**Nonparametric Density Estimation with a Parametric Start**

Nils Lid Hjort; Ingrid K. Glad

*The Annals of Statistics*, Vol. 23, No. 3 (Jun., 1995), 882-904.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199506%2923%3A3%3C882%3ANDEWAP%3E2.0.CO%3B2-E

*The Annals of Statistics* is currently published by Institute of Mathematical Statistics.

---



---

*The Annals of Statistics*
1995, Vol. 23, No. 3, 882–904

# NONPARAMETRIC DENSITY ESTIMATION WITH A PARAMETRIC START

BY NILS LID HJORT AND INGRID K. GLAD

*University of Oslo and Norwegian Institute of Technology*

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.

**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

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

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 $\sigma_K^2 = \int z^2K(z)\,dz$ and $R(K) = \int K(z)^2\,dz$. The basic statistical properties are that

$$\begin{aligned} &\mathbb{E}\tilde{f}(x) \doteq f(x) + \tfrac{1}{2}\sigma_K^2 h^2 f''(x) \quad \text{and} \\ &\operatorname{Var}\tilde{f}(x) \doteq R(K)(nh)^{-1}f(x) - \frac{f(x)^2}{n}. \end{aligned} \tag{1.2}$$

The integrated mean squared error (MISE) 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 and Jones [(1995), Chapter 2] for recent accounts of the theory.

Received February 1994; revised October 1994.
*AMS* 1991 *subject classifications.* Primary 62G07; secondary 62G20.
*Key words and phrases.* Bandwidth selection, correction factor, kernel methods, lowering the bias, semiparametric density estimation, test cases.

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 $f(x, \hat{\theta})$, say, the normal, and then multiply with a nonparametric kernel-type estimate of the correction function $r(x) = f(x)/f(x, \hat{\theta})$. Our proposal is $\hat{r}(x) = n^{-1}\sum_{i=1}^{n} K_h(X_i - x)/f(X_i, \hat{\theta})$, producing

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

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 starting 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 nonrandom starting 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. We have developed comparative formulae for exact analysis of asymptotic mean integrated squared error as well as for exact finite-sample mean integrated squared error. The results are illuminated by working through a list of 15 "test densities" proposed by Marron and Wand (1992), chosen to exhibit a broad range of distributional shapes. Because of space considerations the rather long technical calculations as well as several comparative tables are not included here; these are available in the technical report by Hjort and Glad (1994). The results are, however, briefly reported on in Section 5. The new "nonparametrically corrected normal estimator" outperforms the usual kernel method in 12 of these 15 test cases, and in all the "not drastically unreasonable" cases, in terms of approximate in 12 of these 15 test cases, and in all the "not drastically unreasonable" cases, in terms of approximate mean integrated squared error (AMISE). 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 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 nonnormal 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 multidimensional case, beginning, for example, with a multinormal-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, Section 8.4 spells out the corresponding estimation idea for nonparametric regression, giving (for example) a generalised Nadaraya–Watson estimator.

Our estimators can be viewed as semiparametric in that they combine parametric and nonparametric methods. As such, they are in the same realm as recent methods of Hjort (1995a) and Hjort and Jones (1995a). 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, e.g., Friedman, Stuetzle and Schroeder (1984)] and also to the normal times Hermite expansion method [see, e.g., Hjort (1986), Buckland (1992) and Fenstad and Hjort (1995)]. A somewhat less attractive semiparametric method is that of Schuster and Yakowitz (1985) and Olkin and Spiegelman (1987) [see the discussion in Jones (1993)]. Various semiparametric Bayesian density estimators are proposed in Hjort (1995b).

Another semiparametric technique, perhaps mildly related to our new method, is the transformation idea of Wand, Marron and Ruppert (1991), where data are semiparametrically transformed so as to work well with a nonadaptive 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 having similar intentions. In other words, (1.3) can be seen as being similar in spirit to a suitable semiparametrically adaptive $n^{-1}\sum_{i=1}^{n} K_{h(x,\hat{\theta})}(X_i - x)$. Finally, we mention a recent bias-reduction method due to Jones, Linton and Nielsen (1995). 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 and Nielsen (1995) 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.

Finally we mention the recent work of Efron and Tibshirani (1995), who exploit Poisson regression techniques to put specially designed parametric families "through" the kernel estimator. Efron and Tibshirani also discuss the relations between their methods and those presented here.

**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 $\hat{r}(x) = n^{-1}\Sigma_{=1}^{n} K_h(X_i - x)/f_0(X_i)$, with ensuing estimator

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

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

$$\begin{aligned}\mathbb{E}\hat{r}(x) &= \int K_h(y - x) f_0(y)^{-1} f(y)\, dy \\ &= \int K(z) r(x + hz)\, dz = r(x) + \tfrac{1}{2}\sigma_K^2 h^2 r''(x) + O(h^4)\end{aligned}$$

and

$$\begin{aligned}\operatorname{Var}\hat{r}(x) &= \frac{1}{n}\left[\int \frac{K_h(y - x)^2}{f_0(y)^2} f(y)\, dy - \{\mathbb{E}\hat{r}(x)\}^2\right] \\ &= \frac{R(K)}{nh}\frac{f(x)}{f_0(x)^2} - \frac{r(x)^2}{n} + O\left(\frac{h}{n}\right),\end{aligned}$$

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

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

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 multidimensional parameter $\theta = (\theta_1, \ldots, \theta_p)'$ belongs to some open and connected region in $p$-space. The parametric-start estimate is $f(x, \hat{\theta})$, where for concreteness we

let $\hat{\theta}$ be the maximum likelihood estimator (quite general estimators for $\theta$ are allowed later). Thus $f(x, \hat{\theta})$ 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 $f(\cdot, \theta)$.

