Bernoulli **6**(1), 2000, 169–182

Skewing methods for two-parameter locally parametric density estimation

MING-YEN CHENG,[1] EDWIN CHOI,[1] JIANQING FAN[2] and PETER HALL[1]

[1]*Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia*
[2]*Department of Statistics, University of North Carolina, Chapel Hill NC 27599-3260, USA*

A 'skewing' method is shown to effectively reduce the order of bias of locally parametric estimators, and at the same time retain positivity properties. The technique involves first calculating the usual locally parametric approximation in the neighbourhood of a point x' that is a short distance from the place x where we wish to estimate the density, and then evaluating this approximation at x. By way of comparison, the usual locally parametric approach takes $x' = x$. In our construction, $x' - x$ depends in a very simple way on the bandwidth and the kernel, and not at all on the unknown density. Using skewing in this simple form reduces the order of bias from the square to the cube of bandwidth; and taking the average of two estimators computed in this way further reduces bias, to the fourth power of bandwidth. On the other hand, variance increases only by at most a moderate constant factor.

Keywords: bias reduction; kernel methods; local least squares; local likelihood; local linear methods; score function; weighted least squares

1. Introduction

There is a wide variety of high-order methods for reducing the bias of kernel-type density estimators. Jones and Signorini (1997) have reviewed the class of techniques that have bias of order h^4, where h denotes bandwidth, and showed that most are of one or other of six basic types: classical fourth-order kernel methods (e.g. Bartlett 1963); 'non-negativization' of fourth-order kernel methods (e.g. Terrell and Scott 1980); multiplicative bias correction methods (e.g. Jones *et al.* 1995); nonparametric transformation methods (e.g. Ruppert and Cline 1994); variable bandwidth methods (e.g. Abramson 1982); and variable location methods (e.g. Samiuddin and el-Sayyad 1990).

The approach of Samiuddin and el-Sayyad (1990) involves shifting the location of each data value by an amount proportional to the square of bandwidth, and recomputing the density estimate for the shifted data. Since the amount of shift depends on the unknown density (and its derivative) as well as on the datum and the bandwidth, computation of a pilot density estimator is required. In the present paper we suggest an entirely different shifting technique that does not require a pilot estimator. The new approach is closely allied to contemporary, locally parametric methods (e.g. Hjort and Jones 1996). It may also be regarded as an application to density estimation of 'skewing' methods suggested in the

context of nonparametric regression by Choi and Hall (1998). A potential advantage of the skewing approach is that it may be applied to a great many curve estimation problems, for example to generalized linear models, the Cox proportional likelihood model, and nonparametric estimation of survival functions, as well as to more conventional problems in nonparametric density estimation and regression.

Skewing methods involve calculating a nonparametric curve estimator in a traditional way by locally weighting in a region that is symmetrically placed on either side of the point x of interest; and then computing the final estimator in a skew way, at a point that is slightly to one side or the other of x. If the extent of offset is chosen appropriately (it depends only on the kernel and bandwidth), skewing reduces the order of bias, but incurs only a moderate increase in variance. By averaging two skewed estimators one can reduce bias from $O(h^2)$ to $O(h^4)$, still at the expense of only a constant-factor inflation of variance. In this paper we show that skewing may be used with general two-parameter locally parametric methods for density estimation, including methods based on a local likelihood or on local least squares.

Importantly, skewed estimators are guaranteed to be non-negative, since they are convex combinations of evaluations of non-negative functions $g(\cdot, \theta)$ determined by a parameter vector θ. Therefore, skewing reproduces the bias-reduction effect of high-order density estimation without risking the occurrence of negative estimates.

As a simple example of the method of skewing, let $\hat{f}_{\text{class}}$ be the classical kernel density estimator constructed using the standard normal kernel and bandwidth h. Then, using the skewing method for a local likelihood construction from an exponential model, we obtain the estimators $\tilde{f}_+$ and $\tilde{f}_-$ given by

$$\tilde{f}_\pm(x) = \hat{f}_{\text{class}}(x \pm h) \exp[\tfrac{1}{2} - \tfrac{1}{2}\{1 \pm h(\hat{f}_{\text{class}})'(x \pm h)\hat{f}_{\text{class}}(x \pm h)^{-1}\}^2], \tag{1.1}$$

where the $+$ signs and $-$ signs are chosen respectively. The following results are true: (a) both $\tilde{f}_+$ and $\tilde{f}_-$ have bias of size h^3 as estimators of f; (b) the estimator $\tilde{f} = \frac{1}{2}(\tilde{f}_+ + \tilde{f}_-)$ has bias of size h^4; and (c) each of $\tilde{f}_+$, $\tilde{f}_-$, $\tilde{f}$ has variance of size $(nh)^{-1}$. By way of comparison, $\hat{f}_{\text{class}}$ itself has variance of size $(nh)^{-1}$ but larger bias, of size h^2.

These improvements in performance are available for general kernels and general approaches to locally parametric estimation, for example those based on either local likelihood or local least squares. They allow mean square error to be reduced from order $n^{-4/5}$, in the case of standard kernel or locally parametric methods, to order $n^{-8/9}$, for $\tilde{f}$ and analogous estimators.

Local parametric methods in statistics have a particularly long history, if one includes among them local linear and local polynomial techniques in nonparametric regression. In this context the review paper of Hastie and Loader (1993), and monographs of Wand and Jones (1995) and Fan and Gijbels (1996), should be particularly mentioned. The recent surge of interest in locally parametric fitting for density and regression estimation is largely motivated by work of Copas (1995), Fan *et al.* (1996), Hjort and Jones (1996) and Loader (1996), which in turn prompted the present paper. In our presentation and discussion we have followed the development of Hjort and Jones, which applies in a particularly broad setting. High-order methods in curve estimation include work of Ruppert and Wand (1994), in the context of local high-order polynomial modelling in regression, as well as

contributions by Hjort and Jones (1996) and Loader (1996) to high-order local log-polynomial modelling in density estimation.

Section 2 will outline the methodology of skewing in the context of general locally parametric methods for density estimation. Details of technical arguments which justify the claims made there will be deferred to Section 4. Numerical properties of skewing for locally parametric density estimation will be illustrated in Section 3.

