Maximum likelihood kernel density estimation

M.C. Jones and D.A. Henderson

The Open University, UK, and University of Newcastle, UK

Summary. Methods for improving the basic kernel density estimator include variable locations, variable bandwidths (often called variable kernels) and variable weights. Currently these methods are implemented separately and via pilot estimation of variation functions derived from asymptotic considerations. In this paper, we propose a simple maximum likelihood procedure which allows (in its greatest generality) variation of all these quantities at once, bypasses asymptotics and explicit pilot estimation, and turns out to perform better. This maximum likelihood kernel density estimation, which involves a greatly overparametrised mixture model, works because the overall bandwidth (the geometric mean of individual bandwidths) is fixed. This overall bandwidth, in turn, is the single smoothing parameter of the methodology which has to be chosen separately. And the method has a further advantage: it automatically reduces, where appropriate, to a few-component mixture model which indicates and initialises parametric mixture modelling of the data. We set out simple algorithms, perform a substantial simulation study, give an illustrative example and provide some background theory. For computational and performance reasons we particularly recommend the variable location version of the methodology.

Keywords: Mixture modelling; Normal mixtures; Variable kernel; Variable location.

Address for correspondence: M.C. Jones, Department of Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK.
E-mail: m.c.jones@open.ac.uk

1. Introduction

Let $\mathbf{X} = (X_1, ..., X_n)$ be a random sample from a univariate distribution with unknown density f. Let K be a symmetric probability density function to be used as a kernel and $h > 0$ its scaling parameter, or bandwidth; write $K_h(\cdot) = h^{-1}K(h^{-1}\cdot)$. Then the standard kernel estimator of the density f at a point x is given by

$$\widehat{f}(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)$$

(Silverman, 1986, Wand and Jones, 1995, Simonoff, 1996). Let $\mathbf{m} = (m_1, ..., m_n)$, $\mathbf{w} = (w_1, ..., w_n)$ and $\mathbf{b} = (b_1, ..., b_n)$ with $w_i \geq 0$ and $b_i > 0$, $i = 1, ..., n$. One interpretation of the kernel density estimator is as a special case of a highly parametrised mixture model of the form

$$\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}(x) = \left\{ \sum_{i=1}^{n} w_i \right\}^{-1} \sum_{i=1}^{n} w_i K_{b_i}(x - m_i) \tag{1}$$

where each datapoint X_i is associated with its own mixture component density K with its own location, m_i, scale parameter or bandwidth, b_i, and weight, w_i. Clearly, $\widehat{f}(x) = \widehat{f}_{\mathbf{X},\mathbf{1},h\mathbf{1}}(x)$ where $\mathbf{1} = (1, ..., 1)$. (Even more general mixtures with e.g. skew components could perhaps be envisaged but will not be considered here.) Other special cases of the general formulation include variable location kernel density estimators of the form $\widehat{f}_{\mathbf{m}}(x) \equiv \widehat{f}_{\mathbf{m},\mathbf{1},h\mathbf{1}}(x)$ (Samiuddin and el-Sayyad, 1990, called 'data sharpening' by Choi and Hall, 1999), variable weight kernel density estimators $\widehat{f}_{\mathbf{w}}(x) \equiv \widehat{f}_{\mathbf{X},\mathbf{w},h\mathbf{1}}(x)$ (Hall and Turlach, 1999) and variable bandwidth kernel density estimators $\widehat{f}_{\mathbf{b}}(x) \equiv \widehat{f}_{\mathbf{X},\mathbf{1},\mathbf{b}}(x)$ (Abramson, 1982), the last often being referred to as variable kernel density estimators. These approaches can all achieve 'higher order bias' and as such (most of them) are reviewed and compared with other higher order bias kernel density estimators in Jones and Signorini (1997).

In this paper, we simply fit the general model $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}(x)$ or one of its submodels such as $\widehat{f}_{\mathbf{m}}(x)$ or perhaps $\widehat{f}_{\mathbf{m},\mathbf{b}}(x) \equiv \widehat{f}_{\mathbf{m},\mathbf{1},\mathbf{b}}(x)$ to data by maximum likelihood ... with the crucial proviso that $g(\mathbf{b}) \equiv (\prod_{i=1}^{n} b_i)^{1/n} = h$ for a given overall bandwidth h. It is this single constraint on the scale associated with each component density of the mixture that stops the maximum likelihood

solution from becoming degenerate. The result is some new density estimators, again indexed by a single bandwidth h, the best of which, as we will show empirically, can achieve every bit as good performance as the best of the higher order bias kernel density estimators. Moreover, they achieve this level of performance without resort to the asymptotic arguments that underlie the practical implementation of most higher order bias kernel density estimators. And there is an added advantage of the new approach, at least for variable $\mathbf{m}$ and/or $\mathbf{w}$. In addition to yielding a good density estimator, its form can be used to suggest a much simpler parametric mixture model for the data. This is because, typically, most of the n location/weight values are redundant, reducing to many equal locations and/or many zero weights.

The estimators and associated algorithms are given in Section 2. A fairly substantial simulation study is performed in Section 3, some observations on patterns of locations, weights and bandwidths are made in Section 4 and a real data example is given in Section 5.

The same basic idea can be found particularly for $\widehat{f}_{\mathbf{m},\mathbf{w}}$, i.e. location and weight together with no bandwidth variation, in an exceptional doctoral thesis by Storvik (1999). (Earlier, Roeder, 1990, considered an estimator very like $\widehat{f}_{\mathbf{m},\mathbf{w}}$ except that n was replaced by a general integer; however, she explicitly decided against restricting the value of h and using maximum likelihood.) Storvik concentrates on the single iteration version of the procedure while we tend to prefer the fully converged version, at least in terms of estimation quality as measured by integrated squared error. Storvik mentions the full-blown $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$ too, but ignores the $g(\mathbf{b}) = h$ constraint in another one iteration version of the methodology. Like Storvik, we will find it particularly convenient to concentrate on using the standard normal density as kernel. Storvik also provides other variations on the basic theme that we do not consider here.

We wish to give considerable credit to Storvik (1999). Our additional contributions and/or re-emphases are: (i) to concentrate on the simple but general form (1); (ii) to iterate to convergence rather than to stop after one iteration; (iii) to understand empirically how the resulting estimators fare and to compare these results to those achieved by their asymptotic forms and to other higher order bias estimators considered by Jones and Signorini (1997); (iv) to stress that the resulting estimates often suggest simple parametric mixture models for the data; and (v) to make a case for preferring the particularly simple variable location estimator $\widehat{f}_{\mathbf{m}}$ in practice.

What we are unable to provide is a full theoretical analysis of the per-

formance of any version of $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$ as an estimator of f. Some background theory — again related to that of Storvik, but potentially more general — is given and discussed in Section 6. Conclusions complete the paper in Section 7. It is the very lack of sophistication of the approach we advocate that we feel is its great strength and appeal.

2. The estimators

We shall give the iterative algorithms for the special cases of $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$ for which just $\mathbf{m}$, $\mathbf{w}$ and $\mathbf{b}$ are varied singly, in turn, in Sections 2.1–2.3. In each case, we start out using a general symmetric probability density kernel K until, where it is advantageous to do so, we switch to $K = \phi$, the standard normal density kernel. In the case of varying $\mathbf{m}$ and/or $\mathbf{w}$, the algorithms are EM algorithms (e.g. McLachlan and Peel, 2000) but this is not the case for varying $\mathbf{b}$. Some comments on the cases where more than one of the defining vectors varies are given in Section 2.4 and some practical experiences concerning convergence etc. are given in Section 2.5.

