Exact performance of a semiparametric density estimator for normal mixture truths

Nils Lid Hjort and Ingrid K. Glad

University of Oslo and University of Rome La Sapienza

March 1996

ABSTRACT. A semiparametric density estimator that combines a parametric start estimate with a nonparametric kernel type correction factor was introduced in Hjort and Glad (1995), where also the basic large-sample properties of the new estimator were developed. The present work complements our previous paper by offering an exact analysis of its performance and comparing it to that of the traditional kernel type estimator, for the wide class of all normal mixtures. Using a normal density as the initial parametric start, exact expressions for the leading terms of mean integrated squared error as well as the finite-sample mean integrated squared error are derived and compared to the corresponding expressions for the kernel estimator. A set of 15 normal mixtures whose forms vary from normal and moderately non-normal to extremely non-normal cases is used to illustrate the comparison. It is seen that in almost all these test situations the semiparametric method is superior to the traditional kernel method both in terms of asymptotic and exact finite-sample mean integrated squared errors. It is also observed that in the highly non-normal cases where the initial normal density clearly is an unreasonable start, the semiparametric estimator behaves very similarly to the kernel method.

KEY WORDS: *correction factor, exact mean integrated squared error, finite-sample comparisons, kernel methods, lowering the bias, normal mixtures, semiparametric density estimation, test cases*

1. Introduction and summary. This paper investigates the asymptotic and finite-sample performance of a semiparametric density estimator proposed in Hjort and Glad (1995) when the underlying density is a normal mixture on the real line. The family of such mixtures forms an extremely wide and flexible class of densities and hence is capable of mimicking a broad spectrum of underlying truths f. The class of estimators developed and analysed in Hjort and Glad (1995) combines an initial parametric estimate of f with a nonparametric kernel type estimate of the necessary correction factor. This approach enjoys performance properties that are in general similar to the totally nonparametric kernel method, but indeed better when the true density is in a broad vicinity of the chosen parametric family. The present paper provides a deeper study of such a nonparametric vicinity around the normal distribution, based on exact expressions derived for asymptotic or approximate mean integrated squared error (AMISE) and finite-sample mean integrated squared error (MISE) in the context of normal mixture truths.

Let $X_1, \ldots, X_n$ be independent observations from the unknown density f which is to be estimated. The traditional nonparametric kernel estimator of the unknown density is

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

where $K_h(z) = h^{-1}K(h^{-1}z)$ and $K(z)$ is a kernel function, which is assumed 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

$$\begin{aligned} \mathrm{E}\widetilde{f}(x) &= f(x) + \tfrac{1}{2}\sigma_K^2 h^2 f''(x) + O(h^4) \\ \text{and} \quad \mathrm{Var}\,\widetilde{f}(x) &= R(K)(nh)^{-1} f(x) - f(x)^2/n + O(h/n). \end{aligned} \tag{1.2}$$

Consistency requires both $h \to 0$ and $nh \to \infty$. The MISE is of order $n^{-4/5}$ when h is proportional to $n^{-1/5}$, which is the optimal size. Scott (1992, Chapter 6) and Wand and Jones (1995, Chapter 2) give recent accounts of the theory.

The idea of the semiparametric estimator is to incorporate some possible preference for the shape of the density into the shape impartial kernel estimator, while not violating the nonparametric ability of adaptation to the data. Let $f(x,\theta)$ be a given parametric family of densities, and let the possibly multi-dimensional parameter $\theta = (\theta_1, \ldots, \theta_p)'$ be estimated from the data by some estimator of the form $\widehat{\theta} = T(F_n)$, writing F_n for the empirical distribution of the n data points, and having an influence function with finite covariance matrix. The parametric start estimate is $f(x,\widehat{\theta})$. For example, think of the normal density with maximum likelihood estimates for mean and variance, on which we will focus in the following sections. This initial data summary is not necessarily a serious description of the true density; the method is intended to work well even if f cannot be well approximated by any usual $f(.,\theta)$. The multiplicative correction function $f(x)/f(x,\widehat{\theta})$ is estimated by means of kernel smoothing, $\widehat{r}(x) = n^{-1}\sum_{i=1}^n K_h(X_i - x)/f(X_i,\widehat{\theta})$, giving the semiparametric density estimator

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

The traditional kernel estimator (1.1) corresponds to using a uniform density as the parametric start and can therefore be viewed as a special case of this estimator. We stress that in general any parametric family can be used; one possibility is to choose the parametric family according to some goodness of fit criterion.

Omitting all details and referring to Hjort and Glad (1995) for proofs, the main statistical properties of the (1.3) estimator can be summarised as

$$\begin{aligned} \mathrm{E}\widehat{f}(x) &= f(x) + \tfrac{1}{2}\sigma_K^2 h^2 f_0(x) r''(x) + O(h^4 + h^2/n + n^{-2}) \\ \text{and} \quad \mathrm{Var}\,\widehat{f}(x) &= R(K)(nh)^{-1} f(x) - f(x)^2/n + O(h/n + n^{-2}), \end{aligned} \tag{1.4}$$

where $h \to 0$ while $n \to \infty$. Here $f_0(x) = f(x,\theta_0)$ is the best parametric approximant within the family $f(x,\theta)$ to f (thus θ_0 is the T functional evaluated at the distribution with density f), and $r = f/f_0$. This means that for the same $K(\cdot)$ and h, the variance of the (1.3) estimator is the very same as the variance of the nonparametric kernel method, up to the order of approximation used, while the bias has the same order h^2 as for the kernel estimator, but with another constant.

If the chosen parametric family happens to be the right one, then the correction function r will be 1 and the bias reduces to only $O(h^2/n + n^{-2})$. In this case it is also shown in Hjort and Glad (1995) that the semiparametric estimator shares the advantageous mean squared error order of n^{-1} with the strictly parametric estimators.

In general, by comparing the bias terms or by means of approximate mean squared error (AMSE), the semiparametric method is better than the kernel method in points x for which $|f_0(x)r''(x)| < |f''(x)|$. According to the integrated AMSE, say AMISE, the semiparametric method is better than the kernel method when the 'roughness' functional for the (1.3) estimator,

$$R_{\text{new}}(f) = \int (f_0 r'')^2 \, \mathrm{d}x, \tag{1.5}$$

is smaller than the corresponding functional for the (1.1) estimator,

$$R_{\text{trad}}(f) = \int (f'')^2 \, \mathrm{d}x. \tag{1.6}$$

Hence the new estimator is better for all f in some nonparametric neighbourhood around the parametric family. One might say that the new estimator wins when the parametric family has managed to capture some of f's structural features, leading to a correction factor which is less rough than the original density.