2. Methodology

2.1. Locally parametric methods

Let $X_1, \ldots, X_n$ denote a random sample from a distribution with density f, which we wish to estimate. We follow the general prescription of Hjort and Jones (1996) for locally parametric methods, based on two-parameter fits. Let $g(\cdot, \theta)$ be a family of two-parameter functions, indexed by $\theta = (\theta^{(1)}, \theta^{(2)})^{\mathrm{T}}$, which we wish to fit to data in a neighbourhood of x. Hjort and Jones suggest first defining the parameter estimator $\hat{\theta} = \hat{\theta}(x)$ as the solution in θ of the equation

$$n^{-1} \sum_{i=1}^{n} K_h(x - X_i) v_j(x, X_i, \theta) - \int K_h(x - t) v_j(x, t, \theta) g(t, \theta) \, \mathrm{d}t = 0, \tag{2.1}$$

where $K_h(t) = h^{-1} K(t/h)$, K is the kernel function (here taken to be either the standard normal density or a symmetric, non-negative, compactly supported density), h is the bandwidth, and $v_j(x, t, \theta)$ for $j = 1, 2$ is a generalized two-parameter score function. See (2.5) below for a 'population version' of (2.1). Hjort and Jones (1996), noting similar methods suggested by Copas (1995) and Loader (1996), take their estimator of f to be $\hat{f}(x) = g\{x, \hat{\theta}(x)\}$. Note that $\hat{f}$ need not integrate to one.

One of the very attractive features of Hjort and Jones's approach is its considerable generality, obtained partly through general interpretation of the score function. For example, taking

$$v_j(x, t, \theta) = (\partial / \partial \theta^{(j)}) \log g(t, \theta) \quad \text{or} \quad v_j(x, t, \omega) = (\partial / \partial \theta^{(j)}) g(t, \theta), \tag{2.2}$$

we obtain a local likelihood estimator or a local least-squares estimator, respectively. As these examples suggest, it is typically true that the dependence of $v_j(x, t, \theta)$ on x is degenerate.

Hjort and Jones argue that in this general setting, variance and bias admit the following asymptotic approximations:

$$\operatorname{var}\{\hat{f}(x)\} = (nh)^{-1} \kappa_1 f(x) + o\{(nh)^{-1}\}, \tag{2.3}$$

$$\begin{aligned} \mathrm{E}\{\hat{f}(x)\} - f(x) &= g\{x, \theta_0(x)\} - f(x) + O\{(nh)^{-1}\} \\ &= \tfrac{1}{2} \kappa_2 h^2 [f''(x) - g''\{x, \theta_0(x)\}] + O\{h^4 + (nh)^{-1}\}, \end{aligned} \tag{2.4}$$

as $h \to 0$ and $nh \to \infty$, where $\kappa_1 = \int K^2$, $\kappa_2 = \int t^2 K(t)\,\mathrm{d}t$, $g^{(j)}(x, \theta)$ (or g with j dashes) denotes $(\partial/\partial x)^j g(x, \theta)$, and $\theta_0(y) = \theta_0(y, h)$ is the solution in θ of the equation

$$\int K_h(y - t) v_j(y, t, \theta)\{f(t) - g(t, \theta)\}\,\mathrm{d}t = 0, \qquad \text{for } j = 1, 2. \tag{2.5}$$

We assume that, for each y and all sufficiently small h, $\theta_0(y)$ exists and is unique. When we intend $h = 0$ in $\theta_0(y)$, we write it as $\theta_0(y, 0)$; in all other cases, h is non-zero. Other regularity conditions for (2.3) and (2.4) will be discussed in Section 2.3. They allow regular mathematical expectations to be used, rather than simply expectations in the limiting distribution of $\hat{f}(x)$.

2.2. Skewing

Following standard practice in local curve fitting, Hjort and Jones (and others working on locally parametric methods) compute $\hat{f}$ symmetrically. That is, they weight data on either side of x in a symmetric way, and calculate $\hat{f}$ at the 'centre' of the weights. Skewing involves using symmetric weights at an off-centre point x', but nevertheless calculating the estimator at x. Thus, we replace $\hat{f}(x) = g\{x, \hat{\theta}(x)\}$ by $\hat{f}(x|x') = g\{x, \hat{\theta}(x')\}$. In the general setting of Section 2.1, using this method with $x' = x_\pm \equiv x \pm \kappa_2^{1/2} h$ (for either choice of the $+$ and $-$ signs) produces estimators $\tilde{f}_\pm(x) = g\{x, \hat{\theta}(x_\pm)\}$ whose bias is $O\{h^3 + (nh)^{-1}\}$ rather than $O\{h^2 + (nh)^{-1}\}$. Using the symmetric convex combination $\tilde{f} = \frac{1}{2}(\tilde{f}_+ + \tilde{f}_-)$ reduces bias further, to $O\{h^4 + (nh)^{-1}\}$. More generally, employing the estimator

$$\tilde{f}_\lambda(x) = (2\lambda + 1)^{-1}\{\lambda \hat{f}(x|x + lh) + \hat{f}(x|x) + \lambda \hat{f}(x|x - lh)\},$$

where $0 \leqslant \lambda < \infty$ and

$$l = l(\lambda) = \{(1 + 2\lambda)\kappa_2/(2\lambda)\}^{1/2}, \tag{2.6}$$

also reduces bias to $O\{h^4 + (nh)^{-1}\}$. (Note that $\tilde{f} = \tilde{f}_\infty$.) Thus, we have

$$\mathrm{E}(\tilde{f}_\pm) = f + O\{h^3 + (nh)^{-1}\}, \qquad \mathrm{E}(\tilde{f}) = f + O\{h^4 + (nh)^{-1}\},$$

$$\mathrm{E}(\tilde{f}_\lambda) = f + O\{h^4 + (nh)^{-1}\}. \tag{2.7}$$

The variance remains at order $(nh)^{-1}$ throughout these manipulations. Indeed, under regularity conditions implicit in Hjort and Jones (1996) (see, for example, (2.10) below),

$$\mathrm{var}(\tilde{f}_\pm) \sim (nh)^{-1}(\kappa_1 + \kappa_2^{-1}\kappa_3) f, \qquad \mathrm{var}(\tilde{f}_\lambda) \sim (nh)^{-1} V(\lambda) f, \tag{2.8}$$

as $h \to 0$ and $n \to \infty$ in such a manner that $nh \to \infty$, where $\kappa_3 = \int t^2 K(t)^2\,\mathrm{d}t$ and

$$V(\lambda) = (2\lambda+1)^{-2}\Big[(2\lambda^2+1)\kappa_1 + (6\lambda+1)\int K(u-l)K(u)\,\mathrm{d}u$$

$$+\tfrac{1}{2}(4\lambda+1)^2\int K(u-l)K(u+l)\,\mathrm{d}u$$

$$+\lambda(2\lambda+1)\kappa_2^{-1}\int u^2\{K(u)^2 - K(u-l)K(u+l)\}\,\mathrm{d}u\Big].$$

Formula (2.8) for $\mathrm{var}(\tilde{f}_\lambda)$ holds when $\lambda = \infty$, so that $\mathrm{var}(\tilde{f}) \sim (nh)^{-1}V(\infty)f$.