2.1. Variable locations only

In this first version, $\widehat{f}_{\mathbf{m}}$, set $\mathbf{w} = \mathbf{1}$ and $\mathbf{b} = h\mathbf{1}$. Then

$$\hat{f}_{\mathbf{m}}(x) = n^{-1} \sum_{i=1}^{n} K_h(x - m_i).$$

Maximising likelihood, choose $\mathbf{m}^* = \text{argmax}_{\mathbf{m}} \sum_{k=1}^{n} \log \hat{f}_{\mathbf{m}}(X_k)$. Straight away specifying $K = \phi$, the estimating equations are

$$\sum_{k=1}^{n} \frac{\phi_h'(X_k - m_\ell)}{\hat{f}_{\mathbf{m}}(X_k)} = \sum_{k=1}^{n} \frac{(X_k - m_\ell)\phi_h(X_k - m_\ell)}{\hat{f}_{\mathbf{m}}(X_k)} = 0, \qquad \ell = 1, ..., n,$$

which immediately yields a beautifully simple iterative scheme for computing $\mathbf{m}^*$:

$$m_\ell^{\text{new}} = \left\{ \sum_{k=1}^{n} t_{\ell,k} \right\}^{-1} \sum_{k=1}^{n} X_k t_{\ell,k} \quad \text{with} \quad t_{\ell,k} = \frac{\phi_h(X_k - m_\ell^{\text{old}})}{\hat{f}_{\mathbf{m}^{\text{old}}}(X_k)}.$$

Finally, use $\widehat{f}_{\mathbf{m}*}$. The natural starting point for this iteration, which we always use, is $\mathbf{m} = \mathbf{X}$. The weights are all nonnegative and so each calculated m_ℓ falls within the range of $\mathbf{X}$ at all iterations.

2.2. Variable weights only

Next, set $\mathbf{m} = \mathbf{X}$, $\mathbf{b} = h\mathbf{1}$; we have

$$\hat{f}_{\mathbf{w}}(x) = \{\sum_{i=1}^{n} w_i\}^{-1} \sum_{i=1}^{n} w_i K_h(x - X_i)$$

and $\mathbf{w}^* = \text{argmax}_{\mathbf{w}} \sum_{k=1}^{n} \log \hat{f}_{\mathbf{w}}(X_k)$. Write $K_{i,k} = K_h(X_i - X_k)$. Differentiation simply gives the estimating equations

$$\sum_{k=1}^{n} \frac{K_{k,\ell}}{\sum_{i=1}^{n} w_i K_{k,i}} - \frac{n}{\sum_{i=1}^{n} w_i} = 0, \quad \text{or} \quad n^{-1} \sum_{k=1}^{n} \frac{K_{k,\ell}}{\hat{f}_{\mathbf{w}}(X_k)} = 1, \qquad \ell = 1, ..., n.$$

This, in turn, yields the following simple iterative scheme for computing $\mathbf{w}^*$:

$$w_\ell^{\text{new}} = w_\ell^{\text{old}} n^{-1} \sum_{k=1}^{n} \frac{K_{k,\ell}}{\hat{f}_{\mathbf{w}^{\text{old}}}(X_k)}, \qquad \ell = 1, ..., n.$$

Attractive properties of this iteration are that it maintains the same value for $\sum_{i=1}^{n} w_i$ at all iterations and, provided the initial $\mathbf{w}$ has all nonnegative elements, all elements of iterated $\mathbf{w}$s remain nonnegative. The natural starting point, which we always use, is $\mathbf{w} = \mathbf{1}$. Note that this iterative scheme works for any K, not just ϕ.

2.3. Variable bandwidths only

In the third special case, $\widehat{f}_{\mathbf{b}}$, set $\mathbf{m} = \mathbf{X}$ and $\mathbf{w} = \mathbf{1}$. Then,

$$\hat{f}_{\mathbf{b}}(x) = n^{-1} \sum_{i=1}^{n} K_{b_i}(x - X_i)$$

where $g(\mathbf{b}) = h$. Choose

$$\mathbf{b}^* = \text{argmax}_{\mathbf{b}} \sum_{k=1}^{n} \log \hat{f}_{\mathbf{b}}(X_k) \quad \text{subject to} \quad \sum_{k=1}^{n} \log b_k = n \log h.$$

It is possible to incorporate the constraint directly in to the main objective function, but we have found it speedier to pursue the following alternative course.

Introducing a Lagrange multiplier and differentiating gives estimating equations of the form

$$\sum_{k=1}^{n} \frac{(\partial K_{b_\ell}/\partial b_\ell)(X_k - X_\ell)}{\hat{f}_{\mathbf{b}}(X_k)} + \frac{\lambda}{b_\ell} = 0, \qquad \ell = 1, ..., n.$$

Note that

$$\frac{\partial K_b(u)}{\partial b} = \frac{1}{b} M_b(u), \text{ say,} \quad \text{where} \quad M(u) = -\{K(u) + uK'(u)\}.$$

Progress is again facilitated by reverting to the normal kernel case, $K = \phi$, for which $M(u) = \phi''(u) = (u^2 - 1)\phi(u)$. The estimating equations become of the form

$$\frac{1}{b_\ell} \left\{ \frac{T_2(\ell, \mathbf{b})}{b_\ell^2} - T_0(\ell, \mathbf{b}) \right\} + \frac{\lambda}{b_\ell} = 0$$

where

$$T_0(\ell, \mathbf{b}) = \sum_{k=1}^{n} \frac{\phi_{b_\ell}(X_k - X_\ell)}{\hat{f}_{\mathbf{b}}(X_k)} \quad \text{and} \quad T_2(\ell, \mathbf{b}) = \sum_{k=1}^{n} \frac{(X_k - X_\ell)^2 \phi_{b_\ell}(X_k - X_\ell)}{\hat{f}_{\mathbf{b}}(X_k)}$$

are both positive. If λ were known, the natural iteration arising from this would take

$$b_\ell^{\text{new}} = b_\ell^{\text{new}}(\lambda) = \sqrt{\frac{T_2(\ell, \mathbf{b}^{\text{old}})}{T_0(\ell, \mathbf{b}^{\text{old}}) - \lambda}}.$$

We select a different value for $\lambda = \lambda^{\text{new}}$ at each iteration by choosing it so that $G(\lambda^{\text{new}}) \equiv g(\mathbf{b}^{\text{new}}(\lambda^{\text{new}})) = h$ (noting that the sequence of λs will converge as the $\mathbf{b}$s converge). $G(\lambda)$ is clearly monotone increasing in λ from zero when $\lambda = -\infty$ to ∞ when $\lambda = \min_j\{T_0(j, \mathbf{b})\}$ and so a simple binary procedure is adequate to isolate $-\infty < \lambda^{\text{new}} < \min_j\{T_0(j, \mathbf{b})\}$. The entire iterative process is started from $\mathbf{b} = h\mathbf{1}$.

2.4. Variable combinations of locations, weights and/or bandwidths

All three two-way combinations, $\widehat{f}_{\mathbf{m},\mathbf{w}}$, $\widehat{f}_{\mathbf{m},\mathbf{b}}$ and $\widehat{f}_{\mathbf{w},\mathbf{b}}$ will be considered in what follows along with the full three-way combination $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$. Iterations proceed in essentially the obvious ways combining the iterations above. Most generally for $K = \phi$, let

$$t_{\ell,k} = \frac{\phi_{\mathbf{b}^{\text{old}}}(X_k - m_\ell^{\text{old}})}{\hat{f}^{\text{old}}(X_k)}$$

where $\widehat{f}^{\text{old}} \equiv \widehat{f}_{\mathbf{m}^{\text{old}},\mathbf{w}^{\text{old}},\mathbf{b}^{\text{old}}}$. Then let

$$T_0(\ell) = \sum_{k=1}^{n} t_{\ell,k}, \qquad T_1(\ell) = \sum_{k=1}^{n} X_k t_{\ell,k}, \qquad T_2(\ell) = \sum_{k=1}^{n} (X_k - m_\ell^{\text{old}})^2 t_{\ell,k}.$$

Update each set of parameters simultaneously using

$$m_\ell^{\text{new}} = \frac{T_1(\ell)}{T_0(\ell)}, \qquad w_\ell^{\text{new}} = \frac{1}{n} T_0(\ell) w_\ell^{\text{old}}, \qquad b_\ell^{\text{new}} = \sqrt{\frac{T_2(\ell)}{T_0(\ell) - \lambda}}.$$

When any of the locations, weights and/or bandwidths are not varying, set them to their default values in the definition of $t_{\ell,k}$ and, of course, do not update. The version of this with $\mathbf{b}$ fixed at $h\mathbf{1}$ is a contribution of Storvik (1999) who, in different notation, essentially gave them (and suggested, but did not explore, related formulae for general $\mathbf{b}$).

2.5. Some computational considerations

Because the algorithm with varying $\mathbf{m}$ and/or $\mathbf{w}$ is an EM algorithm it is guaranteed to increase the likelihood at each iteration and therefore to converge to a local maximum of the likelihood. Our bandwidth variation is different, although we have always empirically observed increases in likelihood at each iteration there too.