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

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

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 $\theta_0$, 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 $f(\cdot, \theta)$. 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

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

leading to

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

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 $\hat{\theta} - \theta_0$ 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 $\hat{f}(x)$. 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 $\theta$ of the form $\hat{\theta} = T(F_n)$ with influence

function $I(x) = \lim_{\varepsilon \to 0}\{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)\}/\varepsilon$, writing $\delta_x$ for unit point mass at $x$, and we assume that $\Sigma = \mathbb{E}_f I(X_i)I(X_i)'$ is finite. The best approximant $f_0(x) = f(x, \theta_0)$ to $f(x)$ that $f(x, \hat{\theta})$ aims for is determined by $\theta_0 = T(F)$. Under mild regularity conditions [see, e.g., Huber (1981) or Shao (1991)], one has

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

where $\varepsilon_n = O_p(n^{-1})$ with mean $O(n^{-2})$, that is, $d/n$ is essentially the bias of $\hat{\theta}$. It is generally possible to debias 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 $J = -\mathbb{E}_f\, \partial^2 \log f(X_i, \theta_0)/\partial\theta\, \partial\theta'$.

PROPOSITION 1. *Let $f_0(x) = f(x, \theta_0)$, with $\theta_0 = T(F)$, be the best parametric approximant to $f$, and let $r = f/f_0$. As $n \to \infty$ and $h \to 0$, the semiparametric estimator* (3.1) *has*

$$\mathbb{E}\hat{f}(x) = f(x) + \frac{1}{2}\sigma_K^2 h^2 f_0(x) r''(x) + O\left(\frac{h^2}{n} + h^4 + n^{-2}\right),$$

$$\operatorname{Var}\hat{f}(x) = R(K)(nh)^{-1} f(x) - \frac{f(x)^2}{n} + O\left(\frac{h}{n} + n^{-2}\right).$$

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

$$A_i = K_h(X_i - x)\frac{f_0(x)}{f_0(X_i)},$$

$$B_i = -K_h(X_i - x)\frac{f_0(x)}{f_0(X_i)}\{u_0(X_i) - u_0(x)\},$$

$$C_i = K_h(X_i - x)\frac{f_0(x)}{f_0(X_i)}w(x, X_i),$$

where in fact $w(x, X_i) = v_0(x) - v_0(X_i) + \{u_0(x) - u_0(X_i)\}\{u_0(x) - u_0(X_i)\}'$ and $v_0(x) = \partial^2 \log f(x, \theta_0)/\partial\theta\, \partial\theta$.

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-representa-

tions above, one finds $\mathbb{E}V_n(x) = n^{-1}\mathbb{E}B_i'I_i + n^{-1}(\mathbb{E}B_i)'d + O(n^{-2})$ and $\mathbb{E}W_n(x) = n^{-1}\,\mathrm{Tr}(\mathbb{E}C_i\mathbb{E}I_iI_i') + O(n^{-2})$, using the fact that $I_i = I(X_i)$ has mean zero. However, it is not difficult to see that each of $\mathbb{E}B_i$, $\mathbb{E}B_i'I_i$ and $\mathbb{E}C_i$ is of size $O(h^2)$; for example,

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

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

Next we turn to the variance. The variance of $f^*(x)$ is known from Section 2. From (3.4) and the representation above, one finds $\mathrm{Var}\,V_n(x) = \mathrm{Var}(\bar B_n'\bar I_n) + O(n^{-2}) = n^{-1}(\mathbb{E}B_i)'\Sigma(\mathbb{E}B_i) - \{O(h^2/n)\}^2 + O(n^{-2}) = O(h^4/n + n^{-2})$, and similarly $W_n(x)$ can be seen to have uninfluential variance $O(h^4/n^2)$. Finally, $\mathrm{cov}\{f^*(x), V_n(x)\} = n^{-1}(\mathbb{E}B_i)'\mathbb{E}A_iI_i + O(n^{-2}) = O(h^2/n)$. 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 $\hat\theta$ close to $\theta_0$, but the $\hat f(x)$ estimator uses only $X_i$'s that are close to $x$, making $u_0(X_i)$ close to $u_0(x)$. The story is somewhat different for the correction term $\hat r(x)$ alone (see Section 8.2).

Consistency of the density estimator requires both $h \to 0$ (forcing the bias toward 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 $O(h^2/n + 1/n^2)$.

EXAMPLE 1 (Normal-start estimate). The normal-start estimate is $\hat\sigma^{-1}\phi(\hat\sigma^{-1}(x-\hat\mu))$, where one can use maximum likelihood estimates $\hat\mu = n^{-1}\sum_{i=1}^n X_i$ and $\hat\sigma^2 = n^{-1}\sum_{i=1}^n(X_i-\hat\mu)^2$ (or the debiased version with denominator $n-1$). In view of the generality of Proposition 1, quite general estimators are allowed, without changing the basic structure of bias and variance of $\hat f(x)$. One might, for example, wish to use robust estimates of

mean and standard deviation. In any case the density estimator is

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

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

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

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

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

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 $f(x,\hat{\theta})$, correcting afterward with a $\hat{r}(x)$.