These are the same variances that arise in skewed local linear approximation in nonparametric regression (Choi and Hall, 1998). That is to be expected, given the interpretation of nonparametric density estimation as regression with Poisson-distributed errors. The size of $V(\lambda)$, for $0 \leqslant \lambda \leqslant \infty$, is discussed at length by Choi and Hall. Those authors show that, depending on choice of K and λ, $V(\lambda)$ can actually be smaller than κ_1, although for most values of λ it is larger, up to 39% larger in the case of the normal kernel.

To treat the particular case where $g(y, \theta) = \theta^{(1)}\exp\{(y-x)\theta^{(2)}\}$, first define ψ to be the moment generating function corresponding to the density K, put $A_k(x) = n^{-1}\sum_i K_h(x - X_i)(X_i - x)^k$ and note that $A_0 = \hat{f}_{\mathrm{class}}$ (the classical kernel density estimator), and let v_j be given by the first formula in (2.2). Then $\hat{\theta}^{(1)}$, $\hat{\theta}^{(2)}$ are the solutions of the equations $A_0 = \hat{\theta}^{(1)}\psi(h\hat{\theta}^{(2)})$ and $A_1/A_0 = h\psi'(h\hat{\theta}^{(2)})/\psi(h\hat{\theta}^{(2)})$; and

$$\hat{f}(x|x') = A_0(x')\psi\{h\hat{\theta}^{(2)}(x')\}^{-1}\exp\{(x-x')\hat{\theta}^{(2)}(x')\}.$$

When K is the standard normal kernel we have $\psi(t) = \exp(t^2/2)$ and $\hat{\theta}^{(2)} = (\hat{f}_{\mathrm{class}})'/\hat{f}_{\mathrm{class}}$, and so for each constant c,

$$\hat{f}(x|x+ch) = \hat{f}_{\mathrm{class}}(x+ch)\exp[\tfrac{1}{2}c^2 - \tfrac{1}{2}\{c + h(\hat{f}_{\mathrm{class}})'(x+ch)\hat{f}_{\mathrm{class}}(x+ch)^{-1}\}^2], \qquad (2.9)$$

from which the estimators $\tilde{f}_\pm$, $\tilde{f}_\lambda$ and $\tilde{f} = \tilde{f}_\infty$ may be immediately constructed. Taking $c = 0$ in (2.9) gives the local log-linear density estimator of Hjort and Jones (1996) and Loader (1996).

Hjort and Jones comment that in this example, when $c = 0$ the parameter estimate $\hat{\theta}^{(2)}$ is 'only somewhat silently present'. That cannot be said of the case $c \neq 0$ in which we are interested. Those authors also argue that $\hat{\theta}^{(2)}$ might be computed separately from $\hat{\theta}^{(1)}$, using a larger bandwidth. Following that prescription here would destroy the bias-reduction properties of estimators constructed by skewing.

While the estimator at (2.9) was derived in the special case of the standard normal kernel, it is appropriate much more generally. Indeed, taking $\hat{f}_{\mathrm{class}}$ to be a general kernel estimator computed using a kernel with $\kappa_2 = 1$ (where, here and in the remainder of this paragraph, κ_j is interpreted for the kernel used to compute $\hat{f}_{\mathrm{class}}$), and putting $c = \pm 1$, the estimator $\hat{f}_\pm(x) = \hat{f}(x|x \pm ch)$ (with the right-hand side given by (2.9)) satisfies $\mathrm{E}(\hat{f}_\pm) = f + O\{h^3 + (nh)^{-1}\}$ and $\mathrm{var}(\hat{f}_\pm) \sim (nh)^{-1}(\kappa_1 + \kappa_3)f$. This is the analogue of (2.7) and (2.8) (taken there for $\tilde{f}_\pm$) in the case of $\hat{f}_\pm$. These results may be derived after little more than Taylor expansion. Likewise, the versions of (2.7) and (2.8), for linear combinations of

estimators such as $\hat{f}_{\pm}$ and giving rise to analogues of $\tilde{f}$ and $\tilde{f}_{\lambda}$, may be derived. Similarly, versions of (2.9) that arise for kernels other than the normal may be shown to produce a variety of new estimators which enjoy good bias-reduction properties, provided the kernel is sufficiently smooth. (The smoothness is needed in the Taylor expansion part of the argument.)

As is commonly the case with density estimators derived by locally parametric methods, $\tilde{f}_{+}$, $\tilde{f}_{-}$ and $\tilde{f}_{\lambda}$ do not necessarily integrate to one. Correcting the estimators by dividing by their respective integrals may improve finite-sample performance. For example, in the case of standard normal data, the improvement in mean integrated square error of $\tilde{f}$ is by 10% when $n = 100$, with smaller increases for other densities in our simulation study. Similar results were obtained by Jones *et al.* (1995) and Jones and Signorini (1997).

2.3. Assumptions on g and v_j

The properties required of the parametric model and score functions in the two-parameter case of Hjort and Jones (1996) are not stated explicitly there. Concise conditions are needed if the outline technical arguments in the present paper are to be clear, however, and so we shall be specific about them here.

Any successful candidate for g in a second-order locally parametric method has to be capable of capturing the full range of potential values of both f and its derivative. If g depends on its argument and parameters in a smooth way then this implies that, after a suitable reparametrization, it should be approximately linear in small neighbourhoods of any given point x:

$$g(y, \theta) = \omega^{(1)} + \omega^{(2)}(y - x) + O\{(y - x)^2\} \tag{2.10}$$

as $y \to x$. Furthermore, the transformation which takes θ to $\omega = (\omega^{(1)}, \omega^{(2)})^{\mathrm{T}}$ should be one-to-one and onto the whole of $(0, \infty) \times (-\infty, \infty)$. (The transformation will of course depend on x.) The differentiated forms of (2.10) must also be valid for as many derivatives of g (with respect to y and θ, with x held fixed) as are required for other aspects of the proof. For example, we need $g'(y, \theta) = \omega^{(2)} + O(|y - x|)$ as $y \to x$.

