Nonparametric density estimation with a parametric start

Nils Lid Hjort[1] and Ingrid K. Glad[2]

University of Oslo and

the Norwegian Institute of Technology

ABSTRACT. The traditional kernel density estimator of an unknown density is by construction completely nonparametric, in the sense that it has no preferences and will work reasonably well for all shapes. The present paper develops a class of semiparametric methods that are designed to work better than the kernel estimator in a broad nonparametric neighbourhood of a given parametric class of densities, for example the normal, while not losing much in precision when the true density is far from the parametric class. The idea is to multiply an initial parametric density estimate with a kernel type estimate of the necessary correction factor. This works well in cases where the correction factor function is less rough than the original density itself. Extensive comparisons with the kernel estimator are carried out, including exact analysis for the class of all normal mixtures. The new method, with a normal start, wins quite often, even in many cases where the true density is far from normal. Procedures for choosing the smoothing parameter of the estimator are also discussed. The new estimator should be particularly useful in higher dimensions, where the usual nonparametric methods have problems. The idea is also spelled out for nonparametric regression.

KEY WORDS: *bandwidth selection, correction factor, kernel methods, lowering the bias, semiparametric density estimation, test cases*

1. Introduction and summary. Let $X_1, \ldots, X_n$ be independent observations from an unknown density f on the real line. The traditional nonparametric density estimator is

$$\widetilde{f}(x) = \frac{1}{n}\sum_{i=1}^{n} h^{-1}K(h^{-1}(X_i - x)) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x), \tag{1.1}$$

where $K_h(z) = h^{-1}K(h^{-1}z)$ and $K(z)$ is a kernel function, which is taken here to be a symmetric probability density with finite values of $\sigma_K^2 = \int z^2 K(z)\,\mathrm{d}z$ and $R(K) = \int K(z)^2\,\mathrm{d}z$. The basic statistical properties are that

$$\mathrm{E}\widetilde{f}(x) \doteq f(x) + \tfrac{1}{2}\sigma_K^2 h^2 f''(x) \quad \text{and} \quad \mathrm{Var}\,\widetilde{f}(x) \doteq R(K)(nh)^{-1} f(x) - f(x)^2/n. \tag{1.2}$$

The integrated mean squared error is of order $n^{-4/5}$ when h is proportional to $n^{-1/5}$, which is the optimal size. See Scott (1992, Chapter 6) and Wand & Jones (1994, Chapter 2) for recent accounts of the theory.

Method (1.1) is totally nonparametric and admirably impartial to special types of shapes of the underlying density. The intention of the present paper is to construct competitors to (1.1) with properties that are generally similar but indeed better in the broad vicinity of given parametric families. The basic idea is to start out with a parametric density estimate $f(x, \widehat{\theta})$, say the normal, and then multiply with a nonparametric kernel type estimate of the correction function $r(x) = f(x)/f(x, \widehat{\theta})$. Our proposal is $\widehat{r}(x) = n^{-1}\sum_{i=1}^{n} K_h(X_i - x)/f(X_i, \widehat{\theta})$, producing

$$\widehat{f}(x) = f(x, \widehat{\theta})\widehat{r}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\frac{f(x, \widehat{\theta})}{f(X_i, \widehat{\theta})}. \tag{1.3}$$

We emphasise that the initial parametric estimate is not (necessarily) intended to provide a serious approximation to the true density; our method will often work well even if the parametric description is quite crude. The case of a constant start value for $f(x, \theta)$, corresponding to choosing a uniform distribution as the initial description, gives back the classic kernel estimator (1.1).

The basic bias and variance properties of the new estimator (1.3) are investigated in Section 2, treating the simplest case of a non-random start function $f_0(x)$, and in Section 3, covering a broad class of parametric start estimators. It turns out that the variance of the (1.3) estimator is simply the same as the variance of the traditional (1.1) estimator, to the order of approximation used, while the bias is quite similar in structure to (1.2), and often smaller. Comparisons with the traditional estimator (1.1) are made in Sections 4 and 5. It is seen that the new method generally is the better one in cases where the correction function is less 'rough' than the original density, in a sense made precise in Section 4, and illustrated there in the realm of Hermite expansions around the normal.

Further analysis is provided in Section 5, for the version of (1.3) that starts with the normal, comparing behaviour with the kernel method when the true density belongs to the large class of all normal mixtures. There and in the paper's appendix comparative formulae are developed for exact analysis of asymptotic mean squared error as well as for exact finite-sample mean squared error. The results are illuminated by working through a list of 15 'test densities' proposed by Marron & Wand (1992), chosen to exhibit a broad range of distributional shapes. The new 'nonparametrically corrected normal estimate' outperforms the usual kernel method in 12 of these 15 test cases, and in all the 'not drastically unreasonable' cases, in terms of approximate mean integrated squared error. The same pattern is observed for finite sample sizes.

The bottom line is that (1.3) will be more precise than (1.1) in a broad nonparametric neighbourhood around the parametric family, while at the same time losing surprisingly little, or not at all, when the true density is far from the parametric family. One explanation is that the uniform prior description, which in the light of (1.3) is the implicit start estimator for the kernel estimator (1.1), is overly conservative and less advantageous than say the normal, even in quite non-normal cases.

The problem of selecting a good smoothing parameter is discussed in Section 6, and some solutions are outlined, including versions of plug-in and cross validation. Our method also works well in the multi-dimensional case, starting out for example with a multi-normal start estimate, as demonstrated in Section 7. The method should be particularly useful in the higher-dimensional case since the ordinary nonparametric methods, including the kernel method, are quite imprecise then. Our paper ends with some supplementary comments in Section 8. In particular Remark 8E spells out the corresponding estimation idea for nonparametric regression, giving a generalised Nadaraya–Watson estimator.

Our estimators can be viewed as semiparametric in that they combine parametric and nonparametric methods. They are as such in the same realm as recent methods of Hjort (1993) and Hjort & Jones (1993). These latter methods are quite different but also have the property that the variance is approximately the same as in (1.2) while the bias is similar but sometimes smaller. The (1.3) method is

also similar in spirit to the projection pursuit density estimation methods, see for example Friedman, Stuetzle & Schroeder (1984), and also to the normal times Hermite expansion method, see for example Hjort (1986), Buckland (1992), and Hjort & Fenstad (1994). A somewhat less attractive semiparametric method is that of Schuster & Yakowitz (1985) and Olkin & Spiegelman (1987), see the discussion in Jones (1993). Various semiparametric Bayesian density estimators are proposed in Hjort (1994).

Another semiparametric technique, perhaps mildly related to our new method, is the transformation idea of Wand, Marron & Ruppert (1991), where data are semiparametrically transformed so as to work well with a non-adaptive constant smoothing parameter, and then ending in a back-transformed density estimator. This is a promising way of using an adaptive smoothing parameter, and our estimator can be seen as as having similar intentions. In other words, (1.3) can be seen as being similar in spirit to a suitable semiparametrically adaptive $n^{-1}\sum_{i=1}^n K_{h(x,\widehat\theta)}(X_i-x)$. Finally we mention a recent bias reduction method due to Jones, Linton & Nielsen (1993). Our (1.3) idea is to start with any parametric estimator and then multiply with a nonparametric correction function, and in essence this does not affect the variance but changes the bias. Serendipitously and independently of the present authors Jones, Linton & Nielsen (1993) use essentially the same idea but in a totally nonparametric mode, correcting the initial kernel estimator with a nonparametric correction factor in the (3.1) manner. This typically gives a smaller bias but a somewhat larger variance.

2. Nonparametric correction on a fixed start. Suppose f_0 is a fixed density, perhaps a crude guess of f. Write $f = f_0 r$. The idea is to estimate the nonparametric correction factor r via kernel smoothing. One version of this is $\widehat r(x) = n^{-1}\sum_{i=1}^n K_h(X_i-x)/f_0(X_i)$, with ensuing estimator

$$\widehat f(x) = f_0(x)\widehat r(x) = \frac{1}{n}\sum_{i=1}^n K_h(X_i-x)\frac{f_0(x)}{f_0(X_i)}. \tag{2.1}$$

Note that a constant $f_0(x)$ gives back the ordinary kernel estimator (1.1). We have

$$\begin{aligned}\mathrm{E}\widehat r(x) &= \int K_h(y-x)f_0(y)^{-1}f(y)\,\mathrm{d}y \\ &= \int K(z)r(x+hz)\,\mathrm{d}z = r(x) + \tfrac12\sigma_K^2h^2r''(x) + O(h^4),\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}\,\widehat r(x) &= \frac1n\Big[\int \frac{K_h(y-x)^2}{f_0(y)^2}f(y)\,\mathrm{d}y - \{\mathrm{E}\widehat r(x)\}^2\Big] \\ &= \frac{R(K)}{nh}\frac{f(x)}{f_0(x)^2} - \frac{r(x)^2}{n} + O(h/n),\end{aligned}$$

by a variation of the arguments traditionally used to establish (1.2). This shows that the (2.1) estimator has

$$\text{bias} \doteq \tfrac12\sigma_K^2h^2f_0(x)r''(x) \quad\text{and}\quad \text{variance} \doteq R(K)(nh)^{-1}f(x) - f(x)^2/n. \tag{2.2}$$

In other words, the variance is of the very same size as that of the traditional estimator, to the order of approximation used, and the bias is of the same order h^2, but proportional to $f_0 r''$ rather than to f''. The new estimator is better than the traditional one in all cases where $f_0 r''$ is smaller in size than $f'' = f_0'' r + 2 f_0' r' + f_0 r''$. In cases where f_0 is already a good guess one expects r near constant and r'' small, so this describes a certain neighbourhood of densities around f_0 where the new method is better than the traditional one. This is further discussed and exemplified in Section 4.

3. Nonparametric correction on a parametric start. Let $f(x,\theta)$ be a given parametric family of densities, where the possibly multi-dimensional parameter $\theta = (\theta_1, \ldots, \theta_p)'$ belongs to some open and connected region in p-space. The parametric start estimate is $f(x,\widehat{\theta})$, where we for concreteness let $\widehat{\theta}$ be the maximum likelihood estimator (quite general estimators for θ are allowed later). Thus $f(x,\widehat{\theta})$ could be the estimated normal density, for example, or an estimated mixture of two normals. This initial data summary is not necessarily meant to be a serious description of the true density; the method we will develop is intended to work well even if f cannot be well approximated by any $f(.,\theta)$.

The task is to estimate the necessary correction function $f(x)/f(x,\widehat{\theta})$ by kernel smoothing means. In view of Section 2 $\widehat{r}(x) = n^{-1}\sum_{i=1}^n K_h(X_i - x)/f(X_i,\widehat{\theta})$ is a natural choice. In other words,

$$\widehat{f}(x) = f(x,\widehat{\theta})\frac{1}{n}\sum_{i=1}^n K_h(X_i - x)/f(X_i,\widehat{\theta}). \tag{3.1}$$

In order to understand to what extent the parametric estimation makes this estimator quantitatively different from the cleaner version (2.1), we bring in facts about the behaviour of the maximum likelihood estimator outside model conditions. It aims at a certain θ_0, the least false value according to the Kullback–Leibler distance measure $\int f(x)\log\{f(x)/f(x,\theta)\}\,dx$ from true f to approximant $f(.,\theta)$. Write $f_0(x) = f(x,\theta_0)$ for this best parametric approximant, and let $u_0(x) = \partial \log f(x,\theta_0)/\partial\theta$ be the score function evaluated at this parameter value. A Taylor expansion gives

$$\begin{aligned}\frac{f(x,\widehat{\theta})}{f(X_i,\widehat{\theta})} &= \exp\{\log f(x,\widehat{\theta}) - \log f(X_i,\widehat{\theta})\} \\ &\doteq \frac{f_0(x)}{f_0(X_i)} + \frac{f_0(x)}{f_0(X_i)}\{u_0(x) - u_0(X_i)\}'(\widehat{\theta} - \theta_0),\end{aligned} \tag{3.2}$$

leading to

$$\begin{aligned}\widehat{f}(x) &\doteq \frac{1}{n}\sum_{i=1}^n K_h(X_i - x)\frac{f_0(x)}{f_0(X_i)}\left[1 - \{u_0(X_i) - u_0(x)\}'(\widehat{\theta} - \theta_0)\right] \\ &= f^*(x) + V_n(x),\end{aligned} \tag{3.3}$$

say. Here f^* is as in (2.1), except for the fact that the f_0 function appearing here is not directly visible, and the $V_n(x)$ term stems from the parametric estimation variability.

Representation (3.3), in concert with expressing $\widehat{\theta}-\theta_0$ as an average of i.i.d. zero mean variables plus remainder term, can now be used to establish approximate bias and variance results for $\widehat{f}(x)$. We shall be somewhat more general and allow arbitrary regular estimators having an influence with finite covariance matrix. To define this properly, let F be the true distribution, the cumulative of f, and let F_n be the empirical distribution function. We consider functional estimators of θ of the form $\widehat{\theta} = T(F_n)$ with influence function $I(x) = \lim_{\varepsilon\to 0}\{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)\}/\varepsilon$, writing δ_x for unit point mass at x, and assume that $\Sigma = \mathrm{E}_f I(X_i)I(X_i)'$ is finite. The best approximant $f_0(x) = f(x,\theta_0)$ to $f(x)$ that $f(x,\widehat{\theta})$ aims for is determined by $\theta_0 = T(F)$. Under mild regularity conditions, see for example Huber (1981) or Shao (1991), one has

$$\widehat{\theta} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} I(X_i) + \frac{d}{n} + \varepsilon_n, \tag{3.4}$$

where $\varepsilon_n = O_p(n^{-1})$ with mean $O(n^{-2})$, i.e. d/n is essentially the bias of $\widehat{\theta}$. It is generally possible to de-bias the estimator, for example by jackknifing or bootstrapping, making the d/n term disappear. The maximum likelihood case corresponds to $I(x) = J^{-1}u_0(x)$ where $J = -\mathrm{E}_f\partial^2 \log f(X_i,\theta_0)/\partial\theta\partial\theta'$.

PROPOSITION. *Let $f_0(x) = f(x,\theta_0)$ with $\theta_0 = T(F)$ be the best parametric approximant to f, and let $r = f/f_0$. The semiparametric estimator (3.1) has*

$$\mathrm{E}\widehat{f}(x) = f(x) + \tfrac{1}{2}\sigma_K^2 h^2 f_0(x)r''(x) + O(h^2/n + h^4 + n^{-2})$$
$$\text{and}\quad \mathrm{Var}\,\widehat{f}(x) = R(K)(nh)^{-1}\,f(x) - f(x)^2/n + O(h/n + n^{-2}).$$

PROOF: The detailed proof we present needs a second order Taylor approximation version of the simpler first order Taylor versions (3.2)–(3.3). This more complete approximation becomes $\widehat{f}(x) = f^*(x) + V_n(x) + \frac{1}{2}W_n(x)$, where we write $f^*(x) = \bar{A}_n$, $V_n(x) = \bar{B}_n'(\widehat{\theta}-\theta_0)$, and $W_n(x) = (\widehat{\theta}-\theta_0)'\bar{C}_n(\widehat{\theta}-\theta_0)$. The representations are in terms of averages of i.i.d. variables

$$\begin{aligned}
A_i &= K_h(X_i - x)f_0(x)/f_0(X_i),\\
B_i &= -K_h(X_i - x)\{f_0(x)/f_0(X_i)\}\,\{u_0(X_i) - u_0(x)\},\\
C_i &= K_h(X_i - x)\{f_0(x)/f_0(X_i)\}\,w(x,X_i),
\end{aligned}$$

where in fact $w(x,X_i) = v_0(x) - v_0(X_i) + \{u_0(x) - u_0(X_i)\}\{u_0(x) - u_0(X_i)\}'$.