REMARK 1. The correction factor $f(x,\hat{\theta})/f(X_i,\hat{\theta})$ can occasionally be too influential, in cases where the denominator is too small. This is not a problem for small $h$ since then only $X_i$'s quite close to $x$ contribute to the estimate at that point. The problem might appear in cases where $h$ is of a size that allows $X_i$'s some distance from $x$ to have significant weights, and such an $X_i$ is in the very tail of $f(x,\hat{\theta})$ while $x$ is not. In our experience such situations hardly ever occur for sizes of $h$ corresponding to the selection procedures outlined in Section 6. One might for safety apply a "clipping" procedure similar in spirit to precautions recommended for variable kernel density

estimators [see Abramson (1982) and Terrell and Scott (1992)]. Our suggestion is to truncate suitably, using, for example,

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\text{trunc}\left\{\frac{f(x,\hat{\theta})}{f(X_i,\hat{\theta})}, \left[\frac{1}{10}, 10\right]\right\}, \tag{3.6}$$

clipping below $\frac{1}{10}$ and above 10; this effectively alleviates the problem. Some analysis shows that in the case of (3.5), truncation is almost never necessary when $K$ has bounded support.

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.

REMARK 3. Our estimator does not integrate to precisely 1, but postnormalisation may of course be carried out. For the case of (3.5) the integral can be shown to be of order $1 + \frac{1}{8}\hat{\gamma}_4 h^4/\hat{\sigma}^4$, where $\hat{\gamma}_4$ is the empirical kurtosis [see Hjort and Glad (1994), Remark 8C, for further comments].

**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 Section 4.2 we study 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 fuller detail in Hjort and Glad (1994)] with attention given to the list of 15 test densities chosen by Marron and Wand (1992).

4.1. *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 Proposition 1 (Section 3). We find the following:

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

featuring "roughness" functionals

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

The new estimator is better, in the sense of approximate (leading terms) integrated mean squared error, whenever $R_{\text{new}}(f)$ is smaller than $R_{\text{trad}}(f)$. This defines a nonparametric neighbourhood of densities around the parametric class. When $f$ belongs to this neighbourhood, $\hat{f}$ is better than $\tilde{f}$ 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

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

This is useful for actual inspection of the bias terms for different $f$'s 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 $x$'s, 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 $0 \le g_0''/g'' \le 2$.

4.2. *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 and Jones (1995b) and in Fenstad and Hjort (1995). Let $H_j(x)$ be the $j$th Hermite polynomial, given by $\phi^{(j)}(x) = (-1)^j \phi(x) H_j(x)$. We shall in fact consider two different expansions. The first uses the representation

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

Its mean is $\mu$ and its standard deviation is $\sigma$, and $\gamma_j = \mathbb{E} H_j((X-\mu)/\sigma)$. Note that $\gamma_0 = 1$ and that $\gamma_1 = \gamma_2 = 0$, while

$$\gamma_3 = \frac{\mathbb{E}(X-\mu)^3}{\sigma^3}, \qquad \gamma_4 = \frac{\mathbb{E}(X-\mu)^4}{\sigma^4} - 3,$$

$$\gamma_5 = \frac{\mathbb{E}(X-\mu)^5}{\sigma^5} - 10\mathbb{E}(X-\mu)^3,$$

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 (4.4), through inclusion of enough terms. See Hjort and Jones (1995b) for details pertaining to this and some of the following calculations.

The Hermite expansion (4.4) is of the type encountered in Edgeworth–Cramér expansions. It is pleasing from a theoretical point of view in that it incorporates skewness, kurtosis and so on to refine the normal approximation, but it has shortcomings as well. The coefficients are not always finite and empirical estimates are quite variable and nonrobust. Hjort

and Jones (1995b) and Fenstad and Hjort (1995) give further reasons favouring a second and more robust Hermite expansion, in terms of the polynomials $H_j^*(y) = H_j(\sqrt{2}\,y)$ instead. In this case,

$$f(x) = \phi\Big(\frac{x-\mu}{\sigma}\Big)\frac{1}{\sigma}\sum_{j=0}^{m}\frac{\delta_j}{j!}H_j\Big(\sqrt{2}\,\frac{x-\mu}{\sigma}\Big), \tag{4.5}$$

where the coefficients are determined from $\delta_j = \sqrt{2}\,\mathbb{E}H_j(\sqrt{2}(X-\mu)/\sigma)\exp\{-\frac{1}{2}(X-\mu)^2/\sigma^2\}$. If $f$ is taken as an approximation to a given density $q$ with mean $\mu$ and standard deviation $\sigma$, then the $L_2$-distance $\int(f-q)^2\,dx$ is minimised for exactly these $\delta_j$ [see Hjort and Jones (1995b)].

One may now compare the large-sample behaviour of the traditional and the new normal-corrected estimator (3.5), in situations where the true $f$ is as in (4.4) or (4.5). Here $f = f_0 r_1$ in the first case and $f = f_0 r_2$ in the second case, with $f_0$ being the simple normal approximation, and

$$r_1(x) = 1 + \sum_{j=3}^{m}\frac{\gamma_j}{j!}H_j(y) \quad\text{and}\quad r_2(x) = \sum_{j=0}^{m}\frac{\delta_j}{j!}H_j(\sqrt{2}\,y),$$

writing $y = (x-\mu)/\sigma$. Neighbourhoods around the normal density are described by looking at variations of the $\gamma_j$'s and $\delta_j$'s around 0 (for $j \geq 1$). Expressions for $f''$ and $f_0 r''$ can now be obtained and compared, for each of the two expansions. Experience from this exercise shows, broadly speaking, that the leading bias term for the new estimator is indeed smaller than the leading bias term for the kernel estimator, for a broad range of central $x$-values, provided the true $f$ is in a reasonably sized neighbourhood around the normal. More information is in Hjort and Glad [(1994), Section 4].

It is also fruitful to derive expressions for the roughness quantities (4.2). Because of space considerations we merely give the formulae for the direct Hermite expansion (4.4); a fuller account is in Hjort and Glad [(1994), Section 4]. One finds

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

and

$$f_0(x)r_1''(x) = \sigma^{-3}\phi(y)\sum_{j=2}^{m}j(j-1)\frac{\gamma_j}{j!}H_{j-2}(y),$$

again with $y = (x-\mu)/\sigma$. Some calculations give that $A_{j,k} = \int H_j H_k \phi^2\,dy$ 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 and Jones (1995b)]. This makes it possible to evaluate

$$R_{\mathrm{trad}}(f) = \frac{1}{\sigma^5}\sum_{j,k\leq m}\frac{\gamma_j}{j!}\frac{\gamma_k}{k!}A_{j+2,k+2}$$

and

$$R_{\text{new}}(f) = \frac{1}{\sigma^5} \sum_{2 \le j,\, k \le m} \frac{\gamma_j}{(j-2)!} \frac{\gamma_k}{(k-2)!} A_{j-2,\, k-2}$$

for given values of $m$. As an example, suppose terms corresponding to skewness, kurtosis and pentakosis are included. Then, per similar calculations of Hjort and Jones (1995b),

$$\begin{aligned} R_{\text{trad}} &= \sigma^{-5}(3/8\sqrt{\pi})\big(1 + \tfrac{35}{48}\gamma_4 + \tfrac{35}{32}\gamma_3^2 + \tfrac{385}{1024}\gamma_4^2 + \tfrac{1001}{10{,}240}\gamma_5^2 - \tfrac{77}{128}\gamma_3\gamma_5\big), \\ R_{\text{new}} &= \sigma^{-5}(3/8\sqrt{\pi})\big(\tfrac{2}{3}\gamma_3^2 + \tfrac{1}{4}\gamma_4^2 + \tfrac{5}{72}\gamma_5^2 - \tfrac{1}{3}\gamma_3\gamma_5\big). \end{aligned} \tag{4.6}$$

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

**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 forms a very wide and flexible class of densities. Marron and 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. We have used these as well as further mixtures to compare the new normal-start times correction method with the traditional kernel method. In Section 5.1 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 Section 5.2 we go further and analyse exact finite-sample mean squared errors for the two methods.

5.1. *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

$$r'' = \sum_{i=1}^{k} p_i \exp(g_i - g_0)\big\{g_i'' - g_0'' + (g_i' - g_0')^2\big\}.$$

This leads to

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

while

$$f''(x) = \sum_{i=1}^{k} p_i \phi''_{\sigma_i}(x - \mu_i) = \sum_{i=1}^{k} p_i \left\{ \frac{(x-\mu_i)^2}{\sigma_i^2} - 1 \right\} \frac{f_i(x)}{\sigma_i^2}. \tag{5.3}$$

We used these formulae for visually inspecting $f''$ versus $f_0 r''$ in a number of cases, including the 15 test cases of Marron and Wand (1992); see Hjort and Glad [(1994), Figure 1]. 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.

With some effort (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 Hjort and Glad [(1994), Proposition A.1]. These global criteria have been compared for a collection of normal mixtures. The overall comparison in terms of approximate MISE is in clear favour of the new method. Hjort and Glad [(1994) Table A.1] present appropriately transformed versions of these, for the 15 test cases. The new method has $R_{\text{new}} < R_{\text{trad}}$ in 12 of the 15 cases and loses to the kernel method, and then only very slightly, in the quite extreme cases #12 (the asymmetric claw), #14 (the smooth comb) and #15 (the discrete comb). 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 $R_{\text{new}}$ and $R_{\text{trad}}$. The same picture emerges also when one computes values for the $L_1$-based criteria $\int |f''|$ versus $\int |f_0 r''|$, also given (appropriately transformed to ease interpretation) in Hjort and Glad [(1994), Table A.1]. 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. 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 nonnormal situations.