Of course, (2.10) is satisfied by all standard two-parameter models that are used in practice in locally parametric density estimation. In particular, if g is the log-linear model employed as an example in Section 2.2 then (2.10) holds with $\omega^{(1)} = \theta^{(1)}$ and $\omega^{(2)} = \theta^{(1)}\theta^{(2)}$; and if g is the normal model,

$$g(y, \theta) = (2\pi)^{-1/2}(\theta^{(2)})^{-1} \exp\{-\tfrac{1}{2}(\theta^{(2)})^{-2}(y - x - \theta^{(1)})^2\},$$

then (2.10) is valid with

$$\omega^{(1)} = (2\pi)^{-1/2}(\theta^{(2)})^{-1} \exp\{-\tfrac{1}{2}(\theta^{(1)}/\theta^{(2)})^2\}, \qquad \omega^{(2)} = \omega^{(1)}\theta^{(1)}(\theta^{(2)})^{-2}.$$

In the general formulation of locally parametric methods suggested by Hjort and Jones (1996), no explicit connection is required between the score functions v_j and the model g. Nevertheless, their arguments implicitly ask that

for each x, each of the conditions $v_j\{x, x, \theta_0(x, 0)\} \neq 0$ and $(\partial/\partial t)v_j\{x, t, \theta_0(x, 0)\}|_{t=x} \neq 0$ holds for some $j = j(x)$ (not necessarily the same j in both cases), (2.11)

where, as before, $\theta_0(y, h)$ is defined as the solution of equation (2.5). In particular, without the second part of (2.11), $g'\{x, \hat{\theta}(x)\}$ does not approximate $f'(x)$. Assuming that (2.10) holds and v_j is given by one of the formulae at (2.2), (2.11) is valid if and only if $\omega^{(1)}$, $(\partial/\partial\theta^{(j)})\omega^{(1)}$ and $(\partial/\partial\theta^{(j)})\omega^{(2)}$ are non-zero when evaluated at $\theta = \theta_0(x, 0)$.

If one fits only densities in a uniformly bounded two-parameter class $\mathscr{G} = \{g(\cdot, \theta)\colon \theta \in \Theta\}$, that is, one satisfying

$$\sup_x \sup_{\theta: g(x,\theta)\in\mathscr{G}} g(x, \theta) < \infty, \tag{2.12}$$

then all the bias and variance formulae in Sections 2.1 and 2.2 (for example, (2.7) and (2.8), the latter provided that $f(x) \neq 0$) are correct as they stand, for the *actual* bias and variance. They do not represent simply the bias and variance of asymptotic distributions of $\hat{f}$, $\tilde{f}_\pm$, $\tilde{f}_\lambda$ or $\tilde{f}$. This is in contradistinction to the case of local polynomial methods in nonparametric regression, where the actual bias and variance are typically not well defined.

To establish this result we need a mild additional condition on the bandwidth. It is sufficient to ask that for some $\delta > 0$ and all sufficiently large n, $h(n) \geqslant n^{-1+\delta}$. In company with assumptions already made, for example the condition that K be either compactly supported or the standard normal kernel (see Section 2.1), this may be shown to imply that for all ε, $\lambda > 0$, the event $\mathscr{E} = \{|\tilde{f}_\pm(x) - f(x)| > \varepsilon\}$ satisfies

$$P(\mathscr{E}) = O(n^{-\lambda}). \tag{2.13}$$

Standard arguments that would be employed to establish versions of (2.7) and (2.8) when expectations are taken in asymptotic distributions, may be used to show that (2.7) and (2.8) hold when, on the left-hand sides, the estimator $\tilde{f}_\pm$ (for example) is replaced by $\tilde{f}_\pm I(\tilde{\mathscr{E}})$, where $\tilde{\mathscr{E}}$ denotes the complement of $\mathscr{E}$ and $I(\tilde{\mathscr{E}})$ is the indicator of $\tilde{\mathscr{E}}$. Since, by (2.12), $0 \leqslant \tilde{f}_\pm \leqslant C$ for a finite constant C, then by (2.13), the mean and mean square error (in fact, any finite moment) of $\tilde{f}_\pm - \tilde{f}_\pm I(\tilde{\mathscr{E}}) = \tilde{f}_\pm I(\mathscr{E})$ equal $O(n^{-\lambda})$ for all $\lambda > 0$. This allows us to make the transition from the versions of (2.7) and (2.8) for $\tilde{f}_\pm I(\tilde{\mathscr{E}})$, to the actual formulae (2.7) and (2.8). The cases of $\tilde{f}$ or $\tilde{f}_\lambda$, rather than $\tilde{f}_\pm$, may be treated similarly.

3. Numerical results

The simulation study is only summarized here; further details are available from the authors. A wide range of other high-order kernel-type estimators, with bias $O(h^4)$, is compared numerically by Jones and Signorini (1997), and so we limit ourselves to comparing $\tilde{f}_\pm$ and $\tilde{f}$ with (a) a standard second-order kernel estimator $\hat{f}_{\text{class}}$, using the standard normal kernel ϕ, (b) a fourth-order kernel estimator $\hat{f}_{\text{class},4}$, based on the kernel $K_{(4)}(x) = \frac{1}{2}(3 - x^2)\phi(x)$, and (c) the shift-type estimator proposed by Samiuddin and el-Sayyad (1990). The latter is

arguably the pre-existing method that is most closely related to our own. We shall choose the version of Samiuddin and el-Sayyad's estimator employed by Jones and Signorini (1997),

$$\hat{f}_{\mathrm{SS}}(x) = (nh)^{-1} \sum_{i=1}^{n} K[h^{-1}\{x - X_i - \tfrac{1}{2}h^2\kappa_2(\tilde{f}_{\mathrm{class}})'(X_i)/\hat{f}_{\mathrm{class}}(X_i)\}],$$

where K is taken as ϕ, and $\hat{f}_{\mathrm{class}}$ is as defined in Section 1 – that is, $\hat{f}_{\mathrm{class}}$ is a standard kernel estimator with kernel ϕ and bandwidth h. So as to directly compare $\hat{f}_{\mathrm{class},4}$ and $\hat{f}_{\mathrm{SS}}$ with $\tilde{f}$, we renormalized the latter by dividing by $\int \tilde{f}$. Note that both $\hat{f}_{\mathrm{class},4}$ and $\hat{f}_{\mathrm{SS}}$ integrate to 1. We do not alter the notation $\tilde{f}$, however, in the discussion below. Our simulation results indicate that renormalization has minimal effects on $\tilde{f}_{\pm}$, and hence we do not pursue normalization of $\tilde{f}_{\pm}$ in our study.