In our implementation of these algorithms, we deemed convergence to have occurred when the average absolute difference between the parameter values at the current and previous iterations was less than 10^{-C} for some appropriate value of C. Most of our simulations concern $n = 100$ in which case we took $C = 3$ for $\widehat{f}_{\mathbf{w}}$, $\widehat{f}_{\mathbf{w},\mathbf{b}}$ and $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$ and $C = 5$ otherwise. Typical numbers of iterations to convergence were then 100–500 for $\widehat{f}_{\mathbf{m}}$ and $\widehat{f}_{\mathbf{m},\mathbf{w}}$ and 50–100 for $\widehat{f}_{\mathbf{b}}$. When operating with $C = 5$, $\widehat{f}_{\mathbf{w}}$ required around 30,000 iterations on average and rarely less than 10,000 iterations, hence our adjustment of C, reducing the number of iterations required to several hundreds. The effect on our performance measure of changing C was not large. The slow convergence in $\widehat{f}_{\mathbf{w}}$ transferred to $\widehat{f}_{\mathbf{w},\mathbf{b}}$ and $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$ but not to $\widehat{f}_{\mathbf{m},\mathbf{w}}$ whose convergence rate was comparable to that of $\widehat{f}_{\mathbf{m}}$. In further simulations of the performance of $\widehat{f}_{\mathbf{m}}$ when $n = 500$, we reduced C from 5 to 3. The number of iterations was then in the tens rather than the hundreds.

3. Simulation Study

3.1. Simulation set-up

We pattern our simulation study after that of Jones and Signorini (1997) who in turn utilise the first 10 densities suggested as a simulation testbed by Marron and Wand (1992). The densities, all built from normal mixtures, are referred to as 1: Gaussian, 2: Skewed unimodal, 3: Strongly skewed, 4: Kurtotic unimodal, 5: Outlier, 6: Bimodal, 7: Separated bimodal, 8: Skewed bimodal, 9: Trimodal, 10: Claw.

To deal with the bandwidth selection problem (selection of an appropriate value for the overall value h), we provide a "best possible" analysis in the sense that we empirically approximately compute the value of h that minimised the integrated squared error (ISE) between the estimated density and the true density. In this way, we decouple the potential capability of each density estimator, which is what is considered here, from the thorny issue of empirical bandwidth selection.

All density estimates were calculated on an equally spaced grid of 301 points on $[-3, 3]$. The following rule was used to approximate the ISE:

$$\mathrm{ISE}(\widehat{f}) \simeq \frac{1}{50} \sum_{j=1}^{301} \left\{ \widehat{f}(y_j) - f(y_j) \right\}^2,$$

where $y_j = -3 + (j-1)/50$. Minimisation over h was performed over a grid of values. For each estimator, a different, carefully chosen, grid of values of h was used. Our procedure was: (i) to start with an equally-spaced grid of 12 points from 0.1 to 1.2. (If the optimal h was not in this range, a different grid was tried); (ii) based on the results of the first grid, a second, finer, grid needed to be used for some estimator/density combinations. These finer grids were nearly always between 9 and 13 point grids but not always equally-spaced; (iii) this ad hoc process was repeated — rarely beyond three different grids — until a 'satisfactory' grid of values was obtained and used thereafter.

Our main simulation study concerns samples of size $n = 100$, replicating the experiment 1000 times. This is the subject of Section 3.2. A smaller simulation study when $n = 500$ (also with 1000 replications) is reported in Section 3.3. The normal kernel is used in all cases.

3.2. Simulation results, $n = 100$

The results of our simulations when $n = 100$, which will be found in Table 1, follow Jones and Signorini (1997) in giving the mean and standard error of the minimised ISE ($\times 10^5$) calculated over the 1000 simulated datasets for each estimator/density combination. Also given in Table 1 is the median percentage reduction of the minimised ISE for the particular estimator compared with that of the basic estimator $\widehat{f}$.

* * * Table 1 about here * * *

Let us start our study of Table 1 by comparing the three fully iterated single-vector estimators $\widehat{f}_{\mathbf{m}}$, $\widehat{f}_{\mathbf{w}}$ and $\widehat{f}_{\mathbf{b}}$. There is essentially nothing whatsoever to choose between the performances of $\widehat{f}_{\mathbf{m}}$ and $\widehat{f}_{\mathbf{w}}$ in any case. (For information on the similarity or otherwise of the actual estimates produced by the two methods, see Section 4.) We therefore prefer $\widehat{f}_{\mathbf{m}}$ because it is much the computationally faster of the two, as described in Section 2.5. The variable bandwidth estimator, $\widehat{f}_{\mathbf{b}}$, is decidely inferior to both of these except in the one case tailor-made for variable bandwidth estimation, that of model 3, the strongly skewed density.

A parallel set of results for a variety of different (fourth order bias) estimators is given in Jones and Signorini (1997, Table 1; henceforth JST1). We can compare the results of Table 1 with those of JST1, bearing in mind that the two refer to different sets of simulations, that JST1 uses the biweight kernel while Table 1 uses the normal, and that in the current experiment, we may not have computed the ISE-optimal bandwidth to such a high degree of accuracy. Comparison of results for $\widehat{f}$ in Table 1 and in JST1 show the two to be comparable to the levels of accuracy required to make general claims. In particular, we can compare the performances of $\widehat{f}_{\mathbf{m}}$ and $\widehat{f}_{\mathbf{b}}$ in Table 1 with that of their "asymptotics-based" counterparts, Jones and Signorini's (1997) implementations of the variable location estimator ($\widehat{f}_6$ in JST1) and three implementations of the variable bandwidth approach ($\widehat{f}_{5,1}$, $\widehat{f}_{5,2}$ and $\widehat{f}_{5,3}$ in JST1). The new estimator $\widehat{f}_{\mathbf{m}}$ shows considerably better performance than JST1's $\widehat{f}_6$ except for slightly inferior performance for model 3 and, perhaps, comparable performance for model 8 (the skewed bimodal density). On the other hand, $\widehat{f}_{\mathbf{b}}$ is unable to match the usually superior performance (often considerably so) of the JST1 implementations of the variable bandwidth approach except for the strongly skewed density (and, perhaps, for model 10, the claw).

We also obtained results for the one-iteration versions of variable location, weight and bandwidth estimators, starting from their usual starting points of $\mathbf{X}, \mathbf{1}$ and $h\mathbf{1}$, respectively; call them $\widehat{f}_{\mathbf{m}(1)}$, $\widehat{f}_{\mathbf{w}(1)}$ and $\widehat{f}_{\mathbf{b}(1)}$. We do so to explore the relative performance of these computationally faster estimators preferred by Storvik (1999). For most densities, $\widehat{f}_{\mathbf{m}}$ and $\widehat{f}_{\mathbf{w}}$ clearly outperform $\widehat{f}_{\mathbf{m}(1)}$ and $\widehat{f}_{\mathbf{w}(1)}$, exceptions being slight preferences for the latter at models 3 and 8 and relative parity at models 9 (trimodal) and 10. The one-step variable bandwidth estimator, however, generally improves on its fully iterated counterpart, except for models 3, 4 (kurtotic unimodal) and, at equality, 5 (outlier). Estimator $\widehat{f}_{\mathbf{b}(1)}$ never does better than comparably to the asymptotics-based variable bandwidth estimators in JST1, and usually still somewhat worse. It remains inferior, too, to $\widehat{f}_{\mathbf{m}}$ and $\widehat{f}_{\mathbf{w}}$ except, again, for model 3 and, insignificantly, for model 8.

With the essential equivalence in performance of $\widehat{f}_{\mathbf{m}}$ and $\widehat{f}_{\mathbf{w}}$, it is no surprise to find that $\widehat{f}_{\mathbf{m},\mathbf{w}}$ is essentially equivalent in performance too. We find that $\widehat{f}_{\mathbf{m},\mathbf{w}}$ takes little more time to compute than the very speedy $\widehat{f}_{\mathbf{m}}$ — which gives it an edge over $\widehat{f}_{\mathbf{w}}$ — but, even so, there seems to be no reason to prefer $\widehat{f}_{\mathbf{m},\mathbf{w}}$ to $\widehat{f}_{\mathbf{m}}$. Storvik's estimator $\widehat{f}_{\mathbf{m},\mathbf{w}(1)}$ stands in a similar relation to $\widehat{f}_{\mathbf{m},\mathbf{w}}$ as $\widehat{f}_{\mathbf{m}(1)}$ and $\widehat{f}_{\mathbf{w}(1)}$ do to $\widehat{f}_{\mathbf{m}}$, $\widehat{f}_{\mathbf{w}}$, except to a less pronounced extent resulting in a general preference for $\widehat{f}_{\mathbf{m},\mathbf{w}(1)}$ over $\widehat{f}_{\mathbf{m}(1)}$ and $\widehat{f}_{\mathbf{w}(1)}$.

