mathcal{X}$ a distance $D(\mathcal{X}, \mathcal{Y}_1)$ to $\mathcal{Y}_1$, where*

$$h_0^2 \, D(\mathcal{X}, \mathcal{Y}_1) \to 0 \tag{3.3}$$

with probability 1*. Then when $\mathcal{A}_0$ is used instead of $\mathcal{A}_1$, c satisfies* (3.2) *and $n \to \infty$,*

$$P_f\{Y_i = X_i \text{ for each } i \text{ such that either } X_i \text{ or } Y_i \text{ is in } \mathcal{S}(c)\} \to 1, \tag{3.4}$$

$$\sup_{-\infty<x<\infty} |\hat{f}_{\mathcal{Y}}(x) - \hat{f}_{\mathcal{X}}(x)| = o_p(h_0)\,, \qquad \sup_{-\infty<x<\infty} |\hat{f}'_{\mathcal{Y}}(x) - \hat{f}'_{\mathcal{X}}(x)| = o_p(1)\,. \quad (3.5)$$

The algorithm $\mathcal{A}_0$ is not necessarily uniquely defined, and results such as (3.4) should be interpreted as holding uniformly in all versions of $\mathcal{A}_0$. That is, the probability that there is a version of $\mathcal{A}_0$ for which $Y_i \neq X_i$ for some i such that either X_i or Y_i is in $\mathcal{S}(c)$, converges to zero as $n \to \infty$. Non-uniqueness of $\mathcal{A}_0$ can occur with positive probability when $n = 2$, although we conjecture that when ($\mathrm{C}_{K,1}$) is satisfied and $n \geq 3$, $\mathcal{A}_0$ is uniquely defined with probability 1.

We may paraphrase (3.4) by saying that with probability converging to 1, $\mathcal{A}_0$ leaves unaltered each data value X_i that is neither very close to the true mode nor very far out in the tails. Since K is compactly supported, and (3.4) holding for $\mathcal{Y}_1$ implies the same for $\mathcal{Y}_0$, then if $\mathcal{Y}$ is produced by either $\mathcal{A}_0$ or $\mathcal{A}_1$ we have, for each vector c such that (3.2) holds,

$$P_f\{\hat{f}_{\mathcal{X}}(x) = \hat{f}_{\mathcal{Y}}(x) \text{ for all } x \in \mathcal{S}(c)\} \to 1 \quad (3.6)$$

as $n \to \infty$. Therefore, away from the mode and the tails, $\hat{f}_{\mathcal{Y}}$ and its derivatives have all the weak convergence properties of $\hat{f}_{\mathcal{X}}$ and the latter's derivatives. Hence, in terms of integrated squared error, we cannot expect improved performance of $\hat{f}_{\mathcal{Y}}$ over $\hat{f}_{\mathcal{X}}$ in those regions.

Note that we have assumed only one derivative of f. This explains the somewhat slow convergence rate at (3.5). However, (3.6) implies that if $\mathcal{Y} = \mathcal{Y}_0$ is produced by $\mathcal{A}_0$, and if f has two bounded derivatives, then (3.5) has a more conventional form away from the mode and the tails, as follows. Using the data sharpening algorithm $\mathcal{A}_0$, we have for each vector c such that (3.2) holds, $|\hat{f}_{\mathcal{Y}}(x) - f(x)| = O_p(h_0^2)$ and $|\hat{f}'_{\mathcal{Y}}(x) - f'(x)| = O_p(h_0)$ whenever $x \in \mathcal{S}(c)$, and

$$\sup_{x\in\mathcal{S}(c)} |\hat{f}_{\mathcal{Y}}(x) - f(x)| = O_p(h_0^2\,\ell^{1/2})\,, \qquad \sup_{x\in\mathcal{S}(c)} |\hat{f}'_{\mathcal{Y}}(x) - f'(x)| = O_p(h_0\,\ell^{1/2})\,, \quad (3.7)$$

where $\ell = \log n$. In fact, under mild additional conditions these results also extend to neighbourhoods of the mode, as we show following Theorem 3.2. It is only in the extreme tails, beyond locations that themselves move further out into each tail as n increases, that $\hat{f}_{\mathcal{Y}}$ and $\hat{f}'_{\mathcal{Y}}$ may not enjoy the convergence rates of the conventional estimators $\hat{f}_{\mathcal{X}}$ and $\hat{f}'_{\mathcal{X}}$, respectively.

In principle (3.4) does not exclude the possibility of "cross mappings", where a datum X_i in one tail is mapped to Y_i in the opposite tail or in a close neighbourhood of the mode; or where X_i near the mode is mapped to Y_i in a tail. However, it is easy to see that the probability that this occurs tends to 0 as $n \to \infty$. Indeed, if a cross mapping were to arise then with probability converging to 1 we could produce a lesser value of $D(\mathcal{X}, \mathcal{Y}_0)$ by instead moving $\mathcal{X}$ to a

point in $\mathcal{S}(c)$ and then moving another point in $\mathcal{S}(c)$ to the original image of X_i. However, for any given c satisfying (3.2), result (3.4) states that the probability of this occurring converges to 0.

Result (3.4) implies that when $D(\mathcal{X}, \mathcal{Y})$ is defined by (3.1), and (3.3) holds, the algorithm $\mathcal{A}_0$ is in effect constructed quite separately near the mode and in either tail. In the next two theorems we note that it is possible to find algorithms for the mode and the tails, respectively, such that $\hat{f}_{\mathcal{Y}}$ is unimodal and (3.3) holds. Following Theorem 3.3 we describe how to combine these two algorithms into a single algorithm that satisfies (3.3).

Assume $f \in \mathcal{F}_0$ has two bounded derivatives in a neighbourhood of the mode, f'' is continuous at the mode, and $f''(m_f) < 0$. Call this condition (C_f). Suppose too that K is a compactly supported, symmetric, uniquely unimodal probability density with three bounded derivatives; denote this constraint by ($\mathrm{C}_{K,2}$).

Theorem 3.2. *Assume* (C_f) *and* ($\mathrm{C}_{K,2}$) *hold, and* $h \asymp h_0$. *Then there exists a data sharpening algorithm* $\mathcal{A}_1$ *which, when applied to data on* $(m_f - \epsilon, m_f + \epsilon)$ *for* $\epsilon > 0$ *sufficiently small, ensures that* $\hat{f}_{\mathcal{Y}}$ *is unimodal on* $(m_f - \epsilon, m_f + \epsilon)$ *and also guarantees that for each* $0 < \delta < \epsilon$, *as* $n \to \infty$, $P_f\{Y_i = X_i$ *for each* i *such that either* X_i *or* Y_i *is in* $(m_f - \epsilon, m_f - \delta) \cup (m_f + \delta, m_f + \epsilon)\} \to 1$ *and*

$$\sum_{i\,:\,|X_i - m_f| \le \epsilon} |X_i - Y_i| = O_p(h_0^{-1}). \tag{3.8}$$

Result (3.8) implies a rate of convergence in the neighbourhood of the origin. Indeed, suppose we use the algorithm $\mathcal{A}_0$, instead of $\mathcal{A}_1$, to sharpen data. Assume too that the conditions of Theorems 3.1 and 3.2 hold. Then it follows from (3.4), (3.8) and the fact that the probability of cross-mappings converges to 0, that for any $\epsilon > 0$, (3.8) holds when the Y_i's are produced by $\mathcal{A}_0$ rather than $\mathcal{A}_1$. Now, a Taylor expansion gives