We chose five densities, $f_1, \ldots, f_5$, namely 'Gaussian', 'skewed unimodal', 'bimodal', 'separated bimodal' and 'asymmetric bimodal' as described by Marron and Wand (1992). We used sample sizes $n = 50$, 100, 200, 500 and 1000, although only results for $n = 100$ and for the 'Gaussian' and 'skewed unimodal' densities will be discussed in detail. Results for other values of n and other densities are similar, and their mean integrated square error (MISE) performances are summarized in Figure 3.1. We employed the local log-linear parametric model $g(y, \theta) = \theta^{(1)} \exp\{(y - x)\theta^{(2)}\}$ because of its popularity (e.g. Hjort and Jones 1996; Loader 1996), its simplicity (e.g. the availability of the closed-form estimator (2.9)), and the central position occupied by local linear methods in contemporary curve estimation.

To calculate MISE curves we used a grid of bandwidths consisting of 51 logarithmically equally spaced points in the interval [0.1, 1.0]. Each MISE curve was obtained by averaging 1000 replications of integrated square error (ISE) curves. For each bandwidth h in the grid, we calculated the pointwise square errors of the estimates at 201 equally spaced points on the interval $[-3, 3]$. The trapezoidal rule was employed to evaluate ISE. The MISE curves for $n = 100$ and for the densities f_1 and f_2 are depicted in Figures 3.2 and 3.3 respectively. For the sake of clarity, only bandwidths in the interval [0.15, 1.0] are displayed. Vertical lines are drawn through the minimizers of the MISE curves, and have the same line types as the respective curves.

For the Gaussian and skewed unimodal densities, the estimator $\tilde{f}$ performs better than $\tilde{f}_{\pm}$ in MISE terms throughout the range of bandwidths considered. In the case of small bandwidths, the MISE curves for the standard kernel estimator $\hat{f}_{\mathrm{class}}$ and the standard locally parametric estimator $\hat{f}_0 = \hat{f}(x|x)$ are almost identical, whereas discrepancies are noted for large h. This is to be expected since, as mentioned by Hjort and Jones (1996), for small to moderate h the locally parametric estimator utilizes primarily local properties of the model g, and hence the estimation method is essentially nonparametric. As h increases the method becomes more parametric, and the difference between MISE curves is best explained by errors in approximating the true density by the model. Note, however, that the minimum MISEs for $\hat{f}_{\mathrm{class}}$ and $\hat{f}_0$ are approximately equal in all our simulations.

The performance of $\tilde{f}_{\pm}$ improves on that of $\hat{f}_0$ for large n, although not necessarily for smaller sample sizes. This is illustrated in Figures 3.2 and 3.3, where the minimum MISE for $\hat{f}_0$ is seen to be less than that for $\tilde{f}_{\pm}$ in the case $n = 100$. From the theory, $\tilde{f}_{\pm}$

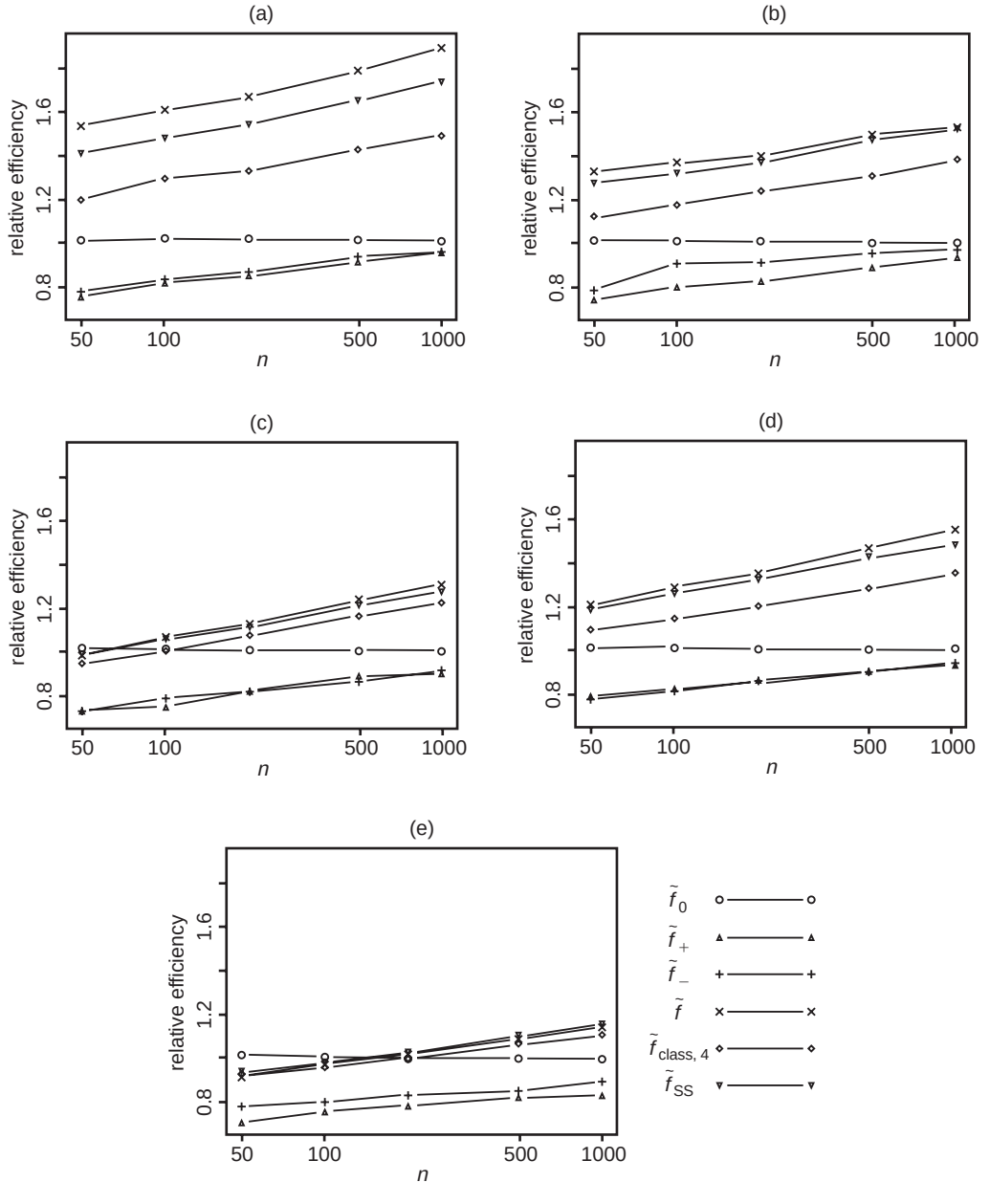

Figure 3.1. Relative efficiencies. The vertical axis gives ratios of the minimum MISE values of $\hat{f}_0$, $\tilde{f}_+$, $\tilde{f}_-$, $\tilde{f}$, $\hat{f}_{\text{class},4}$ and $\hat{f}_{\text{SS}}$ relative to the standard kernel estimator $\hat{f}_{\text{class}}$, calculated for five of the 15 Gaussian mixture densities of Marron and Wand (1992): (a) Gaussian; (b) skewed unimodal; (c) bimodal; (d) separated bimodal; (e) asymmetric bimodal. The vertical axes in all panels have the same range and scale.

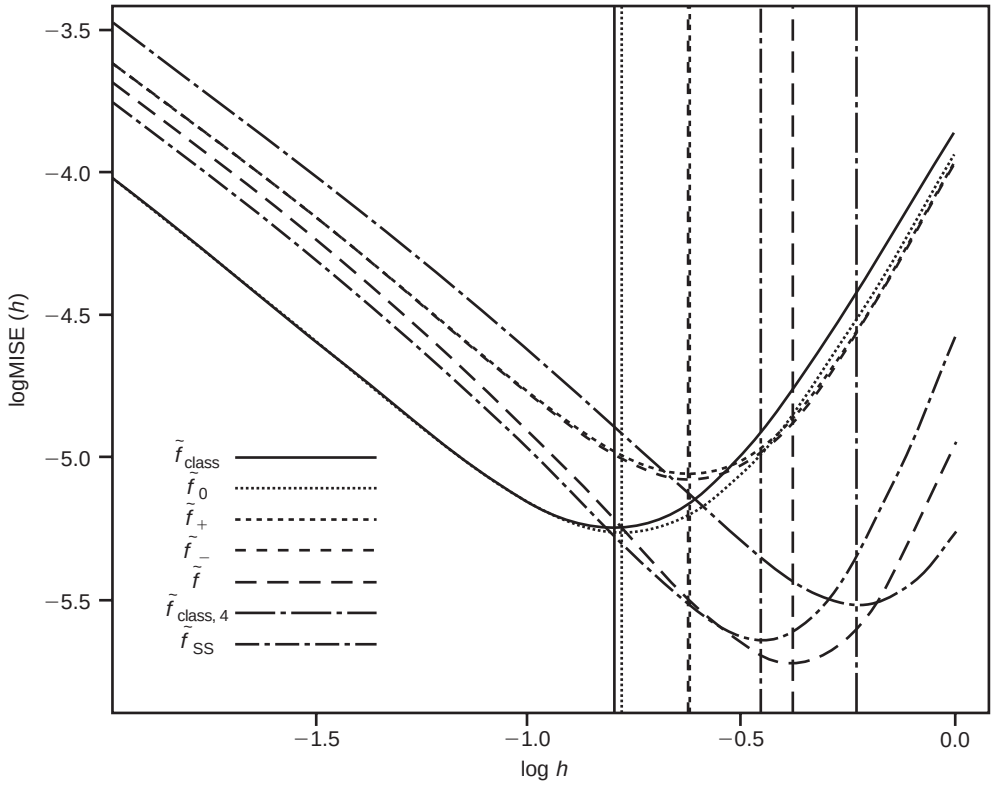

Figure 3.2. Comparison of MISE curves for the Gaussian density f_1 with sample size $n = 100$. The figure is plotted on a log–log scale, for the sake of clarity. The line types in the legend correspond to the estimators $\hat{f}_{\text{class}}$, $\hat{f}_0$, $\tilde{f}_+$, $\tilde{f}_-$, $\tilde{f}$, $\hat{f}_{\text{class,4}}$ and $\hat{f}_{\text{SS}}$.

outperforms $\hat{f}_{\text{class}}$ when the sample size n is large enough, and this is demonstrated by the increasing efficiency as a function of n in Figure 3.1. Nevertheless, the asymmetric quality of $\tilde{f}_\pm$ is reflected clearly in the MISE performance when estimating the skewed unimodal and asymmetric bimodal densities.

The substantial improvements offered by $\tilde{f}$ are clear even for the small sample size $n = 100$. Among the high-order methods we compared, $\tilde{f}$ has a better overall performance than $\tilde{f}_{\text{SS}}$, and both estimators have greater efficiency than $\hat{f}_{\text{class,4}}$ in all cases, as indicated in Figure 3.1. Moreover, our skewed estimator $\tilde{f}$ has advantages over $\hat{f}_{\text{SS}}$ from a computational viewpoint. To appreciate why, assume that the data are not binned, and that the density is estimated at M grid points. Then the number of kernel evaluations needed to compute $\tilde{f}$ is of size $O(nM)$, whereas that for $\hat{f}_{\text{SS}}$ is of size $O(n^2 M)$.

The improvements in performance offered by skewing methods over standard kernel estimators are generally apparent in both the body of the distribution and the tails. For