By and large, $\widehat{f}_{\mathbf{m},\mathbf{b}}$ has performance intermediate to those of $\widehat{f}_{\mathbf{m}}$ and $\widehat{f}_{\mathbf{b}}$. Exceptions are density 8 (skewed bimodal) where $\widehat{f}_{\mathbf{m},\mathbf{b}}$ is the best of the three and density 10 (claw) where $\widehat{f}_{\mathbf{m},\mathbf{b}}$ is the worst of the three. In comparison with $\widehat{f}_{\mathbf{m}}$, $\widehat{f}_{\mathbf{m},\mathbf{b}}$ performs better only for densities 3 (strongly skewed) and 8, and is in most cases considerably less good. Estimator $\widehat{f}_{\mathbf{w},\mathbf{b}}$, on the other hand, while often having performance intermediate to those of $\widehat{f}_{\mathbf{w}}$ and $\widehat{f}_{\mathbf{b}}$, is best a little more often: for densities 3, 4 (kurtotic unimodal) and 8. $\widehat{f}_{\mathbf{w},\mathbf{b}}$ is consistently preferable to $\widehat{f}_{\mathbf{m},\mathbf{b}}$ (except for density 8), but it does not attain as good performance as $\widehat{f}_{\mathbf{m},\mathbf{w}}$ except for being better on densities 3, 4 and 8.

The performance of the "fully fledged" variable kernel density estimator $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$ is very variable relative to that of the other estimators. It exhibits the best performance of all for density 4 (kurtotic unimodal) and very strong performance for densities 3 (strongly skewed) and 8 (skewed bimodal). On the other hand, it exhibits the worst performance of all for density 10 (claw) and very poor performance for density 9 (trimodal). In all other cases, its performance is very much intermediate. Given that the greatest complica-

tion is associated with this estimator, we are inclined to require outstanding performance from it in a considerable number of cases and therefore, not observing this, we do not think it sufficiently successful to be worthy of general recommendation.

In general, it seems that our maximum likelihood approach works well for variable locations and variable weights but not so well for variable bandwidths. Whether the disappointing performance of the latter is due to failings of our particular implementation (although recall that we tried other approaches to less good effect) or is inherent to the approach, we are not sure, although we suspect the latter.

Overall, then, according to Table 1 together with computational considerations, there seems to be a general preference for $\widehat{f}_{\mathbf{m}}$ with the important caveat that it does not seem to work particularly well for the strongly skewed density. How does $\widehat{f}_{\mathbf{m}}$ compare with the best of the single-bandwidth higher order bias estimators in JST1? (There are also two bandwidth versions of some estimators in JST1, but Jones and Signorini, 1997, Section 6, are suspicious of their worth because of the extra difficulties in realising any potential improvements in practice involving data-driven bandwidth selection.) This best estimator appears to be what JST1 calls $\widehat{f}_{3,1}$, the multiplicatively bias corrected density estimator of Jones, Linton and Nielsen (1995). And, consistently across all densities except numbers 3 (barely significantly) and 8 (insignificantly), $\widehat{f}_{\mathbf{m}}$ outperforms $\widehat{f}_{3,1}$, albeit not by huge amounts. This is consistent with Jones and Signorini's observation that the variable location estimator in JST1, $\widehat{f}_6$ is "a little behind $\widehat{f}_{3,1}$" while, from above, "$\widehat{f}_{\mathbf{m}}$ shows considerably better performance than JST1's $\widehat{f}_6$ except for slightly inferior performance for model 3 and, perhaps, comparable performance for model 8".

3.3. Simulation results, $n = 500$

In order to check that our claims of good performance of $\hat{f}_{\mathbf{m}}$ are not confined, for some reason, to $n = 100$, we repeated the above experiment for $n = 500$ but only for $\widehat{f}$ and $\widehat{f}_{\mathbf{m}}$. The results are in Table 2.

* * * Table 2 about here * * *

Whether measured by mean or median reduction, the performance of $\widehat{f}_{\mathbf{m}}$ relative to that of $\widehat{f}$ is enhanced over that for $n = 100$ in every single case. (This even translates a slightly worse relative performance for Model 8, the

skewed bimodal, when $n = 100$ into a slightly better relative performance when $n = 500$.) Also, when $n = 500$, $\widehat{f}_{\mathbf{m}}$ continues to perform generally better than the Jones, Linton and Nielsen estimator (in JST1), perhaps even a little enhanced relatively in most cases, although slightly less well for a few. By the way, we have evidence that the relaxing of the convergence criterion for $n = 500$ mentioned in Section 2.5 has made the ISE values for $\widehat{f}_{\mathbf{m}}$ a little larger than they might be, but this effect is generally small.

4. Patterns of locations, weights and bandwidths

Not only is the performance of our maximum likelihood density estimators very good, they have another important advantage over existing kernel-based estimators: the resulting patterns of locations and/or weights and/or bandwidths are interpretable and, particularly for the best performing estimator $\widehat{f}_{\mathbf{m}}$, suggest a simple parametric form for the distribution in terms, typically, of a few normal mixture components.

To illustrate this, we take a single example dataset extracted from the simulation study above. It belongs to model 2, the skewed unimodal density, and is one for which $\widehat{f}_{\mathbf{m}}$, $\widehat{f}_{\mathbf{w}}$ and $\widehat{f}_{\mathbf{m},\mathbf{w}}$ all achieve approximately their median performance. The ISE-optimal value of h is 0.7. In this case, all three estimates are almost visually indistinguishable, although $\widehat{f}_{\mathbf{m}}$ is marginally best.

* * * Figs 1 and 2 about here * * *

Figs 1 and 2 show the maximum likelihood locations and weights, respectively, arising when they alone are varied. The graph for variable location is particularly attractive. The nonparametric density estimator has automatically reduced to just two location values, 96 points (= 96% of the data) located at 0.86 and the other 4 points at -2.03. This corresponds to a fitted model of $0.96\phi_{0.7}(\cdot - 0.86) + 0.04\phi_{0.7}(\cdot + 2.03)$. A very similar model is suggested by $\widehat{f}_{\mathbf{w}}$. Only 4 of the 100 potentially nonzero weights are nonzero. Three of them are for very close datapoints which it seems natural to combine by adding the weights and using the average of their three locations; the fourth non-zero weight is 0.033 at location -2.08. This gives fitted model $0.967\phi_{0.7}(\cdot - 0.85) + 0.033\phi_{0.7}(\cdot + 2.08)$. Now, the true, skewed unimodal, distribution was constructed by Marron and Wand (1992) as the three component mixture $0.2\phi(\cdot) + 0.2\phi_{2/3}(\cdot - 0.5) + 0.6\phi_{5/9}(\cdot - (13/12))$. It and our

two more parsimonious fitted models are shown in Fig. 3. This figure is essentially the same as looking at a graph of true model and nonparametric estimators $\widehat{f}_{\mathbf{m}}$ and $\widehat{f}_{\mathbf{w}}$, except for a slight discrepancy between $\widehat{f}_{\mathbf{w}}$ and the second simplified fitted model above.

* * * Fig. 3 about here * * *

Resulting locations and weights are given for $\widehat{f}_{\mathbf{m},\mathbf{w}}$ in Fig. 4. The locations follow a similar pattern as they did when varied alone, except that one datapoint transfers between clusters. The weights look quite different to those produced by $\widehat{f}_{\mathbf{w}}$, following simple curves within clusters; this does not reduce to such a parsimonious parametric representation. In Fig. 5, we display the varying bandwidths associated with $\widehat{f}_{\mathbf{b}}$ for the same dataset. The expected pattern of smaller bandwidths in the centre of the distribution and larger bandwidths as one goes out into the tails (e.g. Wand and Jones, 1995, Section 2.10.2) is apparent, but again not so parametrically simple (nor indeed, in this case, so effective).

