Sometimes nonparametrics beat parametrics, even when the model is right

Morten Byholt and Nils Lid Hjort

University of Oslo

October 1996

ABSTRACT. A basic issue in both teaching of and practice of statistics is the interplay between modelling assumptions and inference performance. The general message conveyed is that stronger assumptions lead to better statistical performance of the relevant estimators, tests and confidence intervals, provided that these assumptions hold. On the other hand, fewer assumptions often lead to safer and more robust methods that are good also outside narrow conditions, but not quite as good as specialist methods that exploit such narrower conditions, if these are fulfilled.

This interplay is nicely illustrated in the context of density estimation, where parametric and nonparametric methods can be contrasted. The parametric ones have mean squared errors of size $O(n^{-1})$ in terms of sample size n if the parametric model is right, but are not even consistent outside the model. The nonparametric methods are everywhere consistent and have mean squared errors of size $O(n^{-4/5})$ for broad classes of estimands.

The point we are making here is that this picture is not universally true! We show that a simple kernel density estimator can perform better than a directly estimated parametric density on the latter's home turf, for small sample sizes, in the sense of mean integrated squared error. Our main example is that of estimating an unknown normal density. In the process of developing and discussing this somewhat counter-intuitive and half-paradoxical example we touch on several tangential issues of interest, pertaining to exact small-sample analysis of density estimators.

KEY WORDS: *comparison of performance; density estimators; nonparametric versus parametric; small-sample analysis*

1. Introduction and summary. There is a plethora of examples in statistics that illustrate the following principle: stronger assumptions about the model lead to construction of inference methods with stronger performance, provided the assumptions in question are in force. The other side of the coin is equally important and well understood in many situations: weaker assumptions lead to alternative methods that perform well under a broader range of underlying probability mechanisms than those that require harder assumptions, but these robust methods typically do not perform quite as well, under a narrower model, as the specialist methods that exploit such assumptions. These principles can be formulated and exemplified in contexts of smaller and bigger parametric models, or in situations with simpler and not so simple regression structures, and so on. They are also nicely illustrated in situations where there are both parametric and natural nonparametric competitors for an inference task; one expects the parametric methods to behave worse outside the parametric model, but one surely expects the parametrically based optimum methods to outperform the nonparametric ones on the parametric model's home turf.

The main purpose of this paper is to exhibit a counterexample to the belief that these principles always hold. We shall discuss a natural situation where a simple nonparametric density estimator outperforms the natural method based on normality, even when the normal model is perfect. We hasten to add that this only can happen for small sample sizes, however.

Suppose that $X_1, \ldots, X_n$ are independent and that it is required to estimate its density f. If f is modelled parametrically, say $f(x) = f(x, \theta)$, one would typically plug in natural estimators for the parameters to get a density estimator of the form $\widehat{f}(x) = f(x, \widehat{\theta})$. In the normal case this leads to

$$\widehat{f}(x) = \widehat{\sigma}^{-1}\phi(\widehat{\sigma}^{-1}(x - \widehat{\mu})), \tag{1.1}$$

writing $\phi(x)$ for the standard normal density, where $\widehat{\mu} = n^{-1}\sum_{i=1}^n X_i$ and $\widehat{\sigma}^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \widehat{\mu})^2$; variations in the choice of parameter estimates could of course be considered. If on the other hand nothing is assumed a priori besides some smoothness, one would use a nonparametric estimator, of which there are several well-explored classes. The most important one is the kernel type estimator,

$$\widetilde{f}(x) = n^{-1}\sum_{i=1}^{n} K_h(X_i - x), \tag{1.2}$$

where $K_h(u) = h^{-1}K(h^{-1}u)$ is the scaled version of a suitable symmetric probability density function K called the kernel function K.

The behaviour of such estimators has been studied extensively, see for example recent accounts by Scott (1992) and Wand and Jones (1995). The primary focus has been performance for moderate to large sample sizes, however, partly because of the availability of revealing large-sample approximations. In particular it is well known that the mean squared error (MSE) of these estimators is of size $O(n^{-4/5})$ whereas the parametric ones, exemplified here by (1.1), has MSE rate $O(n^{-1})$. Again, our aim is to show that for small sample sizes the picture is reversed. We include two kernel functions in this little study, namely the standard normal kernel $K = \phi$ and the so-called Yepanechnikov kernel $K_{\text{епан}}(u) = \frac{3}{2}(1 - 4u^2)$ on $[-\frac{1}{2}, \frac{1}{2}]$. The reason is that this kernel function is known to minimise both the pointwise MSE and the integrated mean squared error (MISE) for large samples.

We go through the details of parametric estimators in Section 2 and nonparametric estimators in Section 3 and 4. Then Section 5 gives some comparison results and offers some illustrations. Finally some supplementing comments and results are listed in Section 6.

It should be emphasised that this paper is consistently concerned with exact finite-sample properties of both types of estimators. This requires more attention to technical details than for the more easily available and implementable methods and results about large-sample behaviour. We do however touch on these easier to get results too for illustration and comparison purposes.

2. MSE and MISE for the parametric estimator. Here we calculate exact mean, variance and MISE for the natural normality-based estimator (1.1).

2.1. Large-sample calculations. Approximations in the tradition of large-sample theorems are both reasonably easy to obtain and illustrative. For a general parametric family $f(x,\theta)$, let $u(x,\theta) = \partial \log f(x,\theta)/\partial\theta$ be the score function. It is well known that the maximum likelihood estimator $\widehat\theta$ behaves asymptotically according to $\sqrt{n}(\widehat\theta-\theta)\to_d \mathrm{N}(0,J(\theta)^{-1})$, under model conditions, where $J(\theta)=\mathrm{E}u(X,\theta)u(X,\theta)^{\mathrm{t}}$ is the Fisher information matrix. By the delta method it follows that

$$\sqrt{n}\{f(x,\widehat\theta)-f(x,\theta)\}\to_d f(x,\theta)u(x,\theta)^{\mathrm{t}}V, \quad \text{where } V\sim \mathrm{N}(0,J(\theta)^{-1}).$$

In other words, n times the squared error at x tends to a variable with variance $v(x,\theta)^2 = f(x,\theta)^2u(x,\theta)^{\mathrm{t}}J(\theta)^{-1}u(x,\theta)$, and n times MISE will under mild regularity conditions tend to $\int v(x,\theta)^2\,\mathrm{d}x$. For the normal case this leads to the approximation

$$\mathrm{MSE}(x)=\mathrm{E}\Big\{\frac{1}{\widehat\sigma}\phi\Big(\frac{x-\widehat\mu}{\widehat\sigma}\Big)-\frac{1}{\sigma}\phi\Big(\frac{x-\mu}{\sigma}\Big)\Big\}^2 \doteq \frac{1}{n}\frac{1}{\sigma^2}\phi(y)^2\{y^2+\tfrac12(y^2-1)^2\}, \tag{2.1}$$

in terms of $y=y_x=(x-\mu)/\sigma$, and secondly to

$$\begin{aligned}\mathrm{MISE}&\doteq n^{-1}\sigma^{-2}\int\phi(y_x)^2\{y_x^2+\tfrac12(y_x^2-1)^2\}\,\mathrm{d}x\\ &=n^{-1}\sigma^{-1}\int\phi(y)^2\{y^2+\tfrac12(y^2-1)^2\}\,\mathrm{d}y=\frac{1}{n}\frac{1}{\sigma}\frac{1}{2\sqrt{\pi}}\Big(\frac48+\frac38\Big).\end{aligned} \tag{2.2}$$

The last formula indicates that $4/7$ of the total MISE stems from estimation of unknown μ while estimation of unknown σ is responsible for the remaining $3/7$.