$$|\hat{f}_{\mathcal{X}}(x) - \hat{f}_{\mathcal{Y}}(x)| \le \frac{\sup |K'|}{nh^2} \sum_{i\,:\,|X_i - m_f| \le \epsilon} |X_i - Y_i|\,,$$

uniformly in x satisfying $|x - m_f| \le \delta$, for all sufficiently large n. Therefore, by (3.8), $|\hat{f}_{\mathcal{X}} - \hat{f}_{\mathcal{Y}}| = O_p(h_0^2\, \ell^{1/2})$ uniformly in a neighbourhood of the mode. It follows from this result and the fact that $|\hat{f}_{\mathcal{X}} - f| = O_p(h_0^2\, \ell^{1/2})$ uniformly in a neighbourhood of the mode, that $|\hat{f}_{\mathcal{Y}} - f| = O_p(h_0^2\, \ell^{1/2})$ there. The latter property enables us to replace the supremum over $x \in \mathcal{S}(c)$, in the first result at (3.7), by the supremum over $x \in [c_1, c_4]$, where c_1 and c_4 satisfy (3.2). Similarly it may be proved that an identical change can be made to the second result at (3.7). The particular construction of $\mathcal{A}_1$ that we give in the proof of Theorem 3.2 actually

ensures that $|\hat{f}_{\mathcal{X}} - f| = O_p(h_0^2)$ in an $O_p(h_0)$-neighbourhood of m_f, and $\hat{f}_{\mathcal{X}} = \hat{f}_{\mathcal{Y}}$ outside that neighbourhood and away from the extreme tails.

Finally we discuss a data sharpening algorithm that produces monotone tails and at the same time ensures (3.3). We consider the case of tails that decrease only polynomially fast, of which those of Student's t density are an example. Densities with tails that decrease exponentially quickly are more easily treated. It suffices to consider data sharpening in the right hand tail.

Let $\alpha > 1$ be given and assume K is a symmetric, compactly supported, uniquely unimodal density with $\nu_{\alpha,K}$ bounded derivatives, that $f \in \mathcal{F}_0$ has $\nu_{\alpha,f}$ derivatives on (C, ∞) for some $C > 0$, and that $|f^{(j)}(x)| \asymp x^{-\alpha-j}$ as $x \to \infty$ for $0 \le j \le \nu_{\alpha,f}$, where $\nu_{\alpha,f}, \nu_{\alpha,K} \ge 2$. Call this condition ($\mathrm{C}_\alpha$). The values of $\nu_{\alpha,K}$ and $\nu_{\alpha,f}$ depend only on α; details are given by (5.14) in the proof of Theorem 3.3.

Theorem 3.3. *If* (C_α) *holds for* $\alpha > 8$, *and* $h \asymp h_0$, *then there exists a data sharpening algorithm* $\mathcal{A}_1$ *which, when applied to data on* (C, ∞) *for* $C > 0$ *sufficiently large, ensures* $\hat{f}_{\mathcal{Y}}$ *is decreasing on* (C, ∞) *and also guarantees that for all* $C' > C$, $P_f\{Y_i = X_i$ *for each* i *such that either* X_i *or* Y_i *is in* $(C, C')\} \to 1$ *and* $h_0^2 \sum_{i\,:\,X_i > C} |X_i - Y_i| \to 0$ *in probability as* $n \to \infty$.

The constraint $\alpha > 8$ can be weakened by using a longer argument. Moreover, our numerical work will focus particular attention on performance of data sharpening when α is as small as 3/2.

Assume $f \in \mathcal{F}_0$, and suppose in addition that the conditions of Theorems 3.1–3.3 apply to f and K, with the conditions of Theorems 3.3 holding in both tails, possibly for different values of α. Theorems 3.2 and 3.3 imply that we may construct a single data sharpening algorithm $\mathcal{A}_1$ that produces an estimator $\hat{f}_{\mathcal{Y}}$ which is unimodal, and such that (3.3) and (3.4) both hold.

Indeed, (3.4) itself enables us to splice together the separate data sharpening algorithms for the tails, and that for the mode, into a single algorithm that applies to the whole dataset, in the knowledge that unimodality of $\hat{f}_{\mathcal{Y}}$ will take care of itself at all places that are intermediate between either tail and the mode. This is guaranteed by the fact that $\hat{f}_{\mathcal{Y}} = \hat{f}_{\mathcal{X}}$ in such places; since f' does not vanish there then $\hat{f}'_{\mathcal{X}}$ is monotone at the intermediate places, with high probability. Of course, once (3.3) has been proved for $\mathcal{A}_1$ then we know from Theorem 3.1 that both (3.3) and (3.4) hold for $\mathcal{A}_0$.

4. Numerical Properties

In the work summarised here we drew datasets of size $n = 25$ from the standard Normal distribution (distribution 1), Student's t distribution with three

degrees of freedom (distribution 2), the Normal mixture

$$0.2\,\mathrm{N}(0,1) + 0.2\,\mathrm{N}\Big(\tfrac{1}{2}, (\tfrac{2}{3})^2\Big) + 0.6\,\mathrm{N}\Big(\tfrac{13}{12}, (\tfrac{5}{9})^2\Big)$$

(distribution 3), and the Normal mixture

$$0.35\,\mathrm{N}\Big(-1, (\tfrac{3}{5})^2\Big) + 0.5\,\mathrm{N}\Big(1, (\tfrac{5}{2})^2\Big) + 0.15\,\mathrm{N}\Big(5, (\tfrac{3}{2})^2\Big)$$

(distribution 4). All are unimodal and are depicted in Figure 1. Distributions 1 and 2 were chosen because they represent opposite extremes in terms of tail weight, distribution 3 represents moderately skewed densities (it is density #2 of Marron and Wand (1992)), while distribution 4 is highly skewed with a long and relatively flat part, and therefore presents particular challenges to methods for unimodal density estimation.

Nevertheless, out of the four densities the second is arguably the most difficult for which to construct unimodal estimators. Standard kernel estimators, computed using data from the second distribution, usually have quite a few spurious bumps in the tails. Ideally, these should be removed without increasing mean integrated squared error. We therefore devote much of our attention to the relative performance of methods applied to distribution 2.

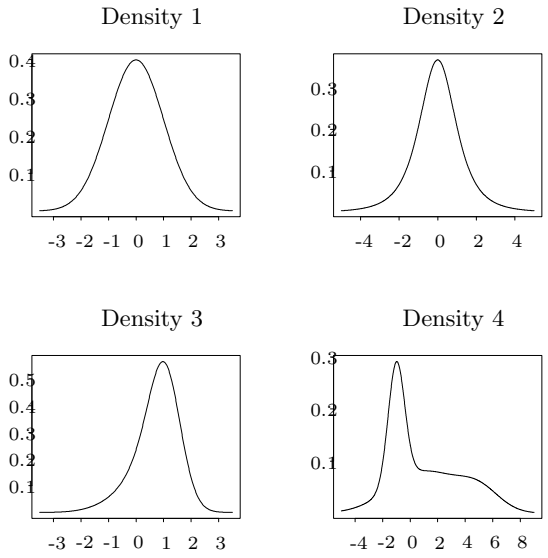

Figure 1. Four true densities. Density 1 is standard normal, density 2 is Student's t with three degrees of freedom, density 3 is moderately skewed, and density 4 is highly skewed with a long, relatively flat portion.