* * * Figs 4 and 5 about here * * *

These results are not at all atypical. For $\widehat{f}_{\mathbf{m}}$, the locations consistently cluster into a relatively small number of components: large h gives rise to a smaller number of components while small h gives rise to a larger number of components. For $\widehat{f}_{\mathbf{w}}$, it often happens that more than a handful of the weights, perhaps 10–15, are nonzero, but these are usually clustered around common values and so can be interpreted as suggesting a more parsimonious mixture model as above. Fig. 4 is typical of the behaviour for $\widehat{f}_{\mathbf{m},\mathbf{w}}$, with the behaviour of the locations dominating that of the weights. Finally, Fig. 5 is fairly typical of a smooth pattern of bandwidths in $\widehat{f}_{\mathbf{b}}$ with low bs in high density regions and higher bs in low density regions. Sometimes, however, one or two datapoints in low density regions have very small bandwidths while neighbouring points have high bandwidths.

5. Example

As an illustrative example of our methodology, we estimate the density of velocities associated with $n = 82$ galaxies in the Corona Borealis region of the sky. These data, due to Postman, Huchra and Geller (1986), are as given and studied by Roeder (1990). (We join Roeder in working in units of

1000km/s.) We use only $\widehat{f}_{\mathbf{m}}$. In the absence of any viable alternative at this stage, we resorted to least squares cross-validation to select h for $\widehat{f}_{\mathbf{m}}$. That is, h was chosen to minimise

$$\int \{\widehat{f}_{\mathbf{m}}(x;h)\}^2 dx - 2n^{-1}\sum_{i=1}^{n} \widehat{f}_{\mathbf{m},-i}(X_i;h)$$

where $\widehat{f}_{\mathbf{m}}(\cdot;h)$ is the density estimate (with bandwidth h) derived from the full data set, and $\widehat{f}_{\mathbf{m},-i}(\cdot;h)$ is the density estimate (with bandwidth h) based on all datapoints except X_i. The cross-validated value of $h = 0.79$. The data and resulting density estimate are shown in Fig. 6.

* * * Fig. 6 about here * * *

The density estimate in Fig. 6 displays six modes. For convenience, call them modes 1 to 6 reading from left to right. In comparison with the least squares cross-validated kernel density estimate given in Roeder (1990, Fig. 1), our estimate is a little less smooth: in Roeder's estimate, modes 3 and 4 are not separated, there is but the vaguest suggestion of mode 2 and none of mode 5, and there is an extra small mode to the left of mode 1. The same kernel density estimate with (smaller) bandwidth selected by the Sheather and Jones (1991) method (not shown) is more similar to Fig. 6; the differences are only less accentuation of modes 3 and 4 and the division of mode 6 into two modes. Fig. 6 is pretty similar to Fig. 5(c) of Roeder (1990) where a parametric six component normal mixture (with equal variances) has been fitted. It differs relatively little from Roeder's (1990, Fig. 7) preferred normal mixture-based density estimate which has five modes, mode 2 not being present. However, Roeder's five mode density corresponds to a mixture model with some 17 components, although she does state that "this is partly an artifact of [her particular implementation of her] algorithm".

However, $\widehat{f}_{\mathbf{m}}$ has, without any prior specification of, or any complex procedure directly selecting, the number of components, reduced to a relatively simple normal mixture model anyway. The normal mixture has eight components; their locations are plotted on Fig. 6 and their weights (proportional to number of datapoints coalescing to each location) and precise locations are given in Table 3. Modes 4 and 6 are each made up of two normal components suggesting skewness. Of course, arguments about mode 6 are meaningless, there being only three datapoints involved, but that of mode 4 is based on 31 of the 82 datapoints and hence may be more interesting.

* * * Table 3 about here * * *

Our illustration of maximum likelihood density estimation, particularly via $\widehat{f}_{\mathbf{m}}$, stops there, but further modelling of the data might start with the eight component mixture model we have found and work parametrically towards a simpler model (e.g. without two components for mode 6).

6. Theoretical background

Prior to realising that what was desired could be achieved through the maximum likelihood fitting of $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$, a fuller asymptotic analysis of $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$ with functions m, w and b explicitly chosen to achieve fourth (and higher) order bias density estimation had been envisaged. To this end, write $k_2 = \int x^2 K(x)dx$ and

$$\begin{aligned} m(x) &= x + M(h^2 k_2/2)\mu(x), \\ w(x) &= 1 + W(h^2 k_2/2)\omega(x), \\ b(x) &= h/(1 + B(\beta(x) - 1)) \end{aligned}$$

and let $\widetilde{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$ be formula (1) employing these functions. Then, using Taylor series approximation incorporating the approach of Hall (1990), it is fairly easy to show (non-rigorously) that the bias of $\widetilde{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}(x)$ looks like

$$\frac{1}{2}h^2 k_2 \left[-M(\mu f)'(x) + W\left\{ f(x)(\omega(x) - \bar{\omega}) \right\} + \left\{ \frac{f(x)}{(1 + B(\beta(x) - 1))^2} \right\}'' \right]$$

where $\bar{\omega} = \int (\omega f)(z)dz$. Such a general formulation for asymptotic location/scale/bandwidth variation has not, by the way, been considered before.

On the other hand, numerous special cases of this formulation have been considered:

(1) $M = W = B = 0$: the ordinary $O(h^2)$ bias kernel density estimator $\widehat{f}$ with bias proportional to $f''(x)$;

(2) $W = B = 0,\ M = 1$: this is the $O(h^4)$ bias special case of variable location. One needs to take μ such that $f''(x) = (\mu f)'(x)$ i.e. $\mu(x) = (f'/f)(x)$, corresponding to Samiuddin and el–Sayyad (1990) and Choi and Hall (1999).

(3) $M = B = 0,\ W = 1$: fourth order bias results for $\omega(x) = -(f''/f)(x)$ (Hall and Turlach, 1999).

(4) $M = W = 0,\ B = 1$: this is the best-known $O(h^4)$ bias special case of the variable bandwidth (kernel) estimator of Abramson (1982). It is clear that $\beta = \sqrt{f}$ will do the trick.

(5), (6) & (7) Here are three ways of achieving $O(h^4)$ bias via combinations of pairs of varying parameters. First, $B = 0, M = W = 1$: any μ and ω such that $f(x)(\omega(x) - \bar{\omega}) + f''(x) = (\mu f)'(x)$. Second, $W = 0, M = B = 1$: any μ and β such that $(f/\beta^2)'(x) = (\mu f)(x)$. And third, $M = 0, W = B = 1$: any ω and β such that $f(x)(\omega(x) - \bar{\omega}) + (f/\beta^2)''(x) = 0$. The second of these, combining variable location with variable bandwidth, is the only one considered before, by Jones, McKay and Hu (1994).

(8)–(13) First, $B = 0, W = 1, M \sim h^2 M_0$, *say*: a special case of this is a major recommendation of Storvik (1999). The five other possibilities here are the permutations of how M, W and B are treated; they need not be explicitly written out. In Storvik's work, M_0 is explicitly, novelly and interestingly used to control, and hopefully reduce, variance, W on its own being enough to reduce bias. Targetting variance as well as bias can also be done in cases (5)–(7) by making e.g. W and M different $O(1)$ terms. But Storvik's version at least makes for simpler formulae e.g. one still needs $\omega(x) = -(f''/f)(x)$.

It is possible to calculate a general $O(h^4)$ bias term for $\widetilde{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$ which ignores the need to estimate the f-dependencies in the variations suggested above. However, explicit pilot estimation of such quantities, employing ordinary kernel estimation with a bandwidth h_1, say, makes for much more difficult calculations, at least if $h_1 \sim h$, which will not be pursued here.

Does this theoretical work have relevance to maximum likelihood density estimation? Well, partially. Further simple Taylor series considerations show that for $\widehat{f}_{\mathbf{w}}$ the first step of our iterative process (starting from $(\mathbf{m}, \mathbf{w}, \mathbf{b}) = (\mathbf{X}, \mathbf{1}, h\mathbf{1})$) results in $E(w_\ell^{\rm new} | X_1, ..., X_{\ell-1}, X_{\ell+1}, ..., X_n) \simeq 1 - (h^2 k_2/2)(f''/f)(X_\ell)$, just as required under (3) above (Hall and Turlach, 1999). Perhaps surprisingly, the same attractive property does not hold for $\widehat{f}_{\mathbf{m}}$ or $\widehat{f}_{\mathbf{b}}$. Instead, each is affected to a lower order than that which might have been expected: $E(m_\ell^{\rm new} | X_1, ..., X_{\ell-1}, X_{\ell+1}, ..., X_n) \simeq X_\ell + o(h^2)$ and, for $\lambda = o(h)$, $E(b_\ell^{\rm new} | X_1, ..., X_{\ell-1}, X_{\ell+1}, ..., X_n) \simeq h(1 + O(h^2))$. So, progress in terms of $\mathbf{m}$ and $\mathbf{b}$ is, in this sense, slower than in terms of $\mathbf{w}$. For $\widehat{f}_{\mathbf{m},\mathbf{w}}$, Storvik (1999) makes the plausible claim that "the asymptotic bias order will be improved in every iteration", raising the possibility that, in asymptotic bias terms at least, our fully iterated estimators might behave like so-called infinite-order kernel estimators.