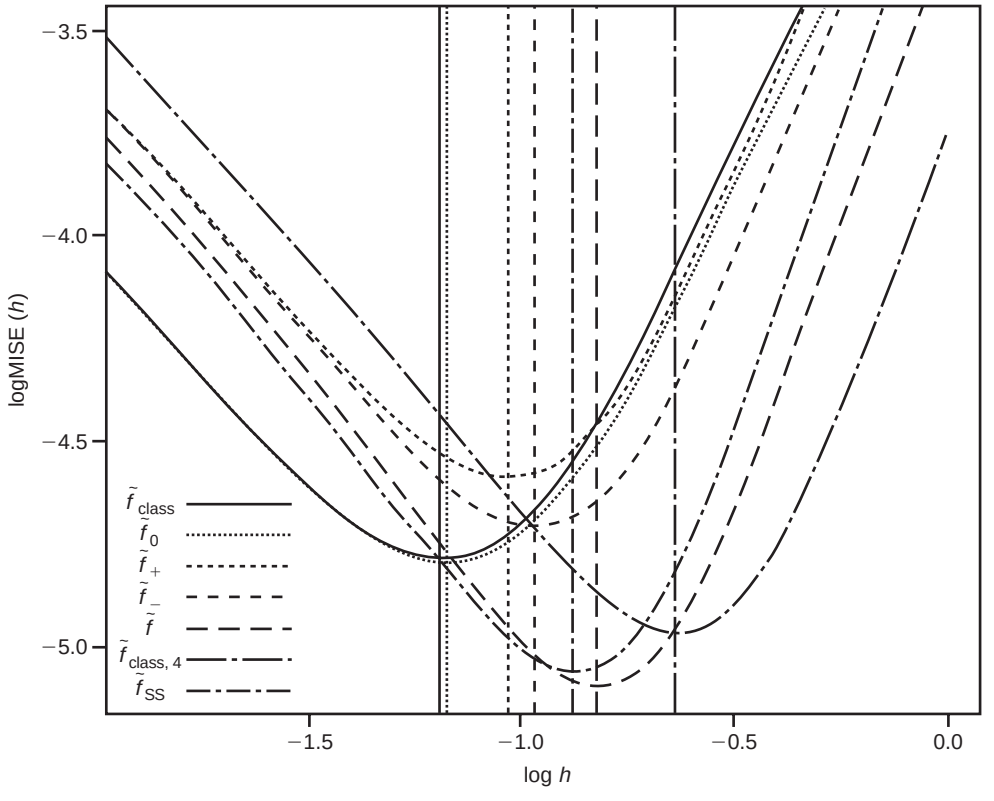

Figure 3.3. Comparison of MISE curves for the skewed unimodal density f_2 with sample size $n = 100$. Again, the figure is plotted on a log–log scale. The line types in the legend correspond to the estimators $\hat{f}_{\text{class}}$, $\hat{f}_0$, $\tilde{f}_+$, $\tilde{f}_-$, $\tilde{f}$, $\hat{f}_{\text{class,4}}$ and $\hat{f}_{\text{SS}}$.

example, if $n = 100$ and the target density is normal (e.g. density f_1); and if we use globally optimal bandwidths in each estimator, then the pointwise mean square error (PMSE) of $\hat{f}_{\text{class}}$ exceeds that of $\tilde{f}$ over the range ± 3 standard deviations from the mean, and the excess of PMSE for $\hat{f}_{\text{class}}$ relative to that for $\tilde{f}$ equals 86%, 36% and 54% at 0, 1 and 2 standard deviations from the mean, respectively. The relatively low figure at 1 standard deviation reflects the fact that the bias of $\hat{f}_{\text{class}}$ is of order h^4 there.

In practice the bandwidth would be selected empirically, but in a comparison of estimators using different bandwidth choice methods, the performance of estimators would be confounded with the performance of bandwidth selectors. A cross-comparison, of both bandwidth selectors and estimators, is beyond the scope of this paper. However, we note that standard bandwidth choice methods, such as cross-validation, are applicable without change to our skewing estimators.

4. Outline of technical arguments

4.1. Biases of skewed estimators

Here we show that $\tilde{f}_{\pm}$, $\tilde{f}_{\lambda}$ and $\tilde{f}$ have biases of orders h^3, h^4 and h^4, respectively. Assume that (2.11) and (2.12) hold, and that f, g, v_1, v_2 have four bounded derivatives with respect to each variable. (Only three derivatives are required to derive the $O\{h^3 + (nh)^{-1}\}$ bias of $\tilde{f}_{\pm}$.) Arguing as in Hjort and Jones (1996), we may deduce that for any constant c the bias of $g\{x, \hat{\theta}(x + ch)\}$, as an estimator of $f(x)$, equals

$$g\{x, \theta_0(x + ch)\} - f(x) + O\{(nh)^{-1}\}, \tag{4.1}$$

where $\theta_0(y)$, assumed uniquely defined in a neighbourhood of x, is the solution of (2.5). Put $\delta = ch$ and Taylor-expand $\gamma(\delta) \equiv g\{x, \theta_0(x + \delta)\}$ around $\delta = 0$, as a power series in δ. The coefficient of δ in the expansion equals

$$\theta^{(1)'}(x) g_{10}\{x, \theta_0(x)\} + \theta^{(2)'}(x) g_{01}\{x, \theta_0(x)\} = (\partial/\partial x) g\{x, \theta_0(x)\} - g'\{x, \theta_0(x)\}, \tag{4.2}$$

where

$$g_{jk}(y, \theta) = \{\partial^{j+k}/(\partial\theta^{(1)})^j (\partial\theta^{(2)})^k\} g(y, \theta).$$

To evaluate the right-hand side of (4.2) observe that, on setting $y = x$ and $\theta = \theta_0(x)$ in (2.5), Taylor-expanding, and differentiating with respect to x the left-hand side, we obtain

$$(\partial/\partial x)(v_j\{x, x, \theta_0(x)\}[f(x) - g\{x, \theta_0(x)\}]) + O(h^2) = 0.$$

Using the product rule to evaluate the differential on the left-hand side, employing (2.4) to prove that the term $f(x) - g\{x, \theta_0(x)\}$ that forms part of the result equals $O(h^2)$, and choosing j so that $v_j\{x, x, \theta_0(x, 0)\} \neq 0$ (see (2.11)), we deduce that

$$f'(x) - (\partial/\partial x) g\{x, \theta_0(x)\} = O(h^2). \tag{4.3}$$

Again Taylor-expanding the left-hand side of (2.5) with $\theta = \theta_0(x) = \theta_0(x, h)$, this time not differentiating but choosing j such that

$$(\partial/\partial t) v_j\{x, t, \theta_0)\}|_{t=x} \neq 0 \tag{4.4}$$

(see (2.11)), we obtain

$$\begin{aligned} v_j\{x, x, \theta_0(x)\}[f(x) - g\{x, \theta_0(x)\}] \\ + \tfrac{1}{2}\kappa_2 h^2 (v_j\{x, x, \theta_0(x)\}[f''(x) - g''\{x, \theta_0(x)\}] \\ + v_j''\{x, x, \theta_0(x)\}[f(x) - g\{x, \theta_0(x)\}] \\ + 2v_j'\{x, x, \theta_0(x)\}[f'(x) - g'\{x, \theta_0(x)\}]) + O(h^4) = 0, \end{aligned}$$

where $v_j^{(k)}(x, t, \theta)$ (or v_j with k dashes) denote $(\partial/\partial t)^k v_j(x, t, \theta)$. Using (2.4), we deduce that the left-hand side equals

$$\kappa_2 h^2 v_j'\{x, x, \theta_0(x)\}[f'(x) - g'\{x, \theta_0(x)\}] + O(h^4),$$

whence it follows from (4.4) that

$$f'(x) - g'\{x, \theta_0(x)\} = O(h^2). \tag{4.5}$$

Combining (4.3) and (4.5) we see that the right-hand side of (4.2) equals $O(h^2)$. Hence, the term in δ in the Taylor expansion of $\gamma(\delta)$ is of size $O(\delta h^2) = O(h^3)$.