In the following sections we will concentrate on the special case of (1.3) that uses a normal density as the initial descriptor in concert with the standard Gaussian kernel $K = \phi$. That is, we focus on

$$\begin{aligned} \widehat{f}(x) &= \frac{1}{n}\sum_{i=1}^{n} \phi_h(X_i - x) \frac{\phi_{\widehat{\sigma}}(x - \widehat{\mu})}{\phi_{\widehat{\sigma}}(X_i - \widehat{\mu})} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\exp\{-\frac{1}{2}(X_i - x)^2/h^2\}}{\sqrt{2\pi}h} \frac{\exp\{-\frac{1}{2}(x - \widehat{\mu})^2/\widehat{\sigma}^2\}}{\exp\{-\frac{1}{2}(X_i - \widehat{\mu})^2/\widehat{\sigma}^2\}}, \end{aligned} \tag{1.7}$$

with $\widehat{\mu}$ and $\widehat{\sigma}$ being the usual maximum likelihood estimates.

The AMISE comparisons in terms of roughness functionals for the normal start estimator (1.7) are investigated in the context of Hermite expansions in Hjort and Glad (1995). By using standard and robust Hermite expansions around the normal, see Fenstad and Hjort (1996) and Hjort and Jones (1996), conditions on the Hermite coefficients of the true f could be found which assure smaller AMISE for the semiparametric estimator than for the kernel method.

In Section 2 below the AMSE and AMISE criteria for the (1.7) estimator are studied further, assuming that the true f belongs to the large class of all normal mixtures. Under this assumption we present exact expressions for the leading terms of AMSE and AMISE for the (1.7) estimator and compare these to those of the traditional kernel estimator. In addition we derive expressions for the main terms of a corresponding L_1 type criterion, based on integrated absolute bias and integrated mean absolute deviation. The comparison of the two estimation methods is illustrated by computing these performance measures for a set of 15 special test cases selected by Marron and Wand (1992).

In Section 3 we go further and develop the exact finite-sample MISE for the (1.7) estimator in this setting. This has been carried out for the kernel estimator for each of the 15 test densities in Marron and Wand (1992). In order to compare exact MISE of our estimator to that of the kernel estimator we use a best case versus best case approach and present the

two best achievable exact MISE values with respect to h for each of the 15 test cases and for sample sizes 25, 50, 100, 200, and 1000.

The comparisons show that our new estimator performs better than the traditional nonparametric method both in significant x-areas and in terms of AMISE and MISE, for almost all the distributional shapes represented by the test densities. In the few cases where this is not the case, the difference in these measures is surprisingly small. To sum up, the main point is that the semiparametric method leads to bias reduction when the true distribution is in a reasonable vicinity of the start density, without sacrificing variance. And even if the chosen parametric model is highly misleading, there is still little to lose in precision compared to the traditional nonparametric method.

For a more detailed presentation of the semiparametric estimation method, also in higher dimensions, along with a careful study of the problem of choosing h, we refer again to Hjort and Glad (1995). A similar idea is explored for regression analysis in Glad (1995). For an overview of and comments on related work, we refer to Hjort and Glad (1995), Efron and Tibshirani (1995), Hjort (1996), Jones and Signorini (1996) and the references therein.

REMARK 1. The (1.3) estimator does not integrate exactly to 1. The additional term is in general of order $O_p(h^2)$ and does not represent a serious drawback, see Remark 3 in Hjort and Glad (1995). Nevertheless, we emphasise here the advantages of normalising (1.3), applying $\widehat{f}_{\rm norm}(x) = \widehat{f}(x)/\int \widehat{f}(s)\,{\rm d}s$. Not only does this version of the estimator integrate to 1; it also has the appealing feature of turning strictly parametric when the smoothing parameter $h \to \infty$. It reproduces, subject to an infinite amount of smoothing, exactly the parametric estimator that we initially applied. This means that the normalised semiparametric estimator provides a continuous bridge between the kernel estimator and the parametric start estimator, between purely nonparametric and purely parametric estimates. The asymptotic properties of the estimator will generally change slightly through this operation, though remaining of the same order as before. For the normal start estimator (1.7), however, the picture simplifies. This estimator has integral $1+O_p(h^4)$, from which it follows that its normalised version has, up to the orders used, the very same asymptotic properties as the (1.7) estimator itself. This normalised estimator explicitly reads

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \phi_h(X_i - x)\frac{\phi_{\widehat{\sigma}}(x-\widehat{\mu})}{\phi_{\widehat{\sigma}}(X_i-\widehat{\mu})} \Big/ \frac{1}{n}\sum_{i=1}^{n} \frac{\phi_{(\widehat{\sigma}^2+h^2)^{1/2}}(X_i-\widehat{\mu})}{\phi_{\widehat{\sigma}}(X_i-\widehat{\mu})}, \tag{1.8}$$

which indeed tends to $\phi_{\widehat{\sigma}}(x-\widehat{\mu})$ as $h \to \infty$.

REMARK 2. We are of course aware of the fact that the ISE and MISE criteria focussed on in our paper are not always the best measures of respectively statistical error and performance. They are in heavy use, for reasons of mathematical convenience and tradition, and they do carry the basic information about squared bias and variance. In Section 2 we also include performance figures using a criterion related to the L_1 distance. Also worth considering, in future work, would be the 'visual-error' type criteria proposed in Marron and Tsybakov (1995). These aim at being more relevant for detecting underlying structure and features of the density.

2. Exact amise analysis for normal mixtures. Let the true density be of the form

$$f = \sum_{i=1}^{k} p_i f_i, \quad \text{where } f_i(x) = \phi_{\sigma_i}(x - \mu_i). \tag{2.1}$$

Our aim is to compare the (1.7) estimator $\widehat{f}$ with the kernel estimator $\widetilde{f}$, which we also take to have the normal kernel $K = \phi$.