However, it is at this stage that we have to admit that colleagues much more mathematically able than the current authors have tried and failed to obtain expressions for the asymptotic bias and variance properties of any fully iterated version of $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$ and thus to shed full light on its theoretical performance.

7. Conclusions

There has been much interest over the years in sophistications of the basic kernel density estimator. A major strand of work has focussed on single adaptations, such as higher order kernels, variable bandwidths, locations or weights, transformations and multiplicative corrections designed to improve performance. Jones and Signorini (1997) reviewed and compared many of these and came to the conclusion that "It remains debatable, however, as to whether even the best methods give worthwhile improvements, at least for small-to-moderate sample exploratory purposes". While they argued at the time that " 'failings' of such estimators thus are often due not to sub-standard methodology, but rather to limitations in the information available in the data", it has since transpired that certain methodological improvements remain possible. Two particular negatives associated with the types of method considered by Jones and Signorini are an overreliance first on asymptotics and second on reducing bias with no particular regard paid to variance (the latter drove much of Storvik's, 1999, work). Maximum likelihood kernel density estimation seeks to sweep these considerations aside in one fell swoop.

Another strand of work has sought to try to reduce the complexity and improve the interpretability of kernel density estimates by (usually complicated) schemes involving removing/reweighting kernels while not changing the density estimate appreciably. Examples include Marchette *et al.* (1996), Priebe and Marchette (2000) and Scott and Szewczyk (2001). We feel that this work, too, is somewhat undermined by the simplicity of the action of maximum likelihood kernel density estimation in automatically producing interpretable solutions of the type desired.

All that said, it does seem to us that the most fully-fledged all-varying version(s) of $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$ are not to be highly recommended for use in practice, because of a mix of performance and computational considerations. (We do not deny that improvements in our algorithm for bandwidth variation

might alter our perception somewhat.) Instead, we currently recommend for practical use the simple location-only maximum likelihood kernel density estimator, $\widehat{f}_{\mathbf{m}}$.

We repeat our admission of a failure to obtain the kinds of theoretical results we would like, either in nonparametric terms of asymptotic bias and variance or indeed in parametric terms of links to mixture modelling. We hope that publication of this paper will inspire such theory to be developed. Not only would it be informative about the quality of maximum likelihood kernel density estimation per se, but it might give assistance concerning the bandwidth selection problem, thereby replacing the computationally clumsy and, we suspect, not highly effective, cross-validation procedure used in the example of Section 5.

To finish, though, we re-emphasise that the general approach to variable location, weights and bandwidths that we have taken via maximum likelihood kernel density estimation is so conceptually simple if not indeed naive — the fixing of h being what avoids unworkable naivety — that it seems to provide an especially promising framework for further practical utilisation and development.

Acknowledgement

This work started on a visit of the first author to the Chinese University of Hong Kong in 2001. We are most grateful to Jianqing Fan for his hospitality at that time and for his continuing interest, encouragement and efforts (along with those of one of his colleagues) since.

References

Abramson, I.S. (1982) On bandwidth variation in kernel estimates — a square root law. *Ann. Statist.*, **9**, 168–176.

Choi, E. and Hall, P. (1999) Data sharpening as a prelude to density estimation. *Biometrika*, **86**, 941–947.

Hall, P. (1990) On the bias of variable bandwidth curve estimators. *Biometrika*, **77**, 529–535.

Hall, P. and Turlach, B.A. (1999) Reducing bias in curve estimation by use of weights. *Comput. Statist. Data Anal.*, **30**, 67–86.

Jones, M.C., Linton, O. and Nielsen, J.P. (1995) A simple and effective bias

reduction method for kernel density estimation. *Biometrika*, **82**, 327–338.

Jones, M.C., McKay, I.J. and Hu, T.C. (1994) Variable location and scale kernel density estimation. *Ann. Inst. Statist. Math.*, **46**, 521–535.

Jones, M.C. and Signorini, D.F. (1997) A comparison of higher order bias kernel density estimators. *J. Amer. Statist. Assoc.*, **92**, 1063–1073.

Marchette, D.J., Priebe, C.E., Rogers, G.W. and Solka, J.L. (1996) Filtered kernel density estimation. *Comput. Statist.*, **11**, 95–112.

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

McLachlan, G.J. and Peel, D. (2000) *Finite Mixture Models*. New York: Wiley.

Postman, M., Huchra, J.P. and Geller, M.J. (1986) Probes of large-scale structures in the Corona Borealis region. *Astronom. J.*, **92**, 1238–1247.

Priebe, C.E. and Marchette, D.J. (2000) Alternating kernel and mixture density estimates. *Comput. Statist. Data Anal.*, **35**, 43–65.

Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Amer. Statist. Assoc.*, **85**, 617–624.

Samiuddin, M. and el-Sayyad, G.M. (1990) On nonparametric kernel density estimates. *Biometrika*, **77**, 865–874.

Scott, D.W. and Szewczyk, W.F. (2001) From kernels to mixtures. *Technometrics*, **43**, 323–335.

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*. London: Chapman and Hall.

Simonoff, J.S. (1996) *Smoothing Methods in Statistics*. New York: Springer.

Storvik, B.E. (1999) Contributions to Nonparametric Curve Estimation. Dr. Scient. thesis, University of Oslo.

Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. London: Chapman and Hall.

Table 1. Means (standard errors in parentheses) of minimized $ISE \times 10^5$ for samples of size $n = 100$ from each of the first ten Marron-Wand densities over 1000 simulations. The median % reduction is given on the second line for each estimator.

Estimator	Density				
	Gaussian	Skewed unimodal	Strongly skewed	Kurtotic unimodal	Outlier
$\widehat{f}$	493 (13)	779 (18)	4165 (49)	4044 (56)	5188 (121)
	—	—	—	—	—
$\widehat{f}_{\mathbf{m}}$	165 (6)	442 (12)	4645 (53)	3526 (48)	2212 (67)
	72.5	45.1	−11.8	11.6	59.5
$\widehat{f}_{\mathbf{w}}$	169 (6)	445 (12)	4623 (52)	3511 (48)	2214 (68)
	72.3	45.3	−10.6	12.9	59.9
$\widehat{f}_{\mathbf{b}}$	860 (17)	1212 (23)	3429 (49)	4384 (64)	5619 (125)
	−87.8	−60.1	17.2	−9.4	−1.2
$\widehat{f}_{\mathbf{m}(1)}$	397 (10)	675 (17)	4066 (49)	3893 (55)	4289 (105)
	16.6	11.6	3.4	3.1	14.6
$\widehat{f}_{\mathbf{w}(1)}$	363 (11)	623 (16)	4263 (50)	3781 (53)	3948 (106)
	31.9	22.7	−2.1	7.3	27.7
$\widehat{f}_{\mathbf{b}(1)}$	613 (14)	942 (20)	4183 (49)	4418 (60)	5609 (125)
	−28.8	−23.5	−0.0	−8.5	−1.2
$\widehat{f}_{\mathbf{m},\mathbf{w}}$	164 (6)	442 (12)	4659 (52)	3523 (48)	2200 (68)
	73.1	46.0	−11.5	12.3	60.1
$\widehat{f}_{\mathbf{m},\mathbf{b}}$	250 (8)	621 (16)	4062 (62)	3760 (77)	5157 (135)
	54.0	24.3	4.6	13.3	10.6
$\widehat{f}_{\mathbf{w},\mathbf{b}}$	203 (7)	593 (15)	3200 (48)	3357 (58)	2831 (78)
	65.3	24.1	26.9	14.5	47.3
$\widehat{f}_{\mathbf{m},\mathbf{w}(1)}$	249 (8)	506 (14)	4360 (52)	3697 (51)	2945 (85)
	56.7	39.0	−3.4	9.6	47.3
$\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$	247 (9)	557 (14)	3693 (58)	2940 (62)	4040 (109)
	57.2	30.3	16.5	33.6	29.7

Table 1 Continued.