For the results presented here the Gaussian kernel was used throughout, since it is a favourite of statisticians in problems involving mode analysis. However,

very similar results were obtained using, for example, the biweight kernel. Indeed, the Gaussian kernel is so light tailed that, even though it is not compactly supported, it produces density estimates which visually satisfy the properties reported in Section 3. That is, except on very fine scales, the constrained density estimates are usually visually indistinguishable, in places away from the modes and the tails, from their conventional, unconstrained counterparts. We return to this point in our discussion of Figure 4 below.

We discuss results for estimators constrained by data sharpening using either L_2 distance or the distance $\Psi_{\tan}$. Similar results are obtained using L_p rather than L_2 distance, for $p > 1$. As expected, the efficiency advantages of $\Psi_{\tan}$ distance decrease as p decreases to 1, although $\Psi_{\tan}$ retains its competitive edge. For comparison we also give results for the standard, unconstrained kernel estimator; for the estimator $\hat{f}_{\text{tilt}}$ obtained by data tilting, this time using the distance measure D_1. See Section 2 for details of the construction of $\hat{f}_{\text{tilt}}$. A comparison with the "kernel rearrangement" method of Fougères (1997) will be made later in this section.

When constructing $\hat{f}_{\mathcal{Y}}$ the constrained optimisation step was usually implemented using the NAG routine E04UCF. This routine failed only rarely (e.g., in fewer than one per cent of simulations in the Normal case), and it can be protected against this problem in those instances where it fails. An alternative is to use an easily-coded simulated annealing algorithm there, similar to that described by Braun and Hall ((2001), Section 3.2), using the penalty

$$p(\mathcal{Y}) = \sum_{\{i\,:\,\xi_i<m\}} |\hat{f}'(\xi_i)|\, I\{\hat{f}'(\xi_i) < 0\} + \sum_{\{i\,:\,\xi_i>m\}} |\hat{f}'(\xi_i)|\, \{\hat{f}'(\xi_i) > 0\}\,,$$

where $\xi_1, \ldots, \xi_\nu$ denote grid points on which the constraints were imposed, and m is a candidate for the mode. In either case, the algorithm can be speeded up by moving outlying data a little closer to the body of the distribution, right at the start. This reflects the theoretical results discussed in Section 3. The simulations reported below used a small subroutine of this type.

When the estimators are applied to data from the standard normal distribution it is found that a graph of MISE for the data sharpening estimator, using distance measure $\Psi_{\tan}$, lies below that for the other three estimators across all values of h. However, although this estimator has a slight advantage over its unconstrained kernel counterpart, at the optimal bandwidth it offers only negligible improvements relative to the other two unimodal estimators. Nevertheless the data tilting estimator, when used with a small suboptimal bandwidth, performs quite poorly relative to both its data sharpening competitors. Analogous results are also obtained for other very light-tailed distributions. As we see below, the

data sharpening estimator based on $\Psi_{\tan}$ really comes into its own when applied to distributions that have heavier tails than the normal.

In particular, Figure 2 plots the logarithm of MISE against the logarithm of bandwidth for all four estimators when the sampled distribution is the heavy-tailed Student's t with three degrees of freedom. The performance advantages of data sharpening using distance measure $\Psi_{\tan}$ are obvious. Note too that, in terms of optimal MISE performance, the unconstrained kernel method lies between the two data sharpening approaches.

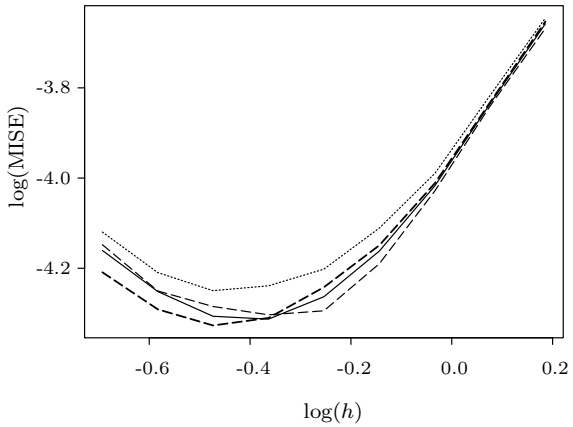

Figure 2. Plots of MISE for distribution 2. The vertical axis gives to the logarithm of MISE, and the horizontal axis gives the logarithm of bandwidth. Solid, bold-dashed, dotted and dashed lines correspond to the unconstrained estimate, data sharpening estimates constrained using distances $\Psi_{\tan}$ and L_2, and the constrained estimate using data tilting, respectively.

Figure 2 also shows that the optimal bandwidth for data sharpening, using $\Psi_{\tan}$, is smaller than that for the standard kernel estimator. This reflects the fact that this form of data sharpening produces an estimator with smaller integrated variance, and slightly larger integrated squared bias. The optimal tradeoff between the two therefore occurs at a smaller bandwidth. The main contribution to reduced integrated variance comes from the tails of the distribution, where data sharpening reduces the considerable stochastic variability of the standard kernel estimator by removing all the bumps there. Variance actually increases in the middle of the distribution; see Figure 3. However, the latter property is not characteristic of data sharpening, which often reduces variance in the vicinity of the mode. In particular this is the case for distributions 1 and 3.

Figure 3, which gives pointwise mean squared error, squared bias and variance in the case of data from distribution 2, shows that in this instance data

sharpening tends to reduce bias towards the centre of the distribution. This is also observed for distribution 1, but for distributions 3 and 4 bias slightly increases at the centre.

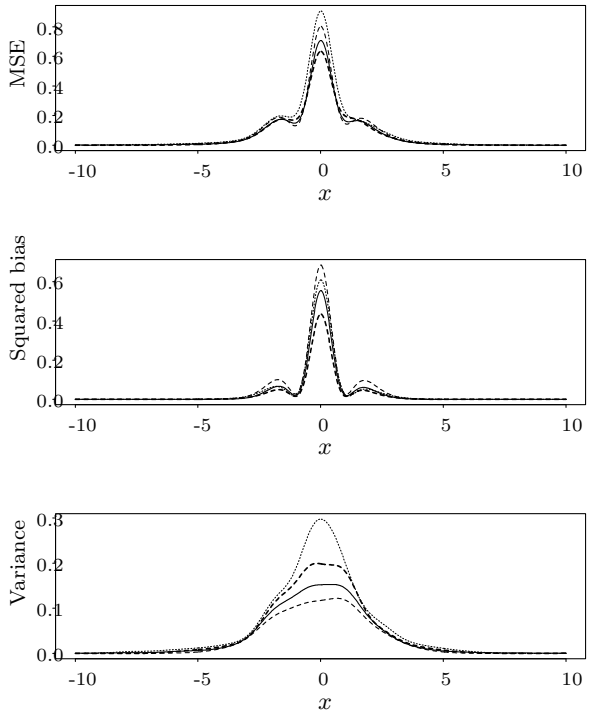

Figure 3. Plots of pointwise mean squared error, squared bias and variance for distribution 2. Line types are as for Figure 2. In each case the bandwidth was chosen to be that which minimises MISE, and the vertical axis is marked in units of 100.