According to (1.2) and (1.4) we need to compare f'' to $f_0 r''$ in order to monitor the bias terms of the two estimators. Here 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

$$f_0(x) r''(x) = \sum_{i=1}^{k} p_i f_i(x)[1/\sigma_0^2 - 1/\sigma_i^2 + \{(x - \mu_i)/\sigma_i^2 - (x - \mu_0)/\sigma_0^2\}^2], \tag{2.2}$$

while

$$f''(x) = \sum_{i=1}^{k} p_i \phi_{\sigma_i}''(x - \mu_i) = \sum_{i=1}^{k} p_i\{(x - \mu_i)^2/\sigma_i^2 - 1\} f_i(x)/\sigma_i^2. \tag{2.3}$$

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

In Figure 1 these formulae are used to visually inspect $f''(x)$ versus $f_0(x)r''(x)$, for each of the 15 test cases of Marron and Wand (1992). (These test densities are normal mixtures originally used for exact MISE analysis of the kernel estimator and are meant to represent different truths f, varying from simple to extremely difficult cases.) First, we observe that in most cases where the initial normal approximation is not totally unreasonable, the new estimator manages to be better than the usual one in significant x-areas. Secondly, we observe that in cases where the initial description is clearly a bad start, the semiparametric method turns almost nonparametric and behaves very similarly to the kernel method.

With some efforts (2.2) and (2.3) also lead to formulae for the roughness values $R_{\text{trad}}(f)$ and $R_{\text{new}}(f)$ for general normal mixtures.

PROPOSITION 2.1. *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 (1.5) and (1.6) can be calculated explicitly;*

$$R_{\text{trad}} = \int (f'')^2 \, dx = \sum_{i,j} p_i p_j (\delta_{i,j}^4 - 6\delta_{i,j}^2 + 3)\phi(\delta_{i,j})/\sigma_{i,j}^5,$$

$$R_{\text{new}} = \int (f_0 r'')^2 \, dx = T_1 + \cdots + T_6,$$

with these latter terms being defined in equation (2.4) below. (The R_{trad} result is also contained in Marron and Wand (1992, Theorem 4.1).)

PROOF: Start out noting that

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

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

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

say, H_r and H_s being the Hermite polynomials defined by $\phi^{(r)}(x) = (-1)^r\phi(x)H_r(x)$. This leads to

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

proving the first and simplest claim. To find $\int(f_0r'')^2\,\mathrm{d}x$, write (2.2) 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 $R_{\mathrm{new}}(f)$, where

$$\begin{aligned}
T_1 &= \sum_{i,j}p_ip_jc_ic_jA_{i,j}^{0,0},\\
T_2 &= 2\sum_{i,j}p_ip_jc_id_j\sigma_j^2A_{i,j}^{0,1},\\
T_3 &= 2\sum_{i,j}p_ip_jc_ia_j^2(\sigma_j^4A_{i,j}^{0,2} + \sigma_j^2A_{i,j}^{0,0}),\\
T_4 &= \sum_{i,j}p_ip_jd_id_j\sigma_i^2\sigma_j^2A_{i,j}^{1,1},\\
T_5 &= 2\sum_{i,j}p_ip_jd_ia_j^2(\sigma_i^2\sigma_j^4A_{i,j}^{2,1} + \sigma_i^2\sigma_j^2A_{i,j}^{1,0}),\\
T_6 &= \sum_{i,j}p_ip_ja_i^2a_j^2\sigma_i^2\sigma_j^2(\sigma_i^2\sigma_j^2A_{i,j}^{2,2} + \sigma_i^2A_{i,j}^{2,0} + \sigma_j^2A_{i,j}^{0,2} + A_{i,j}^{0,0}).
\end{aligned} \tag{2.4}$$

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_{new} and proves the second claim. □

In Table 2.1 below we have chosen to display

$$\rho_{\text{trad}}(f) = \sigma(f) R_{\text{trad}}(f)^{1/5} \qquad \text{and} \qquad \rho_{\text{new}}(f) = \sigma(f) R_{\text{new}}(f)^{1/5} \tag{2.5}$$

rather than R_{trad} and R_{new}, for the 15 test cases of Marron and Wand (1992). Here $\sigma(f)$ is the standard deviation of f. The R_{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 $\widehat{f}$ is

$$\tfrac{5}{4}\sigma(f)^{-1}\{\sigma_K R(K)\}^{4/5}\rho_{\text{new}}(f)/n^{4/5},$$

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. Leading term approximations for these criteria for the two estimators can be shown to be

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

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 Table 2.1 are

$$\rho^1_{\text{trad}}(f) = \Big(\int f^{1/2}\Big)^{4/5}\Big(\int |f''|\Big)^{1/5} \quad \text{and} \quad \rho^1_{\text{new}}(f) = \Big(\int f^{1/2}\Big)^{4/5}\Big(\int |f_0 r''|\Big)^{1/5}. \tag{2.6}$$

This is because the minimal possible value of (iab + imad) for $\widehat{f}$ can be shown to be $\frac{5}{4}(2^3/\pi^2)^{1/5}\{\sigma_K R(K)\}^{2/5}\rho^1_{\text{new}}(f)/n^{2/5}$, and similarly with $\widetilde{f}$. The quantities in (2.6) are scale invariant.

The overall comparison in terms of approximate MISE is in clear favour of the new semiparametric 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 and Wand to exhibit particularities of smoothing parameter problems. And the new method wins in each of the nine worldly cases: the normal, 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 and 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). By the Remark ending Section 1 these favourable comments also automatically apply to the normalised estimator (1.8).

TABLE 2.1. *Values of the global* MISE*-based comparison values* ρ_{trad} *and* ρ_{new}*, given for each of the 15 normal mixture test cases. Also included are the* L_1*-based global comparison values* ρ^1_{trad} *and* ρ^1_{new}*. The normal-start estimator (1.7) 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 + imad it wins in all cases except #3. The* ρ_{new} *and* ρ^1_{new} *figures are also valid for the normalised estimator (1.8).*

Case	ρ_{trad}	ρ_{new}	ρ^1_{trad}	ρ^1_{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

So in terms of AMISE the semiparametric (1.7) 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_{new} and R_{trad}. This picture emerges also when one looks at the values for the L_1-based criteria $\int |f''|$ versus $\int |f_0 r''|$. According to this measure the (1.7) 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. 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. A partial explanation lies in the earlier observation that the kernel estimator can be seen as the special case that starts with a uniform density as its initial description; many non-normal densities are after all better fitted by a normal than by a uniform.