Estimator	Density				
	Bimodal	Separated bimodal	Skewed bimodal	Trimodal	Claw
$\widehat{f}$	706 (13)	1099 (19)	903 (14)	860 (14)	3540 (34)
	—	—	—	—	—
$\widehat{f}_{\mathbf{m}}$	537 (13)	633 (15)	938 (16)	791 (14)	3429 (37)
	28.5	45.5	−1.6	9.1	3.1
$\widehat{f}_{\mathbf{w}}$	540 (13)	638 (16)	939 (16)	798 (14)	3473 (36)
	28.3	44.8	−1.8	8.4	2.0
$\widehat{f}_{\mathbf{b}}$	952 (15)	1579 (22)	1004 (15)	1038 (14)	3654 (33)
	−38.3	−49.0	−11.3	−23.7	−1.9
$\widehat{f}_{\mathbf{m}(1)}$	611 (12)	896 (17)	850 (14)	786 (13)	3477 (34)
	13.8	17.7	6.2	8.3	2.1
$\widehat{f}_{\mathbf{w}(1)}$	681 (13)	933 (19)	909 (15)	847 (14)	3577 (35)
	6.6	16.8	0.3	2.5	−0.8
$\widehat{f}_{\mathbf{b}(1)}$	744 (13)	1302 (21)	900 (14)	890 (13)	3561 (32)
	−7.9	−20.0	−1.4	−5.7	−0.2
$\widehat{f}_{\mathbf{m},\mathbf{w}}$	539 (13)	637 (16)	946 (15)	798 (14)	3485 (36)
	28.8	44.8	−3.1	8.0	1.5
$\widehat{f}_{\mathbf{m},\mathbf{b}}$	829 (16)	878 (18)	825 (17)	1004 (17)	4333 (26)
	−7.4	21.4	16.7	−11.0	−22.4
$\widehat{f}_{\mathbf{w},\mathbf{b}}$	804 (16)	734 (17)	859 (15)	978 (15)	3829 (36)
	−10.1	35.3	5.9	−12.1	−5.5
$\widehat{f}_{\mathbf{m},\mathbf{w}(1)}$	610 (13)	753 (17)	892 (15)	807 (15)	3595 (36)
	17.7	34.2	3.3	8.7	−1.2
$\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$	813 (16)	806 (17)	833 (19)	1053 (18)	4495 (32)
	−1.7	27.3	17.3	−15.1	−25.2

NOTE: The estimators $\widehat{f}, \widehat{f}_{\mathbf{m}}, \widehat{f}_{\mathbf{w}}, \widehat{f}_{\mathbf{b}}, \widehat{f}_{\mathbf{m},\mathbf{w}}$, $\widehat{f}_{\mathbf{m},\mathbf{b}}$, $\widehat{f}_{\mathbf{w},\mathbf{b}}$ and $\widehat{f}_{\mathbf{m},\mathbf{w},\mathbf{b}}$ are the basic and fully iterated variable location, variable weight, variable bandwidth, combined variable location and weight, combined variable location and bandwidth, combined variable weight and bandwidth and combined variable location, weight and bandwidth kernel density estimators, respectively. The subscript (1) indicates alternative versions where the algorithm was stopped after one iteration.

Table 2. Means (standard errors in parentheses) of minimized $ISE \times 10^5$ for samples of size $n = 500$ from each of the first ten Marron-Wand densities over 1000 simulations. The median % reduction is given on the second line for each estimator.

Estimator	Density				
	Gaussian	Skewed unimodal	Strongly skewed	Kurtotic unimodal	Outlier
$\widehat{f}$	158 (3)	247 (5)	1378 (15)	1249 (16)	1634 (35)
	—	—	—	—	—
$\widehat{f}_{\mathbf{m}}$	35 (1)	109 (3)	1439 (14)	988 (12)	571 (16)
	84.6	59.7	−2.7	21.0	67.9
	Bimodal	Separated bimodal	Skewed bimodal	Trimodal	Claw
$\widehat{f}$	230 (4)	325 (5)	301 (5)	284 (4)	1117 (11)
	—	—	—	—	—
$\widehat{f}_{\mathbf{m}}$	128 (3)	132 (3)	275 (5)	258 (4)	884 (10)
	48.5	63.1	10.3	10.3	21.6

NOTE: The estimators $\widehat{f}$ and $\widehat{f}_{\mathbf{m}}$ are the basic and fully iterated variable location kernel density estimators, respectively.

Table 3. The normal mixture components associated with $\widehat{f}_{\mathbf{m}}(\cdot; 0.79)$ for the galaxy velocity data, $n = 82$. Each has variance $0.79^2 = 0.624$.

Component	Weight ($\times 82$)	Mean
1	7	9.710
2	2	16.138
3	36	19.876
4	19	22.507
5	12	23.885
6	3	26.599
7	2	32.561
8	1	34.014

Fig. 1. Locations $m(X_i)$ associated with $\widehat{f}_{\mathbf{m}}$ plotted against X_i, $i = 1, ..., 100$, for an approximately median performance simulation from Model 2, the skewed unimodal density ($h = 0.7$).

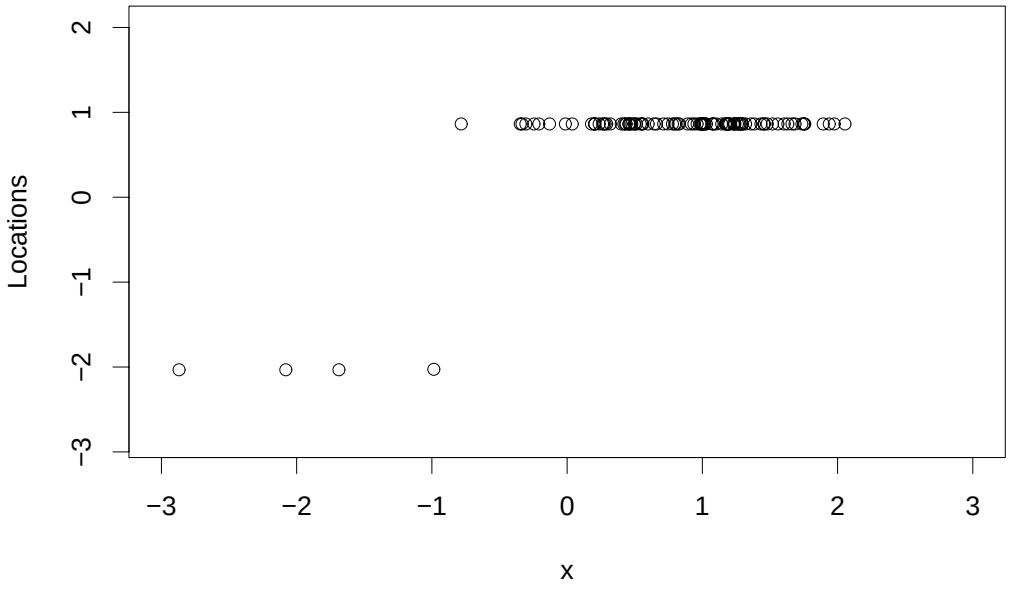

Fig. 2. Weights $w(X_i)$ associated with $\widehat{f}_{\mathbf{w}}$ plotted against X_i, $i = 1, ..., 100$, for the same simulation from Model 2, the skewed unimodal density, as Fig. 1 ($h = 0.7$).