Figure 2 revealed that, while data tilting outperformed L_2 data sharpening for the second distribution, it was not competitive with the unconstrained kernel estimator in MISE terms. Figure 3 indicates why. The only way tilting can remove spurious modes in the tails of the standard kernel estimator is to give outlying data zero weight. This forces relatively high weights to be used near the centre of the distribution, with a corresponding large increase in bias, which is clearly evident from the second panel of Figure 3. The data tilting estimator commonly suffers from this difficulty for distributions with one or more relatively heavy tails.

Sampling from the third distribution, which has the skewed density shown in the third panel of Figure 1, one again finds that the data sharpening estimator based on distance measure $\Psi_{\tan}$ has lowest minimum MISE. For this distribution

the data tilting method has highest MISE of all four estimators. This is again a result of its relatively large bias towards the centre of the distribution, which occurs for the reasons noted in the previous paragraph; in the case of the third distribution the left hand tail is relatively heavy.

Performance advantages of data sharpening are even more clearly evident for the fourth distribution, where the MISE curve for the estimator based on $\Psi_{\tan}$ lies below that for any of the other two estimators over much of its range on either side of its minimum. On this occasion, since the right hand tail of the distribution is relatively heavy, the technique based on data tilting again performs poorly. However, in MISE terms L_2 data sharpening produces an estimator which performs almost as well when sharpening is based on $\Psi_{\tan}$, and which performs better than the unconstrained kernel estimator.

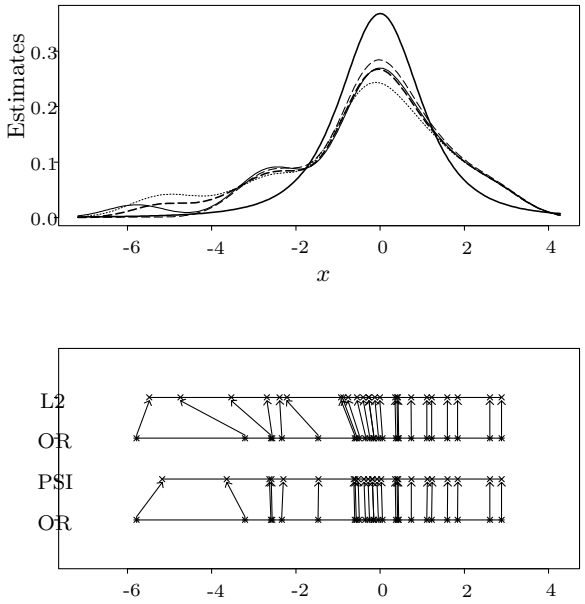

Figure 4. Kernel estimates, and results of data sharpening, for a dataset from the second distribution. Line types are as in Figure 2. The bold solid line represents the true density. In the lower panel the top and bottom graphs show how data are moved in the cases of data sharpening based on L_2 and $\Psi_{\tan}$ distance, respectively.

Figure 4 shows individual curve estimates, and the manner in which data are sharpened, in the case of a dataset drawn from the second distribution. The dataset chosen was that for which ISE was at its 75th percentile, when the bandwidth was chosen to be optimal for minimising MISE of the unconstrained kernel

estimator. This particular dataset produces a markedly asymmetric density estimate regardless of estimator type. There are several outliers in the left hand tail, with the result that the unconstrained estimate has two spurious modes there. However, data in the right hand tail are distributed in a manner which suggests properties of a light tailed distribution, rather than the heavy tailed distribution from which they came. As a result the standard kernel estimate has no unwanted bumps in the right hand tail. Neither does it have any spurious modes in the neighbourhood of the point at which the global maximum is achieved. Therefore, except for the left-hand tail, the estimate based on $\Psi_{\tan}$ coincides almost exactly with its unconstrained counterpart. In the left-hand tail it produces a better fit than data tilting.

The lower panel of Figure 4 shows how data are sharpened to produce a unimodal estimate, when the distance function is either L_2 or $\Psi_{\tan}$. In each instance the data in the right hand tail are virtually unaltered, and in the case of $\Psi_{\tan}$ distance, excepting the two most extreme values, data are altered in only minor ways in the left hand tail. (If the kernel used is compactly supported, for example the biweight, there are no changes at all except for the four data furthest to the left.) However, when distance is measured in L_2 terms the changes to data in the left hand tail are much more widespread.

We also examined the effect of data-driven bandwidth selection on the performance of our data-sharpened estimators. In particular, for each sample of size $n = 25$ we chose the bandwidth using the method suggested by Sheather and Jones (1991), and compared the unconstrained estimates with those constructed by sharpening based on $\Psi_{\tan}$-distance. Mean integrated squared error was approximated by averaging integrated squared error over 200 simulations. Data sharpening reduced MISE by 13%, 18%, 12% and 8% in the cases of distributions 1–4, respectively.

Of course, the main source of improvement comes from the tails and from the vicinity of the mode. As one would expect, given that asymptotic performance of the data-sharpened estimator becomes increasingly like that of its unconstrained counterpart as n increases (see Section 3), the margin of MISE improvement offered by data sharpening narrows for large n.

Figure 5 compares our method with that of Fougères (1997) for distribution 2 (Student's t with three degrees of freedom). Here, and for distribution 4, Fougères' approach has a marginal advantage. However, the method generally gives a relatively rough estimate, especially in sample sizes that are larger than that ($n = 25$) used here. A clearer idea of the roughness problem, for larger sample sizes, is obtainable from the figures of Fougères (1997).

For distribution 1, Fougères' approach gives the largest minimum mean squared error of all shape-constrained methods studied; in the case of distribu-

tion 3, it occupies the rank between the data tilting method and data sharpening, and in particular is behind the latter.

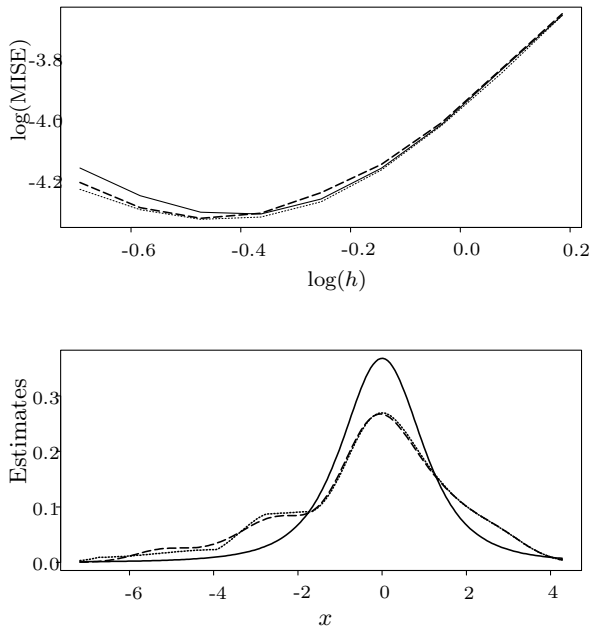

Figure 5. Comparison of Fougères' method with data sharpening using distance $\Psi_{\tan}$. In the context of Figure 2 (i.e., for distribution 2) the upper panel graphs the logarithm of MISE. Kernel estimates, for the same sample as in Figure 4, are shown in the lower panel. In both panels the dotted line corresponds to Fougères' estimate, while the bold-dashed line shows the result for the data-sharpened estimate. In the upper panel the (light) solid line shows results for the unconstrained estimate, while in the lower panel the (bold) solid line depicts the true density.