3. Exact finite-sample mise analysis for normal mixtures. The comparison analysis above is in terms of the Taylor-based approximations to bias and variance. We now go further and analyse exact finite-sample MISE for the two estimation methods, (1.1) and (1.7). For technical reasons we need to treat two situations separately. The first allows known values for mean and variance to be plugged into the (1.7) estimator, and leads to a best case versus best case analysis in Section 3.1. The second situation is technically different but in terms of practical performance almost equivalent, and is the one met in practice, plugging in estimates for mean and variance in (1.7). Section 3.2 investigates this. Our findings confirm the conclusions made after the asymptotic comparison of Section 2, and in particular favour estimator (1.7) over the kernel estimator.

3.1. Best case versus best case analysis. Exact MISE analysis has already been carried out in Marron and Wand (1992) for the kernel method (1.1). A special case of their Theorem

2.1 which we need to record here is that, when f is as in (2.1),

$$\begin{aligned}
\text{MISE}(h) &= \mathrm{E}\int (\widetilde{f}-f)^2\,\mathrm{d}x \\
&= \Big(1-\frac{1}{n}\Big)\sum_{i,j}\frac{p_ip_j}{(\sigma_i^2+\sigma_j^2+2h^2)^{1/2}}\phi\Big(\frac{\mu_j-\mu_i}{(\sigma_i^2+\sigma_j^2+2h^2)^{1/2}}\Big) \\
&\quad + \frac{1}{n}\frac{1}{2\sqrt{\pi}h} - 2\sum_{i,j}\frac{p_ip_j}{(\sigma_i^2+\sigma_j^2+h^2)^{1/2}}\phi\Big(\frac{\mu_j-\mu_i}{(\sigma_i^2+\sigma_j^2+h^2)^{1/2}}\Big) \\
&\quad + \sum_{i,j}\frac{p_ip_j}{(\sigma_i^2+\sigma_j^2)^{1/2}}\phi\Big(\frac{\mu_j-\mu_i}{(\sigma_i^2+\sigma_j^2)^{1/2}}\Big).
\end{aligned} \tag{3.1}$$

Reaching a similar result for the MISE of the normal-start estimator (1.7) is much more demanding. Suppose again that f is as in (2.1). Start out with

$$\text{ISE}(h) = \int (\widehat{f}-f)^2\,\mathrm{d}x = A_h - 2B_h + R(f), \tag{3.2}$$

where

$$R(f) = \int f^2\,\mathrm{d}x = \sum_{i,j} p_ip_j\phi\Big(\frac{\mu_j-\mu_i}{\sigma_{i,j}}\Big)\frac{1}{\sigma_{i,j}}, \tag{3.3}$$

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

$$\int \prod_{j=1}^m \phi_{\sigma_j}(x-\mu_j)\,\mathrm{d}x = \sqrt{2\pi}\widetilde{\sigma}\Big[\prod_{j=1}^m \phi_{\sigma_j}(\mu_j-a)\Big]\exp\Big[\tfrac{1}{2}\widetilde{\sigma}^2\Big\{\sum_{j=1}^m(\mu_j-a)/\sigma_j^2\Big\}^2\Big], \tag{3.4}$$

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. The proof of (3.4) is not very difficult and is omitted. For the first term this identity gives

$$\begin{aligned}
A_h &= \frac{1}{n^2}\sum_{i,j\le n}\int \frac{\phi_h(x-x_i)\phi_h(x-x_j)\phi_{\widehat{\sigma}}(x-\widehat{\mu})^2}{\phi_{\widehat{\sigma}}(x_i-\widehat{\mu})\phi_{\widehat{\sigma}}(x_j-\widehat{\mu})}\,\mathrm{d}x \\
&= \frac{1}{n^2}\sum_{i,j\le n}\phi_{\widehat{\sigma}}(0)^2\sqrt{2\pi}\widetilde{\sigma}\frac{\phi_h(x_i-\widehat{\mu})}{\phi_{\widehat{\sigma}}(x_i-\widehat{\mu})}\frac{\phi_h(x_j-\widehat{\mu})}{\phi_{\widehat{\sigma}}(x_j-\widehat{\mu})}\exp\Big\{\tfrac{1}{2}\widetilde{\sigma}^2\Big(\frac{x_i-\widehat{\mu}+x_j-\widehat{\mu}}{h^2}\Big)^2\Big\},
\end{aligned} \tag{3.5}$$

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,

$$\begin{aligned}
B_h &= \sum_{j=1}^k p_j\frac{1}{n}\sum_{i=1}^n\int \phi_h(x-x_i)\frac{\phi_{\widehat{\sigma}}(x-\widehat{\mu})}{\phi_{\widehat{\sigma}}(x_i-\widehat{\mu})}\phi_{\sigma_j}(x-\mu_j)\,\mathrm{d}x \\
&= \frac{1}{\widehat{\sigma}}\sum_{j=1}^k p_j\widetilde{\sigma}_j\phi_{\sigma_j}(\mu_j-\widehat{\mu})\Big[\frac{1}{n}\sum_{i=1}^n\frac{\phi_h(x_i-\widehat{\mu})}{\phi_{\widehat{\sigma}}(x_i-\widehat{\mu})}\exp\Big\{\tfrac{1}{2}\widetilde{\sigma}_j^2\Big(\frac{x_i-\widehat{\mu}}{h^2}+\frac{\mu_j-\widehat{\mu}}{\sigma_j^2}\Big)^2\Big\}\Big],
\end{aligned} \tag{3.6}$$

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, and this seems forbiddingly difficult. It is however possible to find the exact MISE when the estimator employs true rather than estimated parameter values for μ and σ. This is carried out below and allows the promised best case versus best case comparison with the kernel method to be made. We note that using estimates for μ and σ only has a secondary effect on the performance on the density estimator (1.7); see (1.4) and Sections 2 and 3 in Hjort and Glad (1995), as well as Section 3.2 below.

PROPOSITION 3.1. *Let f be a normal mixture of the form (2.1), with mean $\mu_0 = \sum_{i=1}^k p_i\mu_i$ and variance $\sigma_0^2 = \sum_{i=1}^k p_i\{\sigma_i^2 + (\mu_i - \mu_0)^2\}$. Consider the special normal start times correction estimator*

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

Its exact mean integrated squared error can be expressed as

$$\text{MISE}(h) = (1 - n^{-1})\,\text{E}A_{1,h} + n^{-1}\,\text{E}A_{2,h} - 2\,\text{E}B_h + R(f), \tag{3.7}$$

where $R(f)$ is given in (3.3) and where formulae for the other three terms are

$$\begin{aligned}\text{E}A_{1,h} = \sqrt{2\pi}\sum_{i,j}\frac{p_ip_j}{b_ib_j}\phi_{\sigma_i}(\mu_i)\phi_{\sigma_j}(\mu_j)\frac{1}{c_{i,j}}\exp\Big\{&\tfrac{1}{2}\frac{d_{i,j}^2}{c_{i,j}^2} \\ &+ \tfrac{1}{2}\Big(\frac{\mu_i}{\sigma_i^2} - \frac{\mu_0}{\sigma_0^2}\Big)^2\frac{h^2}{b_i^2} + \tfrac{1}{2}\Big(\frac{\mu_j}{\sigma_j^2} - \frac{\mu_0}{\sigma_0^2}\Big)^2\frac{h^2}{b_j^2}\Big\},\end{aligned} \tag{3.8}$$

$$\text{E}A_{2,h} = \frac{h^{-1}}{\sqrt{2\pi}}\sum_{i=1}^k\frac{p_i}{\sigma_ie_ir_i}\exp\Big\{\tfrac{1}{2}\frac{s_i^2}{r_i^2} - \tfrac{1}{2}\frac{\mu_i^2}{\sigma_i^2} + \tfrac{1}{2}\Big(\frac{\mu_i}{\sigma_i^2} - 2\frac{\mu_0}{\sigma_0^2}\Big)^2\frac{h^2}{e_i^2}\Big\}, \tag{3.9}$$

$$\text{E}B_h = \sqrt{2\pi}\sum_{i,j}p_ip_j\phi_{\sigma_i}(\mu_i)\phi_{\sigma_j}(\mu_j)\frac{1}{b_i}\frac{1}{t_{i,j}}\exp\Big\{\tfrac{1}{2}\Big(\frac{\mu_i}{\sigma_i^2} - \frac{\mu_0}{\sigma_0^2}\Big)^2\frac{h^2}{b_i^2} + \tfrac{1}{2}\frac{u_{i,j}^2}{t_{i,j}^2}\Big\}, \tag{3.10}$$

and where finally b, c, d, e, r, s, t, u with sub-indices are defined in Appendix 1.

The proof of Proposition 3.1 is given in Appendix 1.

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

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

which give, together with (3.7) and (3.3), formula (3.1) for the exact MISE(h) of the kernel estimator, as required.

We use Proposition 3.1 to go through the 15 test densities of Marron and Wand (1992) again, with the natural aim of comparing the minimum possible MISE for the kernel method with the minimum possible MISE for the semiparametric method (1.7). Computer codes for formula (3.1) and the one in Proposition 3.1 allowed us to find for each mixture the minimising value of h and the resulting minimum MISE values, for each of the five sample sizes 25, 50, 100, 200, 1000. These minima, respectively $\text{MISE}^*_{\text{trad}}$ and MISE^*, along with the minimisers h^*_{trad} and h^*, are displayed in Table 3.1 below, as well as the ratio $\text{MISE}^*/\text{MISE}^*_{\text{trad}}$.

TABLE 3.1. *Values are given of the* MISE*-minimising smoothing parameters* h^* *and* h^*_{trad} *for the (1.7) estimator with true parameter values and the kernel estimator, along with the minimum* MISE *values* MISE^* *and* $\text{MISE}^*_{\text{trad}}$. *This is done for each of the 15 test densities of Marron and Wand, for sample sizes 25, 50, 100, 200, 1000. Also included in each case is the ratio* $\text{MISE}^*/\text{MISE}^*_{\text{trad}}$.

n	h^*	MISE^*	h^*_{trad}	$\text{MISE}^*_{\text{trad}}$	MISE-ratio
Case #1, normal:					
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

These numbers support the previous positive conclusions for the new estimator, in its particular form (1.7) with parameters μ_0 and σ_0. The finite-sample MISE-ratio is quite often below 1, and for the rather difficult test densities where the analysis of Section 2 gave very similar values for R_{trad} and R_{new}, Table 3.1 reveals MISE-ratios mostly between 0.99 and 1.01. Even in these highly non-normal situations the new method has, overall, a slight edge also by means of finite-sample MISE comparison.

3.2. MISE*-analysis with estimated mean and variance.* The above results concerned estimator (1.7) with known values of μ and σ. The analysis is now extended to cover the situation met in practice with estimates $\widehat{\mu}$ and $\widehat{\sigma}$. Our method is to turn to the ISE(h) expression in (3.2) and apply random sampling and the law of large numbers to obtain MISE(h).

Let f be a specific mixture, and let the sample size n be fixed. For each h' on a fine grid in a suitable interval, a data-set of size n is drawn 10000 times from the density f. For each realisation $\{x_1, \ldots, x_n\}$ the parameter estimates $\widehat{\mu}$ and $\widehat{\sigma}$ and the values of the two random functions $A_{h'}$ and $B_{h'}$ in (3.5) and (3.6) are evaluated and combined via (3.2) to give ISE(h'). The MISE(h') value for the given h' is estimated by the average of these 10000 realisations of ISE(h'). In order to obtain a precise approximation to the minimum MISE, say MISE**, and the corresponding minimising smoothing parameter h^{**}, a polynomial of degree 4 is fitted to the points $(h', \text{MISE}(h'))$ and minimised. This is done after an initial screening has provided a suitable h' interval containing the minimand. We considered exploiting symmetries in $A_{h'}$ and $B_{h'}$ to make simulation faster, that is, to work instead with reduced variables $A^*_{h'}$ and $B^*_{h'}$ with the same mean values, but then sampling variability increased so much that we chose to go back to the original $A_{h'}$ and $B_{h'}$.

Since computation times did not allow us to produce an equivalent to the whole of Table 3.1 in this way, we selected some mixtures among the test densities that on the basis of Table 3.1 were judged as respectively easy, not so easy, and difficult for the semiparametric estimator, as compared to the kernel method. We included only sample sizes 25, 50 and 100; any differences in performance for the (1.7) method, between using known and estimated parameters, are more likely to be visible for such small and moderate sample sizes, and are guaranteed to disappear as n grows. The resulting MISE** and h^{**} are given in Table 3.2, along with the ratio MISE**/MISE$^*_{\text{trad}}$. Also displayed, for convenience, are the exact minima MISE* and MISE$^*_{\text{trad}}$ for respectively the special estimator of Proposition 3.1 with known values of μ and σ and the kernel method, brought in from Table 3.1.

TABLE 3.2. *The minimum value* MISE** *of the* MISE(h) *curve for the (1.7) estimator, displayed along with the minimising value* h^{**}. *These are found by first estimating* MISE(h') *for each* h' *on a fine grid by averaging 10000 realisations of* ISE(h') *in formula (3.2), and then fitting a 4th order polynomial. The corresponding exact minimal* MISE *values for the version of (1.7) that uses known* (μ_0, σ_0) *and for the kernel estimator,* MISE* *and* MISE$^{*}_{\text{trad}}$, *are also shown, for comparison, as well as the ratio* MISE**/MISE$^{*}_{\text{trad}}$.

n	h^{**}	MISE**	MISE*	MISE$^{*}_{\text{trad}}$	MISE-ratio
Case #1, normal:					
25	0.8239	0.0127	0.0113	0.0137	0.9270
50	0.7247	0.0063	0.0056	0.0087	0.7241
100	0.7344	0.0032	0.0028	0.0054	0.5926
Case #2, skewed unimodal:					
25	0.4626	0.0216	0.0228	0.0211	1.0237
50	0.4008	0.0121	0.0123	0.0134	0.9030
100	0.3618	0.0069	0.0068	0.0083	0.8313
Case #8, skewed bimodal:					
25	0.4884	0.0255	0.0243	0.0222	1.1486
50	0.3802	0.0161	0.0158	0.0151	1.0662
100	0.3157	0.0100	0.0100	0.0097	1.0309
Case #10, claw:					
25	0.5421	0.0676	0.0659	0.0206	1.0629
50	0.4162	0.0586	0.0578	0.0570	1.0281
100	0.1011	0.0372	0.0371	0.0370	1.0054

The first conclusion to be drawn from the MISE-ratios in Table 3.2 is, in agreement with previous statements, that the normal-started semiparametric (1.7) estimator has a better finite-sample performance than the kernel estimator (in terms of MISE) for observations originating from densities that are normal or not too far from normal (#1, #2). Furthermore, the performance is not much violated by the normal start even for seriously non-normal densities (#8, #10).

Moreover, it is interesting to compare the MISE** values with the MISE* values to understand the effect of parameter estimation on the MISE performance. The MISE** values for the normal density (#1) agree very well with the theoretical approximation $\frac{9}{8}(2n\sqrt{\pi})^{-1}$ obtained in Remark 8C under parametric home-turf conditions in Hjort and Glad (1995). As expected, for the normal density itself (#1), the minimum MISE values in Table 3.2 are somewhat larger with estimated μ and σ than without, for small sample sizes n. For the skewed unimodal density (#2) and n small, on the other hand, the minimum MISE is clearly *reduced* by using sample estimates instead of μ_0 and σ_0. In other words, there are cases where the real situation is more favourable than what Table 3.1 suggests. For the other densities, there is only a very slight increase, or none at all, of minimum MISE when passing from known to estimated (μ, σ). Hence the performance of the (1.7) estimator appears praiseworthily stable with respect to parameter estimation even for small sample sizes n, and this stability should carry over to other parameter estimates for (μ, σ) too, like robust alternatives. The complete Table 3.1, with its reliance on Proposition 3.1 about performance of the (μ_0, σ_0)-based estimator, should therefore be considered as presenting good approximations to the corresponding numbers for the bona fide $(\widehat{\mu}, \widehat{\sigma})$-based estimator (1.7). Finally we underline the

strong agreement between results about MISE-ratios obtained through asymptotic analysis and through finite-sample studies in this case, even for sample sizes as small as 25 and 50.

It should be kept in mind here that the list of 15 test densities is not at all constructed to be favourable to using the normal model as start description. For most of them, a more careful choice of parametric model would be likely to have improved the performance additionally. We do believe, however, that the semiparametric density estimator with the simple normal start (1.7) will perform better than the kernel method in many applications, since, statistically speaking, a high proportion of densities actually encountered in real applications are much closer to the normal than each of the test cases #3–#15.

Appendix 1: Proof of Proposition 3.1. 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

$$\mathrm{E}\widehat{f}(x) = \frac{1}{\sqrt{2\pi}}\sum_{i=1}^{k} p_i \frac{1}{\sigma_i b_i}\exp\Big\{-\tfrac12\frac{(x-\mu_i)^2}{\sigma_i^2}+\tfrac12\Big(\frac{x-\mu_i}{\sigma_i^2}-\frac{x-\mu_0}{\sigma_0^2}\Big)^2\frac{h^2}{b_i^2}\Big\}.$$

Indeed this is $f(x)+O(h^2)$. Next consider A_h of (3.2) 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,

$$\begin{aligned}\mathrm{E}A_{1,h} &= \mathrm{E}\int\Big\{\frac{\phi_h(x-X_1)\phi_{\sigma_0}(x-\mu_0)}{\phi_{\sigma_0}(X_1-\mu_0)}\frac{\phi_h(x-X_2)\phi_{\sigma_0}(x-\mu_0)}{\phi_{\sigma_0}(X_2-\mu_0)}\Big\}\,\mathrm{d}x \\ &= \int\{\mathrm{E}\widehat{f}(x)\}^2\,\mathrm{d}x \\ &= \frac{1}{2\pi}\sum_{i,j} p_i p_j \frac{1}{\sigma_i\sigma_j b_i b_j}\int\exp\Big\{-\tfrac12\frac{(x-\mu_i)^2}{\sigma_i^2}-\tfrac12\frac{(x-\mu_j)^2}{\sigma_j^2} \\ &\qquad +\tfrac12\Big(\frac{x-\mu_i}{\sigma_i^2}-\frac{x-\mu_0}{\sigma_0^2}\Big)^2\frac{h^2}{b_i^2}+\tfrac12\Big(\frac{x-\mu_j}{\sigma_j^2}-\frac{x-\mu_0}{\sigma_0^2}\Big)^2\frac{h^2}{b_j^2}\Big\}.\end{aligned}$$

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

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

the result is exactly (3.8).

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

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

to reach the answer (3.9), which 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

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

This time we need

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

and the result is (3.10). This ends our proof. □

Acknowledgements. The authors thank M.C. Jones for helpful comments. Ingrid K. Glad is supported by a grant from the Norwegian Council of Research and a NATO Science Fellowship. She gratefully acknowledges the hospitality of MeMoMat, University of Rome La Sapienza.

References

Efron, B., and Tibshirani, R.J. (1995). Using specially designed exponential families for density estimation. Technical report, Dept. of Statistics, Stanford University.

Fenstad, G.U., and Hjort, N.L. (1996). Comparison of two Hermite expansion density estimators with the kernel method. Manuscript in progress.