2.2. Exact MSE under normality. We may write $\widehat\mu=\mu+(\sigma/\sqrt{n})N$ and $\widehat\sigma^2=\sigma^2Z^2$, say, where N is standard normal and independent of Z^2, a $\chi^2_{n-1}/(n-1)$ variable. Thus the parametric normal density estimator can be expressed as

$$\widehat f(x)=\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma Z}\exp\Big\{-\tfrac12\frac{(N+\sqrt{n}y)^2}{nZ^2}\Big\}, \quad \text{where again } y=(x-\mu)/\sigma.$$

This makes it possible to compute at first conditional moments given $\widehat\sigma$. One needs the fact that

$$\mathrm{E}\exp\{-\tfrac12 t(N+a)^2\}=(1+t)^{-1/2}\exp\{-\tfrac12 a^2t/(1+t)\} \quad \text{for } t>-1, \tag{2.3}$$

which can be verified directly (and is also obtainable from the moment-generating function of a non-central chi squared). Hence

$$\mathrm{E}\{\widehat{f}(x)\,|\,Z\} = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\frac{\sqrt{n}}{(1+nZ^2)^{1/2}}\exp\{-\tfrac{1}{2}y^2n/(1+nZ^2)\} \tag{2.4}$$

and

$$\mathrm{E}\{\widehat{f}(x)^2\,|\,Z\} = \frac{1}{2\pi}\frac{1}{\sigma^2}\frac{\sqrt{n}}{Z(2+nZ^2)^{1/2}}\exp\{-y^2n/(2+nZ^2)\}. \tag{2.5}$$

It follows that the exact mean and variance, and hence the exact $\mathrm{MSE}(x)$, can be arrived at by numerical integration with respect to the density for Z, which is of the form $g_n(z) = \gamma_{n-1}((n-1)z^2)2(n-1)z$, where $\gamma_{n-1}(\cdot)$ is the density of a χ^2_{n-1}. Note that Z for reasonably sized n is quite close to 1 with high probability.

2.3. Exact MISE under normality. Turning next to the exact MISE, note first the technical fact that

$$\int \frac{1}{\sigma_1}\phi\Big(\frac{x-\mu_1}{\sigma_1}\Big)\frac{1}{\sigma_2}\phi\Big(\frac{x-\mu_2}{\sigma_2}\Big)\,\mathrm{d}x = \frac{1}{(\sigma_1^2+\sigma_2^2)^{1/2}}\phi\Big(\frac{\mu_1-\mu_2}{(\sigma_1^2+\sigma_2^2)^{1/2}}\Big);$$

a simple demonstration involves noting that $\int f_1f_2\,\mathrm{d}x$ is the density at zero for a variable of the form $Y_1 - Y_2$, where Y_1 and Y_2 are independent from respectively f_1 and f_2. This leads to

$$\int (\widehat{f}-f)^2\,\mathrm{d}x = \phi(0)\Big[\frac{1}{\sqrt{2}\widehat{\sigma}} + \frac{1}{\sqrt{2}\sigma} - 2\frac{1}{(\widehat{\sigma}^2+\sigma^2)^{1/2}}\exp\Big\{-\frac{1}{2}\frac{(\widehat{\mu}-\mu)^2}{\widehat{\sigma}^2+\sigma^2}\Big\}\Big].$$

Now use $\mathrm{E}\exp(-\frac{1}{2}t\chi^2_1) = (1+t)^{-1/2}$ for the moment-generating function of a squared standard normal, a special case of (2.3). One then finds

$$\mathrm{MISE} = n^{-1}\sigma^{-1}(2\sqrt{\pi})^{-1}a_n, \tag{2.6}$$

where

$$a_n = n\,\mathrm{E}\Big[1 + \frac{1}{Z} - 2\Big\{\frac{2n}{1+n(1+Z^2)}\Big\}^{1/2}\Big].$$

The result is given in this form since a_n is stable as n grows. Note that the formula (2.6) is exact, with a value for a_n that can be computed via numerical integration using the density $g_n(z)$ for Z again (the $1/Z$ term has exact mean equal to $\{\frac{1}{2}(n-1)\}^{1/2}\Gamma(\frac{1}{2}(n-2))/\Gamma(\frac{1}{2}(n-1))$, but the second term does not have a closed-form expectation formula). The a_n sequence goes to 7/8, as indicated by result (2.2), and the convergence is by slow monotone decrease. In fact, somewhat long calculations involving Taylor expansions and moments of the χ^2_{n-1} yield

$$\mathrm{MISE} = n^{-1}\sigma^{-1}(2\sqrt{\pi})^{-1}\{7/8 + (271/96)/n + O(1/n^2)\}.$$

2.4. The UMVU density estimator. There is a unique uniformly minimum variance unbiased estimator of $f(x,\theta)=\sigma^{-1}\phi(\sigma^{-1}(x-\mu))$ under model conditions. We include this in our list of competitors too:

$$\bar f(x)=\frac{k_n}{\widehat\sigma}\Big\{1-\frac{n}{(n-1)^2}\Big(\frac{x-\widehat\mu}{\widehat\sigma}\Big)^2\Big\}^{n/2-2}\quad\text{where } k_n=\frac{\Gamma(\frac12(n-1))}{\sqrt{\pi}\,\Gamma(\frac12(n-2))}\frac{n^{1/2}}{n-1},\tag{2.7}$$

in the region where $|x-\widehat\mu|/\widehat\sigma\le(n-1)/n^{1/2}$. Outside this (random) region the density estimate is zero. See Lehmann (1983, p. 87) for a sketch of a derivation (where the same estimator is given in a slightly different form).

A natural question is the relative comparison between the direct parametric estimator (1.1), which uses UMVU estimates for the parameters, and the UMVU density estimator $\bar f$. It is difficult to obtain a useful expression for the pointwise MSE of $\bar f$; stochastic simulation or two-dimensional numerical integration seem to be called for. However, with some work the exact MISE can be obtained in a form suitable for numerical evaluation. With a standard normal f, start out writing the integrated squared error as

$$\text{ISE}=\frac{k_n^2}{\widehat\sigma}\frac{n-1}{n^{1/2}}\int_{-1}^{1}(1-v^2)^{n-4}\,\mathrm{d}v-2\int\bar f(x)f(x)\,\mathrm{d}x+R(f).$$

By the unbiasedness property of $\bar f$ this leads to

$$\text{MISE}=\Big(\frac{n-1}{2}\Big)^{1/2}\frac{\Gamma(\frac12(n-2))}{\Gamma(\frac12(n-1))}k_n^2\frac{n-1}{n^{1/2}}\frac{\Gamma(n-3)\sqrt{\pi}}{\Gamma(\frac12(2n-5))}-\frac{1}{2\sqrt{\pi}}.\tag{2.8}$$

This can now be compared to the size of (2.6) for various n. It turns out that the UMVU density estimator actually loses, for each n, to the simpler plug-in method. The ratio of (2.8) by (2.6) is always greater than 1, but by an amount that quickly becomes small with n, see the table in Section 5. That the ratio tends to 1 also reflects the fact that the (2.7) estimator is quite close to the plug-in (1.1) estimator for moderate to large n.