Finally, we applied the sharpening approach based on $\Psi_{\tan}$ distance, and data tilting, to the Buffalo snowfall dataset, which is frequently used to explore the effect of bandwidth choice on density estimate shape; see e.g., Silverman (1986, p. 45) and Scott (1992, p. 137). The dataset consists of the annual snowfalls, measured in inches, at Buffalo, New York, from 1910 to 1972. Figure 6 illustrates the unconstrained kernel estimate, and the results of data sharpening and data tilting, applied to these data using the bandwidth $h = 6$. Silverman (1986) discussed the use of this bandwidth, for these data, noting that it gave a trimodal estimate when applied to a standard kernel estimator. He observed that doubling the size of the bandwidth gave a unimodal estimator in the unconstrained case.

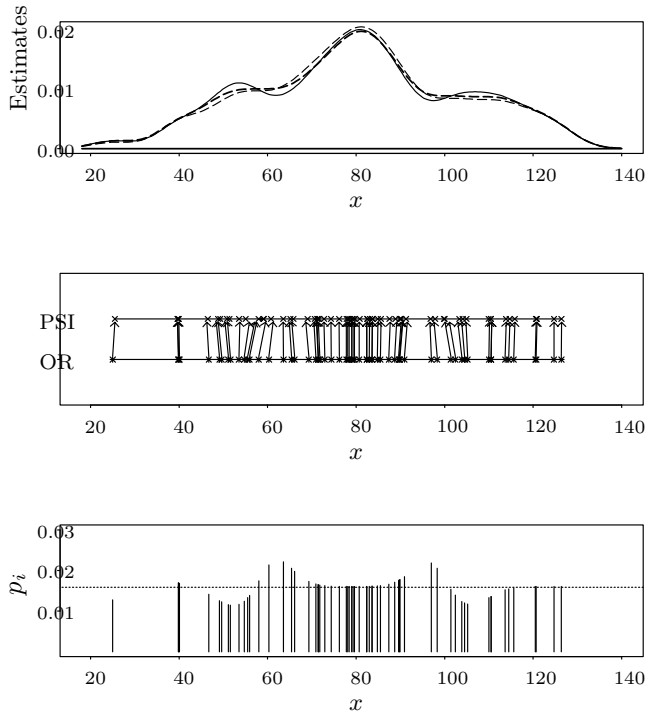

Figure 6. Unconstrained kernel estimates, and results of data sharpening and data tilting, for Buffalo snowfall data. The upper panel graphs kernel estimates whose line types are as in Figure 2. The middle panel indicates how data are moved in the case of data sharpening based on $\Psi_{\tan}$ distance. Data weights for the tilting approach are shown in the lower panel.

5. Technical Arguments

5.1. Proof of Theorem 3.1.

Consider a general data sharpening algorithm in which $\mathcal{X}$ is moved a distance $D(\mathcal{X}, \mathcal{Y})$ to $\mathcal{Y}$. By a Taylor expansion,

$$\begin{aligned} |\hat{f}'_{\mathcal{Y}}(x) - \hat{f}'_{\mathcal{X}}(x)| &\le \frac{1}{nh^2} \sum_{i=1}^{n} \Big| K'\Big(\frac{x - X_i}{h} + \frac{X_i - Y_i}{h}\Big) - K'\Big(\frac{x - X_i}{h}\Big) \Big| \\ &\le \frac{\sup |K''|}{nh^3}\, D(\mathcal{X}, \mathcal{Y})\,. \end{aligned} \tag{5.1}$$

Given a particular algorithm $\mathcal{A}_j$, let $\mathcal{Y}_j$ denote the corresponding sharpened dataset, and let $\hat{f}_{Yj}$ be the version of $\hat{f}_{\mathcal{Y}}$ produced from $\mathcal{Y}_j$. By assumption, we can construct $\mathcal{A}_1$ such that $\hat{f}_{Y1}$ is unimodal and $h_0^2\, D(\mathcal{X}, \mathcal{Y}_1) \to 0$ in probability

as $n \to \infty$. Hence, by the definition of $\mathcal{A}_0$, $\hat{f}_{Y0}$ is unimodal and

$$h_0^2 \, D(\mathcal{X}, \mathcal{Y}_0) \to 0 \tag{5.2}$$

in probability.

Let $c = (c_1, \ldots, c_4)$ and $c' = (c_1', \ldots, c_4')$ be vectors such that

$$a_f < c_1' < c_1 < c_2 < c_2' < m_f < c_3' < c_3 < c_4 < c_4' < b_f \, .$$

Then both c and c' satisfy (3.2). Consider the algorithm $\mathcal{A}_2$ constructed from $\mathcal{A}_0$ by putting $Y_i = X_i$ for $X_i \in \mathcal{S}(c)$ and taking Y_i to be defined by $\mathcal{A}_0$ otherwise. Then

$$D(\mathcal{X}, \mathcal{Y}_2) \le D(\mathcal{X}, \mathcal{Y}_0) \, . \tag{5.3}$$

It is readily proved that for $f \in \mathcal{F}_0$,

$$\sup_{-\infty < x < \infty} |\hat{f}_{\mathcal{X}}'(x) - f'(x)| = o_p(1) \, . \tag{5.4}$$

Since $h \to 0$ and K is compactly supported then for all sufficiently large n, $\hat{f}_{Y2} = \hat{f}_{Y0}$ outside $\mathcal{S}(c')$; call this result (R_1). Properties (5.1), employed in the case $\mathcal{Y} = \mathcal{Y}_2$, (5.3) and (5.4), and the fact that $h_0^2 \, D(\mathcal{X}, \mathcal{Y}_0) \to 0$ in probability, imply that $\sup |\hat{f}_{Y2}' - f'| \to 0$ in probability. It follows from this result and the fact that $|f'|$ is bounded away from 0 on an open set containing $\mathcal{S}(c')$, that with probability converging to 1 as $n \to \infty$, $\hat{f}_{Y2}'$ is positive on $[c_1', c_2']$ and negative on $[c_3', c_4']$. In conjunction with (R_1) this implies that with probability tending to 1, $\hat{f}_{Y2}$ is unimodal on the whole real line; call this result (R_2).

It follows from (R_2) that if the probability that the inequality (5.3) is strict does not converge to 0, then with strictly positive probability, for arbitrarily large n, the algorithm $\mathcal{A}_2$ produces a unimodal density estimator and at the same time strictly reduces the distance $D(\mathcal{X}, \mathcal{Y}_0)$. However, by definition of $\mathcal{A}_0$ this is impossible. Therefore, with probability converging to 1 the inequality at (5.3) is actually an identity. It follows from this result and the definition of $\mathcal{A}_2$ that the probability that $\mathcal{A}_0$ fixes X_i for each i such that $X_i \in \mathcal{S}(c)$, converges to 1 as $n \to \infty$.

This is equivalent to $P_f\{Y_i = X_i \text{ for each } i \text{ such that } X_i \in \mathcal{S}(c)\} \to 1$, which is a weaker form of (3.4). To obtain the full force of (3.4), observe that if $\mathcal{A}_0$ involves mapping $X_i \notin \mathcal{S}(c)$ to $Y_i \in \mathcal{S}(c)$, then with probability converging to 1 we may produce a lesser value of $D(\mathcal{X}, \mathcal{Y})$ by instead mapping X_i to a point in $\mathcal{S}(c') \backslash \mathcal{S}(c)$, while retaining unimodality of $\hat{f}_{\mathcal{Y}}$. However, this contradicts the definition of $\mathcal{A}_0$ as an algorithm that minimises $D(\mathcal{X}, \mathcal{Y})$, and so the probability that $X_i \notin \mathcal{S}(c)$ is mapped to $Y_i \in \mathcal{S}(c)$ must converge to 0.