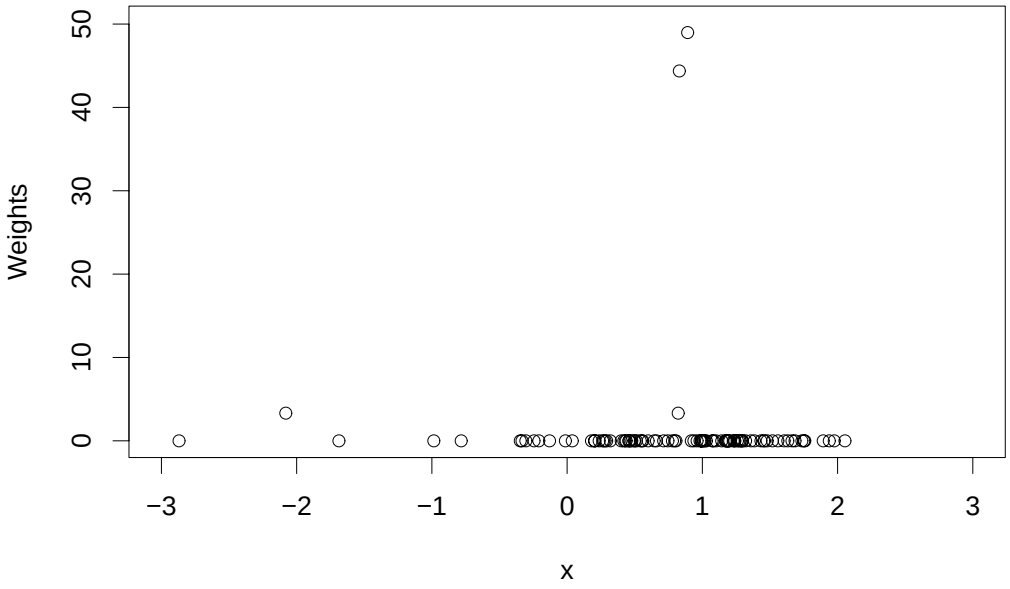

Fig. 3. Model 2 density (solid line) and two component normal mixtures suggested by $\widehat{f}_{\mathbf{m}}$ (dashed line) and $\widehat{f}_{\mathbf{w}}$ (dotted line) of Figs 1 and 2.

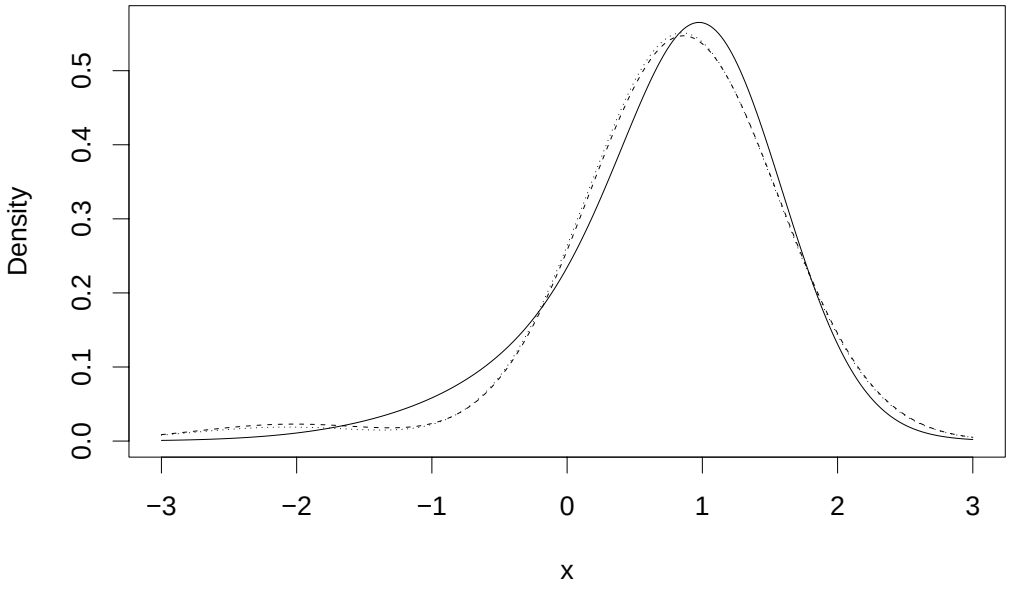

Fig. 4. Locations $m(X_i)$ and weights $w(X_i)$ associated with $\widehat{f}_{\mathbf{m},\mathbf{w}}$ plotted against X_i, $i = 1, ..., 100$, for the same simulation from Model 2, the skewed unimodal density, as Figs 1, 2 and 3 ($h = 0.7$).

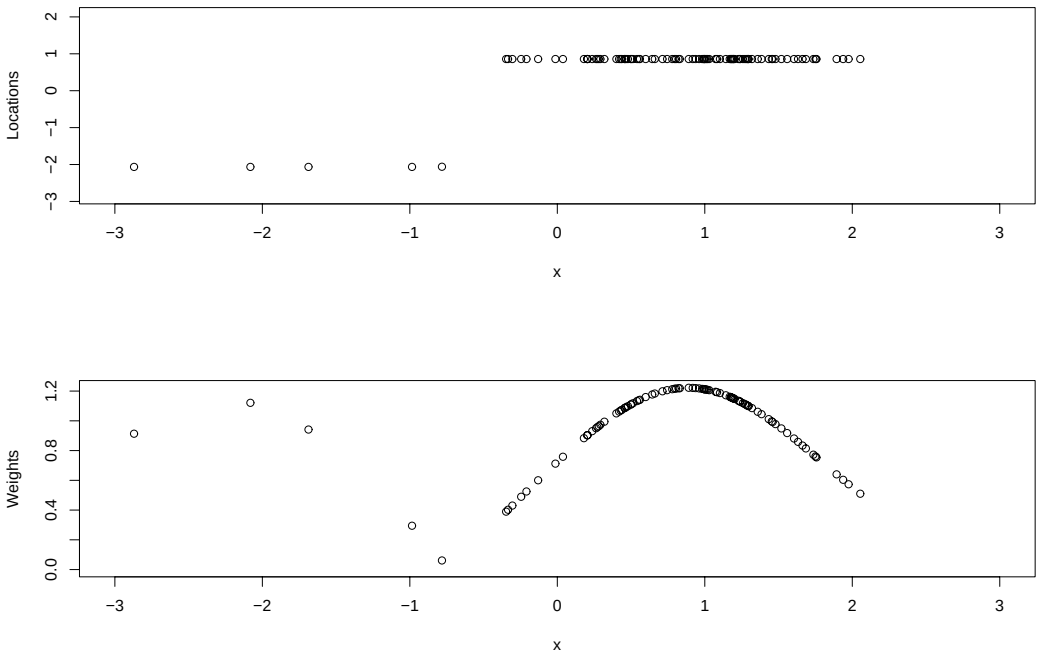

Fig. 5. Bandwidths $b(X_i)$ associated with $\widehat{f}_{\mathbf{b}}$ plotted against X_i, $i = 1, ..., 100$, for the same simulation from Model 2, the skewed unimodal density, as Figs 1, 2, 3 and 4 ($h = 0.7$).

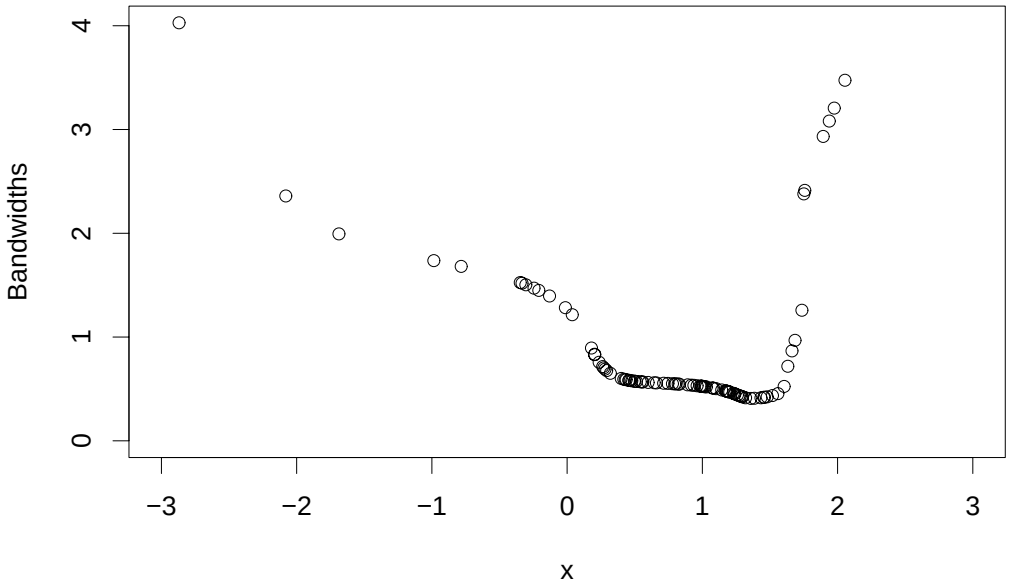

Fig. 6. The variable location density estimate $\widehat{f}_{\mathbf{m}}$ with $h = 0.79$ for the galaxy velocity data. The datapoints are plotted as vertically jittered circles. The eight distinct locations involved in $\widehat{f}_{\mathbf{m}}(\cdot; 0.79)$ are shown by crosses.