3. MSE and MISE for the nonparametric estimator. This section works out the necessary exact expressions for the bias and variance of the kernel estimator (1.2).

3.1. Large-sample approximations. Again our focus is on the finite-sample case to follow afterwards, but it is illuminating to compare those results to the traditional approximations suggested by well-known asymptotics arguments.

The leading terms of the bias and variance are respectively $\frac12 k_2h^2f''(x)$ and $(nh)^{-1}R(K)f(x)-n^{-1}f(x)^2$, see for example Scott (1992, Ch. 6). Here $k_2=\int u^2K(u)\,\mathrm{d}u$ while $R(q)$ is generic notation for $\int q(x)^2\,\mathrm{d}x$. Hence the MISE can be approximated by $(nh)^{-1}R(K)+\frac14k_2^2h^4R(f'')-n^{-1}R(f)$, as h goes to zero with nh tending to infinity. The minimiser of this approximate MISE is $h_a=\{R(K)/k_2^2\}^{1/5}R(f'')^{-1/5}/n^{1/5}$, leading to a minimum approximate MISE of the form $d_n/n^{4/5}-R(f)/n$, where $d_n=\frac54\{R(K)k_2^{1/2}\}^{4/5}R(f'')^{1/5}$.

These formulae can be suitably translated to the situation where f is normal, for our two choices $K = \phi$ and $K = K_{\text{епан}}$. For the normal kernel case,

$$h_a \doteq 1.0592\,\sigma/n^{1/5} \quad \text{and} \quad \min_h \textsc{mise} \doteq 0.3329\,\sigma^{-1}/n^{4/5} - 0.2821\,\sigma^{-1}/n.$$

For the Yepanechnikov kernel case,

$$h_a \doteq 4.6898\,\sigma/n^{1/5} \quad \text{and} \quad \min_h \textsc{mise} \doteq 0.3198\,\sigma^{-1}/n^{4/5} - 0.2821\,\sigma^{-1}/n.$$

3.2. Exact finite-sample MSE *performance.* The mean and variance of the (1.2) estimator can generally be expressed as respectively

$$e_h(x) = \mathrm{E}K_h(X_i - x) = \int K(u)f(x+hu)\,\mathrm{d}u \tag{3.1}$$

and

$$v_{n,h}(x)^2 = n^{-1}\mathrm{Var}K_h(X_i - x) = (nh)^{-1}a_h(x) - n^{-1}e_h(x)^2, \tag{3.2}$$

where $a_h(x) = \int K(u)^2 f(x+hu)\,\mathrm{d}u$. Our task involves computing, displaying and comparing these function when f is normal. It suffices to take care of the standard normal case, giving certain $e_h^0(x)$ and $a_h^0(x)$ functions, say, in that the case of a normal (μ, σ^2) density then will lead to respectively

$$e_h(x) = \sigma^{-1}e^0_{h/\sigma}(y) \quad \text{and} \quad a_h(x) = \sigma^{-2}a^0_{h/\sigma}(y), \tag{3.3}$$

in terms of $y = (x-\mu)/\sigma$.

Consider first the case $K = \phi$. We there find

$$e_h^0(x) = \phi\Big(\frac{x}{(1+h^2)^{1/2}}\Big)\frac{1}{(1+h^2)^{1/2}},$$
$$a_h^0(x) = \frac{1}{2\sqrt{\pi}}\phi\Big(\frac{x}{(1+h^2/2)^{1/2}}\Big)\frac{1}{(1+h^2/2)^{1/2}}.$$

We may now compute $\textsc{mse}(x)$ for any x, via formulae (3.1)–(3.3).

More troublesome is finding the MSE for the Yepanechnikov kernel $K_{\text{епан}}$. It is convenient to define

$$N_{j,h}(x) = h^{-1}\int_{x-h/2}^{x+h/2} v^j\phi(v)\,\mathrm{d}v \quad \text{for } j = 0,1,2,3,4.$$

Note that $N_{j,h}(x) = x^j\phi(x) + O(h^2)$ for small h. Values of these can be found exactly in terms of the ϕ and Φ functions, the density and cumulative of the standard normal, in that $\int_a^b v\phi(v)\,\mathrm{d}v = \phi(a) - \phi(b)$, $\int_a^b v^2\phi(v)\,\mathrm{d}v = a\phi(a) - b\phi(b) + \Phi(a,b)$, while

$$\int_a^b v^j\phi(v)\,\mathrm{d}v = a^{j-1}\phi(a) - b^{j-1}\phi(b) + (j-1)\int_a^b v^{j-2}\phi(v)\,\mathrm{d}v \quad \text{for } j \geq 3. \tag{3.4}$$

These formulae are found by simple partial integration. Thus

$$\begin{aligned}e_h^0(x) &= (3/2)h^{-1}\int_{x\pm h/2}\Big\{1-4\Big(\frac{v-x}{h}\Big)^2\Big\}\phi(v)\,\mathrm{d}v\\ &= (3/2)\big[N_{0,h}(x)-4h^{-2}\{N_{2,h}(x)-2xN_{1,h}(x)+x^2N_{0,h}(x)\}\big]\end{aligned}$$

and similarly

$$\begin{aligned}a_h^0(x) &= \frac{9}{4}h^{-1}\int_{x\pm h/2}\Big\{1-4\Big(\frac{v-x}{h}\Big)^2\Big\}^2\phi(v)\,\mathrm{d}v\\ &= \frac{9}{4}\big[N_{0,h}(x)-8h^{-2}\{N_{2,h}(x)-2xN_{1,h}(x)+x^2N_{0,h}(x)\}\\ &\quad+16h^{-4}\{N_{4,h}(x)-4xN_{3,h}(x)+6x^2N_{2,h}(x)-4x^3N_{1,h}(x)+x^4N_{0,h}(x)\}\big].\end{aligned}$$

Again, the exact bias, variance and MSE can now be computed for any x, via (3.1)–(3.3).

3.3. Exact finite-sample MISE *calculations.* An exact formula for the MISE for the case of a normal kernel and a normal estimand is not difficult to obtain, and was first exhibited by Fryer (1976). It also follows below via other arguments. Presently we also need such a formula when $K = K_{\text{епан}}$, which is a more demanding technical exercise. In general it turns out to be helpful to work in terms of $g(y) = \int f(x)f(x+y)\,\mathrm{d}x$, the density of a difference $X_i - X_j$, and with the corresponding $g_K(y) = \int K(u)K(u+y)\,\mathrm{d}u$. By suitably interchanging the order of integration in a couple of multiple integrals, using Fubini's theorem, one finds that $\int a_h(x)\,\mathrm{d}x = R(K)$, that $\int e_h(x)^2\,\mathrm{d}x = \int g_K(u)g(hu)\,\mathrm{d}u$, and that $\int e_h(x)f(x)\,\mathrm{d}x = \int K(u)g(hu)\,\mathrm{d}u$. Hence

$$\mathrm{MISE}(h) = (nh)^{-1}R(K) + (1-n^{-1})\int g_K(u)g(hu)\,\mathrm{d}u - 2\int K(u)g(hu)\,\mathrm{d}u + g(0) \quad (3.5)$$

is a generally valid formula for the exact MISE, with $g(0) = R(f)$.