The second part of (3.5) follows from the three conditions (5.1), employed in the case $\mathcal{Y} = \mathcal{Y}_0$, (5.2) and (5.4). The first part of (3.5) may be derived similarly.

5.2. Proof of Theorem 3.2.

Write $Y_i = X_i - h^3\Delta_i$ and observe that by a Taylor expansion,

$$\hat f'_{\mathcal{Y}}(x) = \hat f'_{\mathcal{X}}(x) + \frac{1}{n}\sum_{i=1}^n \Delta_i\, K''\Big(\frac{x - X_i}{h}\Big) + O_p\Big(\frac{h^2}{n}\sum_{i=1}^n \Delta_i^2\Big)$$

where, here and at (5.5) below, the remainder is of the stated form uniformly in x. Applying Theorem 3 of Komlós, Major and Tusnády (1975) we deduce that

$$\hat f'_{\mathcal{X}}(x) = E\{\hat f'_{\mathcal{X}}(x)\} + (nh^3)^{-1/2}\, U(x) + O_p\{(nh^2)^{-1}\log n\}\,, \tag{5.5}$$

where $U(x) = h^{-1/2}\int W_0\{F(x - hy)\}\, K''(y)\, dy$, W_0 is a standard Brownian bridge whose construction relative to the data depends on n, and F denotes the distribution function for which f is the density. Therefore, noting that $h \asymp h_0$, and assuming for the present that

$$\sum_{i=1}^n \Delta_i^2 = O_p(nh)\,, \tag{5.6}$$

we deduce that

$$\hat f'_{\mathcal{Y}}(x) = E\{\hat f'_{\mathcal{X}}(x)\} + (nh^3)^{-1/2}\, U(x) + \frac{1}{n}\sum_{i=1}^n \Delta_i\, K''\Big(\frac{x - X_i}{h}\Big) + O_p(h_0^3\log n) \tag{5.7}$$

uniformly in x.

Without loss of generality the true mode is 0. Let L be a bounded, compactly supported function with a continuous derivative, put $\Delta_i = L(X_i/h)$, and define $t = x/h$ and $a(t) = \int K''(u)\, L(t + u)\, du$. Then (5.6) holds and

$$\frac{1}{n}\sum_{i=1}^n \Delta_i\, K''\Big(\frac{x - X_i}{h}\Big) = h\, a(t)\, f(x) + o_p(h_0)$$

uniformly in x. Furthermore, (3.8) follows from the construction of the data sharpening algorithm. It remains to show that L can be chosen such that $\hat f_{\mathcal{Y}}$ is unimodal.

To this end, observe that $E\{\hat f'_{\mathcal{X}}(x)\} = f'(x) + o(h) = ht\, f''(0) + o(h)$ and $f(x) = f(0) + o(1)$ uniformly in x such that $|x| \le C_1 h$, for any $C_1 > 0$. Combining (5.7) and the results in this paragraph we see that

$$\hat f'_{\mathcal{Y}}(x) = \begin{cases} ht\, f''(0) + h\, a(t)\, f(0) + (nh^3)^{-1/2}\, U(x) + o_p(h_0) \\ \qquad\qquad \text{uniformly in } |x| \le C_1 h \\ E\{\hat f'_{\mathcal{X}}(x)\} + h\, a(t)\, f(x) + (nh^3)^{-1/2}\, U(x) + o_p(h_0) \\ \qquad\qquad \text{uniformly in } x\,. \end{cases} \tag{5.8}$$

Note too that for any $C_2 > 0$,

$$\Big\{ \sup_{|x| \le y} |U(x)| \Big\} \Big/ \Big(\max\left[1, \{\log_+(y/h)\}^{1/2}\right] \Big) = O_p(1) \tag{5.9}$$

for $0 \le y \le C_2$.

Since $K'(0) = 0$, $K' \le 0$ on the positive real line, and L has a bounded derivative, then

$$a(t) = \int_0^\infty \{L'(t+u) - L'(t-u)\} \, |K'(u)| \, du \, .$$

Let the support of K be $[-C_3, C_3]$, let $C_4 > 0$ be much larger than C_3, and let $C_5 > 0$ be much larger than C_4. Given $-\infty < A_1 < \infty$ and $A_2 < 0$, choose L such that $L(u) = A_1 u^2/4 + A_2 u^3/12$ for $u \in [-C_4, C_4]$, and L redescends very slowly, and with very small values of the first three derivatives, to 0 to the left of $-C_4$ and to the right of C_4, vanishing outside $[-C_5, C_5]$. Put $B_j = A_j \int_{u>0} u \, |K'(u)| \, du$. Then $a(t) = B_1 + B_2 t$ on $[-C_4 + C_3, C_4 - C_3]$, and a redescends slowly to 0 to the left of $-C_4 + C_3$ and to the right of $C_4 - C_3$, vanishing outside $[-C_5 - C_3, C_5 + C_3]$.

As A_2 increases, so too does the curvature of $\hat{f}_{\mathcal{Y}}$ near its global maximum, modulo the effect of the additive noise term $(nh^3)^{-1/2} \, U(x)$ at (5.8). However, in neighbourhoods of width $O(h)$ of the origin, the noise is of the same order as the gradient of $\hat{f}'_{\mathcal{Y}}$; but the size of noise does not depend on A_2, whereas the absolute value of the gradient increases with A_2. Arguing thus it follows from (5.8) and (5.9) that if $\eta > 0$ is given, then A_1, B_1, C_4 and C_5 may be chosen such that the probability that $\hat{f}_{\mathcal{Y}}$ is unimodal on $(-\epsilon, \epsilon)$ exceeds $1 - \eta$ for all sufficiently large n. Taking L to be a random function, with $O_p(1)$ bounds applying to $|L'|$ and the support of L, we deduce that the probability that $\hat{f}_{\mathcal{Y}}$ is unimodal converges to 1 as $n \to \infty$. This is sufficient to ensure existence of the claimed algorithm $\mathcal{A}_1$; in cases where this particular construction does not produce a unimodal $\hat{f}_{\mathcal{Y}}$ we instead translate all data to a single point and rely on the unimodality of K to ensure that of $\hat{f}_{\mathcal{Y}}$. Note too that $\hat{f}_{\mathcal{Y}} = \hat{f}_{\mathcal{X}}$ outside an $O_p(h)$-neighbourhood of the origin.

5.3. Proof of Theorem 3.3.

Let $m_f < C < \infty$, put $\beta = 2/(\alpha + 2)$, and let $\delta \in (0, (4/\alpha) - \beta)$ be a small positive number. We take $Y_i = X_i$ for $C \le X_i \le \xi_1 \equiv Z_1 h_0^{-\beta}$, where $Z_1 = Z_1(n)$ denotes a random function of the data with the property

$$\lim_{\epsilon \to 0, \ \lambda \to \infty} \liminf_{n \to \infty} P(\epsilon < Z_1 < \lambda) = 1 \, . \tag{5.10}$$

Data $X_i \in (\xi_1, \xi_2]$, where $\xi_2 = h_0^{\delta-(4/\alpha)}$, will be sharpened respectively to $Y_i = F^{-1}(i/n)$, while data $X_i > \xi_2$ will be sharpened in a manner that we shall describe three paragraphs below. We call ξ_1 and ξ_2 the first and second breakpoints, respectively.