Next we deal with the coefficient of $\frac{1}{2}\delta^2$, which may be shown by an analogue of the argument leading to (4.2) to equal

$$(\partial/\partial x)^2 g\{x, \theta_0(x)\} + g''\{x, \theta_0(x)\} - 2(\partial/\partial x)g'\{x, \theta_0(x)\}. \tag{4.6}$$

Formally, differentiating (2.4), we deduce that $(\partial/\partial x)^2[g\{x, \theta_0(x)\} - f(x)] = O(h^2)$. This result may be obtained rigorously by making minor modifications to arguments of Hjort and Jones (1996). Refining the argument leading to (4.5), we may identify the right-hand side and show that, after one differentiation, it is still of order $O(h^2)$. Therefore, $f''(x) - (\partial/\partial x)g'\{x, \theta_0(x)\} = O(h^2)$. Combining the last two results, we see that the quantity at (4.6) equals $g''\{x, \theta_0(x)\} - f''(x) + O(h^2)$. From this formula for the coefficient of $\frac{1}{2}\delta^2$ in the Taylor expansion of $\gamma(\delta)$, and from the result in the previous paragraph for the coefficient of δ, we deduce that

$$g\{x, \theta_0(x + \delta)\} - g\{x, \theta_0(x)\} = \tfrac{1}{2}\delta^2[g''\{x, \theta_0(x)\} - f''(x)] + O(h^3). \tag{4.7}$$

Using (2.4) and (4.7), we find that the quantity at (4.1) (equal to the bias of $g\{x, \hat{\theta}(x + ch)\}$) equals

$$\tfrac{1}{2}h^2(\kappa_2 - c^2)[f''(x) - g''\{x, \theta_0(x)\}] + O\{h^3 + (nh)^{-1}\}.$$

Since $\tilde{f}_\pm$ is defined by taking $c = \pm\kappa_2^{1/2}$ in $g\{x, \hat{\theta}(x + ch)\}$ then, for either choice of the $+$ and $-$ signs, its bias equals simply $O\{h^3 + (nh)^{-1}\}$.

Appealing to symmetry properties when evaluating Taylor expansions, it may be proved by a similar but longer argument than that leading to (4.7) that

$$g\{x, \theta_0(x + ch)\} + g\{x, \theta_0(x - ch)\} - 2f(x) = h^2(\kappa_2 - c^2)[f''(x) - g''\{x, \theta_0(x)\}] + O(h^4).$$

From this formula and (2.4) we deduce that

$$\begin{aligned}(2\lambda + 1)^{-1}&(\lambda[g\{x, \theta_0(x + ch)\} + g\{x, \theta_0(x - ch)\}] + g\{x, \theta_0(x)\}) - f(x)\\ &= \frac{h^2}{2}(2\lambda + 1)^{-1}\{(2\lambda + 1)\kappa_2 - 2\lambda c^2\}[f''(x) - g''\{x, \theta_0(x)\}] + O(h^4).\end{aligned} \tag{4.8}$$

The left-hand side equals the bias of $\hat{f}_\lambda$, up to terms of order $(nh)^{-1}$. Taking $c = l$, where l is defined by (2.6), the right-hand side of (4.8) equals $O(h^4)$. Hence, $\mathrm{E}(\tilde{f}_\lambda) - f = O\{h^4 + (nh)^{-1}\}$. Similarly, we may prove that the bias of $\tilde{f} = \tilde{f}_\infty$ equals $O\{h^4 + (nh)^{-1}\}$.

4.2. Variance of skewed estimators

We assume (2.10), and also (without loss of generality) that the original parametrization was $\theta = \omega$. We further assume (2.10) and its differentiated form, and (2.12). Then, following the argument in Section 4.2 of Hjort and Jones (1996), the variance of $\tilde{f}_\pm$ is seen to be asymptotic to $(nh)^{-1}\tau(K)^2 f(x)$, where, in place of Hjort and Jones's formula for $\tau(K)^2$, one has $\tau(K)^2 = w^{\mathrm{T}} M_1^{-1} M_2 M_1^{-1} w$, with $w^{\mathrm{T}} = (1 + o(1),\ ch + o(h))$, $M_1 = \mathrm{diag}(1,\ h^2\kappa_2)$, $M_2 =$

diag(κ_1, $h^2\kappa_3$) and $c = \pm\kappa_2^{1/2}$. It follows that $\tau(K)^2 \sim \kappa_1 + c^2\kappa_2^{-2}\kappa_3 = \kappa_1 + \kappa_2^{-1}\kappa_3$, as had to be proved. Formulae for the variances of $\tilde{f}_\lambda$ and $\tilde{f}$ may be derived by similar but more elaborate arguments, which are detailed in the ANU Ph.D. thesis of E. Choi (1998).

Acknowledgement

Hall gratefully acknowledges financial support from the Institute of Mathematical Sciences of the Chinese University of Hong Kong, and Choi and Hall express their gratitude for the hospitality of the Department of Statistics at CUHK, where the work described in this paper was commenced. The helpful comments of two reviewers are also gratefully acknowledged.

References

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

Bartlett, M.S. (1963) Statistical estimation of density functions. *Sankhyā Ser. A*, **25**, 245–254.

Choi, E. (1998) Some problems in curve and surface estimation. Ph.D. Thesis, Australian National University.

Choi, E. and Hall, P. (1998) On bias reduction in local linear smoothing. *Biometrika*, **85**, 333–346.

Copas, J.B. (1995) Local likelihood based on kernel censoring. *J. Roy. Statist. Soc. Ser. B*, **57**, 221–235.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and its Applications*. London: Chapman & Hall.

Fan, J., Farmen, M. and Gijbels, I. (1996) A blueprint of local maximum likelihood estimation. To appear.

Hastie, T. and Loader, C. (1993) Local regression: automatic kernel carpentry (with discussion.) *Statist. Sci.*, **8**, 120–143.

Hjort, N.L. and Jones, M.C. (1996) Locally parametric nonparametric density estimation. *Ann. Statist.*, **24**, 1619–1647.

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

Jones, M.C., Linton, O. and Nielsen, J.P. (1995) A simple bias reduction method for density estimation. *Biometrika*, **82**, 327–338.

Loader, C.R. (1996) Local likelihood density estimation. *Ann. Statist.*, **24**, 1602–1618.

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

Ruppert, D. and Cline, B.H. (1994) Bias reduction in kernel density estimation by smoothed empirical transformations. *Ann. Statist.*, **22**, 185–210.

Ruppert, D. and Wand, M.P. (1994) Multivariate weighted least squares regression. *Ann. Statist.*, **22**, 1346–1370.

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

Terrell, G.R. and Scott, D.W. (1980) On improving convergence rates for nonnegative kernel density estimators. *Ann. Statist.*, **8**, 1160–1163.

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

Received September 1997 and revised February 1998