For $K = \phi$, $g(u)$ becomes $\phi(u/\sqrt{2})/\sqrt{2}$, and the calculations become easy when f is standard normal. They give

$$\mathrm{MISE}^0(h) = (2\sqrt{\pi})^{-1}\{(nh)^{-1} + (1-n^{-1})(1+h^2)^{-1/2} - 2(1+\tfrac{1}{2}h^2)^{-1/2} + 1\}, \quad (3.6)$$

where the '0' superscript indicates that the f in question is $N(0,1)$. For a general $N(\mu,\sigma^2)$ one finds

$$\mathrm{MISE}(h) = \sigma^{-1}\mathrm{MISE}^0(\sigma^{-1}h). \quad (3.7)$$

This can be minimised over h for each sample size n to find the optimal possible performance; see Section 4.

Next turn to the $K_{\text{епан}}$ kernel, for which answers require more efforts. Rather than integrating the exact MSE expressions found above over x, which is possible but rather involved, we prefer using the (3.5) identity. It requires finding g_K for the $K_{\rm епан}$ kernel. One sees that $g_K(u)$ is zero outside $[-1,1]$. For $0 \le u \le 1$ somewhat long calculations give

$$\begin{aligned} g_K(u) &= \int K(x - \tfrac12 u)K(x + \tfrac12 u)\,\mathrm{d}x \\ &= \int_{-(1-u)/2}^{(1-u)/2} \{1 - 4(x - \tfrac12 u)^2\}\{1 - 4(x + \tfrac12 u)^2\}\,\mathrm{d}x \\ &= (6/5)(1 - 5u^2 + 5u^3 - u^5), \end{aligned}$$

while $g_K(u) = (6/5)(1 - 5u^2 - 5u^3 + u^5)$ for $-1 \le u \le 0$, by symmetry. Let again f be standard normal; the (3.7) scale identity also holds for the $K_{\rm епан}$ kernel. Further calculations give that $\int K(u)g(hu)\,\mathrm{d}u$ is equal to

$$\begin{aligned} \mathrm{I}(h) &= h^{-1}\int K(\sqrt{2}v/h)\phi(v)\,\mathrm{d}v \\ &= h^{-1}3\int_0^{h/2\sqrt{2}} (1 - 4\cdot 2v^2/h^2)\phi(v)\,\mathrm{d}v \\ &= 3h^{-1}\{(1 - 8/h^2)\Phi(0, h/2\sqrt{2}) + 2\sqrt{2}h^{-1}\phi(h/2\sqrt{2})\}. \end{aligned}$$

Similarly, $\int g_K(u)g(hu)\,\mathrm{d}u$ is equal to

$$\begin{aligned} \mathrm{II}(h) &= h^{-1}\int g_K(\sqrt{2}v/h)\phi(v)\,\mathrm{d}v \\ &= h^{-1}(12/5)\int_0^{h/\sqrt{2}} (1 - 5\cdot 2v^2/h^2 + 5\cdot 2\sqrt{2}v^3/h^3 - 4\sqrt{2}v^5/h^5)\phi(v)\,\mathrm{d}v \\ &= (12/5)h^{-1}[(1 - 10/h^2)\Phi(0, h/\sqrt{2}) + (20\sqrt{2}/h^3 - 32\sqrt{2}/h^5)\phi(0) \\ &\qquad + (\sqrt{2}/h - 12\sqrt{2}/h^3 + 32\sqrt{2}/h^5)\phi(h/\sqrt{2})]. \end{aligned}$$

This suffices, via (3.5), to produce an exact MISE formula for the kernel estimator with the $K_{\rm епан}$ kernel:

$$\mathrm{MISE}^0(h) = (6/5)(nh)^{-1} + (1 - n^{-1})\,\mathrm{II}(h) - 2\,\mathrm{I}(h) + 1/(2\sqrt{\pi}). \tag{3.8}$$

4. The real MISE when a random bandwidth is used. We have not yet been explicit about which h value to use in the kernel estimator. The results above give relevant guidelines and lead to natural data-based proposals. This also necessitates some extra analysis about the real MISE.

4.1. Deciding on the bandwidth. Knowing that the optimal size of the h is $O(n^{-1/5})$, we consider the MISE curves via bandwidths of the form $h = b/n^{1/5}$, focusing on the b parameter. For the normal kernel case (3.6) leads to

$$\mathrm{MISE}^0\Big(\frac{b}{n^{1/5}}\Big) = \frac{1}{2\sqrt{\pi}}\Big[\frac{1}{bn^{4/5}} + \Big(1 - \frac{1}{n}\Big)\frac{1}{(1 + b^2/n^{2/5})^{1/2}} - \frac{2}{(1 - \frac12 b^2/n^{2/5})^{1/2}} + 1\Big]. \tag{4.1}$$

According to this and (3.7), the best theoretical bandwidth to use for estimation of a normal estimand is $h_{\rm mise} = b_n\sigma/n^{-1/5}$, where b_n is the minimiser of (4.1). The theory reviewed in Section 3.1 predicts that the minimum should occur around $b = 1.059$ when n is large. Some exact values are given in the table of Section 5. One should actually stretch the bandwidth a little bit in comparison with the famous $1.059\,\sigma/n^{1/5}$ rule of thumb, thereby earning a couple of percent in MISE. The actual rule we use in practice is $\widetilde h = b_n\widehat\sigma/n^{1/5}$.

Similarly, for the Yepanechnikov kernel case, consider bandwidths of the form $h = c/n^{1/5}$, with

$$\mathrm{MISE}^0(c/n^{1/5}) = (6/5)c^{-1}n^{-4/5} + (1-n^{-1})\,\mathrm{II}(cn^{-1/5}) - 2\,\mathrm{I}(cn^{-1/5}) + (2\sqrt{\pi})^{-1}, \quad (4.2)$$

in which

$$\mathrm{I}\Big(\frac{c}{n^{1/5}}\Big) = 3\frac{n^{1/5}}{c}\Big[\Big(1-\frac{8n^{2/5}}{c^2}\Big)\Phi\Big(0,\frac{c}{2\sqrt{2}n^{1/5}}\Big) + 2\sqrt{2}\frac{n^{1/5}}{c}\phi\Big(\frac{c}{2\sqrt{2}n^{1/5}}\Big)\Big],$$

$$\begin{aligned}\mathrm{II}\Big(\frac{c}{n^{1/5}}\Big) = \frac{12}{5}\frac{n^{1/5}}{c}\Big[\Big(1-\frac{10n^{2/5}}{c^2}\Big)\Phi\Big(0,\frac{c}{\sqrt{2}n^{1/5}}\Big) + \Big(\frac{20\sqrt{2}n^{3/5}}{c^3} - \frac{32\sqrt{2}n}{c^5}\Big)\phi(0)\\ + \Big(\frac{\sqrt{2}n^{1/5}}{c} - \frac{12\sqrt{2}n^{3/5}}{c^3} + \frac{32\sqrt{2}n}{c^5}\Big)\phi\Big(\frac{c}{\sqrt{2}n^{1/5}}\Big)\Big].\end{aligned}$$

The theoretically best bandwidth for a normal estimand is $h_{\rm mise} = c_n\sigma/n^{1/5}$, where c_n minimises (4.2). The rule we use here is therefore $\widetilde h = c_n\widehat\sigma/n^{1/5}$. By results of Section 3.1, c_n should be close to 4.690 for large n. Exact values are provided in Section 5's table for some sample sizes. The convergence of c_n towards its limiting value 4.690 is slow and from above.