The size of the first breakpoint is that of values x at which the order of the stochastic error of $\hat{f}'_{\mathcal{X}}(x)$ is the same as the order of the quantity $f'(x)$ being estimated; note that the bias of $\hat{f}'_{\mathcal{X}}(x)$ is always of strictly smaller order than $f'(x)$. To appreciate why the critical size is $h_0^{-\beta}$, note that the variance of $\hat{f}'_{\mathcal{X}}(x)$ is asymptotic to $v_n(x) \equiv h_0^2\, x^{-\alpha}$ as both n and x increase, in the sense that the ratio of the variance and $v_n(x)$ is bounded away from zero and infinity. Moreover, $|f'(x)|/x^{-\alpha-1}$ is bounded away from zero and infinity as $x \to \infty$. Therefore the relative stochastic error of the estimator $\hat{f}'_{\mathcal{X}}(x)$ of $f'(x)$ is of order 1 for values x such that $v_n(x)$ is of size $x^{-2(\alpha+1)}$, i.e., for x such that $x \asymp h_0^{-\beta}$. More concisely, using methods developed during the proof of Theorem 3.2, particularly the argument based on the Komlós-Major-Tusnády (1975) approximation, it may be proved that if $Z_2 = Z_2(n)$ denotes the supremum of values z such that $\hat{f}'_{\mathcal{X}}(x) < 0$ for $C \le x \le z$, then (5.10) holds with Z_1 replaced there by Z_2.

Therefore, at the very least, values of X_i that exceed $Z_2 h_0^{-\beta}$ must be sharpened if the density estimator $\hat{f}_{\mathcal{Y}}$ is to have negative gradient at all points to the right of C. We take $Z_1 < Z_2$ in order to effect a slight taper, to ensure the gradient remains negative on both sides of the first breakpoint.

The size of the second breakpoint is chosen for technical reasons that will become clear in the proof at (5.15) below. Tapering beyond the second breakpoint, used to ensure that the density estimator has negative gradient there, does not require any data value X_i to be moved by more than $O_p(X_i)$, uniformly in i. For example it is sufficient, modulo minor tapering in the neighbourhood of ξ_2, to equally space the Y_i's to the right of ξ_2, using the spacing $F^{-1}(i/n) - F^{-1}\{(i-1)/n\}$ where $F^{-1}(i/n)$ is the supremum of values $F^{-1}(j/n) \le \xi_2$. The density of data at any point x, where $x \ge C$, is by assumption bounded by a constant multiple of $x^{-\alpha}$. Therefore the total contribution to $h_0^2\, D(\mathcal{X}, \mathcal{Y})$ from sharpening those X_i's for which $X_i > \xi_2$, equals

$$O_p\biggl(n h_0^2 \int_{\xi_2}^{\infty} x \cdot x^{-\alpha}\, dx\biggr) = O_p(n h_0^2\, \xi_2^{2-\alpha}) = o_p(1)\,, \tag{5.11}$$

provided $\alpha > 8$ and δ in the definition of ξ_2 is sufficiently small.

The sum of the distances, multiplied by h_0^2, through which data X_i that lie between the first and second breakpoints are sharpened, equals

$$h_0^2 \sum_{X_i \in (\xi_1, \xi_2]} |X_i - F^{-1}(i/n)| = O_p\biggl\{n h_0^2 \int_{\xi_1}^{\xi_2} (x^\alpha/n)^{1/2}\, x^{-a}\, dx\biggr\}$$

$$= O_p(n^{1/2} h_0^2\, \xi_1^{1-(\alpha/2)})$$
$$= O_p(h_0^{(\alpha-6)/2(\alpha+2)}) = o_p(1),$$

since $\alpha > 6$. Combining this result with (5.11) we deduce that $h_0^2\, D(\mathcal{X}, \mathcal{Y}) = o_p(1)$.

It remains to show that the data sharpening step that takes X_i to $Y_i = F^{-1}(i/n)$, for $X_i \in (\xi_1, \xi_2]$ modulo the tapers, produces a monotone decreasing estimator, at least for all sufficiently large n. We may confine attention to deriving monotonicity in $x \in (\xi_1 + C_2 h, \xi_2 - C_2 h)$, where $C_2 > 0$ is so large that $[-C_2, C_2]$ contains the support of K. The case where x lies outside $(\xi_1 + C_2 h, \xi_2 - C_2 h)$ may be accommodated by modifying the tapers.

Put $G = F^{-1}$. Assume K has $s+2$ bounded derivatives, and f has $r+s+1$ bounded derivatives in the far right hand tail. We shall prove that the function

$$\psi_1(x) = \frac{1}{nh^2} \sum_{i=1}^{n} K'\Big(\frac{x - G(i/n)}{h}\Big)$$

may be approximated by

$$\psi_2(x) = \int K(u)\, f'(x - hu)\, du = \frac{1}{nh^2} \sum_{i=1}^{n} \int_{-1/2}^{1/2} K'\Big(\frac{x - G\{(i+u)/n\}}{h}\Big)\, du\,, \tag{5.12}$$

in the sense that

$$|\psi_1(x) - \psi_2(x)| = O(h_0^{\alpha\delta(s+1)-1} + h_0^{r-2}) \tag{5.13}$$

uniformly in $x = O(\xi_2)$. Now, $\psi_2(x) < 0$ and $|\psi_2(x)| \asymp x^{-\alpha-1}$ as $x \to \infty$. Furthermore, provided

$$\min\{\alpha\delta(s+1) - 1, r-2\} \geq 4(\alpha+1)/\alpha\,, \tag{5.14}$$

$h_0^{\alpha\delta(s+1)-1} + h_0^{r-2}$ is of strictly smaller order than $x^{-\alpha-1}$ uniformly in $C_3 \leq x \leq C_4\, \xi_2$, for any $0 < C_3, C_4 < \infty$. It follows that $\psi_1(x) < 0$, uniformly in $C_3 \leq x \leq C_4\, \xi_2$ for all sufficiently large n, as had to be proved.

We conclude by deriving (5.13). Note that by a Taylor expansion, and for $|u| \leq \frac{1}{2}$, $G\{(i+u)/n\} = G(i/n) + S_i(u) + (u/n)^{r+1}\, R_{i1}(u)$, where

$$S_i(u) = \sum_{j=1}^{r} \frac{(u/n)^j}{j!}\, G^{(j)}(i/n)\,,$$
$$R_{i1}(u) = \frac{1}{(r-1)!} \int_0^1 t^{r-1}(1-t)\, G^{(r+1)}\{(i+tu)/n\}\, dt\,.$$

Now, $|G^{(j)}(i/n)| = O(x^{j\alpha-j+1})$ uniformly in indices i such that $|x-(i/n)| \le C_5 h$, and in $x \ge C_6 > C$, where $C_5 > 0$ may be arbitrarily large. Hence for $C_7, C_8 > 0$, and uniformly in $|u| \le 1/2$, in $x \in [C_6, \xi_2]$, and in i such that $|x - (i/n)| \le C_5 h$, we have

$$\Big|\frac{G\{(i+u)/n\} - G(i/n)}{h}\Big| \le \frac{C_7\, x^\alpha}{nh} \le C_8\, h_0^{\alpha\delta}\,. \tag{5.15}$$