5.2. *Exact finite-sample comparison.* The comparison analysis above was in terms of the Taylor-based approximations to bias and variance. We have also ventured further and analysed exact finite-sample MISE for the two methods. Such analysis was carried out in Marron and Wand (1992) for the kernel method (1.1). Their Theorem 2.1 gives a formula for $\text{MISE}(h) = \mathbb{E} \int (\tilde{f} - f)^2 \, dx$ for $f$ of the form (5.1). Reaching a similar result for the MISE of the normal-start estimator (3.5) is much more demanding. Proposition A.2 in Hjort and Glad (1994) 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, $\text{MISE}^*_{\text{trad}}$ and $\text{MISE}^*$. We pro-

grammed the two MISE formulae and went through a collection of normal mixtures, including once more the list of 15 test densities, 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 findings were summarised in Hjort and Glad [(1994), Table A.2] featuring also the ratio $\text{MISE}^*/\text{MISE}^*_{\text{trad}}$. These numbers supported 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 summarised in Section 5.1 gave very similar values for $R_{\text{trad}}$ and $R_{\text{new}}$, the table gave MISE-ratios mostly between 0.99 and 1.01. Even in these highly nonnormal situations the new method has, overall, a slight edge. The findings also indicated that choosing the same bandwidth for the new method as for the kernel method will be quite acceptable in most of the definitely nonnormal 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 starting 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).

6.1. *Minimising AMISE.* From (4.1) it is seen that the $h$-parameter minimising approximate integrated mean squared error for $\hat{f}$ is

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

The resulting minimal AMISE is $\frac{5}{4}\{\sigma_K R(K)\}^{4/5} R_{\text{new}}^{1/5} n^{-4/5}$. The same $\{\sigma_K R(K)\}^{4/5}$ factor appears also in a similar expression for the theoretically best pointwise 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, e.g., Scott (1992), Chapter 6]. 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 $R_{\text{new}}$ 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 $R_{\text{new}}$ available in Hjort and Glad [(1994), Appendix] 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 $\gamma_3$, $\gamma_4$ and $\gamma_5$. This leads to an estimate of $R_{\text{new}}$ via formula (4.6), and in the case of the normal kernel $K = \phi$ to a practical formula of the form

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

Details are furnished in Hjort and Glad [(1994), Sections 4 and 6]. 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 and Jones (1995b). In any case (6.2) may be somewhat unstable, particularly for small to moderate sample sizes, since the $\hat{\gamma}_j$ statistics are unstable. The alternative robust Hermite expansion described in Section 4.2 should be safer, using (4.5) as the point of departure rather than (4.4). Hjort and Glad (1994) show that $R_{\text{new}} = \sigma^{-5}(2/\sqrt{\pi})\sum_{j=0}^{m-2}\sigma_{j+2}^2/j!$, a formula simpler than the analogous (4.6). When the standard normal kernel is used this leads to using

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

for example, featuring the automatically robust estimates

$$\hat{\delta}_j = \frac{1}{n}\sum_{i=1}^{n}\sqrt{2}\,H_j\left(\sqrt{2}\,\frac{X_i - \hat{\mu}}{\hat{\sigma}}\right)\exp\left\{-\frac{1}{2}\left(\frac{X_i - \hat{\mu}}{\hat{\sigma}}\right)^2\right\}$$

(the summands are bounded in $X_i$). Again bias should ideally be deducted when plugging in squared estimates. See analogous comments in Hjort and Jones (1995b).

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

$$\begin{aligned}\hat{R}_{\text{new}} &= \int\left\{f(x,\hat{\theta})\hat{r}''(x)\right\}^2 dz \\ &= \frac{1}{n^2}\frac{1}{h^6}\sum_{i,j}\int\frac{f(x,\hat{\theta})}{f(X_i,\hat{\theta})}\frac{f(x,\hat{\theta})}{f(X_j,\hat{\theta})}K''\left(h^{-1}(x - X_i)\right)K''\left(h^{-1}(x - X_j)\right)dx\end{aligned}$$

as its starting point. Explicit expressions for the integral here can be worked out for most choices of $K$ [see again the Appendix of Hjort and Glad (1994)]. Somewhat lengthy calculations, involving Taylor series expansions and other techniques, can be furnished to reach

$$\mathbb{E}\hat{R}_{\text{new}} = \frac{n-1}{n}\int(f_0 r'')^2\,dx + \frac{1}{nh^5}\left\{R(K'') + O(h^2)\right\},$$

where $R(K'') = \int (K'')^2\,dz$. See Hjort and Glad [(1994), Section 6] for the details. 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 and Terrell (1987)]. This invites

$$\frac{n}{n-1}\left\{\hat{R}_{\text{new}} - \frac{R(K'')}{nh^5}\right\}$$

to be used as a corrected estimate. One version of the plug-in method is therefore as follows: select a starting value for $h$ in a reasonable way, perhaps using (6.3). Then compute $\hat{R}_{\text{new}}$ and its debiased 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 endpoints of its support; in particular the Yepanechnikov kernel is not allowed in this operation.

6.2. *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

$$\text{(6.4)}\quad \widehat{\text{AMISE}}(h) = \text{bcv}(h) = \frac{1}{4}\sigma_K^4 h^4\left\{\hat{R}_{\text{new}}(h) - \frac{R(K'')}{nh^5}\right\} + \frac{R(K)}{nh},$$

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 $h_{\text{os}}$, the "oversmoothing" bandwidth. Scott and 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 formulawise similarity to unbiased cross-validation (see below) and the desire to estimate the biased approximation AMISE to the true MISE.