4.2. Determining the real MISE. We used formulae (4.1) and (4.2) to decide on methods for choosing h, but as the $\widetilde h$ used depend on data we cannot use the MISE formulae to evaluate the real MISE incurred. We should consider the expected value of the integrated squared error of the real estimator, that is,

$$\mathrm{MISE} = \mathrm{E}\,\mathrm{ISE}(\widetilde f(\cdot,\widetilde h)) = \mathrm{E}\int\{\widetilde f(x,\widetilde h) - f(x)\}^2\,\mathrm{d}x. \quad (4.3)$$

Jones (1991) calls this EISE to make the distinction clear; see his paper for a general discussion of the performance criteria used. If $\widehat\sigma$ is reasonably precise for σ, then $\widetilde h$ is in the vicinity of $h_{\rm mise}$, and the MISE(h) curves are reasonably flat around their minimands. Hence the additional uncertainty in $\widetilde h$ is bound to be a secondary effect, not likely to give a significant increase in the MISE level compared to MISE$(h_{\rm mise})$, at least when n is moderate or big. However, precise numbers are called for in order to finish our comparison study for small n. Two methods are now described for obtaining these numbers. We still assume that the true density is normal, and may take it to be in standard form.

The first method is exact, but requires numerical integration. Start out writing the ISE of a kernel estimator as

$$\frac{1}{n^2}\sum_{i,j}\int K_{\widetilde{h}}(x-X_i)K_{\widetilde{h}}(x-X_j)\,\mathrm{d}x - 2\frac{1}{n}\sum_{i=1}^n\int K_{\widetilde{h}}(x-X_i)f(x)\,\mathrm{d}x + R(f)$$
$$= \frac{R(K)}{n\widetilde{h}} + \frac{1}{n^2}\sum_{i\neq j}\frac{1}{\widetilde{h}}g_K\Big(\frac{X_i-X_j}{\widetilde{h}}\Big) - 2\frac{1}{n}\sum_{i=1}^n\int K(u)f(X_i+\widetilde{h}u)\,\mathrm{d}u + R(f),$$

which has mean value

$$\text{MISE} = \mathrm{E}\frac{R(K)}{n\widetilde{h}} + \Big(1-\frac{1}{n}\Big)\mathrm{E}\frac{1}{\widetilde{h}}g_K\Big(\frac{X_1-X_2}{\widetilde{h}}\Big) - 2\,\mathrm{E}\int K(u)f(X_1+\widetilde{h}u)\,\mathrm{d}u + R(f).$$

This is valid if only the $\widetilde{h}$ operation is symmetric in the n data points. Now focus on $\widetilde{h}=a\widehat{\sigma}$ for a known value of a, like for the rules described above. Let $p_n(r)$ and $q_n(s)$ be the densities of respectively $R=(X_1-\widehat{\mu})/\widehat{\sigma}$ and $S=(X_1-X_2)/\widehat{\sigma}$. These are ancillary statistics and as such are independent of $(\widehat{\mu},\widehat{\sigma})$, by the so-called Basu Lemma, cf. Lehmann (1983, Ch. 1.5). Hence

$$\begin{aligned}\text{MISE} &= R(K)(na)^{-1}\mathrm{E}\,\widehat{\sigma}^{-1} + (1-n^{-1})\mathrm{E}\,\widehat{\sigma}^{-1}\,\mathrm{E}a^{-1}g_K(a^{-1}S)\\ &\quad - 2\,\mathrm{E}\int K(u)f(\widehat{\mu}+\widehat{\sigma}(R+au))\,\mathrm{d}u + R(f),\end{aligned} \tag{4.4}$$

and this can be evaluated exactly. The inverse mean of $\widehat{\sigma}$ was mentioned in Section 2.3 and $\mathrm{E}a^{-1}g_K(a^{-1}S) = \int a^{-1}g_K(a^{-1}s)q_n(s)\,\mathrm{d}s$. Furthermore, for the -2 term, some calculations give

$$\mathrm{E}\int\int K(u)f(\widehat{\mu}+\widehat{\sigma}(r+au))p_n(r)\,\mathrm{d}r\,\mathrm{d}u$$
$$= \int\int K(u)\frac{1}{(2\pi)^{1/2}}\Big(\frac{n}{n+1}\Big)^{1/2}\Big\{1+\frac{n}{n+1}\frac{(r+au)^2}{n-1}\Big\}^{-\frac{1}{2}(n-1)}p_n(r)\,\mathrm{d}r\,\mathrm{d}u.$$

Finally needed are the densities of R and S, and with the required stamina one finds $p_n(r) = k_n\{1-nr^2/(n-1)^2\}^{\frac{1}{2}(n-4)}$ for $|r|\le(n-1)/n^{1/2}$, where k_n is as in (2.7), and $q_n(s)=\lambda_n\{1-s^2/2(n-1)\}^{\frac{1}{2}(n-4)}$ for $|s|\le\{2(n-1)\}^{1/2}$, with $\lambda_n = k_n(n-1)^{1/2}/(2n)^{1/2}$.

The above is sufficient to evaluate the real MISE for the two kernel estimators, as described above, with $a=b_n/n^{1/5}$ for $K=\phi$ and $a=c_n/n^{1/5}$ for the Yepanechnikov kernel.

Our second method is simpler to execute but relies on simulation and computer power, and is hence not quite as good as the exact solution. Write the MISE as

$$\begin{aligned}\text{MISE} &= \mathrm{E}\int f(x)\{\widetilde{f}(x,\widetilde{h})/f(x)^{1/2} - f(x)^{1/2}\}^2\,\mathrm{d}x\\ &= \mathrm{E}_{n+1}\{\widetilde{f}(X_{n+1},\widetilde{h})/f(X_{n+1})^{1/2} - f(X_{n+1})^{1/2}\}^2,\end{aligned}$$

where the E_{n+1} indicates expectation with respect to $n+1$ variables $X_1, \ldots, X_{n+1}$ independently drawn from f; of course $X_1, \ldots, X_n$ are present in the construction of $\widetilde{f}(x, \widetilde{h})$. The great advantage is that a numerical value for MISE can now be determined by simulation, without any integration, regardless of the complexities of the data-dependent rule used for $\widetilde{h}$. It is also possible and sometimes advantageous, with respect to computer time and required precision, to evaluate the estimator constructed from $X_1, \ldots, X_n$ in more than one new point, using say $\text{MISE} = \mathrm{E}_{n+m} Z_{n,m}$, where

$$Z_{n,m} = m^{-1} \sum_{j=1}^{m} \{\widetilde{f}(X_{n+j}, \widetilde{h})/f(X_{n+j})^{1/2} - f(X_{n+j})^{1/2}\}^2.$$

The MISE is obtained by simulating a large enough number of such variables. The accuracy is essentially determined by the size of B, the number of times the $(X_1, \ldots, X_n)$ set is sampled to form the density estimator, but is also helped by m, the number of sampled extra points at which to evaluate the constructed estimator. Some calculations show that the variance of the simulation based MISE figure has the approximate form $\{\tau_1^2/(Bm) + \tau_2^2/B\}/n^{8/5}$, where τ_1 and τ_2 are certain constants. The MISEs themselves go down to zero with rate $n^{-4/5}$, and we need precision on the relative and not absolute scale, so the point is to get a low $\tau_1^2/(Bm) + \tau_2^2/B$.