Starting with the expected value, we already know that f^* has mean $f(x) + \frac{1}{2}\sigma_K^2 h^2 f_0(x)r''(x) + O(h^4)$. Through (3.4) and the averages representations above one finds $\mathrm{E}V_n(x) = n^{-1}\mathrm{E}B_i'I_i + n^{-1}(\mathrm{E}B_i)'d + O(n^{-2})$ and $\mathrm{E}W_n(x) = n^{-1}\mathrm{Tr}(\mathrm{E}C_i\,\mathrm{E}I_iI_i') + O(n^{-2})$, using the fact that $I_i = I(X_i)$ has mean zero. But it is not difficult to see that each of $\mathrm{E}B_i$, $\mathrm{E}B_i'I_i$, $\mathrm{E}C_i$ is of size $O(h^2)$; for example,

$$\begin{aligned}
\mathrm{E}B_i'I_i &= -\int K_h(y-x)\frac{f_0(x)}{f_0(y)}\{u_0(y) - u_0(x)\}'I(y)f(y)\,dy\\
&= -\int K(z)f_0(x)\{u_0(x+hz) - u_0(x)\}'(Ir)(x+hz)\,\mathrm{d}z\\
&= -h^2\sigma_K^2 f_0(x)\{u_0'(x)'(Ir)'(x) + \tfrac{1}{2}u_0''(x)'(Ir)(x)\} + O(h^4).
\end{aligned}$$

One can also see that the remainder of the second order Taylor approximation used, involving $(\widehat{\theta}_i - \theta_{0,i})^3$ terms, is of size $O_p(n^{-2})$. Thus the bias of $\widehat{f}(x)$ is $\frac{1}{2}\sigma_K^2 h^2 f_0(x) r''(x) + (h^2/n)b(x) + O(h^4 + n^{-2})$, for a certain $b(x)$ function.

Next turn to the variance. The variance of $f^*(x)$ is known from Section 2. From (3.4) and the representation above one finds $\mathrm{Var}\, V_n(x) = \mathrm{Var}(\bar{B}_n' \bar{I}_n) + O(n^{-2}) = n^{-1}(\mathrm{E}B_i)'\Sigma(\mathrm{E}B_i) - \{O(h^2/n)\}^2 + O(n^{-2}) = O(h^4/n + n^{-2})$, and similarly $W_n(x)$ can be seen to have uninfluential variance $O(h^4/n^2)$. Finally $\mathrm{cov}\{f^*(x), V_n(x)\} = n^{-1}(\mathrm{E}B_i)'\mathrm{E}A_i I_i + O(n^{-2}) = O(h^2/n)$. This combines to give the necessary variance expression. □

The result is remarkable in its simplicity; the sizes of bias and variance are only affected by parametric estimation noise to the quite small $O(h^2/n + n^{-2})$ order. The reason lies with (3.2); not only is $\widehat{\theta}$ close to θ_0, but the $\widehat{f}(x)$ estimator uses only X_is that are close to x, making $u_0(X_i)$ close to $u_0(x)$. The story is somewhat different for the correction term $\widehat{r}(x)$ alone, see Remark 8B.

Consistency of the density estimator requires both $h \to 0$ (forcing the bias towards zero) and $nh \to \infty$ (making the variance go to zero). The optimal size of h will later be seen to be proportional to $n^{-1/5}$. These observations match the traditional facts for the classic (1.1) estimator. Note also that if the parametric model happens to be accurate, then the r function is equal to 1, and the bias is only $O(h^4 + h^2/n)$.

EXAMPLE 1: NORMAL START ESTIMATE. The normal start estimate is of the form $\widehat{\sigma}^{-1}\phi(\widehat{\sigma}^{-1}(x - \widehat{\mu}))$, where one can use maximum likelihood estimates $\widehat{\mu} = n^{-1}\sum_{i=1}^n X_i$ and $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \widehat{\mu})^2$ (or the de-biased version with denominator $n-1$). In view of the generality of the proposition above quite general estimators are allowed, without changing the basic structure of bias and variance of $\widehat{f}(x)$. One might for example wish to use robust estimates of mean and standard deviation. In any case the density estimator is

$$\begin{aligned} \widehat{f}(x) &= \frac{1}{\widehat{\sigma}}\phi\Big(\frac{x-\widehat{\mu}}{\widehat{\sigma}}\Big)\frac{1}{n}\sum_{i=1}^n K_h(X_i - x)\Big/\frac{1}{\widehat{\sigma}}\phi\Big(\frac{X_i-\widehat{\mu}}{\widehat{\sigma}}\Big) \\ &= \frac{1}{n}\sum_{i=1}^n K_h(X_i - x)\frac{\exp\{-\frac{1}{2}(x-\widehat{\mu})^2/\widehat{\sigma}^2\}}{\exp\{-\frac{1}{2}(X_i-\widehat{\mu})^2/\widehat{\sigma}^2\}}. \end{aligned} \tag{3.5}$$

Note that its implementation is straightforward.

EXAMPLE 2: LOG-NORMAL START ESTIMATE. One option for positive data is to start with a log-normal approximation and then multiply with a correction factor. The result is

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^n K_h(X_i - x)\frac{\exp\{-\frac{1}{2}(\log x - \widehat{\mu})^2/\widehat{\sigma}^2\}}{\exp\{-\frac{1}{2}(\log X_i - \widehat{\mu})^2/\widehat{\sigma}^2\}}\frac{X_i}{x}.$$

EXAMPLE 3: GAMMA START ESTIMATE. A version of the general method which should work well for positive data from perhaps unimodal and right-skewed distributions is to start with a gamma distribution approximation. The final estimator is then of the form

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^n K_h(X_i - x)(x/X_i)^{\widehat{\alpha}-1}\exp\{-\widehat{\beta}(x - X_i)\},$$

for example with moment estimates for the gamma parameters.

EXAMPLE 4: NORMAL MIXTURE START ESTIMATE. We believe proper use of the three special cases mentioned now would work satisfactorily in many applications. Most unimodal densities would be approximable with either a normal, a log-normal or a gamma, perhaps after a transformation. Cases where still other tactics might prove superior include densities exhibiting two or more bumps. One method in such cases would be to fit a normal mixture first and use that as the $f(x,\widehat{\theta})$, correcting afterwards with a $\widehat{r}(x)$.

REMARK 1. The correction factor $f(x,\widehat{\theta})/f(X_i,\widehat{\theta})$ can occasionally be too influential, in cases where the denominator is too small. This is not a problem for small h since then only X_is quite close to x matter, but it can happen for moderate h and for lonely data points. An effective safety procedure is to replace the original $f(x,\widehat{\theta})$ function with a somewhat adjusted $\bar{f}(x,\widehat{\theta})$, bounding it suitably away from zero. In the normal case we advocate putting $\bar{f}(x,\theta)$ equal to $f(\widehat{\mu}\pm 2.5\,\widehat{\sigma},\widehat{\mu},\widehat{\sigma}) = (2\pi)^{-1/2}\widehat{\sigma}^{-1}\exp(-2.5^2/2)$ for $|x-\widehat{\mu}|\geq 2.5\,\widehat{\sigma}$.

REMARK 2. We have developed a method that can be used for any given parametric model. It is intuitively clear that the method works best in cases where the model employed is not too far from covering the truth (and this is borne out by precise analysis in the following sections). One could think of ways of automatising the choice of the parametric vehicle model, through suitable goodness of fit measures, thereby obtaining an overall adaptive density estimator, but this is not pursued here.

4. Comparison with the traditional kernel density estimator. In this and the following section the performance of the new estimator is compared to that of the usual (1.1) estimator. We look into a couple of 'test areas', that is, classes of densities for which comparison of behaviour can be carried out. In 4B and 4C below we study two versions of Hermite expansions around the normal density. The calculations we give for these turn out to be useful also in connection with the problem of choosing the bandwidth parameter h, see Section 6. The second test area is that of finite normal mixtures, studied in Section 5 and in the Appendix, with attention given to the list of 15 test densities chosen by Marron & Wand (1992).

4A. GENERAL MSE AND MISE COMPARISON. Expressions can be found for the leading terms of the integrated mean squared errors of the usual kernel estimator (1.1) and the new estimator (3.1), using respectively (1.2) and the proposition of Section 3. We find

$$\begin{aligned}\text{amise for } \widetilde{f} &= \tfrac{1}{4}\sigma_K^4 h^4 R_{\rm trad}(f) + R(K)(nh)^{-1},\\ \text{amise for } \widehat{f} &= \tfrac{1}{4}\sigma_K^4 h^4 R_{\rm new}(f) + R(K)(nh)^{-1},\end{aligned} \tag{4.1}$$

featuring 'roughness' functionals

$$R_{\rm trad}(f) = \int\{f''(x)\}^2\,{\rm d}x \quad\text{and}\quad R_{\rm new}(f) = \int\{f_0(x)r''(x)\}^2\,{\rm d}x. \tag{4.2}$$

The new estimator is better, in the sense of approximate (leading terms) integrated mean squared error, whenever $R_{\rm new}(f)$ is smaller than $R_{\rm trad}(f)$. This defines a nonparametric neighbourhood of densities around the parametric class. When f belongs to this neighbourhood, $\widehat{f}$ is better than $\widetilde{f}$ when the same K and the same h

are used in the two estimators. In such a case the new estimator can be made even better by choosing an appropriate h, see Section 6.

It is also of interest to see in which x-regions the new estimator is better than the traditional one. Write $f = \exp(g)$ and $f_0 = \exp(g_0)$. Then

$$f'' = f\{g'' + (g')^2\} \quad \text{while} \quad f_0 r'' = f\{g'' - g_0'' + (g' - g_0')^2\}. \tag{4.3}$$

This is useful for actual inspection of the bias terms for different fs, and is attractive in that it clearly exhibits the roles of the first and second log-derivatives. Note in particular that if the parametric model used is good enough to secure $|g' - g_0'| \le |g'|$ and $|g'' - g_0''| \le |g''|$, for a region of relevant xs, then that clearly suffices for the new method to be better than the traditional one. These requirements can also be written $0 \le g_0'/g' \le 2$ and $0 \le g_0''/g'' \le 2$.

4B. First test-bed: Hermite expansions. A test area where these matters can be explored is in the context of the Hermite expansions considered (for other purposes) in Hjort & Jones (1994) and in Hjort & Fenstad (1994). Let $H_j(x)$ be the jth Hermite polynomial, given by $\phi^{(j)}(x) = (-1)^j \phi(x) H_j(x)$. Consider the Hermite expansion representation

$$f(x) = \phi\Big(\frac{x-\mu}{\sigma}\Big)\frac{1}{\sigma}\Big\{1 + \sum_{j=3}^{m} \frac{\gamma_j}{j!} H_j\Big(\frac{x-\mu}{\sigma}\Big)\Big\} = g\Big(\frac{x-\mu}{\sigma}\Big)\frac{1}{\sigma}, \tag{4.4}$$

writing $g(y) = \phi(y)\{1 + \sum_{j=3}^{m}(\gamma_j/j!)H_j(y)\}$. Its mean is μ and its standard deviation is σ, and $\gamma_j = \mathrm{E}H_j((X-\mu)/\sigma)$. Note that $\gamma_0 = 1$ and that $\gamma_1 = \gamma_2 = 0$, while

$$\gamma_3 = \mathrm{E}\Big(\frac{X-\mu}{\sigma}\Big)^3,\ \gamma_4 = \mathrm{E}\Big(\frac{X-\mu}{\sigma}\Big)^4,\ \gamma_5 = \mathrm{E}\Big(\frac{X-\mu}{\sigma}\Big)^5 - 10\,\mathrm{E}\Big(\frac{X-\mu}{\sigma}\Big)^3,$$

and so on, featuring skewness, kurtosis, pentakosis and so on, all of which are zero for the normal density. Any density with finite moments can be approximated with one of the form (4.6), through inclusion of enough terms. See Hjort & Jones (1994) for details pertaining to this and some of the following calculations.

Assume that the true f is as in (4.4) and that the normal-corrected estimator (3.5) is used, so that $f = f_0 r$ with f_0 being the simple normal approximation and $r(x) = r_0(y)$, where $r_0(y) = \sum_{j=0}^{m}(\gamma_j/j!)H_j(y)$, writing $y = (x-\mu)/\sigma$. Then

$$f''(x) = \sigma^{-3} g''(y) = \sigma^{-3}\phi(y)\sum_{j=0}^{m}(\gamma_j/j!)H_{j+2}(y),$$

$$f_0(x)r''(x) = \sigma^{-3}\phi(y) r_0''(y) = \sigma^{-3}\phi(y)\sum_{j=2}^{m}(\gamma_j/j!)j(j-1)H_{j-2}(y).$$

Calculations give that $A_{j,k} = \int H_j H_k \phi^2\,\mathrm{d}y$ is zero when $j+k$ is odd and equal to $(-1)^{j+p}(2\sqrt{\pi})^{-1}(2p)!/(p!2^{2p})$ when $j+k = 2p$, see Hjort & Jones (1994). This makes it possible to evaluate

$$R_{\mathrm{trad}}(f) = \sigma^{-5}\sum_{j,k\le m}(\gamma_j/j!)(\gamma_k/k!)\,A_{j+2,k+2}$$

$$\text{and } R_{\mathrm{new}}(f) = \sigma^{-5}\sum_{2\le j,k\le m}\frac{\gamma_j}{(j-2)!}\frac{\gamma_k}{(k-2)!}A_{j-2,k-2}$$

for given values of m. As an example, suppose terms corresponding to skewness, kurtosis and pentakosis are included. Then

$$\begin{aligned} R_{\rm new} &= \sigma^{-5}\{\gamma_3^2 A_{1,1} + (\gamma_4^2/4)A_{2,2} + (\gamma_5^2/36)\gamma_5^2 + 2(\gamma_3\gamma_5/6)A_{1,3}\} \\ &= \sigma^{-5}\frac{1}{2\sqrt{\pi}}\{(1/2)\gamma_3^2 + (3/16)\gamma_4^2 + (5/96)\gamma_5^2 - (1/4)\gamma_3\gamma_5\} \\ &= \sigma^{-5}\frac{3}{8\sqrt{\pi}}\Big(\frac{2}{3}\gamma_3^2 + \frac{1}{4}\gamma_4^2 + \frac{5}{72}\gamma_5^2 - \frac{1}{3}\gamma_3\gamma_5\Big), \end{aligned} \tag{4.5}$$

while Hjort & Jones (1994) finds

$$R_{\rm trad} = \sigma^{-5}\frac{3}{8\sqrt{\pi}}\Big(1 + \frac{35}{48}\gamma_4 + \frac{35}{32}\gamma_3^2 + \frac{385}{1024}\gamma_4^2 + \frac{1001}{10240}\gamma_5^2 - \frac{77}{128}\gamma_3\gamma_5\Big).$$

This indicates that the new estimator is better than the traditional one for all cases in a large neighbourhood around the normal distribution.

One might also use this test-bed to see where $f_0(x)r''(x)$ is smaller in size than $f''(x)$, say for moderate values of γ_3, γ_4, γ_5. This would be analogous to the experiments described in Section 5A for normal mixtures.