6.3. *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 $\text{MISE}(h) - R(f) = \mathbb{E}\{\int \hat{f}^2\,dx - 2\int \hat{f}f\,dx\}$ with

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

Here $h$ is included in the notation, for clarity, and $\hat{f}_{h,(i)}$ is the estimator

constructed from the diminished data set that excludes $X_i$. The function to compute is

$$\frac{1}{n^2}\sum_{i,j}\frac{1}{f(X_i,\hat\theta)f(X_j,\hat\theta)}\int f(x,\hat\theta)^2K_h(x-X_i)K_h(x-X_j)\,dx$$
$$-\frac{2}{n(n-1)}\sum_{i,j}K_h(X_i-X_j)\frac{f(X_i,\hat\theta_{(i)})}{f(X_j,\hat\theta_{(j)})},$$

where $\hat\theta_{(i)}$ 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 Hjort and Glad [(1994), Appendix].

It turns out that ucv($h$) is nearly but not exactly unbiased for MISE($h$) − $R(f)$, as discussed in suitable detail in Hjort and Glad [(1994), Section 6]. The difference is minuscule, however, and choosing $h$ to minimise the UCV($h$) function, among $h\le h_{\rm os}$ for a suitable oversmoothing upper limit, remains a useful and honestly nonparametric option.

6.4. *Other techniques.* Other techniques can also be proposed, for example, trying to adapt recent methods of Sheather and Jones (1991) and of Hall, Sheather, Jones and 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 multidimensional case.** Standard nonparametric methods like the kernel method have severe difficulties in the vector case, due to the curse of dimensionality. Our multiplicative correction factor method is easy to implement also in the vector case, as is now briefly indicated. We speculate that its potential for improving on standard nonparametric methods is larger in higher dimensions, through the use of a reasonable parametric-start description.

The setting is that $d$-dimensional i.i.d. vectors $\mathbf{X}_1,\dots,\mathbf{X}_n$ are observed from a density $f$. The traditional kernel estimator uses a kernel density function $K(z_1,\dots,z_d)$, 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 $\mathbf{x}=(x_1,\dots,x_d)'$ is $\tilde f(\mathbf{x})=n^{-1}\sum_{i=1}^n K_{\mathbf{h}}(\mathbf{X}_i-\mathbf{x})$, where $K_{\mathbf{h}}(z_1,\dots,z_d)=(h_1\cdots h_d)^{-1}\times K(h_1^{-1}z_1,\dots,h_d^{-1}z_d)$ [see, e.g., Scott (1992), Chapter 6, or Wand and Jones (1995), Chapter 4]. Our parametric start with a multiplicative correction method is now

$$\hat f(\mathbf{x})=f(\mathbf{x},\hat\theta)\frac{1}{n}\sum_{i=1}^n\frac{K_{\mathbf{h}}(\mathbf{X}_i-\mathbf{x})}{f(\mathbf{X}_i,\hat\theta)}. \tag{7.1}$$

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 $f(\mathbf{x}, \hat{\theta})$. Clipping of the (3.6) type should again be applied to safeguard in general. The most important case is that of a multinormal-start density [see (7.3)].

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. First, the variance of the (7.1) estimator is found to be $R(K_1)\cdots R(K_d)(nh_1\cdots h_d)^{-1}f(\mathbf{x}) - n^{-1}f(\mathbf{x})^2$, which is exactly equal to the variance for the traditional (7.1) estimator, to the order of approximation used. Second, the bias is of the form

$$\frac{1}{2}\sum_{j=1}^{d}\sigma(K_j)^2h_j^2f_0(\mathbf{x})r''_{jj}(\mathbf{x}) + O\left(\sum_{j=1}^{d}\left(h_j^4 + \frac{h_j^2}{n}\right) + n^{-2}\right), \tag{7.2}$$

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 $\hat{\theta}$, for example (cf. the discussion of Section 3). The corresponding bias formula for the kernel estimator is of the form $\frac{1}{2}\Sigma_{j=1}^{d}\sigma(K_j)^2h_j^2f''_{jj}(\mathbf{x})$. Method (7.1) can therefore be expected to perform well in all situations where the $f_0r''_{jj}$ 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 multinormal-start method, for example, to work better than the traditional estimator for densities in a broad nonparametric vicinity of the multinormal.

A particular scheme, starting out with a multinormal estimate and a Gaussian kernel function, is further discussed in Hjort and Glad [(1994), Section 7]. It takes the form

$$\begin{aligned}\hat{f}(\mathbf{x}) = {} & \frac{1}{n}\sum_{i=1}^{n}\frac{\exp\{-(1/2)(\mathbf{x}-\mathbf{X}_i)'\hat{\Sigma}^{-1}(\mathbf{x}-\mathbf{X}_i)/h^2\}}{(2\pi)^{d/2}h^d} \\ & \times\frac{\exp\{-(1/2)(\mathbf{x}-\hat{\mu})'\hat{\Sigma}^{-1}(\mathbf{x}-\hat{\mu})\}}{\exp\{-(1/2)(\mathbf{X}_i-\hat{\mu})'\hat{\Sigma}^{-1}(\mathbf{X}_i-\hat{\mu})\}}.\end{aligned} \tag{7.3}$$

An initial sphering transformation has been used to create an estimator with only one smoothing parameter (as opposed to one for each of the $d$ directions), and a recipe for setting this smoothing parameter is also proposed in Hjort and Glad (1994), generalising the (6.3) rule.

## 8. Supplementary remarks.

8.1. *How close is the new estimator to the old?* In the sum that defines $\hat{f}(x)$ of (3.1), the ratios $f(x, \hat{\theta})/f(X_i, \hat{\theta})$ are close to 1 for small values of $h$ since then the $X_i$'s quite close to $x$ are those given significant weights. In

other words, $\hat{f}(x)$ cannot be very different from the traditional kernel estimator $\tilde{f}(x)$ of (1.1) when $h$ is small. A Taylor analysis is informative:

$$\frac{f(x,\hat{\theta})}{f(X_i,\hat{\theta})} \doteq 1 - a(x,\hat{\theta})(X_i - x) + \frac{1}{2}\Big\{a(x,\hat{\theta})^2 - b(x,\hat{\theta})\Big\}(X_i - x)^2,$$

where $a(x,\theta)$ and $b(x,\theta)$ are the two first $x$-derivatives of $\log f(x,\theta)$. Hence,

$$\hat{f}(x) \doteq \tilde{f}(x) - a(x,\hat{\theta})e_1(x) + \tfrac{1}{2}\Big\{a(x,\hat{\theta})^2 - b(x,\hat{\theta})\Big\}e_2(x), \tag{8.1}$$

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 $O(h^3 n^{-1} f(x))$. The new method attempts to make a bias correction of size $O(h^2)$ for densities in the vicinity of the parametric model, by taking information about first and second derivatives into account.

8.2. *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 $\hat{r}(x)$, 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

$$\mathbb{E}\hat{r}(x) \doteq r(x) + \frac{1}{2}\sigma_K^2 h^2 r''(x) - n^{-1}r(x)u_0(x)'\{I(x) + d\},$$

$$\operatorname{Var}\hat{r}(x) \doteq (nh)^{-1}R(K)\frac{r(x)}{f_0(x)} - n^{-1}r(x)^2\{1 + 2u_0(x)'I(x) - u_0(x)'\Sigma U_0(x)\},$$

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 $\log\hat{r}(x)$, to see how far from zero it is. Bias and variance results for this curve are also developed in Hjort and Glad (1994). A simple graphical goodness-of-fit method emerges from these results: plot

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

against $x$, or perhaps 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 $\pm 1.96$ about 95% of the time.

8.3. *Parametric home-turf conditions.* If model conditions $f(x) = f(x,\theta)$ can be trusted, the natural estimator is simply $f(x,\hat{\theta})$, for example, with the

maximum likelihood estimator. Well-known regularity conditions on the parametric model secure $\sqrt{n}(\hat\theta - \theta) \to_d Z \sim \mathcal{N}_p\{0, J(\theta)^{-1}\}$, where $J(\theta) = \int f(x,\theta)u(x,\theta)u(x,\theta)'\,dx$ is the information matrix, writing $u(x,\theta) = \partial \log f(x,\theta)/\partial\theta$ for the score function. This fact, combined with the delta method and some extra arguments, yield

$$n\,\mathrm{ISE}_n = n\int\{f(x,\hat\theta) - f(x,\theta)\}^2\,dx \to_d \int f(x,\theta)^2 u(x,\theta)'ZZ'u(x,\theta)\,dx,$$

a variable with expected value $\int f(x,\theta)^2 u(x,\theta)'J(\theta)^{-1}u(x,\theta)\,dx$. It turns out that the MISE of the new nonparametric (3.1) estimator, computed when $f(x) = f(x,\theta)$ really belongs to the parametric family in question, typically is only slightly larger than this. Using the $\hat f(x) \doteq f^*(x) + \bar B_n(x)'(\hat\theta - \theta) + \frac12(\hat\theta - \theta)'\bar C_n(x)(\hat\theta - \theta)$ representation used to prove Proposition 1 (Section 3), one may show that $\sqrt{n}\{\hat f(x) - f(x,\theta)\}$ tends to a certain zero-mean Gaussian process, when $h$ stays fixed, and that $n\,\mathrm{ISE}_n$ has a well-defined limit distribution. Its expected value, after somewhat arduous calculations, is shown to be

$$h^{-1}\int K(z)^2\left\{\int \frac{f(x,\theta)^2}{f(x+hz,\theta)}\,dx\right\}dz - \int f(x,\theta)^2\,dx$$

$$-\int f(x,\theta)^2 g_h(x,\theta)'J(\theta)^{-1}g_h(x,\theta)\,dx$$

$$-2\int f(x,\theta)^2 g_h(x,\theta)'J(\theta)^{-1}u(x,\theta)\,dx,$$

where $g_h(x,\theta) = \int K(z)\{u(x+hz,\theta) - u(x,\theta)\}\,dz$. First of all this shows that the nonparametric estimator can share with the parametric methods the favourable $O(n^{-1})$ MISE rate, under the home-turf conditions of the parametric vehicle model, by letting $h$ stay away from zero. Algebraic calculations for the two-parameter normal model lead to a parametric MISE of size $\frac78(2\sqrt{\pi}\,\sigma)^{-1}/n$ (which is the large-sample approximation to the exact MISE, for which an exact formula also can be found). The (3.5) estimator with $K = \phi$, on the other hand, has $n$ times MISE tending to

$$\left\{a^{-1}(1-a^2)^{-1/2} - 1 - \tfrac12 a^4 + \tfrac12 a^2\right\}(2\sqrt{\pi}\,\sigma)^{-1} \quad \text{where } a = h/\sigma \in (0,1).$$

This is minimized for $h^* = \sigma/\sqrt{2}$ (the best bandwidth under normality, regardless of sample size), with minimum value $\frac98(2\sqrt{\pi}\,\sigma)^{-1}$, only $\frac97$ times the optimal parametric achievement.