In our calculations both methods were employed. In the summary below only the exact numbers, arrived at using numerical integration, are given. These have a clear edge in precision; even with a million simulations per sample size one cannot be fully confident about the third decimal place when giving MISE numbers on appropriate relative scale.

5. Comparisons and illustrations. This section offers numerical results and two figures to illustrate our findings.

The table provides full MISE-comparison information about the two parametric and two nonparametric competitors under discussion, for a selection of sample sizes. It is convenient to view the MISE obtained by the parametric method as a benchmark, then dividing other MISE numbers with this benchmark, to avoid staring at numbers that all go to zero. Such a ratio below one means outperforming the direct parametric plug-in method, and vice versa. One immediate fact to note is that the parametric UMVU method always loses to the plug-in method, but with a relative amount that becomes smaller with n. The MISE for the UMVU method for $n = 3$ is actually infinite.

From the table we see that the kernel method with a normal kernel wins over the parametric method for $n \leq 14$, if the optimal bandwidth is available, and with the random bandwidth used in practice the verdict is $n \leq 9$. Corresponding findings for the Yepanechnikov kernel method are $n \leq 15$ and $n \leq 10$. The table in particular sheds light on the extra estimation noise for the kernel methods caused by having to estimate the best bandwidth from data.

Also of general interest are the implicit finite-sample corrections to the well-known 'normal reference rules' for bandwidth selection. Rather than using $1.059\,\widehat{\sigma}/n^{1/5}$ with a normal kernel one should stretch it further to the appropriate $b_n\,\widehat{\sigma}/n^{1/5}$; this will lead to a small reduction in MISE.

TABLE. *The* MISE *behaviour when estimating a normal density is summarised here, for two parametric and two nonparametric estimators. For each sample size n we give first the 'benchmark'* MISE *for the parametric plug-in estimator (column 2). The* MISE *numbers for the other methods are then given as ratios, dividing with the benchmark* MISE. *Column 3 gives this ratio for the UMVU estimator. Columns 5 and 6 give the* MISE*-ratio for the kernel method with the normal kernel, with respectively best possible and estimated best possible bandwidth. Similarly columns 8 and 9 give these ratios for the kernel method with the Yepanechnikov kernel, again with best possible and estimated best possible bandwidth. Columns 4 and 7 display the best constants b_n and c_n involved in the bandwidth selection procedures for the two kernel estimators.*

	parametric:		*normal kernel:*			*Yepanechnikov kernel:*		
n	*plug-in*	*UMVU*	b_n	*ratio*$_1$	*ratio*$_2$	c_n	*ratio*$_1$	*ratio*$_2$
3	0.23230	∞	1.2871	0.208	0.699	5.2822	0.209	0.727
4	0.11829	1.5095	1.2628	0.350	0.727	5.2177	0.349	0.747
5	0.07969	1.2110	1.2458	0.459	0.773	5.1737	0.455	0.786
6	0.06016	1.1223	1.2331	0.548	0.824	5.1411	0.541	0.830
7	0.04835	1.0822	1.2230	0.623	0.874	5.1156	0.614	0.874
8	0.04042	1.0602	1.2148	0.689	0.922	5.0949	0.677	0.918
9	0.03472	1.0466	1.2080	0.748	0.967	5.0776	0.733	0.960
10	0.03044	1.0375	1.2021	0.801	1.010	5.0628	0.784	0.9997
11	0.02710	1.0312	1.1970	0.849	1.050	5.0500	0.830	1.037
12	0.02441	1.0264	1.1925	0.894	1.088	5.0388	0.872	1.072
13	0.02222	1.0229	1.1885	0.935	1.124	5.0288	0.911	1.106
14	0.02038	1.0201	1.1849	0.973	1.157	5.0198	0.948	1.137
15	0.01883	1.0178	1.1816	1.009	1.189	5.0117	0.982	1.167
16	0.01749	1.0160	1.1786	1.043	1.220	5.0043	1.015	1.195
17	0.01633	1.0145	1.1759	1.075	1.249	4.9975	1.045	1.223
18	0.01532	1.0132	1.1734	1.106	1.276	4.9913	1.074	1.248
19	0.01443	1.0121	1.1711	1.135	1.303	4.9855	1.102	1.273
20	0.01363	1.0112	1.1689	1.163	1.328	4.9801	1.128	1.297
50	0.00513	1.0032	1.1368	1.694	1.824	4.8996	1.631	1.764
100	0.00252	1.0014	1.1190	2.150	2.259	4.8540	2.064	2.175
1000	0.00025	1.0001	1.0842	4.163	4.214	4.7617	3.983	4.032

For simplicity and convenience the figures are shown on pages 17 and 18

While the table gives numbers related to an overall comparison, Figures 1 and 2 give supplementary information on the pointwise level. Figure 1 conveys the essentials of our story. It compares the behaviour of the parametric plug-in and the kernel method with the Yepanechnikov kernel, for sample size 14, when estimating an unknown normal density. We see that the kernel method has a much bigger bias than the parametric one (the latter is almost zero), but that it nevertheless wins overall via its lower variance, in the central x-area. The figure also shows that the MISE-based findings are no artifacts of the particular overall criterion used; the kernel method wins in terms of non-controversial mean squared error in the essential central region $(-1.53, 1.53)$, except in the small interval $(-0.17, 0.17)$, and then only barely.

Figure 2 is concerned with another theme, that of comparing two kernel estimators. Both of them are popular in practice, and perhaps in particular the normal kernel one because of ease of computations and guaranteed smoothness. The only real claim of the Yepanechnikov kernel choice is asymptotic in nature; it minimises both pointwise MSE and overall MISE when n is large. It does have a slight advantage for most finite n too, as illustrated here, outperforming its smoother sister very modestly in terms of MSE in the central area.

Our calculations and results also make clear one point which perhaps has been unanswered in the literature: there cannot be a universal finite-sample theorem about the optimality of the Yepanechnikov kernel. The normal kernel does better for $n = 2$ and $n = 3$, in terms of best possible performance, and indeed for all sample sizes up to $n = 7$ in terms of performance with estimated bandwidths.