4C. SECOND TEST-BED: ROBUST HERMITE EXPANSIONS. The Hermite expansion (4.4) is of the type encountered in Edgeworth–Cramér expansions. It is pleasing from a theoretic point of view in that it incorporates skewness, kurtosis etc. to refine the normal approximation, but it has shortcomings as well. The coefficients are not always finite, and empirical estimates are quite variable and non-robust. Hjort & Jones (1994) and Hjort & Fenstad (1994) give further reasons favouring another and more robust Hermite expansion, in terms of the polynomials $H_j^*(y) = H_j(\sqrt{2}y)$ instead. In this case

$$f(x) = \phi\Big(\frac{x-\mu}{\sigma}\Big)\frac{1}{\sigma}\sum_{j=0}^{m}\frac{\delta_j}{j!}H_j^*\Big(\frac{x-\mu}{\sigma}\Big), \tag{4.6}$$

where the coefficients are determined from $\delta_j = \sqrt{2}{\rm E}H_j(\sqrt{2}(X-\mu)/\sigma)\exp\{-\frac{1}{2}(X-\mu)^2/\sigma^2\}$. If f is taken as an approximation to a given density q with mean μ and standard deviation σ, then the L_2 distance $\int(f-q)^2\,{\rm d}x$ is minimised for exactly these δ_j, see Hjort & Jones (1994). For this expansion, $f_0(x)r''(x) = \sigma^{-3}\phi(y)\sum_{j=2}^m 2j(j-1)(\delta_j/j!)H_{j-2}^*(y)$. It follows from this that

$$R_{\rm new}(f) = \sigma^{-5}\frac{2}{\sqrt{\pi}}\sum_{j=0}^{m-2}\delta_{j+2}^2/j!\,. \tag{4.7}$$

An expression for $R_{\rm trad}(f)$ for this robust Hermite expansion is in Hjort & Jones (1994). For illustration consider the 4th order case, where terms having $\delta_0,\ldots,\delta_4$ are included. Then

$$R_{\rm trad}(f) = \frac{\sigma^{-5}}{8\sqrt{\pi}}\Big(3\delta_0^2 + 15\delta_1^2 + \frac{39}{2}\delta_2^2 + \frac{25}{2}\delta_3^2 + \frac{41}{8}\delta_4^2 - 12\delta_0\delta_2 - 20\delta_1\delta_3 - 14\delta_2\delta_4 + 2\delta_0\delta_4\Big),$$

while $R_{\rm new}(f) = (2\sigma^{-5}/\sqrt{\pi})(\delta_2^2 + \delta_3^2 + \frac{1}{2}\delta_4^2)$. Again this indicates superiority of the (3.5) estimator in a broad neighbourhood around the normal.

5. Exact analysis for normal mixtures. Consider a normal mixture

$$f(x) = \sum_{i=1}^{k} p_i f_i(x), \quad \text{where } f_i(x) = \phi_{\sigma_i}(x - \mu_i), \tag{5.1}$$

writing $\phi_\sigma(u) = \sigma^{-1}\phi(\sigma^{-1}u)$. The family of such mixtures form a very wide and flexible class of densities. Marron & Wand (1992) studied such mixtures and in particular singled out 15 different 'test densities', covering a broad spectrum of not so difficult to extremely difficult cases, see the figure. These will now be used by us to compare the new normal-start times correction method with the traditional kernel method. In 5A the asymptotic mean squared errors of the two methods are compared, involving the leading terms of the Taylor-based approximations to bias and variance. In 5B we go further and analyse exact finite-sample mean squared errors for the two methods.

5A. EXACT AMISE ANALYSIS. To monitor the two bias terms we should compare f'' to $f_0 r''$, where f_0 is the best approximating normal, with $\mu_0 = \sum_{i=1}^k p_i\mu_i$ and $\sigma_0^2 = \sum_{i=1}^k p_i\{\sigma_i^2 + (\mu_i - \mu_0)^2\}$. Write $f_i = \exp(g_i)$ and $f_0 = \exp(g_0)$. Then $r = f/f_0 = \sum_{i=1}^k p_i \exp(g_i - g_0)$ and $r'' = \sum_{i=1}^k p_i \exp(g_i - g_0)\{g_i'' - g_0'' + (g_i' - g_0')^2\}$. This leads to

$$f_0(x)r''(x) = \sum_{i=1}^{k} p_i f_i(x)[1/\sigma_0^2 - 1/\sigma_i^2 + \{(x - \mu_i)/\sigma_i^2 - (x - \mu_0)/\sigma_0^2\}^2], \tag{5.2}$$

while

$$f''(x) = \sum_{i=1}^{k} p_i \phi_{\sigma_i}''(x - \mu_i) = \sum_{i=1}^{k} p_i\{(x - \mu_i)^2/\sigma_i^2 - 1\}f_i(x)/\sigma_i^2. \tag{5.3}$$

With some efforts (5.2) and (5.3) also lead to formulae for the roughness values $R_{\text{trad}}(f)$ and $R_{\text{new}}(f)$, cf. (4.2). Exact expressions are given in Proposition A.1 in our Appendix I.

In the figure these formulae are used to visually inspect $f''(x)$ versus $f_0(x)r''(x)$, for each of the 15 test cases. There are two immediate points to note. The first is that in most cases where the initial normal approximation is not very unreasonable, the new estimator manages to be better than the usual one, in significant x-areas. The second observation is that in cases where the initial description is clearly a bad start, the new semiparametric method turns almost nonparametric and behaves almost like the kernel method.

For simplicity the figure is placed at the end of our report

FIGURE. *The 15 test densities (left hand side) presented together with the bias factor functions f'' (solid line, for the kernel method) and $f_0 r''$ (dotted line, for the new method).*

We have also computed the global criteria $R_{\text{trad}}(f)$ and $R_{\text{new}}(f)$, for each of the 15 test densities; see the formulae and Table A.1 in Appendix I. The overall comparison in terms of approximate mise is in clear favour of the new method. Roughly speaking the first nine test cases are the not drastically unreasonable ones, whereas cases 10–15 probably originate from another planet and were chosen by Marron & Wand to exhibit particularities of smoothing parameter problems. And the new method wins in each of the nine worldly cases: the Gaußian, the skewed

unimodal, the strongly skewed, the kurtotic unimodal, the outlier, the bimodal, the separated bimodal, the skewed bimodal, the trimodal. It is also better for the claw density (#10 in Marron & Wand), the double claw (#11), and even for the asymmetric double claw (#13). It only loses to the traditional kernel method, and then only very slightly, in cases #12 (the asymmetric claw), #14 (the smooth comb), and #15 (the discrete comb).

So in terms of approximate mise the semiparametric (3.5) estimator wins over the kernel method in 12 out of 15 cases. It is fair to add that only about half of these victories are clear-cut, and that the remaining cases are almost draws, with surprisingly similar values for $R_{\rm new}$ and $R_{\rm trad}$. This picture emerges also when one computes values for the L_1-based criteria $\int |f''|$ versus $\int |f_0 r''|$, also given in the table of Appendix I. According to this measure the (3.5) estimator wins in 14 out of 15 cases.

We also inspected separately the case of two components in the normal mixture. Only in quite extreme cases does the kernel method win in approximate mise, and then only slightly. And the new method always wins when the two standard deviation parameters in question are equal. It is mildly surprising that a nonparametric correction on a normal start performs better than the kernel method even in such highly non-normal situations.

5B. EXACT FINITE-SAMPLE COMPARISON. The comparison analysis above was in terms of the Taylor-based approximations to bias and variance. Now we go further and analyse exact finite-sample mise for the two methods. Such analysis was carried out in Marron & Wand (1992) for the kernel method (1.1). Their Theorem 2.1 implies that if f is as in (5.1), then

$$\begin{aligned}
\text{mise}(h) &= \mathrm{E}\int (\widetilde{f}-f)^2\,\mathrm{d}x \\
&= \Big(1-\frac{1}{n}\Big)\sum_{i,j}\frac{p_ip_j}{(\sigma_i^2+\sigma_j^2+2h^2)^{1/2}}\phi\Big(\frac{\mu_j-\mu_i}{(\sigma_i^2+\sigma_j^2+2h^2)^{1/2}}\Big) \\
&\quad + \frac{1}{n}\frac{1}{2\sqrt{\pi}h} - 2\sum_{i,j}\frac{p_ip_j}{(\sigma_i^2+\sigma_j^2+h^2)^{1/2}}\phi\Big(\frac{\mu_j-\mu_i}{(\sigma_i^2+\sigma_j^2+h^2)^{1/2}}\Big) \\
&\quad + \sum_{i,j}\frac{p_ip_j}{(\sigma_i^2+\sigma_j^2)^{1/2}}\phi\Big(\frac{\mu_j-\mu_i}{(\sigma_i^2+\sigma_j^2)^{1/2}}\Big).
\end{aligned} \tag{5.4}$$

Reaching a similar result for the mise of the normal-start estimator (3.5) is much more demanding. Proposition A.2 in Appendix II delivers such a formula. It simplifies the comparison quest to care only about 'best case versus best case', which means comparing the two best achievable mise values, say $\text{mise}^*_{\rm trad}$ and mise^*. We programmed formula (5.4) and the one in Proposition A.2 and went through the list of the 15 test densities again, and found for each the minimising value of h and the resulting minimum mise values, for each of the five sample sizes 25, 50, 100, 200, 1000. The results are displayed in Table A.2 of Appendix II, along with the ratio $\text{mise}^*/\text{mise}^*_{\rm trad}$. These numbers support the previous positive conclusions for the new estimator, in its particular form (3.5). The mise-ratio is quite often below 1, and for the quite difficult test densities, where the analysis of 5A gave very similar values for $R_{\rm trad}$ and $R_{\rm new}$, Table A.2 yields mise-ratios mostly between 0.99 and

1.01. Even in these highly non-normal situations the new method has, overall, a slight edge. The table also illustrates that choosing the same bandwidth for the new method as for the kernel method will be quite acceptable in most of the definitely non-normal situations. In a broad vicinity of the normal it should pay to use a little larger bandwidth than what is optimal for the kernel method, however.

It should be kept in mind that the list of 15 test densities is not at all constructed to be favourable to using the normal model as start description. Statistically speaking we believe that a high proportion of densities actually encountered in real life are closer to the normal than each of cases #3–#15. In other words, the new method will win quite often.

6. Choosing smoothing parameter. Our method is defined in terms of a kernel function K and a bandwidth or smoothing parameter h. Choosing h is the more crucial problem, and methods for doing this parallel but by necessity become harder than the well-developed ones for the traditional (1.1) estimator (which is the special case of a constant initial estimator).

6A. MINIMISING AMISE. From (4.1) it is seen that the h parameter minimising approximate integrated mean squared error for $\widehat{f}$ is

$$h = h^* = \{R(K)/\sigma_K^4\}^{1/5} R_{\rm new}(f)^{-1/5}\, n^{-1/5}. \tag{6.1}$$

The resulting minimal amise is $\frac{5}{4}\{\sigma_K R(K)\}^{4/5} R_{\rm new}^{1/5} n^{-4/5}$. The same $\{\sigma_K R(K)\}^{4/5}$ factor appears also in a similar expression for the theoretically best point-wise mean squared error, so the efficiency of the kernel choice lies entirely with this number. This is very similar to what happens with the traditional estimator (1.1), see Scott (1992, Chapter 6), for example. The best possible kernel in this sense is the Yepanechnikov kernel $K_0(z) = \frac{3}{2}(1-4z^2)$ supported on $[-\frac{1}{2}, \frac{1}{2}]$ (or any other scaled version).

A 'plug-in rule' for h is to estimate the roughness $R_{\rm new}$ of (4.2) and insert this into (6.1). We outline three methods for doing this.

The first method is in the parametric 'rule of thumb' tradition and fits the data initially to a normal mixture, say of two or three components, using likelihood-based methods. The idea is then to use the formula for $R_{\rm new}$ in Appendix I to estimate h^* of (6.1). This would work well in many cases.

The second method is to exploit the Hermite expansions of Section 4 as approximations to the true f. An approximation to f that takes the first five moments into account is (4.4), with empirical estimates inserted for γ_3, γ_4, γ_5. This leads to an estimate of $R_{\rm new}$ via (4.5). The result, in the case of the normal kernel $K = \phi$, becomes

$$\widehat{h}_1 = (4/3)^{1/5}\{(2/3)\widehat{\gamma}_3^2 + (1/4)\widehat{\gamma}_4^2 + (5/72)\widehat{\gamma}_5^2 - (1/3)\widehat{\gamma}_3\widehat{\gamma}_5\}^{-1/5}\widehat{\sigma}\, n^{-1/5}. \tag{6.2}$$

One should preferably use robust estimates for the parameters, and one should ideally also deduct for bias when plugging in squared estimates, as explained in Hjort & Jones (1994). In any case (6.2) may be somewhat unstable, particularly for small to moderate sample sizes, since the empirical $\widehat{\gamma}_j$ statistics are unstable.

The alternative robust Hermite expansion described in 4D should be safer, using (4.6)–(4.7) instead of (4.4)–(4.5). It uses the automatically robust estimates

$$\widehat{\delta}_j = \frac{1}{n}\sum_{i=1}^{n} \sqrt{2} H_j\Big(\sqrt{2}\frac{X_i - \widehat{\mu}}{\widehat{\sigma}}\Big) \exp\Big\{-\tfrac{1}{2}\Big(\frac{X_i - \widehat{\mu}}{\widehat{\sigma}}\Big)^2\Big\}$$

(the summands are bounded in X_i) and

$$\widehat{h}_2 = (1/4)^{1/5}(\widehat{\delta}_2^2 + \widehat{\delta}_3^2 + \widehat{\delta}_4^2/2 + \widehat{\delta}_5^2/6)^{-1/5}\widehat{\sigma}\, n^{-1/5}, \tag{6.3}$$

for example. Again bias should ideally be deducted when plugging in squared estimates. See analogous comments in Hjort & Jones (1994).

While this second method can be seen as a semiparametric way of getting hold of $R_{\rm new}$, the third plug-in method is nonparametric on this account and takes the natural statistic

$$\begin{aligned}\widehat{R}_{\rm new} &= \int \{f(x,\widehat{\theta})\widehat{r}''(x)\}^2\,{\rm d}x \\ &= \frac{1}{n^2}\frac{1}{h^6}\sum_{i,j}\int \frac{f(x,\widehat{\theta})}{f(X_i,\widehat{\theta})}\frac{f(x,\widehat{\theta})}{f(X_j,\widehat{\theta})}K''(h^{-1}(x-X_i))K''(h^{-1}(x-X_j))\,{\rm d}x\end{aligned}$$

as its starting point. Explicit expressions for the integral here can be worked out for most choices of K; see formula (A.6) in Appendix II. Using (3.2) and the techniques of Section 3 one can show that

$$\begin{aligned}R^*_{\rm new} &= \int \{f_0(x)(r^*)''(x)\}^2\,{\rm d}x \\ &= \frac{1}{n^2}\frac{1}{h^6}\sum_{i,j}\int \frac{f_0(x)}{f_0(X_i)}\frac{f_0(x)}{f_0(X_j)}K''(h^{-1}(x-X_i))K''(h^{-1}(x-X_j))\,{\rm d}x,\end{aligned}$$

in which $f_0(x) = f(x,\theta_0)$ and $r^*(x) = n^{-1}\sum_{i=1}^n K_h(X_i - x)/f_0(X_i)$, is a good approximation to $\widehat{R}_{\rm new}$; in particular the mean of $\widehat{R}_{\rm new}$ is only $O(h^2/n + n^{-2})$ away from the mean of $R^*_{\rm new}$. Now somewhat long calculations, involving Taylor expansions, can be furnished to reach

$${\rm E}R^*_{\rm new} = \frac{n-1}{n}\int (f_0 r'')^2\,{\rm d}x + \frac{1}{nh^5}\{R(K'') + O(h^2)\},$$

where $R(K'') = \int (K'')^2\,{\rm d}z$. Since nh^5 is stable this shows that there is a fixed amount of overshooting. This is similar to but more involved than the corresponding result for the traditional kernel estimator (1.1) (which is the special case where $f_0(x)$ is constant), see Scott & Terrell (1987). This invites $\frac{n}{n-1}\{\widehat{R}_{\rm new} - R(K'')/(nh^5)\}$ to be used as a corrected estimate. One version of the plug-in method is therefore as follows: Select a start value for h in a reasonable way, perhaps using (6.3). Then compute $\widehat{R}_{\rm new}$ and its de-biased version, and insert in (6.1). One might also iterate this scheme further.

It is required that K here is smooth with vanishing derivatives at the end points of its support; in particular the Yepanechnikov kernel is not allowed in this operation.

6B. MINIMISING ESTIMATED AMISE. A useful idea related to the previous calculations is to estimate the approximate mise of (4.1) directly, that is, producing the curve

$$\widehat{\text{amise}}(h) = \text{bcv}(h) = \tfrac{1}{4}\sigma_K^4 h^4 \Big\{ \widehat{R}_{\text{new}}(h) - \frac{R(K'')}{nh^5} \Big\} + \frac{R(K)}{nh}, \tag{6.4}$$

including for emphasis h in the notation for the roughness estimate. This function must now be computed for a range of h-values, up to some upper limit h_{os}, the 'over-smoothing' bandwidth. Scott & Terrell (1987) and Scott (1992) call this strategy (for the traditional estimator) 'biased cross validation', although nothing seems to be cross validated per se. The bcv name derives rather from formula-wise similarity to unbiased cross validation, see below, and the desire to estimate the biased approximation amise to the true mise.

6C. NEARLY UNBIASED CROSS VALIDATION. A popular technique for the traditional kernel estimator is that of unbiased least squares cross validation, minimising an unbiased estimate of the exact mise as a function of bandwidth. A version of this idea can be carried through for our new estimator as well. The crux is to estimate $\text{mise}(h) - R(f) = \text{E}\{\int \widehat{f}^2 \,\text{d}x - 2\int f\widehat{f}\,\text{d}x\}$ with

$$\text{ucv}(h) = \int \widehat{f}_h(x)^2 \,\text{d}x - \frac{2}{n}\sum_{i=1}^n \widehat{f}_{h,(i)}(X_i). \tag{6.5}$$

Here h is included in the notation for clarity, and $\widehat{f}_{h,(i)}$ is the estimator constructed from the diminished data set that excludes X_i. The function to compute is

$$\text{ucv}(h) = \frac{1}{n^2}\sum_{i,j} \frac{1}{f(X_i,\widehat{\theta})f(X_j,\widehat{\theta})} \int f(x,\widehat{\theta})^2 K_h(x-X_i)K_h(x-X_j)\,\text{d}x$$
$$- \frac{2}{n(n-1)}\sum_{i,j} K_h(X_i - X_j)\frac{f(X_i,\widehat{\theta}_{(i)})}{f(X_j,\widehat{\theta}_{(j)})},$$

where $\widehat{\theta}_{(i)}$ is computed without X_i. In the case of the normal start method (3.5) with normal kernel $K = \phi$ a formula for the first term here is given in (A.6) in the Appendix.

It turns out that $\text{ucv}(h)$ is nearly but not exactly unbiased for $\text{mise}(h) - R(f)$. We have

$$\text{E}\int f\widehat{f}\,\text{d}x = \text{E}\int f(x)K_h(X_1 - x)\frac{f(x,\widehat{\theta})}{f(X_1,\widehat{\theta})}\,\text{d}x$$
$$= \text{E}\int\int f(x)f(y)K_h(y-x)\frac{f(x,\widehat{\theta}(y,X_2,\ldots,X_n))}{f(y,\widehat{\theta}(y,X_2,\ldots,X_n))}\,\text{d}x\,\text{d}y,$$

which is subtly different from

$$\text{E}\frac{1}{n}\sum_{i=1}^n \widehat{f}_{(i)}(X_i) = \text{E}K_h(X_2 - X_1)\frac{f(X_1,\widehat{\theta}(X_2,\ldots,X_n))}{f(X_2,\widehat{\theta}(X_2,\ldots,X_n))}$$
$$= \text{E}\int\int f(x)f(y)K_h(y-x)\frac{f(x,\widehat{\theta}(y,X_3,\ldots,X_n))}{f(y,\widehat{\theta}(y,X_3,\ldots,X_n))}\,\text{d}x\,\text{d}y.$$

The difference is minuscule, however, and choosing h to minimise the ucv(h) function, among $h \leq h_{os}$ for a suitable over-smoothing upper limit, remains a useful and honestly nonparametric option.

6D. OTHER TECHNIQUES. Other techniques can also be proposed, for example trying to adapt recent methods of Sheather & Jones (1991) and of Hall, Sheather, Jones & Marron (1991) to the present situation. One could also look into possible advantages of using a variable h. These matters are not pursued here. In our somewhat limited experience the (6.3) method has been satisfactory.

7. The multi-dimensional case. Our multiplicative correction factor method works well also in the vector case, as is now briefly explained. The setting is that d-dimensional i.i.d. vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are observed from a density f.

7A. THE TRADITIONAL AND THE NEW ESTIMATOR. The traditional kernel estimator uses a kernel density function $K(z_1, \ldots, z_d)$, usually symmetric about zero in each direction and often of product form $K_1(z_1) \cdots K_d(z_d)$. Its value at the point $\mathbf{x} = (x_1, \ldots, x_d)'$ is

$$\widetilde{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{h}}(X_{i,1} - x_1, \ldots, X_{i,d} - x_d), \tag{7.1}$$

where $K_{\mathbf{h}}(z_1, \ldots, z_d) = (h_1 \cdots h_d)^{-1} K(h_1^{-1} z_1, \ldots, h_d^{-1} z_d)$; see for example Scott (1992, Chapter 6) or Wand & Jones (1994, Chapter xx). In the product kernel case the basic bias and variance behaviour is described by

$$\begin{aligned} \text{bias} &\doteq \tfrac{1}{2} \sum_{j=1}^{d} \sigma(K_j)^2 h_j^2 f''_{jj}(\mathbf{x}), \\ \text{variance} &\doteq R(K_1) \cdots R(K_d)(n h_1 \cdots h_d)^{-1} f(\mathbf{x}) - n^{-1} f(\mathbf{x})^2, \end{aligned} \tag{7.2}$$

where $\sigma(K_j)^2 = \int z^2 K_j(z)\, \mathrm{d}z$ and $R(K_j) = \int K_j(z)^2\, \mathrm{d}z$. Furthermore f''_{jj} is the second partial derivative of f in direction x_j.

Our parametric start with a multiplicative correction method is now

$$\widehat{f}(\mathbf{x}) = f(\mathbf{x}, \widehat{\theta}) \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{h}}(\mathbf{X}_i - \mathbf{x}) / f(\mathbf{X}_i, \widehat{\theta}). \tag{7.3}$$

This is the appropriate vector version of (1.3), employing any parametric family $f(\mathbf{x}, \theta)$ and any reasonable parameter estimation method to produce the initial $f(\mathbf{x}, \widehat{\theta})$. The most important case is that of a multinormal start density, in which case the new estimator is

$$\widehat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{h}}(\mathbf{X}_i - \mathbf{x}) \frac{\exp\{-\frac{1}{2}(\mathbf{x} - \widehat{\mu})' \widehat{\Sigma}^{-1} (\mathbf{x} - \widehat{\mu})\}}{\exp\{-\frac{1}{2}(\mathbf{X}_i - \widehat{\mu})' \widehat{\Sigma}^{-1} (\mathbf{X}_i - \widehat{\mu})\}}, \tag{7.4}$$

and with some computational simplifications possible if a Gaußian kernel is used.

One may now go through the theory developed in Sections 2 and 3 and generalise results there to the present d-dimensional state of affairs. We omit details and merely

present the result. Firstly, the variance of the (7.2) estimator is found to be exactly equal to the variance noted above for the traditional (7.1) estimator, to the order of approximation used. Secondly, the bias is of the form

$$\tfrac{1}{2}\sum_{j=1}^{d}\sigma(K_j)^2 h_j^2 f_0(\mathbf{x}) r''_{jj}(\mathbf{x}) + O\big(\sum_{j=1}^{d}(h_j^4 + h_j^2/n) + n^{-2}\big),$$

involving the best parametric approximant $f_0(\mathbf{x}) = f(\mathbf{x}, \theta_0)$ and the ensuing correction factor $r(\mathbf{x}) = f(\mathbf{x})/f_0(\mathbf{x})$. Again the result is remarkably resistant to the actual parameter estimation used to obtain $\widehat{\theta}$, for example, cf. the discussion of Section 3.

Method (7.3) can therefore be expected to perform well in all situations where the $f_0 r''_{jj}$ functions are smaller in size than the f''_{jj} functions. This essentially says that the correction factor r should have smaller sized curvature than f itself, which again means that the initial parametric description should capture the main features of the density. Special cases can be inspected as explained in Sections 4 and 5. We expect the attractive (7.4) method, for example, in which case f_0 becomes the multinormal with parameters equal to the true mean and true covariance matrix for f, to work better than the traditional (7.1) estimator, for densities in a broad nonparametric vicinity of the multinormal.

7B. A PARTICULAR SCHEME. We speculate that the new methods could prove to be particularly useful in higher dimensions, since the traditional estimators, like (7.1), have quite slow convergence rates then. Implementation of the (7.4) estimator is straightforward, but the smoothing parameters remain to be specified. This is a harder problem than in the one-dimensional case. For completeness we briefly describe one particular solution here. It is practical and should work well in many situations, but does not claim optimality.

Start out considering the density $g(\mathbf{y})$ of $\mathbf{Y}_i = \Sigma^{-1/2}(\mathbf{X}_i - \mu)$, where μ and Σ are mean vector and covariance matrix for $\mathbf{X}_i$. Since these 'sphered' variables have mean zero and covariance matrix the identity a natural start description of g is g_0, the standard multi-normal, and furthermore it appears reasonable to smooth with the same amount in each direction, for example using the standard multi-normal $K_h(\mathbf{z}) = h^{-d}(2\pi)^{-d/2}\exp(-\frac{1}{2}\|\mathbf{z}\|^2/h^2)$. An estimate of g would consequently be of the form $\widehat{g}(\mathbf{y}) = g_0(\mathbf{y})\, n^{-1}\sum_{i=1}^{n} K_h(\mathbf{Y}_i - \mathbf{y})/g_0(\mathbf{Y}_i)$. After estimating mean and covariance matrix this amounts to an estimated multinormal start for f and leads to

$$\begin{aligned}\widehat{f}(\mathbf{x}) &= \widehat{g}(\widehat{\Sigma}^{-1/2}(\mathbf{x} - \widehat{\mu}))|\widehat{\Sigma}|^{-1/2} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\exp\{-\frac{1}{2}(\mathbf{x} - \mathbf{x}_i)'\widehat{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}_i)/h^2\}}{(2\pi)^{d/2}h^d}\,\frac{\exp\{-\frac{1}{2}(\mathbf{x} - \widehat{\mu})'\widehat{\Sigma}^{-1}(\mathbf{x} - \widehat{\mu})\}}{\exp\{-\frac{1}{2}(\mathbf{X}_i - \widehat{\mu})'\widehat{\Sigma}^{-1}(\mathbf{X}_i - \widehat{\mu})\}}.\end{aligned} \tag{7.5}$$

This estimator can also be motivated directly without $\mathbf{Y}_i$s. The main reason for using the g-representation is however that it can be used to find a suitable h, as follows. By previous results the approximate mise is $R(\phi)^d(nh^d)^{-1} + \frac{1}{4}h^4 R_{\text{new}}(g)$, featuring $R_{\text{new}}(g) = \int\{g_0(\mathbf{y})r''(\mathbf{y})\}^2\,\mathrm{d}\mathbf{y}$. Its minimiser $h^* = \{dR(\phi)^d\}^{1/(d+4)}R_{\text{new}}(g)^{-1/(d+4)}$ $n^{-1/(d+4)}$ can be estimated in various ways, and one feasible solution, aiming to generalise (4.7) and (6.3), is to approximate r using an expansion with products

$H^*_{j_1}(y_1)\cdots H^*_{j_d}(y_d)$ as basis functions, where again $H^*_j(y) = H_j(\sqrt{2}y)$. We omit the many necessary details here but report that

$$R_{\text{new}}(g) = \frac{4}{(2\sqrt{\pi})^d} \sum_{j_1,\ldots,j_d} (\delta_{j_1+2,\ldots,j_d} + \cdots + \delta_{j_1,\ldots,j_d+2})^2/(j_1!\cdots j_d!),$$

where

$$\delta_{j_1,\ldots,j_d} = 2^{d/2}\mathrm{E}_g \exp(-\tfrac{1}{2}\|\mathbf{Y}\|^2)\, H^*_{j_1}(Y_1)\cdots H^*_{j_d}(Y_d).$$

The procedure is as with (4.7) and (6.3), namely estimating the first few of these using

$$\widehat{\delta}_{j_1,\ldots,j_d} = 2^{d/2}\frac{1}{n}\sum_{i=1}^n \exp\{-\tfrac{1}{2}(\mathbf{X}_i - \widehat{\mu})'\widehat{\Sigma}^{-1}(\mathbf{X}_i - \widehat{\mu})\}\, H^*_{j_1}(Y_{i,1})\cdots H^*_{j_d}(Y_{i,d}),$$

where now $\mathbf{Y}_i = \widehat{\Sigma}^{-1/2}(\mathbf{X}_i - \widehat{\mu})$, and finally calculating

$$\begin{aligned}\widehat{h} &= \{d/(2\sqrt{\pi})^d\}^{1/(d+4)}\widehat{R}_{\text{new}}^{-1/(d+4)}n^{-1/(d+4)} \\ &= (d/4)^{1/(d+4)}\Big\{\sum_{j_1,\ldots,j_d}\frac{(\widehat{\delta}_{j_1+2,\ldots,j_d} + \cdots + \widehat{\delta}_{j_1,\ldots,j_d+2})^2}{j_1!\cdots j_d!}\Big\}^{-1/(d+4)} n^{-1/(d+4)}.\end{aligned}$$

8. Supplementing remarks.

8A. How close is the new estimator to the old? For simplicity of presentation consider $\widehat{f}(x)$ in the form (2.1) in terms of a basis function $f_0(x)$ rather than with estimated parameters. In the sum that defines $\widehat{f}(x)$ the ratios $f_0(x)/f_0(X_i)$ are close to 1 for small values of h since then the X_is quite close to x are those given significant weights. In other words, $\widehat{f}(x)$ cannot be very different from the traditional kernel estimator $\widetilde{f}(x)$ of (1.1) when h is small. A Taylor analysis is informative:

$$\frac{f_0(x)}{f_0(X_i)} \doteq 1 - a_0(x)(X_i - x) + \tfrac{1}{2}\{a_0(x)^2 - b_0(x)\}(X_i - x)^2,$$

where $a_0(x)$ and $b_0(x)$ are the two first x-derivatives of $\log f_0(x)$. Hence

$$\widehat{f}(x) \doteq \widetilde{f}(x) - a_0(x)e_1(x) + \tfrac{1}{2}\{a_0(x)^2 - b_0(x)\}e_2(x), \tag{8.1}$$

where $e_q(x) = n^{-1}\sum_{i=1}^n K_h(X_i - x)(X_i - x)^q$. One can now show that $e_1(x)$ has mean $\sigma_K^2 h^2 f'(x) + O(h^4)$ and small variance $O(hn^{-1}f(x))$, while $e_2(x)$ has mean $\sigma_K^2 h^2 f(x) + O(h^4)$ with even smaller variance $O(h^3 n^{-1} f(x))$. Thus the difference is of size $O(h^2)$. If $f_0(x)$ is the standard normal, for example, then $\widehat{f}(x) - \widetilde{f}(x) = h^2\sigma_K^2\{xf'(x) + \frac{1}{2}(x^2+1)f(x)\} + O_p(h^4)$.

8B. Accuracy of the estimated correction factor. Our machinery can also be used for model exploration purposes, by inspecting the correction factor against x for various potential models. A model's adequacy could be inspected by looking at a plot of $\widehat{r}(x)$, perhaps with a pointwise confidence band, to see if $r(x) = 1$

is reasonable. In the notation of Sections 2 and 3, and using techniques from these sections, one can establish that

$$\begin{aligned}
\mathrm{E}\,\widehat{r}(x) &\doteq r(x) + \tfrac{1}{2}\sigma_K^2 h^2 r''(x) - n^{-1} r(x) u_0(x)' \{I(x) + d\}, \\
\mathrm{Var}\,\widehat{r}(x) &\doteq (nh)^{-1} R(K) r(x)/f_0(x) - n^{-1} r(x)^2 \{1 + 2u_0(x)' I(x) - u_0(x)' \Sigma u_0(x)\},
\end{aligned}$$

with some simplification in the maximum likelihood case, for which $I(x) = J^{-1}u_0(x)$ and $\Sigma = J^{-1}$.

It is also informative to plot the log-correction factor $\log \widehat{r}(x)$, to see how far from zero it is. The bias and variance results for this curve are

$$\begin{aligned}
\mathrm{E}\log\widehat{r}(x) &\doteq \log r(x) + \tfrac{1}{2}\sigma_K^2 h^2 r''(x)/r(x) \\
&\qquad - \tfrac{1}{2} R(K)(nh)^{-1}\{r(x)f_0(x)\}^{-1} - \tfrac{1}{8}\sigma_K^4 h^4 r''(x)^2/r(x)^2, \\
\mathrm{Var}\log\widehat{r}(x) &\doteq R(K)(nh)^{-1}\{r(x)f_0(x)\}^{-1} - n^{-1}\{1 + 2u_0(x)'I(x) - u_0(x)'\Sigma u_0(x)\}.
\end{aligned}$$

A nice graphical goodness of fit method emerges: plot

$$Z(x) = \frac{\log\widehat{r}(x) + \frac{1}{2}R(K)(nh)^{-1} f(x,\widehat{\theta})^{-1}}{\{R(K)(nh)^{-1} f(x,\widehat{\theta})^{-1}\}^{1/2}} \tag{8.2}$$

against x, possibly with a more accurate denominator. Under model conditions this should be approximately distributed as a standard normal for each x, that is, the $Z(x)$ curve should stay within ± 1.96 about 95% of the time.

8C. THE INTEGRAL. Our estimator does not integrate to precisely 1. The normal-based version (3.5), for example, when the Gaußian kernel $K = \phi$ is used, has

$$\int \widehat{f}\,\mathrm{d}x = (1 + h^2/\widehat{\sigma}^2)^{-1/2} \frac{1}{n}\sum_{i=1}^n \exp\{\tfrac{1}{2}h^2(X_i - \widehat{\mu})^2/\{\widehat{\sigma}^2(\widehat{\sigma}^2 + h^2)\}\},$$

which after Taylor expansions is found to be equal to $1 + \frac{1}{8}\widehat{\gamma}_4 h^4/\widehat{\sigma}^4$, where $\widehat{\gamma}_4 = n^{-1}\sum_{i=1}^n\{(X_i - \widehat{\mu})/\widehat{\sigma}\}^4 - 3$ is the estimated kurtosis. Dividing the original estimate with this amount does not lead to superior performance in terms of mise, however. In the general case, in the notation of (2.1), for example, one finds

$$\int \widehat{f}\,\mathrm{d}x = 1 + \tfrac{1}{2}h^2\sigma_K^2 \frac{1}{n}\sum_{i=1}^n \frac{f_0''(X_i)}{f_0(X_i)} + \tfrac{1}{24}h^4 \mathrm{E}_K Z^4 \frac{1}{n}\sum_{i=1}^n \frac{f_0^{(4)}(X_i)}{f_0(X_i)}$$

via Taylor expansions. The h^2 term vanishes in the normal case.

8D. PARAMETRIC HOME-TURF CONDITIONS. If model conditions $f(x) = f(x,\theta)$ can be trusted the natural estimator is simply $f(x,\widehat{\theta})$, for example with the maximum likelihood estimator. From $\sqrt{n}(\widehat{\theta} - \theta) \to_d \mathcal{N}_p\{0, J^{-1}\}$ where J is the information matrix, combined with the delta method and some extra arguments, follows

$$n\mathrm{E}\int\{f(x,\widehat{\theta}) - f(x,\theta)\}^2\,\mathrm{d}x \to_d \int f(x,\theta)^2 u(x,\theta)' J^{-1} u(x,\theta)\,\mathrm{d}x,$$

where $u(x,\theta) = \partial \log f(x,\theta)/\partial\theta$ is the score function. Algebraic calculations for the normal model lead to a parametric mise of size $\frac{7}{8}(2\sqrt{\pi}\sigma)^{-1}/n$. (This is the large-sample approximation to the exact mise, for which an exact formula also can be found.)

It turns out that the mise of the new nonparametric (3.5) estimator, computed under Gaußian home turf conditions, is only slightly larger than this. Going through formulae (A.8)–(A.10) of Appendix II one finds

$$\text{mise}(h) = \frac{1}{2\sqrt{\pi}}\Big\{\Big(1-\frac{1}{n}\Big)\frac{1}{\sigma} + \frac{1}{n}\frac{1}{h}\frac{\sigma}{(\sigma^2-h^2)^{1/2}} - \frac{2}{\sigma}\Big\}.$$

It is of separate interest to note that this is minimised for $h^* = \sigma/\sqrt{2}$, regardless of sample size; cf. Case #1 in Table A.2. The minimum value is $\text{mise}^* = (2\sqrt{\pi}\sigma)^{-1}/n$, only 14% larger than the parametric mise. Of course one should use an estimated $\widehat{\sigma}/\sqrt{2}$ in practice, but this can be seen to alter the minimum mise only to second order terms $O(n^{-2})$. This is shown via an exact formula for $\text{ise}(\widehat{\sigma}/\sqrt{2})$, using results appearing after (A.6) in Appendix II.

8E. NONPARAMETRIC REGRESSION WITH A PARAMETRIC START. The basic estimation idea of our paper works well also in other areas of curve smoothing. An important such area is that of nonparametric regression. Assume that i.i.d. pairs (x_i, y_i) are observed from a smooth bivariate density $f(x,y) = f(x)g(y \mid x)$, and that interest focuses on the conditional mean function $m(x) = \text{E}(Y \mid x)$. A standard method is the Nadaraya–Watson estimator $\widetilde{m}(x) = \sum_{i=1}^n y_i K_h(x-x_i)/\sum_{i=1}^n K_h(x-x_i)$, see for example Scott (1992, Chapter 8) and Wand & Jones (1994, Chapter xx). Taylor expansion analysis and somewhat lengthy calculations lead to

$$\begin{aligned} \text{E}\,\widetilde{m}(x) &\doteq m(x) + \tfrac{1}{2}\sigma_K^2 h^2\{m''(x) + 2m'(x)f'(x)/f(x)\}, \\ \text{Var}\,\widetilde{m}(x) &\doteq R(K)(nh)^{-1}\sigma(x)^2/f(x) + O(h/n). \end{aligned} \tag{8.3}$$

This is a somewhat more complete version of calculations in Scott (1992, p. 223–224). Our calculations are also mildly more general, in that we took care here not to assume merely a constant value for $\sigma(x)^2 = \text{Var}(Y \mid x)$, for reasons appearing below.

A semiparametric estimator can now be constructed as follows. Start out with a parametric initial description, say $m(x,\widehat{\beta})$, perhaps the simple linear $\widehat{\beta}_1 + \widehat{\beta}_2 x$. This start estimator aims really at $m(x,\beta_0)$, say, the best parametric approximant. A multiplicative correction factor, aiming at $r(x) = m(x)/m(x,\beta_0)$, can be given as a Nadaraya–Watson estimator using $y_i/m(x_i,\widehat{\beta})$. This leads to a generalised Nadaraya–Watson estimator

$$\begin{aligned} \widehat{m}(x) &= m(x,\widehat{\beta})\frac{\sum_{i=1}^n \{y_i/m(x_i,\widehat{\beta})\}K_h(x-x_i)}{\sum_{i=1}^n K_h(x-x_i)} \\ &= \frac{\sum_{i=1}^n y_i\{m(x,\widehat{\beta})/m(x_i,\widehat{\beta})\}K_h(x-x_i)}{\sum_{i=1}^n K_h(x-x_i)}. \end{aligned} \tag{8.4}$$

Calculations involving the above result, using $y_i/m(x_i,\beta_0)$ with conditional variance $\sigma(x_i)^2/m(x_i,\beta_0)^2$, and Taylor expansions of $\widehat{\beta}$ around β_0, as in Section 3, lead in the end to

$$\text{E}\,\widehat{m}(x) \doteq m(x) + \tfrac{1}{2}\sigma_K^2 h^2\{m(x,\beta_0)r''(x)^2 + 2m_0(x,\beta)r'(x)f'(x)/f(x)\}, \tag{8.5}$$

with approximation error of size at most $O(h^4 + h^2/n + n^{-2})$, and to a variance being of the very same size as that in (8.3), to the order of approximation used. In many

cases this will mean a genuine reduction of mise, and hence that the generalised Nadaraya–Watson estimator (8.4) is better than the usual estimator. This idea could be particularly useful in situations with several covariates.

Hjort (1993, final section) gives yet another example of the type (3.1) construction, in the realm of nonparametric hazard rate estimation. The result is once again that a bias reduction vis-à-vis the traditional estimator is possible in a broad neighbourhood of the parametric model used, without sacrificing variance.

Appendix I: Roughness measures for normal mixtures. Here formulae are provided for roughnesses $R_{\rm trad}$ and $R_{\rm new}$ for a general normal mixture, results that were used in Section 4B. A table comparing the performance of the normal-start semiparametric method with that of the usual kernel estimator, for each of 15 test cases, is also given.

PROPOSITION A.1. *For a normal mixture* $f(x) = \sum_{i=1}^k p_i\phi_{\sigma_i}(x-\mu_i)$, *let* $\sigma_{i,j}^2 = \sigma_i^2 + \sigma_j^2$ *and* $\delta_{i,j} = (\mu_j - \mu_i)/\sigma_{i,j}$. *The roughness functionals defined in (4.2) can be calculated explicitly;*

$$R_{\rm trad} = \int (f'')^2\,{\rm d}x = \sum_{i,j} p_ip_j(\delta_{i,j}^4 - 6\delta_{i,j}^2 + 3)\phi(\delta_{i,j})/\sigma_{i,j}^5,$$

$$R_{\rm new} = \int (f_0r'')^2\,{\rm d}x = T_1 + \cdots + T_6,$$

with these terms being defined in equation (A.1) below. The $R_{\rm trad}$ *result is also proved in Marron & Wand (1992, Theorem 4.1).*

PROOF: Start out noting that

$$\int \phi_{\sigma_i}(x-\mu_i)\phi_{\sigma_j}(x-\mu_j)\,{\rm d}x = \phi(\delta_{i,j})/\sigma_{i,j} = A_{i,j}(\mu_i,\mu_j),$$

say. Taking derivatives with respect to μ_i and μ_j gives in general that

$$\int H_r\Big(\frac{x-\mu_i}{\sigma_i}\Big)H_s\Big(\frac{x-\mu_j}{\sigma_j}\Big)f_i(x)f_j(x)\,{\rm d}x = \sigma_i^r\sigma_j^s\frac{\partial^{r+s}}{\partial\mu_i^r\partial\mu_j^s}A_{i,j} = \sigma_i^r\sigma_j^sA_{i,j}^{r,s},$$

say, H_r and H_s again being the Hermite polynomials. This leads to

$$R_{\rm trad}(f) = \sum_{i,j}p_ip_j\int\phi_{\sigma_i}''(x-\mu_i)\phi_{\sigma_j}''(x-\mu_j)\,{\rm d}x = \sum_{i,j}p_ip_j\phi^{(4)}(\delta_{i,j})/\sigma_{i,j}^5,$$

proving the first and simplest assertion. To find $\int(f_0r'')^2\,{\rm d}x$, write (4.5) as

$$f_0(x)r''(x) = \sum_{i=1}^k p_if_i(x)\{c_i + d_i(x-\mu_i) + a_i^2(x-\mu_i)^2\},$$

where $a_i = 1/\sigma_i^2 - 1/\sigma_0^2$, $b_i = (\mu_i-\mu_0)/\sigma_0^2$, $c_i = b_i^2 - a_i$, and $d_i = -2a_ib_i$. Somewhat strenuous calculations yield in the end the sought-for six-term expression $T_1+\cdots+T_6$

for $R_{\rm new}(f)$, where

$$\begin{aligned}
T_1 &= \sum_{i,j} p_i p_j c_i c_j A_{i,j}^{0,0}, \\
T_2 &= 2\sum_{i,j} p_i p_j c_i d_j \sigma_j^2 A_{i,j}^{0,1}, \\
T_3 &= 2\sum_{i,j} p_i p_j c_i a_j^2 (\sigma_j^4 A_{i,j}^{0,2} + \sigma_j^2 A_{i,j}^{0,0}), \\
T_4 &= \sum_{i,j} p_i p_j d_i d_j \sigma_i^2 \sigma_j^2 A_{i,j}^{1,1}, \\
T_5 &= 2\sum_{i,j} p_i p_j d_i a_j^2 (\sigma_i^2 \sigma_j^4 A_{i,j}^{2,1} + \sigma_i^2 \sigma_j^2 A_{i,j}^{1,0}), \\
T_6 &= \sum_{i,j} p_i p_j a_i^2 a_j^2 \sigma_i^2 \sigma_j^2 (\sigma_i^2 \sigma_j^2 A_{i,j}^{2,2} + \sigma_i^2 A_{i,j}^{2,0} + \sigma_j^2 A_{i,j}^{0,2} + A_{i,j}^{0,0}).
\end{aligned} \tag{A.1}$$

It is furthermore the case that $A_{i,j}^{r,s} = (-1)^r \phi^{(r+s)}(\delta_{i,j})/\sigma_{i,j}^{r+s+1}$. Hence

$$\begin{aligned}
A_{i,j}^{0,0} &= \phi(\delta_{i,j})/\sigma_{i,j}, \\
A_{i,j}^{1,0} &= \delta_{i,j}\phi(\delta_{i,j})/\sigma_{i,j}^2 = -A_{i,j}^{0,1}, \\
A_{i,j}^{2,0} &= (\delta_{i,j}^2 - 1)\phi(\delta_{i,j})/\sigma_{i,j}^3 = A_{i,j}^{0,2} = -A_{i,j}^{1,1}, \\
A_{i,j}^{2,1} &= (\delta_{i,j}^3 - 3\delta_{i,j})\phi(\delta_{i,j})/\sigma_{i,j}^4 = -A_{i,j}^{1,2}, \\
A_{i,j}^{2,2} &= (\delta_{i,j}^4 - 6\delta_{i,j}^2 + 3)\phi(\delta_{i,j})/\sigma_{i,j}^5.
\end{aligned}$$

This delivers a programmable formula for $R_{\rm new}$ and proves the second assertion. □

In the table below we have chosen to display

$$\rho_{\rm trad}(f) = \sigma(f) R_{\rm trad}(f)^{1/5} \qquad \text{and} \qquad \rho_{\rm new}(f) = \sigma(f) R_{\rm new}(f)^{1/5} \tag{A.2}$$

rather than $R_{\rm trad}$ and $R_{\rm new}$, for the 15 test cases chosen in Marron & Wand (1992). The $R_{\rm trad}$ values in raw form range wildly from 0.212 to 70730, for example, and are not easily interpretable. The ρ-numbers are scale invariant and are directly tied to the best possible approximate mise; the minimum amise for $\widehat{f}$ can be derived from (4.1) and is $\frac{5}{4}\sigma(f)^{-1}\{\sigma_K R(K)\}^{4/5}\rho_{\rm new}(f)/n^{4/5}$, with a similar expression for $\widetilde{f}$.

We have also included similar 'difficulty measures' based on integrated absolute bias plus integrated mean absolute deviation. This is a statistically meaningful criterion which is also a simple upper bound on the expected L_1-distance. The parallel to (4.1) can be shown to be

$$\begin{aligned}
(\text{iab} + \text{imad})(\widetilde{f}) &\doteq \tfrac{1}{2}\sigma_K^2 \int |f''|\,\mathrm{d}x + (2/\pi)^{1/2} R(K)^{1/2} (nh)^{-1/2} \int f^{1/2}\,\mathrm{d}x, \\
(\text{iab} + \text{imad})(\widehat{f}) &\doteq \tfrac{1}{2}\sigma_K^2 \int |f_0 r''|\,\mathrm{d}x + (2/\pi)^{1/2} R(K)^{1/2} (nh)^{-1/2} \int f^{1/2}\,\mathrm{d}x,
\end{aligned}$$

so the values to compute and compare are primarily $\int |f_0 r''|\,\mathrm{d}x$ and $\int |f''|\,\mathrm{d}x$. We have carried out numerical integrations to obtain these numbers, again for each of the 15 test cases. Displayed in the table are

$$\rho_{\rm trad}^1(f) = \Big(\int f^{1/2}\Big)^{4/5} \Big(\int |f''|\Big)^{1/5} \quad \text{and} \quad \rho_{\rm new}^1(f) = \Big(\int f^{1/2}\Big)^{4/5} \Big(\int |f_0 r''|\Big)^{1/5}. \tag{A.3}$$

This is because the minimal possible value of iab + imad for $\widehat{f}$ can be shown to be $\frac{5}{4}(2^3/\pi^2)^{1/5}\{\sigma_K R(K)\}^{2/5}\rho^1_{\rm new}(f)/n^{2/5}$, and similarly with $\widetilde{f}$. The quantities in (A.3) are scale invariant.

TABLE A.1. *Values of the global mise-based comparison values $\rho_{\rm trad}$ and $\rho_{\rm new}$, given for each of the 15 normal mixture test cases. Also included are the L_1-based global comparison values $\rho^1_{\rm trad}$ and $\rho^1_{\rm new}$. The normal-start estimator (3.5) wins in approximate mise over the kernel method for all cases except #12, 14, 15, where it loses very slightly. In terms of approximate iab plus imad it wins in all cases except #3.*

Case	$\rho_{\rm trad}$	$\rho_{\rm new}$	$\rho^1_{\rm trad}$	$\rho^1_{\rm new}$
1	0.7330	0	1.8933	0
2	0.8921	0.6739	2.0343	1.7910
3	5.6070	5.5985	3.4988	3.5202
4	3.8664	3.8354	3.5512	3.5369
5	2.3201	2.2088	2.9388	2.9042
6	1.1183	1.0615	2.1786	2.0575
7	2.0215	1.9579	2.4701	2.4177
8	1.3753	1.3468	2.3095	2.1998
9	1.5600	1.5335	2.4608	2.3763
10	3.5571	3.5421	3.8812	3.8674
11	12.4450	12.4447	5.5611	5.5590
12	6.4350	6.4382	4.0978	4.0909
13	11.1149	11.1147	4.9481	4.9465
14	14.6610	14.6615	4.8733	4.8703
15	9.6259	9.6261	4.3863	4.3821

Appendix II: Exact mise comparisons. The task considered in the following is that of computing the exact mise(h) for the (3.5) estimator, for normal mixtures. The point is to facilitate comparison with the kernel method, for which exact mise-calculations are known (and much easier).

Suppose again that $f(x) = \sum_{i=1}^k p_i\phi_{\sigma_i}(x-\mu_i)$ is a normal mixture. Start out with

$$\text{ise}(h) = \int (\widehat{f}-f)^2\,{\rm d}x = A_h - 2B_h + R(f), \tag{A.4}$$

where

$$R(f) = \int f^2\,{\rm d}x = \sum_{i,j} p_ip_j\phi\Big(\frac{\mu_j-\mu_i}{\sigma_{i,j}}\Big)\frac{1}{\sigma_{i,j}}, \tag{A.5}$$

again using $\sigma_{i,j} = (\sigma_i^2+\sigma_j^2)^{1/2}$. To give useful expressions for A_h and B_h we note the technical fact that

$$\int \prod_{j=1}^m \phi_{\sigma_j}(x-\mu_j)\,{\rm d}x = \sqrt{2\pi}\widetilde{\sigma}\Big[\prod_{j=1}^m \phi_{\sigma_j}(\mu_j-a)\Big]\exp\Big[\tfrac12\widetilde{\sigma}^2\Big\{\sum_{j=1}^m(\mu_j-a)/\sigma_j^2\Big\}^2\Big], \tag{A.6}$$

where $1/\widetilde{\sigma}^2 = \sum_{j=1}^m 1/\sigma_j^2$. The value of a is arbitrary and can be chosen for the occasion. Proving (A.6) is not very difficult and we omit the details. For the first term this identity gives

$$\begin{aligned} A_h &= \frac{1}{n^2}\sum_{i,j\le n}\int \frac{\phi_h(x-x_i)\phi_h(x-x_j)\phi_{\widehat{\sigma}}(x-\widehat{\mu})^2}{\phi_{\widehat{\sigma}}(x_i-\widehat{\mu})\phi_{\widehat{\sigma}}(x_j-\widehat{\mu})}\,{\rm d}x \\ &= \frac{1}{n^2}\sum_{i,j\le n}\frac{1}{2\sqrt{\pi}}\frac{\widetilde{\sigma}}{\widehat{\sigma}^2}\frac{\phi_h(x_i-\widehat{\mu})}{\phi_{\widehat{\sigma}}(x_i-\widehat{\mu})}\frac{\phi_h(x_j-\widehat{\mu})}{\phi_{\widehat{\sigma}}(x_j-\widehat{\mu})}\exp\Big\{\tfrac12\widetilde{\sigma}^2\Big(\frac{x_i-\widehat{\mu}+x_j-\widehat{\mu}}{h^2}\Big)^2\Big\}, \end{aligned}$$

where

$$\widetilde{\sigma}^2 = \Big(\frac{2}{\widehat{\sigma}^2} + \frac{2}{h^2}\Big)^{-1} = \tfrac{1}{2}\frac{\widehat{\sigma}^2}{\widehat{\sigma}^2 + h^2}h^2.$$

And for the second term,

$$\begin{aligned} B_h &= \sum_{j=1}^k p_j \frac{1}{n}\sum_{i=1}^n \int \phi_h(x - x_i)\frac{\phi_{\widehat{\sigma}}(x - \widehat{\mu})}{\phi_{\widehat{\sigma}}(x_i - \widehat{\mu})}\phi_{\sigma_j}(x - \mu_j)\,\mathrm{d}x \\ &= \frac{1}{\widehat{\sigma}}\sum_{j=1}^k p_j \widetilde{\sigma}_j \phi_{\sigma_j}(\mu_j - \widehat{\mu})\Big[\frac{1}{n}\sum_{i=1}^n \frac{\phi_h(x_i - \widehat{\mu})}{\phi_{\widehat{\sigma}}(x_i - \widehat{\mu})}\exp\Big\{\tfrac{1}{2}\widetilde{\sigma}_j^2\Big(\frac{x_i - \widehat{\mu}}{h^2} + \frac{\mu_j - \widehat{\mu}}{\sigma_j^2}\Big)^2\Big\}\Big], \end{aligned}$$

where this time

$$\widetilde{\sigma}_j^2 = \Big(\frac{1}{h^2} + \frac{1}{\widehat{\sigma}^2} + \frac{1}{\sigma_j^2}\Big)^{-1} = \frac{\widehat{\sigma}^2\sigma_j^2}{\widehat{\sigma}^2\sigma_j^2 + h^2(\widehat{\sigma}^2 + \sigma_j^2)}h^2.$$

Finding further exact expressions for the mise involves finding the exact means of A_h and B_h. This would depend on the parameter estimation method used, and in any case seems forbiddingly difficult. The formulae for A_h and B_h can however be used to compute their mean values, and hence the mise, via stochastic simulation, for each given mixture and each given sample size. At this stage we are content to find the exact mise for the estimator that employs true parameter values μ_0 and σ_0 for mean and standard deviation. This allows a 'best case versus best case' comparison with the kernel method to be made, and the extra variability caused by using parameter estimates for μ and σ is in any case of second order importance.

PROPOSITION A.2. *Consider the normal start times correction estimator with the normal kernel,*

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^n \phi_h(X_i - x)\frac{\phi_{\sigma_0}(x - \mu_0)}{\phi_{\sigma_0}(X_i - \mu_0)} = \frac{1}{n}\sum_{i=1}^n \phi_h(X_i - x)\frac{\exp\{-\frac{1}{2}(x - \mu_0)^2/\sigma_0^2\}}{\exp\{-\frac{1}{2}(X_i - \mu_0)^2/\sigma_0^2\}}.$$

Its exact mean integrated squared error, in the case when f is a normal mixture $\sum_{i=1}^k p_i\phi_{\sigma_i}(x - \mu_i)$, can be expressed as

$$\mathrm{mise}(h) = (1 - n^{-1})\,\mathrm{E}A_{1,h} + n^{-1}\,\mathrm{E}A_{2,h} - 2\,\mathrm{E}B_h + R(f), \tag{A.7}$$

where $R(f)$ is given in (A.5) and where formulae for the other three terms appear in equations (A.8–10) below.

PROOF: We start out finding an exact expression for the expected value:

$$\begin{aligned} \mathrm{E}\widehat{f}(x) &= \int \phi_h(y - x)\frac{\phi_{\sigma_0}(x - \mu_0)}{\phi_{\sigma_0}(y - \mu_0)}f(y)\,\mathrm{d}y \\ &= \phi_{\sigma_0}(x - \mu_0)\sum_{i=1}^k p_i \int \phi(z)\frac{\phi_{\sigma_i}(x - \mu_i + hz)}{\phi_{\sigma_0}(x - \mu_0 + hz)}\,\mathrm{d}z, \end{aligned}$$

using the $z = (y - x)/h$ substitution. Expanding the exponent and collecting z^2 terms, and using

$$b_i = \Big(1 + \frac{h^2}{\sigma_i^2} - \frac{h^2}{\sigma_0^2}\Big)^{1/2},$$

one finds

$$\mathrm{E}\widehat{f}(x) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{k} p_i \frac{1}{\sigma_i b_i} \exp\Big\{ -\tfrac{1}{2} \frac{(x-\mu_i)^2}{\sigma_i^2} + \tfrac{1}{2} \Big(\frac{x-\mu_i}{\sigma_i^2} - \frac{x-\mu_0}{\sigma_0^2} \Big)^2 h^2 b_i^2 \Big\}.$$

Indeed this is $f(x) + O(h^2)$. Next consider A_h of (A.4) and its mean value. Splitting A_h into non-diagonal and diagonal terms leads to $\mathrm{E}A_h = (1 - n^{-1})\,\mathrm{E}A_{1,h} + n^{-1}\,\mathrm{E}A_{2,h}$, leaving us the task of calculating $\mathrm{E}A_{1,h}$ and $\mathrm{E}A_{2,h}$ by integration. First,

$$\begin{aligned}
\mathrm{E}A_{1,h} &= \mathrm{E} \int \Big\{ \frac{\phi_h(x - X_1)\phi_{\sigma_0}(x-\mu_0)}{\phi_{\sigma_0}(X_1 - \mu_0)} \frac{\phi_h(x - X_2)\phi_{\sigma_0}(x-\mu_0)}{\phi_{\sigma_0}(X_2 - \mu_0)} \Big\}\, \mathrm{d}x \\
&= \int \{\mathrm{E}\widehat{f}(x)\}^2\, \mathrm{d}x \\
&= \frac{1}{2\pi} \sum_{i,j} p_i p_j \frac{1}{\sigma_i \sigma_j b_i b_j} \int \exp\Big\{ -\tfrac{1}{2} \frac{(x-\mu_i)^2}{\sigma_i^2} - \tfrac{1}{2} \frac{(x-\mu_j)^2}{\sigma_j^2} \\
&\qquad + \tfrac{1}{2} \Big(\frac{x-\mu_i}{\sigma_i^2} - \frac{x-\mu_0}{\sigma_0^2} \Big)^2 h^2 b_i^2 + \tfrac{1}{2} \Big(\frac{x-\mu_j}{\sigma_j^2} - \frac{x-\mu_0}{\sigma_0^2} \Big)^2 h^2 b_j^2 \Big\}.
\end{aligned}$$

Collecting x^2 terms and transforming to the standard normal, employing

$$\begin{aligned}
c_{i,j} &= \Big\{ \frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2} - \Big(\frac{1}{\sigma_i^2} - \frac{1}{\sigma_0^2} \Big)^2 h^2 b_i^2 - \Big(\frac{1}{\sigma_j^2} - \frac{1}{\sigma_0^2} \Big)^2 h^2 b_j^2 \Big\}^{1/2}, \\
d_{i,j} &= \frac{\mu_i}{\sigma_i^2} + \frac{\mu_j}{\sigma_j^2} - \Big(\frac{1}{\sigma_i^2} - \frac{1}{\sigma_0^2} \Big) \Big(\frac{\mu_i}{\sigma_i^2} - \frac{\mu_0}{\sigma_0^2} \Big) h^2 b_i^2 - \Big(\frac{1}{\sigma_j^2} - \frac{1}{\sigma_0^2} \Big) \Big(\frac{\mu_j}{\sigma_j^2} - \frac{\mu_0}{\sigma_0^2} \Big) h^2 b_j^2,
\end{aligned}$$

the result is

$$\begin{aligned}
\mathrm{E}A_{1,h} = \sqrt{2\pi} \sum_{i,j} \frac{p_i p_j}{b_i b_j} \phi_{\sigma_i}(\mu_i) \phi_{\sigma_j}(\mu_j) \frac{1}{c_{i,j}} \exp\Big\{ \tfrac{1}{2} \frac{d_{i,j}^2}{c_{i,j}^2} \\
+ \tfrac{1}{2} \Big(\frac{\mu_i}{\sigma_i^2} - \frac{\mu_0}{\sigma_0^2} \Big)^2 h^2 b_i^2 + \tfrac{1}{2} \Big(\frac{\mu_j}{\sigma_j^2} - \frac{\mu_0}{\sigma_0^2} \Big)^2 h^2 b_j^2 \Big\}.
\end{aligned} \tag{A.8}$$

Similar and somewhat arduous calculations yield the mean of $A_{2,h}$. The starting point is

$$\begin{aligned}
\mathrm{E}A_{2,h} &= \mathrm{E} \int \Big\{ \phi_{\sigma_0}(x - \mu_0) \frac{\phi_h(X_i - x)}{\phi_{\sigma_0}(X_i - \mu_0)} \Big\}^2\, \mathrm{d}x \\
&= \int \phi_{\sigma_0}(x-\mu_0)^2 \Big\{ \sum_{i=1}^{k} p_i \int \frac{\phi_h(y-x)^2}{\phi_{\sigma_0}(y-\mu_0)^2} \phi_{\sigma_i}(y - \mu_i)\, \mathrm{d}y \Big\}\, \mathrm{d}x \\
&= h^{-1} \sum_{i=1}^{k} p_i \int \phi_{\sigma_0}(x-\mu_0)^2 \Big\{ \int \phi(z)^2 \frac{\phi_{\sigma_i}(x - \mu_i + hz)}{\phi_{\sigma_0}(x - \mu_0 + hz)^2}\, \mathrm{d}z \Big\}\, \mathrm{d}x.
\end{aligned}$$

Again z^2 terms have to be collected for the inner integral and then x^2 terms to do the rest. We need to introduce

$$\begin{aligned}
e_i &= \Big(2 + \frac{h^2}{\sigma_i^2} - 2\frac{h^2}{\sigma_0^2} \Big)^{1/2}, \\
f_i &= \Big\{ \frac{1}{\sigma_i^2} - \Big(\frac{1}{\sigma_i^2} - \frac{2}{\sigma_0^2} \Big)^2 h^2 e_i^2 \Big\}^{1/2}, \\
g_i &= \frac{\mu_i}{\sigma_i^2} - \Big(\frac{1}{\sigma_i^2} - \frac{2}{\sigma_0^2} \Big) \Big(\frac{\mu_i}{\sigma_0^2} - 2\frac{\mu_0}{\sigma_0^2} \Big) \frac{h^2}{e_i^2}.
\end{aligned}$$

The answer is

$$\mathrm{E}A_{2,h} = \frac{h^{-1}}{\sqrt{2\pi}} \sum_{i=1}^{k} \frac{p_i}{\sigma_i e_i f_i} \exp\Big\{ \tfrac{1}{2}\frac{g_i^2}{f_i^2} - \tfrac{1}{2}\frac{\mu_i^2}{\sigma_i^2} + \tfrac{1}{2}\Big(\frac{\mu_i}{\sigma_i^2} - 2\frac{\mu_0}{\sigma_0^2}\Big)^2 \frac{h^2}{e_i^2} \Big\}. \tag{A.9}$$

This is close to $h^{-1}(2\sqrt{\pi})^{-1}$ when h is small.

It remains only to find the mean of $B_h = \int f\widehat{f}\,\mathrm{d}x$. By our earlier result about the exact mean of $\widehat{f}$ this is equal to

$$\begin{aligned} \mathrm{E}B_h &= \int f(x)\,\mathrm{E}\widehat{f}(x)\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \sum_{i,j} p_i p_j \frac{1}{\sigma_i b_i} \int \exp\Big\{ \tfrac{1}{2}\Big(\frac{x-\mu_i}{\sigma_i^2} - \frac{x-\mu_0}{\sigma_0^2}\Big)^2 \frac{h^2}{b_i^2} \\ &\qquad - \tfrac{1}{2}\frac{(x-\mu_i)^2}{\sigma_i^2} \Big\} \phi_{\sigma_j}(x-\mu_j)\,\mathrm{d}x. \end{aligned}$$

This time we need

$$\begin{aligned} k_{i,j} &= \Big\{ \frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2} - \Big(\frac{1}{\sigma_i^2} - \frac{1}{\sigma_0^2}\Big)^2 \frac{h^2}{b_i^2} \Big\}^{1/2}, \\ l_{i,j} &= \frac{\mu_i}{\sigma_i^2} + \frac{\mu_j}{\sigma_j^2} - \Big(\frac{1}{\sigma_i^2} - \frac{1}{\sigma_0^2}\Big)\Big(\frac{\mu_i}{\sigma_i^2} - \frac{\mu_0}{\sigma_0^2}\Big)\frac{h^2}{b_i^2}, \end{aligned}$$

and the result is

$$\mathrm{E}B_h = \sqrt{2\pi} \sum_{i,j} p_i p_j \phi_{\sigma_i}(\mu_i)\phi_{\sigma_j}(\mu_j) \frac{1}{b_i}\frac{1}{k_{i,j}} \exp\Big\{ \tfrac{1}{2}\Big(\frac{\mu_i}{\sigma_i^2} - \frac{\mu_0}{\sigma_0^2}\Big)^2 \frac{h^2}{b_i^2} + \tfrac{1}{2}\frac{l_{i,j}^2}{k_{i,j}^2} \Big\}. \tag{A.10}$$

This ends our proof. □

Consider the limiting case where $\sigma_0 \to \infty$. Then our estimator is nothing but the usual kernel estimator. Somewhat strenuous algebraic calculations yield

$$\begin{aligned} \mathrm{E}A_{1,h} &= \sum_{i,j} p_i p_j \phi_{(\sigma_i^2+\sigma_j^2+2h^2)^{1/2}}(\mu_j - \mu_i), \\ \mathrm{E}A_{2,h} &= (2\sqrt{\pi}h)^{-1}, \\ \mathrm{E}B_h &= \sum_{i,j} p_i p_j \phi_{(\sigma_i^2+\sigma_j^2+h^2)^{1/2}}(\mu_j - \mu_i), \end{aligned}$$

which with (A.7) and (A.5) again quite satisfactorily give the (5.4) formula for the exact mise(h) of the kernel estimator.

We used these results to go through the 15 test densities of Marron & Wand (1992), with the natural aim of comparing the minimum possible mise for the kernel method with the minimum possible mise for the new method (3.5). These minima, respectively $\mathrm{mise}^*_{\mathrm{trad}}$ and mise^*, were found, along with the minimisers h^*_{trad} and h^*, for sample sizes $n = 25, 50, 100, 250, 1000$. See the discussion of Section 5B.

TABLE A.2. *Values are given of the mise-minimising smoothing parameters h^* and h^*_{trad} for the (3.5) estimator and the kernel estimator, along with the minimum mise values* mise^* *and* $\mathrm{mise}^*_{\mathrm{trad}}$. *This is done for each of the 15 test densities of Marron & Wand, for sample sizes 25, 50, 100, 200, 1000. Also included in each case is the ratio* $\mathrm{mise}^*/\mathrm{mise}^*_{\mathrm{trad}}$.

n	h^*	mise*	h^*_{trad}	mise$^*_{\text{trad}}$	mise-ratio
Case #1, Gaußian:					
25	0.7071	0.0113	0.6094	0.0137	0.8217
50	0.7071	0.0056	0.5199	0.0087	0.6492
100	0.7071	0.0028	0.4455	0.0054	0.5215
200	0.7071	0.0014	0.3830	0.0033	0.4245
1000	0.7071	0.0003	0.2723	0.0010	0.2740
Case #2, skewed unimodal:					
25	0.3928	0.0228	0.4251	0.0211	1.0772
50	0.3787	0.0123	0.3591	0.0134	0.9173
100	0.3544	0.0068	0.3054	0.0083	0.8250
200	0.3209	0.0040	0.2611	0.0051	0.7767
1000	0.2381	0.0012	0.1841	0.0016	0.7396
Case #3, strongly skewed:					
25	0.0728	0.1456	0.1481	0.1032	1.4107
50	0.0720	0.0786	0.1082	0.0682	1.1523
100	0.0720	0.0444	0.0827	0.0435	1.0208
200	0.0655	0.0270	0.0654	0.0270	0.9996
1000	0.0415	0.0084	0.0414	0.0084	0.9989
Case #4, kurtotic unimodal:					
25	0.1252	0.1098	0.1241	0.1101	0.9972
50	0.0976	0.0688	0.0967	0.0691	0.9949
100	0.0791	0.0421	0.0784	0.0424	0.9937
200	0.0656	0.0253	0.0650	0.0255	0.9930
1000	0.0445	0.0075	0.0441	0.0076	0.9922
Case #5, outlier:					
25	0.0634	0.1433	0.0646	0.1424	1.0062
50	0.0562	0.0862	0.0548	0.0890	0.9690
100	0.0487	0.0523	0.0468	0.0548	0.9549
200	0.0420	0.0317	0.0402	0.0334	0.9492
1000	0.0299	0.0096	0.0285	0.0102	0.9462
Case #6, bimodal:					
25	0.5568	0.0197	0.6028	0.0182	1.0792
50	0.4559	0.0123	0.4721	0.0119	1.0342
100	0.3823	0.0075	0.3854	0.0075	1.0067
200	0.3247	0.0045	0.3217	0.0046	0.9888
1000	0.2278	0.0013	0.2208	0.0014	0.9663
Case #7, separated bimodal:					
25	0.3701	0.0303	0.3661	0.0306	0.9881
50	0.3136	0.0183	0.3082	0.0187	0.9813
100	0.2674	0.0110	0.2616	0.0112	0.9768
200	0.2291	0.0065	0.2235	0.0067	0.9738
1000	0.1620	0.0019	0.1575	0.0020	0.9700
Case #8, skewed bimodal:					
25	0.5136	0.0243	0.5549	0.0222	1.0953
50	0.3903	0.0158	0.4085	0.0151	1.0507
100	0.3112	0.0100	0.3179	0.0097	1.0251
200	0.2554	0.0061	0.2572	0.0061	1.0099
1000	0.1712	0.0019	0.1697	0.0019	0.9924

Case #9, trimodal:					
25	0.5373	0.0224	0.5889	0.0206	1.0840
50	0.4331	0.0144	0.4551	0.0138	1.0435
100	0.3509	0.0091	0.3588	0.0089	1.0193
200	0.2858	0.0057	0.2874	0.0056	1.0052
1000	0.1848	0.0018	0.1829	0.0018	0.9910
Case #10, claw:					
25	0.4930	0.0659	0.5101	0.0636	1.0372
50	0.4267	0.0578	0.4034	0.0570	1.0145
100	0.0955	0.0371	0.0959	0.0370	1.0033
200	0.0774	0.0224	0.0775	0.0224	1.0007
1000	0.0517	0.0067	0.0516	0.0067	0.9979
Case #11, double claw:					
25	0.5556	0.0212	0.6018	0.0197	1.0748
50	0.4550	0.0138	0.4717	0.0134	1.0318
100	0.3817	0.0090	0.3851	0.0089	1.0073
200	0.3242	0.0060	0.3215	0.0061	0.9925
1000	0.2248	0.0028	0.2176	0.0029	0.9854
Case #12, asymmetric claw:					
25	0.7289	0.0363	0.6657	0.0359	1.0121
50	0.6044	0.0312	0.5231	0.0309	1.0079
100	0.1989	0.0232	0.2016	0.0229	1.0115
200	0.1428	0.0161	0.1436	0.0160	1.0073
1000	0.0675	0.0064	0.0678	0.0064	1.0043
Case #13, asymmetric double claw:					
25	0.5254	0.0254	0.5620	0.0241	1.0532
50	0.4315	0.0174	0.4428	0.0171	1.0188
100	0.3608	0.0123	0.3612	0.0123	1.0008
200	0.3021	0.0091	0.2971	0.0091	0.9937
1000	0.1030	0.0045	0.1030	0.0045	1.0010
Case #14, smooth comb:					
25	0.2866	0.0678	0.2858	0.0675	1.0037
50	0.2035	0.0488	0.2031	0.0487	1.0026
100	0.1434	0.0348	0.1434	0.0347	1.0021
200	0.1015	0.0245	0.1016	0.0244	1.0018
1000	0.0439	0.0101	0.0439	0.0101	1.0007
Case #15, discrete comb:					
25	0.2459	0.0704	0.2469	0.0702	1.0033
50	0.2016	0.0493	0.2014	0.0493	1.0007
100	0.1638	0.0362	0.1630	0.0362	0.9998
200	0.0815	0.0266	0.0816	0.0266	1.0016
1000	0.0422	0.0087	0.0423	0.0087	1.0006

Acknowledgements. We are grateful for useful and encouraging comments from M.C. Jones and for discussions with Grete Fenstad. A part of this work was carried out while one of was visiting Oxford University with a grant from the Royal Norwegian Research Council.

References

Buckland, S.T. (1992). Maximum likelihood fitting of Hermite and simple polynomial densities. *Applied Statistics* **41**, 241–266.

Friedman, J.H., Stuetzle, W., and Schroeder, A. (1984). Projection pursuit density estimation. *Journal of the American Statistical Association* **79**, 599–608.

Hall, P.G., Sheather, S.J., Jones., M.C., and Marron, S.J. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78**, 263–269.

Hjort, N.L. (1986). *Statistical Symbol Recognition.* Research monograph, Norwegian Computing Centre, Oslo.

Hjort, N.L. (1993). Dynamic likelihood hazard rate estimation. *Biometrika*, to appear.

Hjort, N.L. (1994). Bayesian approaches to semiparametric density estimation. Invited paper, in progress, to be published in the proceedings of the Fifth Valencia International Meeting on Bayesian Statistics.

Hjort, N.L. and Jones, M.C. (1993). Locally parametric nonparametric density estimation. Statistical Research Report, Department of Mathematics, University of Oslo. Submitted for publication.

Hjort, N.L. and Jones, M.C. (1994). Better rules of thumb for choice of smoothing parameter in density estimation. In progress.

Hjort, N.L. and Fenstad, G.U. (1994). Hermite versus Kernel. In progress.

Huber, P.J. (1981). *Robust Statistics.* Wiley, New York.

Jones, M.C. (1993). Kernel density estimation when the bandwidth is large. *Australian Journal of Statistics*, to appear.

Jones, M.C., Linton, O. and Nielsen, J.P. (1993). A simple and effective bias reduction method for density and regression estimation. Manuscript.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1993). Progress in data-based bandwidth selection for kernel density estimation. Working Paper 92–014, Australian Graduate School of Management, University of New South Wales.

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

Olkin, I. and Spiegelman, C.H. (1987). A semiparametric approach to density estimation. *Journal of the American Statistical Association* **82**, 858–865.

Schuster, E. and Yakowitz, S. (1985). Parametric/nonparametric mixture density estimation with application to flood-frequency analysis. *Water Resources Bulletin* **21**, 797–804.

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

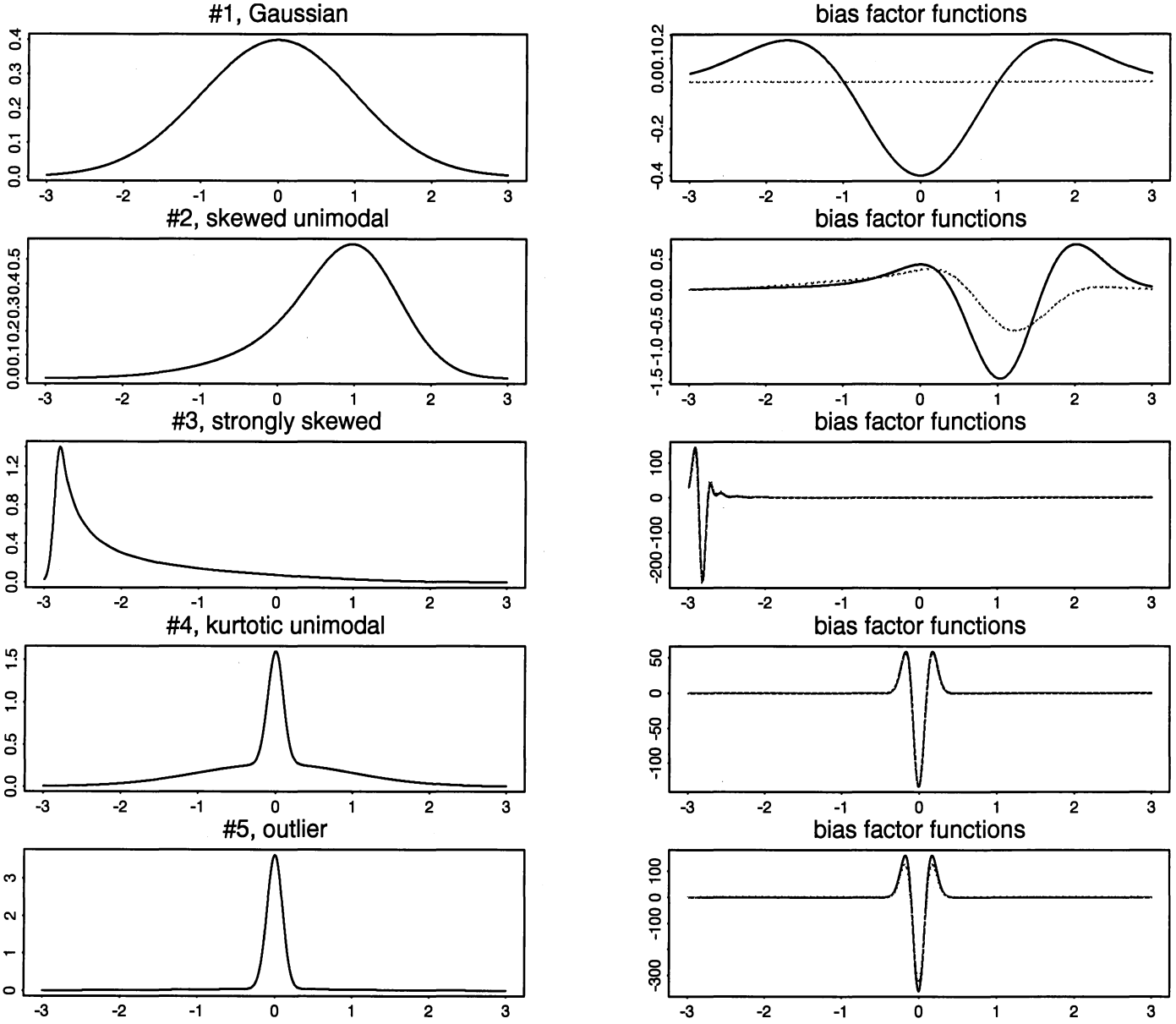
Scott, D.W. and Terrell, G.R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association* **82**, 1131–1146.

Shao, J. (1991). Second-order differentiability and jackknife. *Statistica Sinica* **1**, 185–202.

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Association* **B 53**, 683–690.

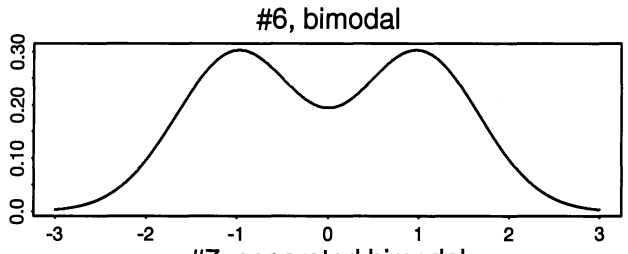
Wand, M.P. and Jones, M.C. (1994). *Kernel Smoothing.* Chapman & Hall, London. To exist.

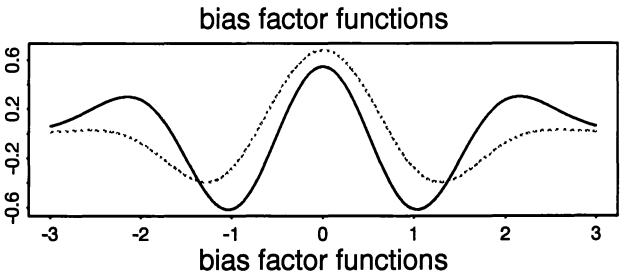
Wand, M.P., Marron, J.S., and Ruppert, D. (1991). Transformations in density estimation [with discussion contributions]. *Journal of the American Statistical Association* **86**, 343–361.

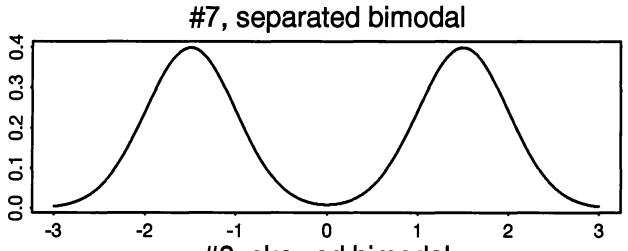


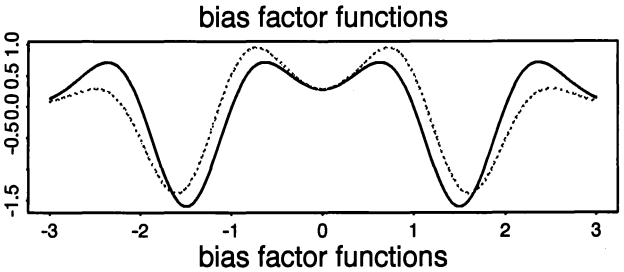
This figure belongs to the discussion of Section 5A

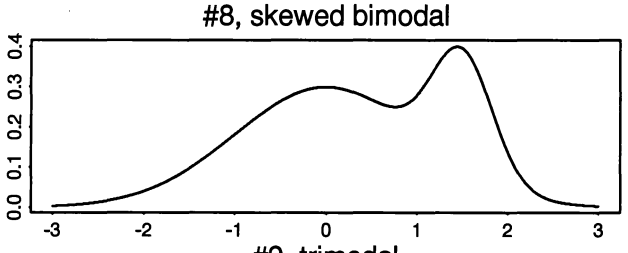
FIGURE. *The 15 test densities (left hand side) presented together with the bias factor functions f'' (solid line, for the kernel method) and $f_0 r''$ (dotted line, for the new method).*

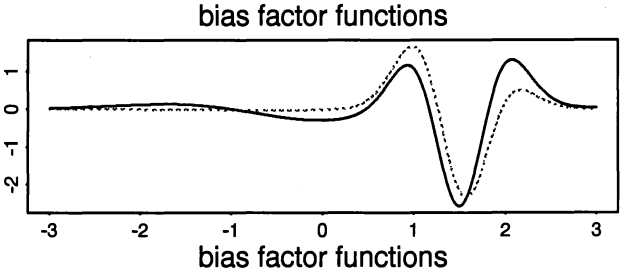

#6, bimodal
0.0 0.10 0.20 0.30
-3 -2 -1 0 1 2 3

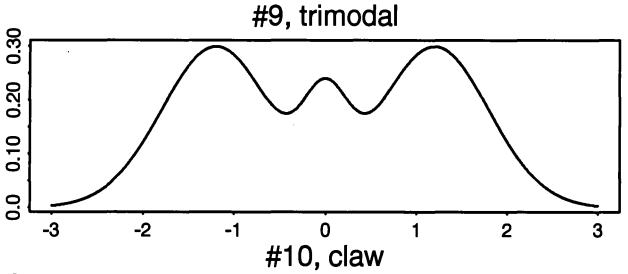

bias factor functions
-0.6 -0.2 0.2 0.6
-3 -2 -1 0 1 2 3

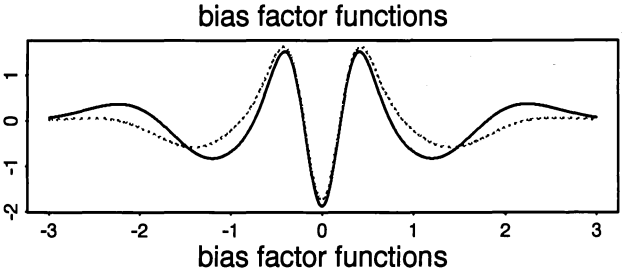

#7, separated bimodal
0.0 0.1 0.2 0.3 0.4
-3 -2 -1 0 1 2 3

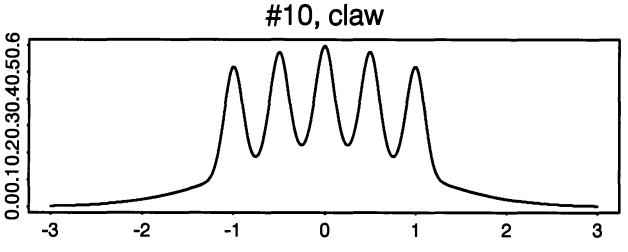

bias factor functions
-1.5 -0.5 0.0 0.5 1.0
-3 -2 -1 0 1 2 3

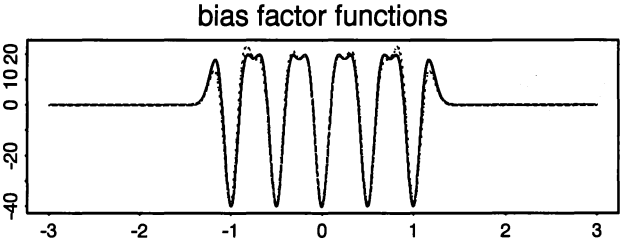

#8, skewed bimodal
0.0 0.1 0.2 0.3 0.4
-3 -2 -1 0 1 2 3

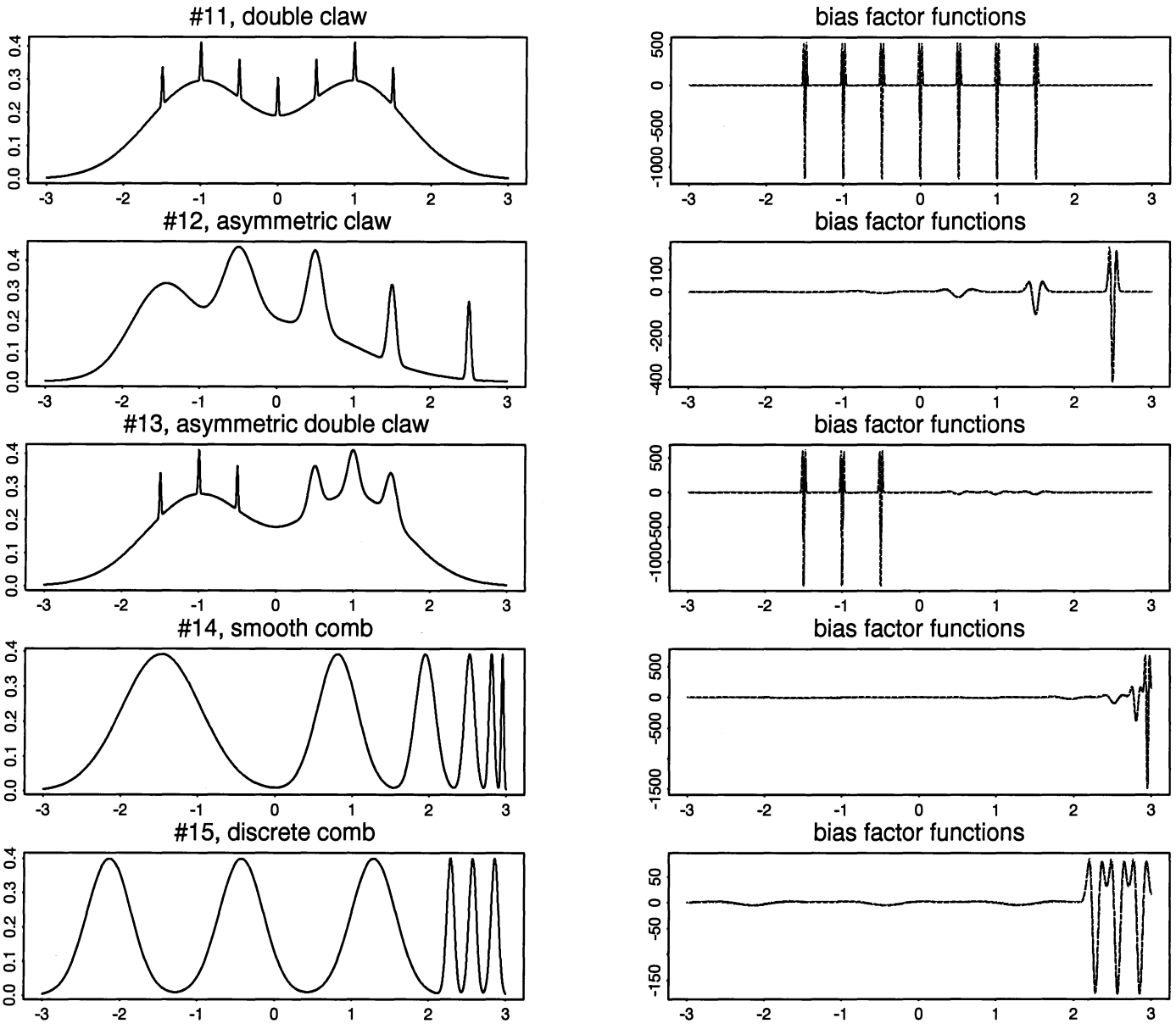

bias factor functions
-2 -1 0 1
-3 -2 -1 0 1 2 3


#9, trimodal
0.0 0.10 0.20 0.30
-3 -2 -1 0 1 2 3


bias factor functions
-2 -1 0 1
-3 -2 -1 0 1 2 3


#10, claw
0.00.10.20.30.40.50.6
-3 -2 -1 0 1 2 3


bias factor functions
-40 -20 0 10 20
-3 -2 -1 0 1 2 3


#11, double claw
bias factor functions
#12, asymmetric claw
bias factor functions
#13, asymmetric double claw
bias factor functions
#14, smooth comb
bias factor functions
#15, discrete comb
bias factor functions

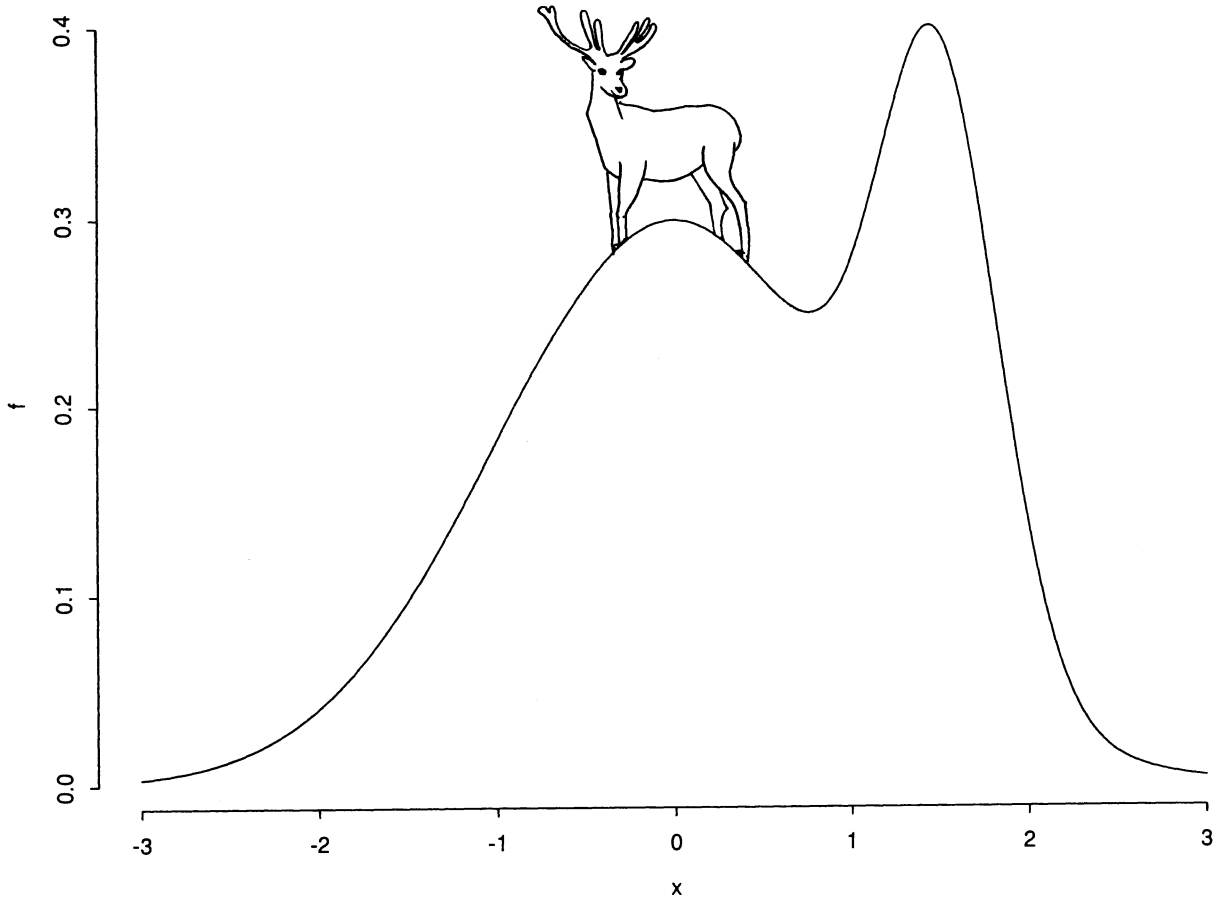


FIGURE 2. *The Glad Hjort method, nonparametrically improving on a preliminary normal density estimate:*

$$\widehat{f}(x) = \frac{1}{\widehat{\sigma}}\phi\Big(\frac{x-\widehat{\mu}}{\widehat{\sigma}}\Big)\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\Big/\frac{1}{\widehat{\sigma}}\phi\Big(\frac{X_i-\widehat{\mu}}{\widehat{\sigma}}\Big)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2\sqrt{\pi}}\frac{1}{h}\exp\{-\tfrac{1}{2}(X_i-x)^2/h^2\}\frac{\exp\{-\frac{1}{2}(x-\widehat{\mu})^2/\widehat{\sigma}^2\}}{\exp\{-\frac{1}{2}(X_i-\widehat{\mu})^2/\widehat{\sigma}^2\}}$$