8.4. *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 $f(x,y) = f(x)g(y|x)$ and that interest focuses on the conditional mean function $m(x) = \mathbb{E}(Y|x)$. A standard method is $\tilde m(x) = \sum_{i=1}^n y_i K_h(x - x_i)/\sum_{i=1}^n K_h(x - x_i)$, the Nadaraya–Watson estimator [see, e.g., Scott (1992), Chapter 8, and Wand

and Jones (1995), Chapter 5]. Taylor-expansion analysis and somewhat strenuous calculations lead to

$$\begin{aligned} \mathbb{E}\tilde{m}(x) &\doteq m(x) + \frac{1}{2}\sigma_K^2 h^2 \left\{ m''(x) + 2m'(x)\frac{f'(x)}{f(x)} \right\}, \\ \operatorname{Var}\tilde{m}(x) &\doteq R(K)(nh)^{-1}\frac{\sigma(x)^2}{f(x)} + O(h/n). \end{aligned} \tag{8.3}$$

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

A semiparametric estimator can now be constructed as follows. Start out with a parametric initial description, say, $m(x, \hat{\beta})$, perhaps the simple linear $\hat{\beta}_1 + \hat{\beta}_2 x$. 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 $y_i/m(x_i, \hat{\beta})$. This leads to a generalised Nadaraya–Watson estimator

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

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 $\hat{\beta}$ around $\beta_0$, as in Section 3, lead in the end to

$$\begin{aligned} \mathbb{E}\hat{m}(x) \doteq m(x) + \frac{1}{2}\sigma_K^2 h^2 \Big\{ & m(x, \beta_0)r''(x) \\ & + 2m(x, \beta_0)r'(x)\frac{f'(x)}{f(x)} \Big\}, \end{aligned} \tag{8.5}$$

with approximation error of size at most $O(h^4 + h/n + n^{-2})$, 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. See Glad (1995) for further discussion, including semiparametric extensions of local linear regression methods.

Hjort [(1995a), 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.

**Acknowledgments.** The authors are grateful for useful and encouraging comments from M. C. Jones and for discussions with Grete Fenstad.

## REFERENCES

ABRAMSON, I. S. (1982). On bandwidth variation in kernel estimates—a square root law. *Ann. Statist.* **10** 1217–1223.

BUCKLAND, S. T. (1992). Maximum likelihood fitting of Hermite and simple polynomial densities. *J. Roy. Statist. Soc. Ser. C* **41** 241–266.

EFRON, B. and TIBSHIRANI, R. (1995). Using specially designed exponential families for density estimation. Technical report, Dept. Statistics, Stanford Univ.

FENSTAD, G. U. and HJORT, N. L. (1995). Comparison of two Hermite expansion density estimators with the kernel method. Unpublished manuscript.

FRIEDMAN, J. H., STUETZLE, W. and SCHROEDER, A. (1984). Projection pursuit density estimation. *J. Amer. Statist. Assoc.* **79** 599–608.

GLAD, I. K. (1995). Nonparametric correction of parametric regression estimates. Unpublished manuscript.

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. (1995a). Dynamic likelihood hazard rate estimation. *Biometrika*. To appear.

HJORT, N. L. (1995b). Bayesian approaches to semiparametric density estimation. *Bayesian Statistics 5. Proceedings of the Fifth Valencia International Meeting on Bayesian Statistics* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford University Press. To appear.

HJORT, N. L. and GLAD, I. K. (1994). Nonparametric density estimation with a parametric start. Statistical research report, Dept. Mathematics, Univ. Oslo.

HJORT, N. L. and JONES, M. C. (1995a). Locally parametric nonparametric density estimation. *Ann Statist.* To appear.

HJORT, N. L. and JONES, M. C. (1995b). Better rules of thumb for choice of smoothing parameter in density estimation. Unpublished manuscript.

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

JONES, M. C. (1993). Kernel density estimation when the bandwidth is large. *Austral. J. Statist.* **35** 319–326.

JONES, M. C., LINTON, O. and NIELSEN, J. P. (1995). A simple and effective bias reduction method for density and regression estimation. *Biometrika*. To appear.

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, Univ. New South Wales.

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

OLKIN, I. and SPIEGELMAN, C. H. (1987). A semiparametric approach to density estimation. *J. Amer. Statist. Assoc.* **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. *J. Amer. Statist. Assoc.* **82** 1131–1146.

SHAO, J. (1991). Second-order differentiability and jackknife. *Statist. Sinica* **1** 185–202.
SHEATHER, S. J. and JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53** 683–690.
TERRELL, G. R. and SCOTT, D. W. (1992). Variable kernel density estimation. *Ann. Statist.* **20** 1236–1265.
WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman & Hall, London.
WAND, M. P., MARRON, J. S. and RUPPERT, D. (1991). Transformations in density estimation (with discussion). *J. Amer. Statist. Assoc.* **86** 343–361.

DEPARTMENT OF MATHEMATICS
DIVISION OF STATISTICS
UNIVERSITY OF OSLO
P.O. BOX 4053 BLINDERN
N-0316 OSLO
NORWAY

NORWEGIAN INSTITUTE OF TECHNOLOGY
INSTITUTE OF MATHEMATICAL SCIENCES
N-7034 TRONDHEIM
NORWAY