Therefore, assuming K has $s+2$ derivatives, we may Taylor-expand the argument of K in the integral on the right hand side of (5.12), obtaining

$$\begin{aligned}
&\int_{-1/2}^{1/2} K'\Big(\frac{x - G\{(i+u)/n\}}{h}\Big)\, du \\
&= \sum_{k=0}^{s} \frac{(-1)^k}{k!}\, K^{(k+1)}\Big(\frac{x - G(i/n)}{h}\Big) \int_{-1/2}^{1/2} \Big(\frac{S_i(u) + (u/n)^{r+1}\, R_{i1}(u)}{h}\Big)^k du + R_{2i}\,,
\end{aligned}$$

where $|R_{2i}| \le C_9\, h_0^{\alpha\delta(s+1)}\, I\{|x - G(i/n)| \le C_{10} h\}$. It follows that

$$\begin{aligned}
&\int_{-1/2}^{1/2} K'\Big(\frac{x - G\{(i+u)/n\}}{h}\Big)\, du \\
&= K'\Big(\frac{x - G(i/n)}{h}\Big) + \sum_{k=1}^{s} \frac{(-1)^k}{k!}\, K^{(k+1)}\Big(\frac{x - G(i/n)}{h}\Big) \\
&\qquad \times \int_{-1/2}^{1/2} \{S_i(u)/h\}^k\, du + R_{3i}
\end{aligned} \tag{5.16}$$

where, uniformly in i such that $|x-(i/n)| \le C_5 h$, and in x for which $C_6 \le x \le \xi_2$,

$$|R_{3i}| \le C_{11}\, \{h_0^{\alpha\delta(s+1)} + h_0^{-1}\, n^{-(r+1)}\, x^{(r+1)\alpha - r}\}\, I\{|x - G(i/n)| \le C_{12} h\}\,. \tag{5.17}$$

Carrying out the integration over u on the right hand side of (5.16), we see that the kth term of the series in k there may itself be expressed as a series in terms of the form

$$h^{-k}\, C(k; j_1, \ldots, j_k)\, n^{-(j_1 + \cdots + j_k)}\, K^{(k+1)}\Big(\frac{x - G(i/n)}{h}\Big)\, G^{(j_1)}(i/n) \cdots G^{(j_k)}(i/n)\,,$$

where each $j_\ell \in [1, r]$ and the constant $C(k; j_1, \ldots, j_k)$ depends only on its arguments, not on either i or n. Now divide this quantity by nh^2 and sum over i; we obtain $h^{-k}\, C(k; j_1, \ldots, j_k)\, n^{-(j_1 + \cdots + j_k)}\, S(k; j_1, \ldots, j_k)$, where

$$S(k; j_1, \ldots, j_k) = \frac{1}{nh^2} \sum_{i=1}^{n} K^{(k+1)}\Big(\frac{x - G(i/n)}{h}\Big)\, G^{(j_1)}(i/n) \cdots G^{(j_k)}(i/n)\,.$$

Approximate this series by the corresponding integral,

$$\frac{1}{nh^2} \sum_{i=1}^{n} \int_{-1/2}^{1/2} K^{(k+1)}\Big(\frac{x - G\{(i+u)/n\}}{h}\Big) G^{(j_1)}\{(i+u)/n\} \cdots G^{(j_k)}\{(i+u)/n\} du,$$

using the method in the previous paragraph. When performing the integration, use integration by parts to reduce the derivative exponent $k+1$ of $K^{(k+1)}$ to 0.

Iterating this procedure we see that the error in the approximation at (5.13) is reduced by the factor x^α/n for each occasion on which the function G is differentiable. At the same time it is increased by a factor h^{-k}, appearing in the denominator of the integral on the right hand side of (5.16), but this is altered to $h^{-k} \times h^{k+1} = h$ after $k+1$ subsequent integrations by parts, provided G has at least $r+s+1$ derivatives. Thus we may deduce from (5.16) and (5.17) that

$$\begin{aligned} &\frac{1}{nh^2} \sum_{i=1}^n \int_{-1/2}^{1/2} K'\Big(\frac{x - G\{(i+u)/n\}}{h}\Big)\, du \\ &= \frac{1}{nh^2} \sum_{i=1}^n K'\Big(\frac{x - G(i/n)}{h}\Big) + O\{h_0^{\alpha\delta(s+1)-1} + h_0^{-3}\,(x^\alpha/n)^{r+1}\}\,, \end{aligned} \tag{5.18}$$

uniformly in $x \in [C_6, \xi_2]$. Since $x = O(\xi_2) = O(h_0^{\delta-(4/\alpha)})$, and $n = h_0^{-5}$, then (5.18) implies (5.13).

Acknowledgements

The research of the second author was mostly done when he was working for Centre for Mathematics and its Applications at the Australian National University, and supported partially by an Australian Research Council grant and by KOSEF through Statistical Research Center for Complex Systems at Seoul National University. Both authors are grateful to two referees and an Associate Editor for their helpful comments.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton University Press, Princeton, New Jersey.

Birgé, L. (1997). Estimation of unimodal densities without smoothness assumptions. *Ann. Statist.* **25**, 970-980.

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

Braun, W. J. and Hall, P. (2001). Data sharpening for nonparametric inference subject to constraints. *J. Comput. Graph. Statist.* **4**, 786-806.

Cheng, M.-Y., Gasser, T. and Hall, P. (1999). Nonparametric density estimation under unimodal 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.

Delecroix, M., Simioni, M. and Thomas-Agnan, C. (1995). A shape constrained smoother: simulation study. *Computat. Statist.* **10**, 155-175.

Fougères, A.-L. (1997). Estimation de densités unimodales. *Canad. J. Statist.* **25**, 375-387.

Friedman, J. H. and Tibshirani, R. J. (1984). The monotone smoothing of scatterplots. *Technometrics* **26**, 243-250.

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

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

Hall, P. and Huang, L.-S. (2002). Unimodal density estimation using kernel methods. *Statist. Sinica* **12**, 965-990.

Hall, P. and Presnell, B. (1999). Density estimation under constraints. *J. Computat. Graph. Statist.* **39**, 259-277.

Hardy, G., Littlewood, J. and Pólya, G. (1952). *Inequalities.* Cambridge Univ. Press, Cambridge.

Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose-response curves and assessment of synergism. *Biometrics* **46**, 1071-1085.

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.

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

Mammen, E. and Thomas-Agnan, C. (1999). Smoothing splines and shape restrictions. *Scand. J. Statist.* **26**, 239-252.

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

Qian, S. (1994). Generalization of least-square isotonic regression. *J. Statist. Plann. Inference* **38**, 389-397.

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

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York.

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. (1986). *Density Estimation For Statistics and Data Analysis.* Chapman and Hall, London.

Tantiyaswasdikul, C. and Woodroofe, M. (1994). Isotonic smoothing splines under sequential designs. *J. Statist. Plann. Inference* **38**, 75-87.

Wang, Y. (1995). The L1-theory of estimation of monotone and unimodal densities. *J. Nonparam. 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 Statistics, Hankuk University of Foreign Studies, Yongin, Kyounggi 449-791, Korea.

E-mail: khkang@hufs.ac.kr

(Received May 2003; accepted April 2004)