6. Concluding comments.

6.1. Other parametric families. We have demonstrated that the simple-minded nonparametric method can win, against odds, over the normal density estimator. The same would typically happen for other low-dimensional parametric families, and particularly so if the model is complex, because of the sampling variability in the parameter estimates. Any model containing the normal, with three or more parameters, would produce density estimators that a fortiori would have larger MSE's and MISE's than estimator (1.1) under normal assumptions, that is, the nonparametric (1.2) would win for even more sample sizes.

As an example, consider the estimator based on a natural three-parameter skew-extended normal family,

$$\widehat{f}(x) = \widehat{\gamma}\Phi\Big(\frac{x-\widehat{a}}{\widehat{b}}\Big)^{\widehat{\gamma}-1}\phi\Big(\frac{x-\widehat{a}}{\widehat{b}}\Big)\frac{1}{\widehat{b}},$$

with estimates coming from maximum likelihood. The motivation for this model is that $T_\gamma((X-a)/b)$ is standard normal for appropriate (a,b,γ), where $T_\gamma(y) = \Phi^{-1}(\Phi(y)^\gamma)$ is a natural skewness-inducing family of transformations. When $\gamma = 1$ things are normal. In general terms, via arguments given in Section 2.1,

$$n\,\mathrm{MISE} \to \int f(x,\theta)^2 u(x,\theta)^{\mathrm{t}} J(\theta)^{-1} u(x,\theta)\,\mathrm{d}x = \mathrm{Tr}\{J(\theta)^{-1}L(\theta)\}$$

as n grows, where $L(\theta) = \int f(x,\theta)^2 u(x,\theta)u(x,\theta)^{\mathrm{t}}\,\mathrm{d}x$. For this particular family some calculations show that the MISE is approximately $0.342/(n\sigma)$ for normal estimands, which is about 1.386 times bigger than the value $0.247/(n\sigma)$ for the normal density estimator (1.1). Extrapolating from a fuller version of the table of Section 5 leads one to the educated guess that the kernel methods will outperform the parametric density estimator for sample sizes at least up to 30.

6.2. Shrinking the parametric estimator. Observing that high variance is the main reason for losing the competition, one might consider shrinking the parametric estimator, that is, using $k\widehat{f}(x)$ instead for suitable k. This somewhat artificial trick can indeed give smaller MISE. We illustrate this working with the shrunk UMVU estimator $k\bar{f}(x)$, since calculations are a bit easier then. Its MISE becomes $k^2\text{MISE} + (k-1)^2 R(f)$, in terms of the MISE for the $\bar{f}$ estimator, and this is minimised for $k = R(f)/\{\text{MISE} + R(f)\}$. This is independent of (μ, σ) and so can be used for normal estimands, resulting in the smaller

$$\text{MISE}_k = \text{MISE}\, R(f)/\{R(f) + \text{MISE}\}.$$

As MISE goes down as $1/n$ the effect is small, only improving the parametric performance slightly.

6.3. Other bandwidth selectors. More ambitious h choice methods abound in the density estimation literature, and most of them will be more statistically variable than the normality-based rule-of-thumb type $a_n\widehat{\sigma}/n^{1/5}$ that we have employed. Thus higher exact MISE's than those arrived at in Section 4 will result. It is perfectly fair to use these rules-of-thumb here, however, when one aims at good performance for all normal estimands.

6.4. Other situations where nonparametric wins against odds. We have worked with density estimators to make our basic educational point, since the contrast between the two camps is so pedagogically conspicuous there. However, there are other and simpler situations where the same phenomenon, nonparametrics outperforming parametrics, can be observed.

Let $X_1, \ldots, X_n$ be i.i.d. from a lognormal (a, b^2) distribution, and suppose it is required to estimate the mean μ. Then the parametrically based proposal is $\widehat{\mu}_{\text{pm}} = \exp(\widehat{a} + \frac{1}{2}\widehat{b}^2)$, in that $\mu = \exp(a + \frac{1}{2}b^2)$ under model conditions. Here $\widehat{a}$ and $\widehat{b}$ are the natural estimators, say $\widehat{a} = \bar{A}$ and $\widehat{b}^2 = (n-1)^{-1}\sum_{i=1}^n (A_i - \bar{A})^2$, in terms of $A_i = \log X_i$, and these are normal (a, b^2). But a competitor is the nonparametric $\widehat{\mu}_{\text{nonpm}} = \bar{X}$.

The well-known asymptotic optimality result for maximum likelihood estimation entails in particular that the parametric must be better than $\bar{X}$ for all large n, and this is nicely illustrated here, in that calculations show

$$\frac{\text{Var}\,\bar{X}}{\text{Var}\,\widehat{\mu}_{\text{pm}}} \to \frac{\exp(b^2) - 1}{b^2 + \frac{1}{2}b^4} = \frac{b^2 + \frac{1}{2}b^4 + \frac{1}{6}b^6 + \cdots}{b^2 + \frac{1}{2}b^4} > 1.$$

But for moderate to small n the picture turns out to be different. We have $\text{MSE}(\bar{X}) = n^{-1}\exp(2a + b^2)\{\exp(b^2) - 1\}$, while calculations give

$$\text{MSE}(\widehat{\mu}_{\text{pm}}) = \exp(2a + b^2/n)\Big[\exp(b^2/n)\Big(1 - \frac{2b^2}{n-1}\Big)^{-(n-1)/2} - \Big(1 - \frac{b^2}{n-1}\Big)^{-(n-1)}\Big].$$

Here one quickly notices that $b < \{\frac{1}{2}(n-1)\}^{1/2}$ is required for the variance and MSE to exist at all for the parametric estimator, so for small n the nonparametric estimator is bound to win. This continues to be true also beyond the $n-1 > 2b^2$ requirement, as some analysis reveals. Thus for each b there is a $n_0(b)$ such that the nonparametric estimator wins over the parametric one, for each $n \le n_0(b)$, while the parametric indeed eventually wins, namely for each $n > n_0(b)$. Values of $n_0(b)$ for $b = 0.2, 0.4, 0.6, 0.8, 1.0, 1.2$ are for example found to be respectively 312, 87, 45, 31, 25, 22.

6.5. Biasing can be worthwhile. One may view our examples as part of a more general theme, which is that procedures that are biased but perhaps less variable sometimes are fruitful and strong competitors to the approximately unbiased ones. Cases in point include Stein estimation and more general shrinkage type methods. See also for example Hjort (1994, 1997) for analysis of procedures based on narrower versus wider parametric models.

6.6. Semiparametric attempts. Our results might serve as a warning against putting too much trust in intuition and large-sample results when judging the relative strengths of parametric and nonparametric methods. The findings may also have consequences for some of several recent semiparametric constructions that in various ways try to balance between the parametric and nonparametric frameworks. One relevant example is the mixture type estimator $f^*(x) = \widehat{p}\, f(x, \widehat{\theta}) + (1-\widehat{p})\, \widetilde{f}(x)$, worked with by Schuster and Yakowitz (1985), Olkin and Spiegelman (1987) and others; more general methods with similar characteristics have been developed from a Bayesian perspective in Hjort (1996) and Byholt (1996). The crux is to have a good procedure for choosing $\widehat{p}$, potentially also depending on x, from data. Intuition says perhaps that $\widehat{p}$ should be one or close to one if the parametric model looks good, and indeed such methods have been proposed based on goodness of fit criteria. But this issue is more involved in view of the findings of this paper; even if the model is perfect we should not have $\widehat{p}$ close to one, for small sample sizes.