Glad, I.K. (1995). Parametrically guided nonparametric regression. In Ph.D. Thesis **1995: 59**, Norwegian Institute of Technology.

Hjort, N.L. (1996). Performance of Efron and Tibshirani's semiparametric density estimator. Submitted for publication.

Hjort, N.L. and Glad, I.K. (1995). Nonparametric density estimation with a parametric start. *Annals of Statistics* **23**, 882–904.

Hjort, N.L. and Jones, M.C. (1996). Better rules of thumb for choice of smoothing parameter in density estimation. Manuscript in progress.

Jones, M.C. and Signorini, D.F. (1996). A comparison of higher order bias kernel density estimators. *Journal of the American Statistical Association*, to appear.

Marron, J.S. and Tsybakov, A.B. (1995). Visual error criteria for qualitative smoothing. *Journal of the American Statistical Association* **90**, 499–507.

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

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

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

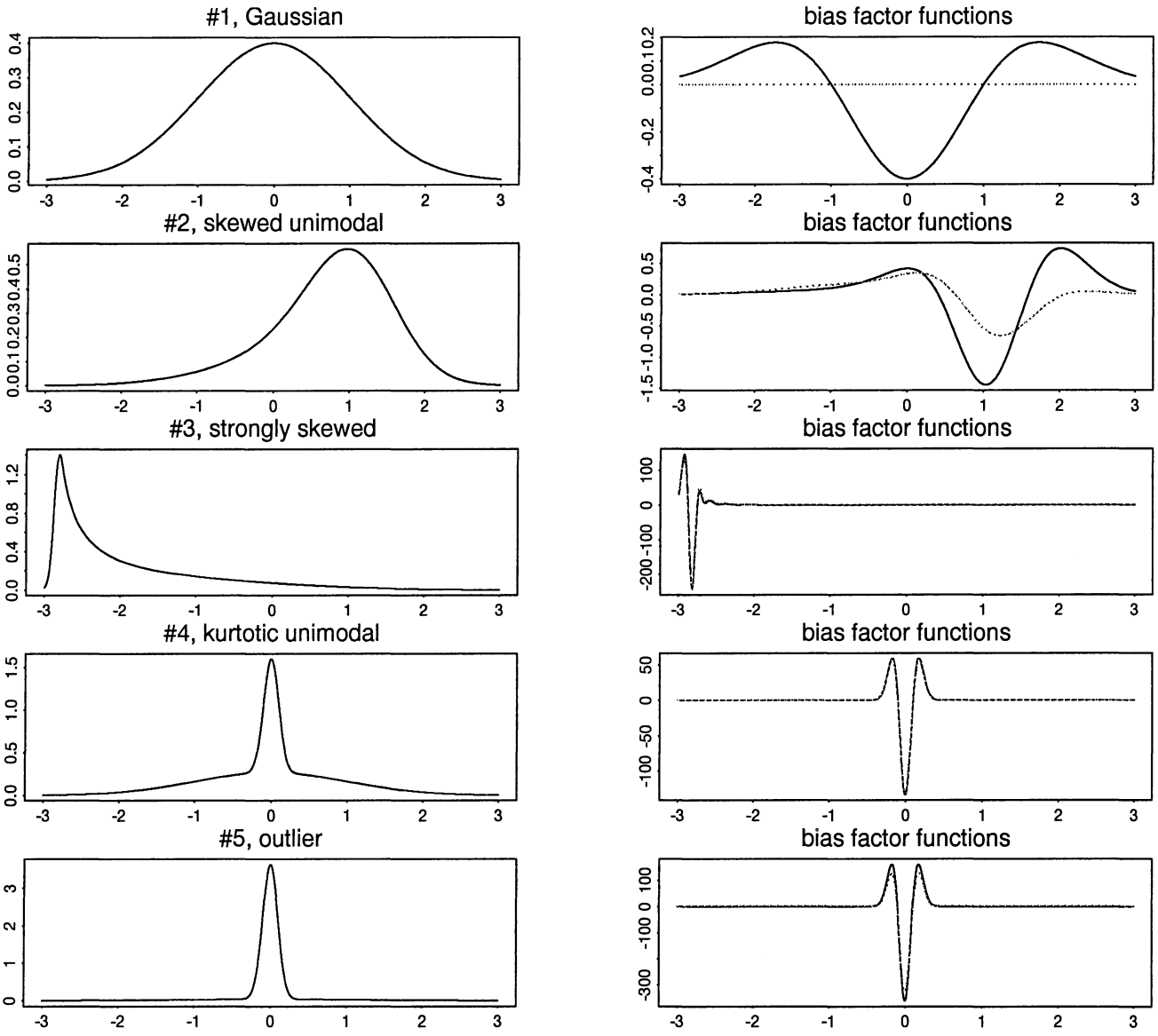
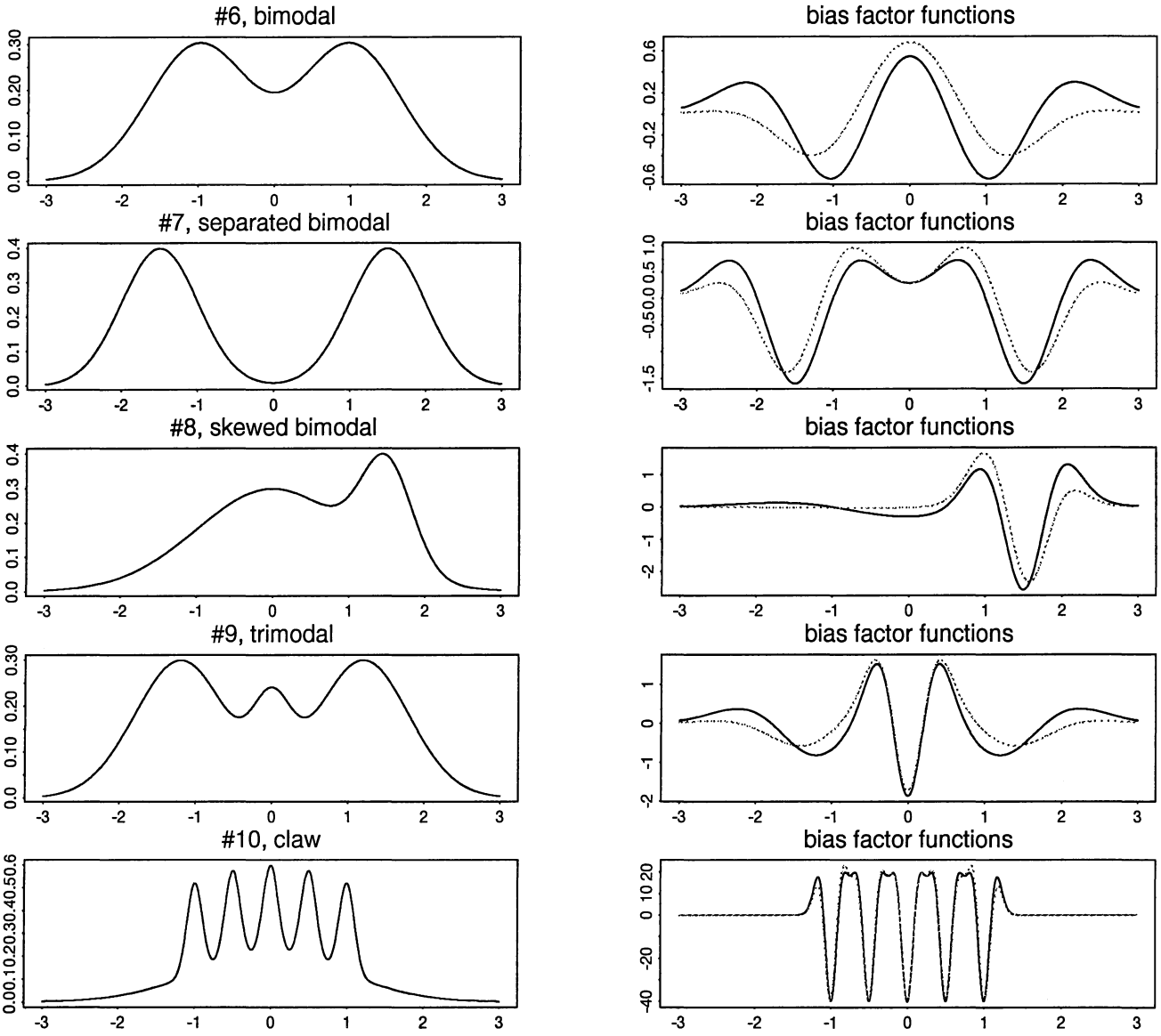
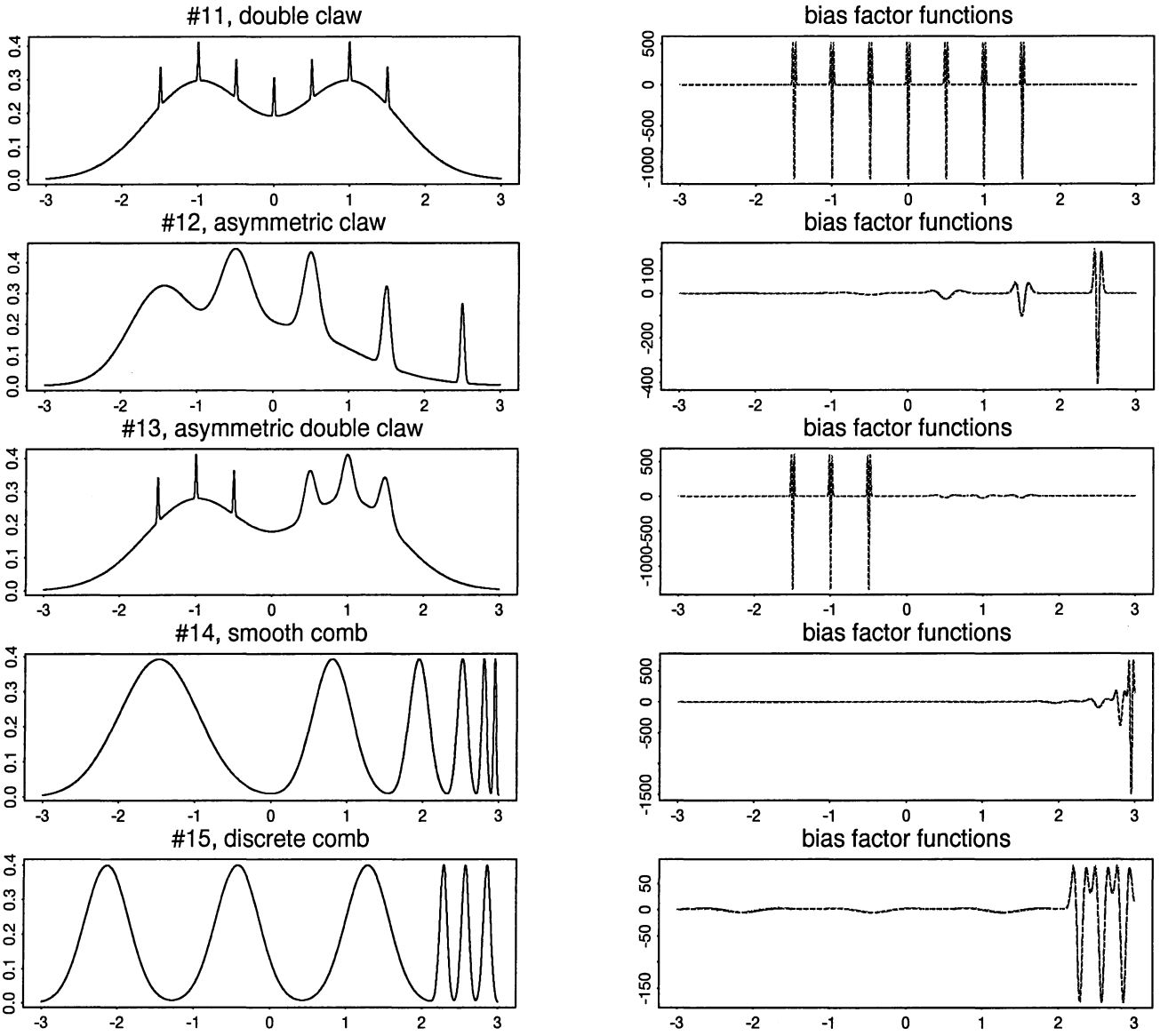


FIGURE 1: The 15 test densities (left hand side) presented together with the bias factor functions f'' (solid line, for the kernel method) and $f_0 r''$ (dotted line, for the new method).


#6, bimodal
bias factor functions
#7, separated bimodal
bias factor functions
#8, skewed bimodal
bias factor functions
#9, trimodal
bias factor functions
#10, claw
bias factor functions


#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