References

Byholt, M. (1996). Nonparametric density estimation with a local likelihood. Cand. scient.-thesis, Department of Mathematics, University of Oslo.

Fryer, M.J. (1976). Some errors associated with the non-parametric estimation of density functions. *Journal of the Institute of Mathematics and its Applications* **18**, 371–380.

Hjort, N.L. (1994). The exact amount of t-ness that the normal model can tolerate. *Journal of the American Statistical Association* **89**, 665–675.

Hjort, N.L. (1996). Bayesian approaches to semiparametric density estimation (with discussion contributions). In *Bayesian Statistics 5*, proceedings of the Fifth International València Meeting on Bayesian Statistics (eds. J. Berger, J. Bernardo, A.P. Dawid, A.F.M. Smith), 223–253. Oxford University Press.

Hjort, N.L. (1997). Estimation in moderately misspecified models. *Journal of the American Statistical Association*, to appear.

Jones, M.C. (1991). The roles of ISE and MISE in density estimation. *Statistics and Probability Letters* **12**, 51–56.

Lehmann, E.L. (1983). *Theory of Point Estimation.* Wiley, New York.

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

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

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

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

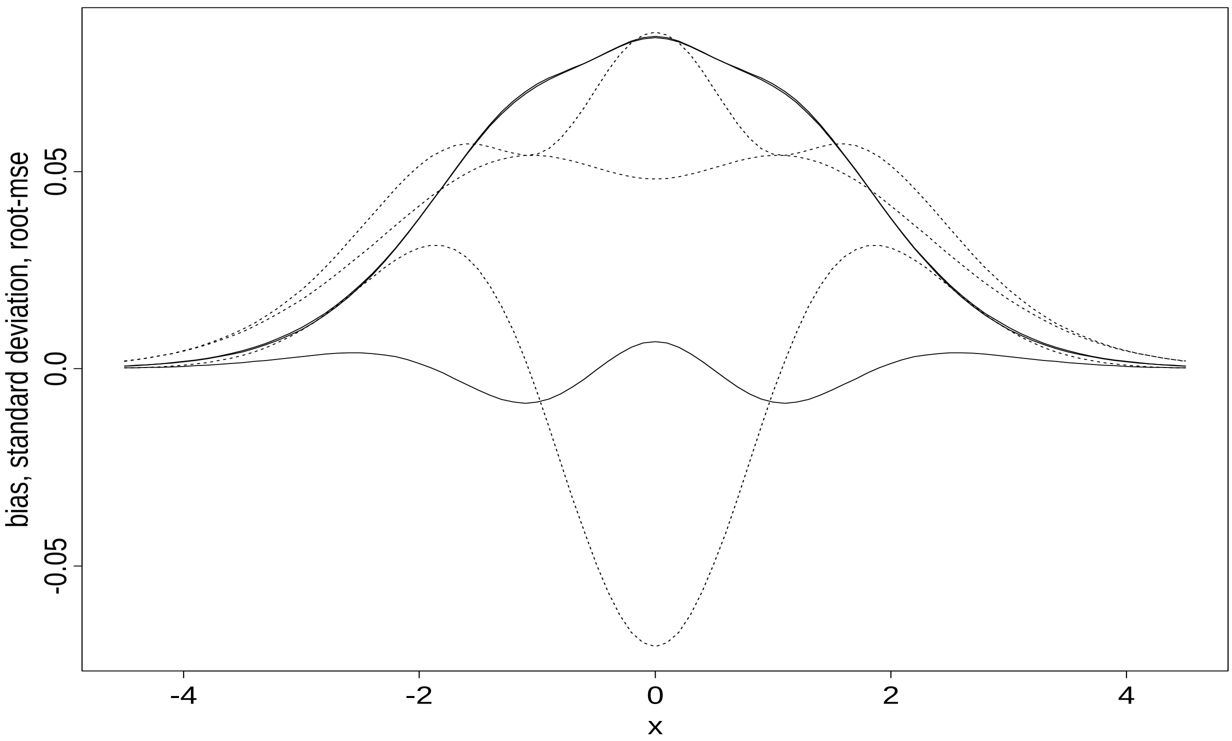

Figure 1:

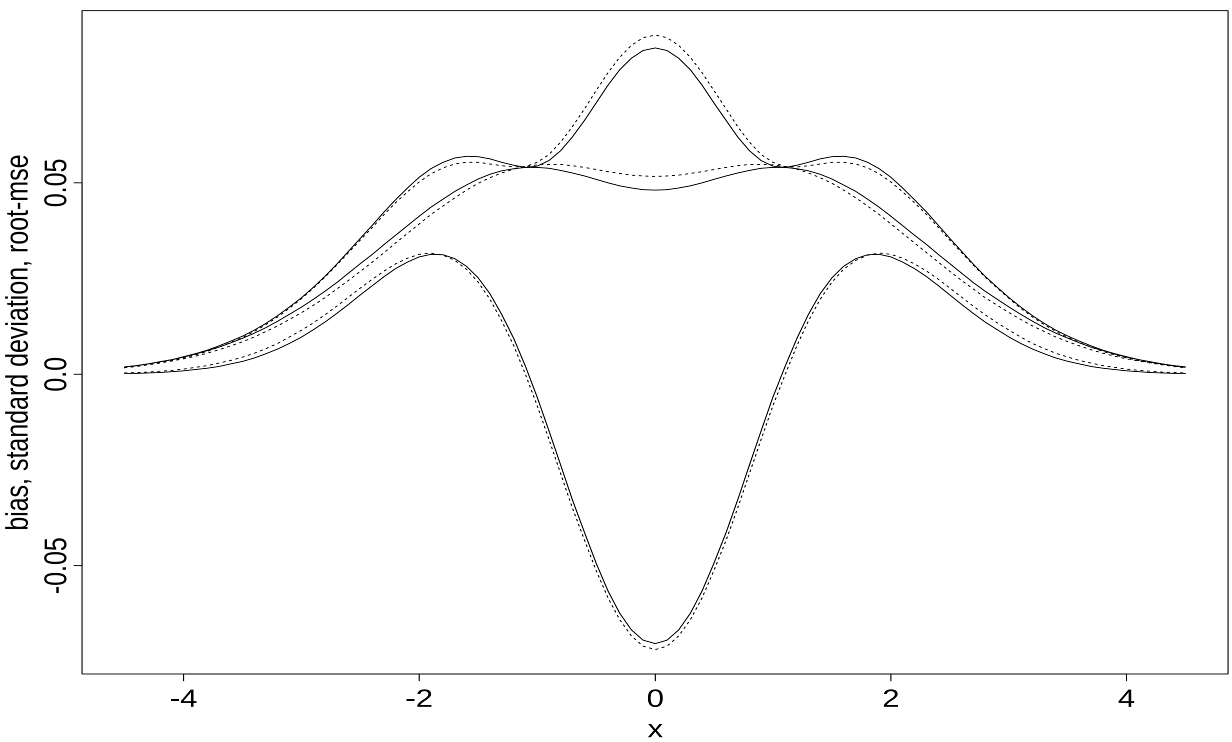

Figure 2: