The Annals of Statistics
1996, Vol. 24, No. 4, 1619–1647

LOCALLY PARAMETRIC NONPARAMETRIC DENSITY ESTIMATION

BY N. L. HJORT AND M. C. JONES

University of Oslo and The Open University

This paper develops a nonparametric density estimator with parametric overtones. Suppose $f(x, \theta)$ is some family of densities, indexed by a vector of parameters θ. We define a local kernel-smoothed likelihood function which, for each x, can be used to estimate the best local parametric approximant to the true density. This leads to a new density estimator of the form $f(x, \widehat{\theta}(x))$, thus inserting the best local parameter estimate for each new value of x. When the bandwidth used is large, this amounts to ordinary full likelihood parametric density estimation, while for moderate and small bandwidths the method is essentially nonparametric, using only local properties of data and the model. Alternative ways more general than via the local likelihood are also described. The methods can be seen as ways of nonparametrically smoothing the parameter within a parametric class.

Properties of this new semiparametric estimator are investigated. Our preferred version has approximately the same variance as the ordinary kernel method but potentially a smaller bias. The new method is seen to perform better than the traditional kernel method in a broad nonparametric vicinity of the parametric model employed, while at the same time being capable of not losing much in precision to full likelihood methods when the model is correct. Other versions of the method are approximately equivalent to using particular higher order kernels in a semiparametric framework. The methodology we develop can be seen as the density estimation parallel to local likelihood and local weighted least squares theory in nonparametric regression.

1. Introduction and summary. Let $X_1, \ldots, X_n$ be independent and identically distributed with density f. The traditional kernel estimator of f is $\widetilde{f}(x) = n^{-1} \sum_{i=1}^{n} K_h(x_i - x)$, where $K_h(z) = h^{-1}K(h^{-1}z)$ and $K(\cdot)$ is some chosen unimodal density, symmetric about zero. The basic properties of $\widetilde{f}$ are well known, and under smoothness assumptions these include

$$\begin{aligned} \mathrm{E}\widetilde{f}(x) &= f(x) + \tfrac{1}{2}\sigma_K^2 h^2 f''(x) + O(h^4), \\ \mathrm{Var}\,\widetilde{f}(x) &= R(K)(nh)^{-1} f(x) - n^{-1} f(x)^2 + O(h/n), \end{aligned} \tag{1.1}$$

where $\sigma_K^2 = \int z^2 K(z)\,\mathrm{d}z$ and $R(K) = \int K(z)^2\,\mathrm{d}z$. See Scott [(1992), Chapter 6] or Wand and Jones [(1995), Chapter 2], for example.

Our aim in this paper is to propose and investigate a class of semiparametric competitors which have precision comparable to that of $\widetilde{f}$, but sometimes better. For any given parametric family, $f(\cdot, \theta) = f(\cdot, \theta_1, \ldots, \theta_p)$, and, for each

Received March 1994; revised November 1995.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20.

Key words and phrases. Bias reduction, density estimation, kernel smoothing, local likelihood, local modelling, parameter smoothing, semiparametric estimation.

given x, we will present ways of estimating the locally best approximant to f and then use

$$\widehat{f}(x) = f(x, \widehat{\theta}_1(x), \ldots, \widehat{\theta}_p(x)). \tag{1.2}$$

Thus the estimated density at x employs a parameter value which depends on x and whose choice is to be tailored to good estimation at x. In other words, the method amounts to a version of nonparametric parameter smoothing within the given parametric class.

1.1. *Local likelihood for densities.* A central idea in our paper is the construction of a *local likelihood function* for density estimation. Local likelihood ideas have been employed in non- and semiparametric regression for some time (see Section 1.2), but the concept is far less immediate in the present context of density estimation. Around each given x we define the local log-likelihood to be

$$\begin{aligned} L_n(x, \theta) &= \int K_h(t - x)\{\log f(t, \theta)\, \mathrm{d}F_n(t) - f(t, \theta)\, \mathrm{d}t\} \\ &= n^{-1} \sum_{i=1}^{n} K_h(x_i - x) \log f(x_i, \theta) - \int K_h(t - x) f(t, \theta)\, \mathrm{d}t, \end{aligned} \tag{1.3}$$

writing F_n for the empirical distribution function. When h is large, this is close to the constant $K(0)h^{-1}$ times the ordinary, normalized log-likelihood function $n^{-1} \sum_{i=1}^{n} \log f(x_i, \theta) - 1$, and maximizing the (1.3) function with respect to the parameters becomes equivalent to ordinary full maximum likelihood estimation. When h is moderate or small, however, maximizing $L_n(x, \theta)$ will be seen to be a fruitful way of obtaining an estimate of the best local approximant to f. This is made clear in Section 2.

A related and in fact more general apparatus is as follows. Decide on suitable weight functions $v_j(x, t, \theta)$, $j = 1, \ldots, p$, guidelines for which will be discussed later, and let $\widehat{\theta}(x)$ be defined as the solution to the p equations

$$\begin{aligned} V_n(x, \theta) &= \int K_h(t - x) v(x, t, \theta)\{\mathrm{d}F_n(t) - f(t, \theta)\, \mathrm{d}t\} \\ &= n^{-1} \sum_{i=1}^{n} K_h(x_i - x) v(x, x_i, \theta) \\ &\quad - \int K_h(t - x) v(x, t, \theta) f(t, \theta)\, \mathrm{d}t = 0. \end{aligned} \tag{1.4}$$

Maximizing the (1.3) function amounts to solving (1.4) with $v(x, t, \theta) = u(t, \theta) = (\partial/\partial\theta) \log f(t, \theta)$, the $p \times 1$ score function of the model, with one component $u_j(x, \theta)$ per parameter. The generalization is analogous to that of M-estimation over maximum likelihood estimation in ordinary estimation theory.

This strategy, with (1.4) or its special case (1.3), gives $\widehat{\theta}(x)$ and in the end (1.2). We call this *local parametric estimation* of the density f, hence the title

of our paper. An attractive motivation for this approach is that as $h \to \infty$, $\widehat{f}$ tends to a global parametric fit of the model $f(\cdot, \theta)$. As in other attempts at semiparametric density estimation (cf. references mentioned below), our methodology should be particularly useful when f exhibits small or moderate departures from a standard parametric form. However, $f(\cdot, \theta)$ need not even be a crude model for the data because, if not, h will be chosen small, and local properties of $\widehat{f}$ will largely be divorced from global properties of $f(\cdot, \theta)$. Thus we view our method as a "continuous bridge" between fully parametric and fully nonparametric options.

The local likelihood function is more fully motivated—in several ways—in Section 2, and a connection is also established to the dynamic likelihood methods for nonparametric hazard rate estimation of Hjort (1991, 1997). Apart from the local likelihood connection, we note that the (1.4)-type approach is natural in that a weighted difference of $\mathrm{d}F_n(t) - f(t, \theta)\,\mathrm{d}t$, which in the limit is a weighted difference of $\{f(t) - f(t, \theta)\}\,\mathrm{d}t$, is set to zero.

The new estimator can and will be motivated also on the grounds of performance, of course. We start our investigation of the large sample properties of $\widehat{f}(x)$ in Section 3, with concentration on one-parameter local fits. This is extended in Section 4 to the multiparameter case, with particular focus on two parameters. The two-parameter case affords an attractive simplification of $O(h^2)$ bias and forms our favored class of locally parametric density estimators. It turns out that the bias and variance properties of $\widehat{f}$ are remarkably comparable to those of the classic estimator $\widetilde{f}$. For many situations it will be seen that

$$\begin{aligned} \mathrm{E}\widehat{f}(x) &= f(x) + \tfrac{1}{2}\sigma_K^2 h^2 b(x) + O(h^4 + (nh)^{-1}), \\ \mathrm{Var}\,\widehat{f}(x) &= R(K)(nh)^{-1} f(x) - n^{-1} f(x)^2 + O(h/n), \end{aligned} \tag{1.5}$$

just as in (1.1), but with a bias factor function $b(x)$ related to but different from $f''(x)$, with characteristics inherited from the parametric class and the weight functions used. To the order of approximation used, the variance is simply the same, regardless of parametric family and of $v(x, t, \theta)$. The statistical advantage will be that for many fs, typically those lying in a broad nonparametric neighborhood of the parametric $f(\cdot, \theta)$, $b(x)$ will be smaller in size than $f''(x)$ for most x. It should also be the case that in aiming for improved performance by choice of $f(\cdot, \theta)$ we will rarely lose too much in performance terms in the sense that $|b(x)|$ should not be too much greater than $|f''(x)|$ on occasions when $f(\cdot, \theta)$ is a totally inappropriate global model.

In Section 4 it is also shown that a bias of the potentially smaller size $O(h^4)$ is attainable if the vehicle model has three or four parameters and the underlying true density is sufficiently smooth. This is achieved without having to (explicitly) resort to higher order kernels. Our method is, however, in its kernel-dependent quantities, asymptotically equivalent to a particular class of higher order kernels which are of the form a suitable polynomial times K. The same higher order kernels arise in local polynomial regression (see Sec-

tion 1.2 below), but we stress that this result is consequent on the number of parameters fitted and not on using any particular form of local parameterization (which shows up only in the bias factor). We conjecture that the same is true in the local least squares regression context. Thus locally smoothing a three- or four-parameter model leads to a superior asymptotic performance. We nevertheless favor two-parameter families for their comparative simplicity conceptually and computationally, and with experience of higher order kernels raising doubts about the transfer of such asymptotic advantages to finite sample practice [Marron and Wand (1992)].

A variety of particular examples are discussed in Section 5. These are not practical examples but rather features and properties of interesting special cases of our methodology. Particular attention is given to the case of an estimated "running normal" density and to estimates that incorporate local modelling of level, slope and curvature. Sections 6 and 7 provide further extensions of the earlier theory. In Section 6, we present results on the boundary behavior of our estimators and note attractive properties thereof. Section 7 indicates extensions to the multivariate case, where the new method could prove to be particularly useful, since the ordinary methods are problematic in higher dimensions. In Section 8 we discuss some other issues such as automatic bandwidth selection and inspection of "running parameters," while our conclusions are offered in Section 9. Our focus throughout this paper is on intuitive and theoretical considerations. Implementation issues and comparative work are left to future studies.

1.2. *Related work.* In nonparametric regression, there has been much recent interest in fitting polynomial functions locally. Relevant references include Fan (1992, 1993), Fan and Gijbels (1992, 1996), Hastie and Loader (1993) and Ruppert and Wand (1994), building on earlier work of Stone (1977) and Cleveland (1979). This has been done by local least squares, which is a normal error distribution version of local likelihood fitting; see Tibshirani and Hastie (1987), Staniswalis (1989), Jones and Hjort (1994) and Fan, Heckman and Wand (1995). Local linear fitting is particularly attractive. It affords asymptotic bias depending only on the second derivative of the regression function, without sacrificing anything in terms of variance [this is not at all trivial to achieve; cf. Jones, Davies and Park (1994)]. It also automatically has very good boundary properties. Higher degree polynomials behave rather like "higher order" kernels. In the large bandwidth limit, the parametric form approached is, of course, a global polynomial regression. Given the large impact of these methods in regression, it is natural to ask if parallel methods can be invented for density estimation. It is indeed an aim of this paper to provide such a methodology.

At around the same time as we were developing our ideas, Loader (1996) independently proposed a version of local likelihood density estimation. A key component is specification of an appropriate likelihood function, and Loader's definition is indeed similar to our (1.3). Loader uses his definition to fit local polynomials to the log density, perhaps the most immediate analogue of the

regression work. Our motivation differs from Loader's in preferring to work with more general local parametric models, seeking semiparametric density estimators, with standard parametric models as limiting cases. However, our methodology covers interesting nonstandard parametric forms and other local estimation methods as well. We arrived at (1.3) and its relative (1.4) partly via the hazard rate case, for which local likelihood specification is more immediate [see Hjort (1991, 1997)], and partly via local weighting of the $\mathrm{d}F_n(t)-f(t,\theta)\,\mathrm{d}t$ difference; see Section 2.

Some semiparametric density estimators already exist. Our approach has similar intentions to that of Copas (1995), but ours appears to be both simpler and more general. A semiparametric method which works by multiplying an initial parametric description with a nonparametric kernel-type estimate of the necessary correction factor is developed in Hjort and Glad (1995). Their estimator also has properties (1.5), but with yet another $b(x)$ bias factor function. Another similarly spirited method consists of using an estimated orthogonal expansion for this multiplicative correction factor; see, for example, Hjort [(1986), Chapter 5], Buckland (1992) and Fenstad and Hjort (1996). An initial nonparametric estimator "corrected" towards the parametric is the topic of recent work of Efron and Tibshirani (1996). These authors also note the role of backfitting as in Hastie and Tibshirani (1990) in a similar context. Various semiparametric density estimators of Bayesian flavor are discussed in Hjort (1995). Earlier work, somewhat less attractively involving an extra parameter in a linear combination of parametric and nonparametric estimators, includes Schuster and Yakowitz (1985) and Olkin and Spiegelman (1987). Jones (1993a) argues that (the natural variance-corrected version of) the kernel density estimator can itself be thought of as a semiparametric estimator.

2. Local likelihood for density estimation. This section gives support for the local parametric estimation method of (1.2) and (1.3). It first relates the method to a well-defined local statistical Kullback–Leibler-type distance function from the true density to the parametric approximant. This is followed by a connection to similar concepts for hazard rate estimation in survival data. Finally, included in this section are alternative motivations, also of others, for considering the same definition of local likelihood.

2.1. *Local parametric approximation.* To explain why maximizing (1.3) is a good idea, note first that

$$L_n(x,\theta) \to_p \lambda(x,\theta) = \int K_h(t-x)\{f(t)\log f(t,\theta) - f(t,\theta)\}\,\mathrm{d}t$$

as n grows. The maximizer $\widehat{\theta}(x)$ hence aims at the parameter value $\theta_0(x)$ that maximizes $\lambda(x,\theta)$. This is a well-defined statistical quantity in that it minimizes the distance

$$d[f, f(\cdot,\theta)] = \int K_h(t-x)\left[f(t)\log\frac{f(t)}{f(t,\theta)} - \{f(t)-f(t,\theta)\}\right]\mathrm{d}t \tag{2.1}$$

between true density (which need not belong to the parametric class under consideration) and approximating parametric density. Noting that the Kullback–Leibler distance from f to f_θ can be written

$$\int f(t)\log\{f(t)/f(t,\theta)\}\,\mathrm{d}t = \int\left[f(t)\log\frac{f(t)}{f(t,\theta)} - \{f(t)-f(t,\theta)\}\right]\mathrm{d}t,$$

we see that (2.1) is a version of the same, locally weighted around x. These arguments show that using (1.2) with (1.3), which is (1.4) with weight function chosen to be the score function $u(t,\theta)$, aims at the best local parametric approximant to the true f. Note also that if f is not far from $f(\cdot,\theta)$, then $d[f, f(\cdot,\theta)] \simeq \frac{1}{2}\int K_h(t-x)\{f(t)-f(t,\theta)\}^2/f(t)\,\mathrm{d}t$. An alternative L_2-based local distance measure is discussed briefly in Section 5.6.

2.2. *The hazard connection.* For a moment, consider survival data on $[0,\infty)$, and switch attention from density $f(t,\theta)$ and cumulative distribution $F(t,\theta)$ to survival function $S(t,\theta) = 1 - F(t,\theta)$ and, particularly, hazard function $\alpha(t,\theta) = f(t,\theta)/S(t,\theta)$. The likelihood is $\prod_{i=1}^n \alpha(t_i,\theta)S(t_i,\theta)$, so that the log-likelihood, after a little manipulation, and disregarding a multiplier of n, takes the form $\int\{\log\alpha(t,\theta)\,\mathrm{d}F_n(t) - S_n(t)\alpha(t,\theta)\,\mathrm{d}t\}$, where $S_n(t) = 1 - F_n(t)$ is the proportion of individuals still at risk just prior to time t. The kernel-smoothed local log-likelihood for the model at location x is, therefore,

$$L_{0,n}(x,\theta) = \int K_h(t-x)\{\log\alpha(t,\theta)\,\mathrm{d}F_n(t) - S_n(t)\alpha(t,\theta)\,\mathrm{d}t\}. \tag{2.2}$$

This local likelihood for hazard models is well motivated and explored in Hjort (1991, 1997). Note that

$$L_{0,n}(x,\theta) \to_p \lambda_0(x,\theta) = \int K_h(t-x)\left\{f(t)\log\frac{f(t,\theta)}{S(t,\theta)} - S(t)\frac{f(t,\theta)}{S(t,\theta)}\right\}\mathrm{d}t.$$

Maximizing $L_{0,n}(x,\theta)$ aims at the best local approximant in the sense of minimizing the local distance function

$$d_0[f, f(\cdot,\theta)] = \int K_h(t-x)\left[f(t)\left\{\log\frac{f(t)}{S(t)} - \log\frac{f(t,\theta)}{S(t,\theta)}\right\} - S(t)\left\{\frac{f(t)}{S(t)} - \frac{f(t,\theta)}{S(t,\theta)}\right\}\right]\mathrm{d}t.$$

This underlies the theory of locally parametric nonparametric hazard rate estimation, and is as in Hjort [(1997), Sections 2 and 3], but now suitably reexpressed as a distance between densities and not hazards.

To see a connection from this context to density estimation, put in $\alpha(t,\theta) = f(t,\theta)/S(t,\theta)$ to see

$$L_{0,n}(x,\theta) = \int K_h(t-x)[\{\log f(t,\theta) - \log S(t,\theta)\}\,\mathrm{d}F_n(t) - S_n(t)f(t,\theta)/S(t,\theta)\,\mathrm{d}t].$$

Now replace $S(t, \theta)$ here with the estimate $S_n(t)$ (this step will be discussed in Section 2.3). This leads to

$$\int K_h(t-x)[\{\log f(t,\theta) - \log S_n(t)\}\,\mathrm{d}F_n(t) - f(t,\theta)\,\mathrm{d}t],$$

and since the $\log S_n(t)$ term is immaterial, this is the same as $L_n(x, \theta)$ of (1.3). We point out that the hazard connection makes it clear how censoring can be coped with also; see Hjort (1997).

2.3. *Justification of $L_n(x, \theta)$ as local log-likelihood.* We think of (1.3) as the *local log-likelihood*, or local kernel-smoothed log-likelihood, for the model at x. The main justification for this is via the best local approximation framework laid out in Section 2.1 above, combined with the appealing feature that large bandwidths lead back to global likelihood analysis, and not least with the fact that the method works, as this paper demonstrates. We also know of four additional justifications for the (1.3) construction.

The first completes the argument of Section 2.2. One can argue that the insertion of $S_n(t)$ for $S(t, \theta)$ here should not alter things very much since $S_n(t)$ is a more precise estimate than is any local parameter estimate (or hence local density estimate) for its population version. Indeed, $S_n(t)$ has mean squared error of order n^{-1}, which is insignificant compared with the mean squared error of our density estimate which, it will turn out, will be $O(n^{-4/5})$.

However, what of a more direct local likelihood argument? The naive local log-likelihood $\int K_h(t-x)\log f(t,\theta)\,\mathrm{d}F_n(t)$ does not work, as inspection in the normal case pedagogically reveals, for example. Similarly the naive nonparametric log-likelihood $\int \log f(t)\,\mathrm{d}F_n(t)$ has problems, whether kernel smoothed or not; it can be made infinite by putting infinite spikes at the data points. Loader (1996) argues that the log-likelihood is truly $\int \log f(t)\,\mathrm{d}F_n(t) - \int f(t)\,\mathrm{d}t$ (think of likelihood estimation of a Poisson intensity function), but the final term is usually discarded since it takes the value 1. Leaving the second term in and then localizing by kernels yields precisely (1.3) again.

Another argument stems from personal communication with J. B. Copas. Note first that the derivative of the simplistic $\int K_h(t-x)\log f(t,\theta)\,\mathrm{d}F_n(t)$ is $\int K_h(t-x)u(t,\theta)\,\mathrm{d}F_n(t)$, which does not have expectation zero, even under model conditions. To remedy this, subtract its expectation, which is $\int K_h(t-x)u(t,\theta)f(t)\,\mathrm{d}t$. Alternatively, at least, if we approximate this last $f(t)$ by $f(t,\theta)$, we obtain the score function case of $V_n(x,\theta)$ of (1.4), and hence motivate $L_n(x,\theta)$ at (1.3) once more. [Copas' (1995), suggestion differs from this. The current version replaces Copas' expression (7), $w(x)\log f(x,\theta) + \{1 - w(x)\}\log B(\theta)$ (in Copas' notation) by $w(x)\log f(x,\theta) + B(\theta) - 1$.]

Comments from a referee triggered the following fourth justification of (1.3). This is interesting in that it connects the density estimation problem to the more well-developed local likelihood methodology for nonparametric regression. It is based on a discretization argument: split the data region into small intervals $D_1, \ldots, D_m$ of lengths $d_1, \ldots, d_m$ and let $s_1, \ldots, s_m$ be the number of points falling in each. Modelling the s_js as independent Poisson variables with

parameters $\gamma\pi_j(\theta)$, where $\pi_j(\theta) = \int_{D_j} f(x, \theta)\,\mathrm{d}x$, gives (omitting an additive constant) the log-likelihood $\sum_{j=1}^m \{-\gamma\pi_j(\theta) + s_j \log\gamma + s_j \log\pi_j(\theta)\}$. Conditioning the Poisson model on $\sum_{j=1}^m s_j = n$, which is also the maximum likelihood estimate of γ, corresponds to the multinomial model for the s_j counts. This formal equivalence to the Poisson model was exploited in Lindsey (1974) and more recently in Efron and Tibshirani (1996).

The present point is that there is a well-established way of localizing such a likelihood [see Tibshirani and Hastie (1987), Jones and Hjort (1994), Fan, Heckman and Wand (1995) and Fan and Gijbels (1996)], since it has been made to belong to nonparametric smoothing of Poisson parameters rather than density estimation. This gives

$$\widetilde{L}_n(x, \theta) = \sum_{j=1}^m K_h(x - x_{(j)})[-\gamma\pi_j(\theta) + s_j \log\{\gamma\pi_j(\theta)\}],$$

where $x_{(j)}$ is a convenient point in D_j. Taking a fine limit, via $\pi_j(\theta) \simeq d_j f(x_{(j)}, \theta)$, leads to

$$\widetilde{L}_n(x, \theta) \simeq -\gamma \int K_h(x-t) f(t, \theta)\,\mathrm{d}t + (\log\gamma)\, n\widetilde{f}(x) + \sum_{i=1}^n K_h(x - x_i) \log f(x_i, \theta).$$

Putting $\gamma = n$ here, as suggested by the original Poisson connection, gives (1.3) again. The connection is not quite as clear-cut, however, since the maximizer is $\widehat{\gamma}_\theta = n\widetilde{f}(x)/(K_h * f_\theta)(x)$, which still depends on θ, and this delivers yet another proposal, namely, the profile log-likelihood

$$L_n^*(x, \theta) = -n\widetilde{f}(x)\, \log\left\{\int K_h(x - t) f(t, \theta)\,\mathrm{d}t\right\} + \sum_{i=1}^n K_h(x - x_i) \log f(x_i, \theta).$$

We would still have $\widehat{\gamma}_\theta = n(1 + O_p(h^2))$ for small h and for the θs of interest, however, leading again to (1.3). See Jones (1995) for more on discretized forms of local likelihood.

3. Large sample properties.

3.1. *h fixed, large n.* Let θ be p-dimensional in this subsection. Estimating θ by solving (1.4) is like M-estimation, with the extra complication that we do not assume the true f to belong to the parametric $f(\cdot, \theta)$ class. For simplicity, suppress the fixed x and write $v(t, \theta) = v(x, t, \theta)$ for the p weight functions. Assume that

$$V(x, \theta) = \int K_h(t - x) v(t, \theta)\{f(t) - f(t, \theta)\}\,\mathrm{d}t = 0 \tag{3.1}$$

has a unique solution $\theta_0 = \theta_0(x)$ (which also depends on h, held fixed here). This essentially says that $f(t)$ should be within reach of $f(t, \theta)$ as θ varies and that the p functions $v_j(t, \theta)$ should be functionally independent; see the examples of Section 5. That $V_n(x, \theta_0)$ has mean zero plays a role in developing the following facts. First, $\widehat{\theta}(x)$ converges to this best local parameter $\theta_0(x)$

in probability. In the score function case $v = u$ this is also the parameter minimizing (2.1). Second,

$$(nh)^{1/2}\{\widehat{\theta}(x) - \theta_0\} \to_d \mathcal{N}_p\{0, J_h^{-1} M_h (J_h^{\mathrm{t}})^{-1}\}, \tag{3.2}$$

where

$$\begin{aligned} J_h &= \int K_h(t-x)\big[v(t,\theta_0)u(t,\theta_0)^{\mathrm{t}} f(t,\theta_0) + v^*(t,\theta_0)\{f(t,\theta_0) - f(t)\}\big]\,\mathrm{d}t, \\ M_h &= \mathrm{Var}_f\{h^{1/2}K_h(X_i - x)v(X_i,\theta_0)\} \\ &= \int hK_h(t-x)^2 v(t,\theta_0)v(t,\theta_0)^{\mathrm{t}} f(t)\,\mathrm{d}t - h\xi_h\xi_h^{\mathrm{t}}, \end{aligned}$$

and $\xi_h = \int K_h(t-x)v(t,\theta_0)f(t)\,\mathrm{d}t$. Again $u(t,\theta)$ is the model's score function while $v^*(t,\theta)$ is the $p \times p$ matrix of derivatives of the $v_j(t,\theta)$ functions. Proving these claims is not very difficult, using variations of arguments used to prove asymptotic normality of M-estimators; see Section 8.4 for relevant details and an additional result. By the delta method,

$$\begin{aligned} &(nh)^{1/2}\{\widehat{f}(x) - f(x,\theta_0)\} \\ &\qquad \to_d \mathcal{N}\{0, f(x,\theta_0)^2 u(x,\theta_0)^{\mathrm{t}} J_h^{-1} M_h (J_h^{\mathrm{t}})^{-1} u(x,\theta_0)\}. \end{aligned} \tag{3.3}$$

3.2. *Decreasing h.* The (3.3) result is valid for a fixed positive h. We are also interested in being increasingly fine-tuned about h as n grows. Observe that, as $h \to 0$,

$$\int K_h(t-x)g(t)\,\mathrm{d}t = g(x) + \tfrac{1}{2}\sigma_K^2 h^2 g''(x) + O(h^4) \tag{3.4}$$

for each smooth g function, by a standard simple Taylor series argument. Using this in conjunction with (3.1) shows that $f(x,\theta_0(x)) - f(x) = O(h^2)$ in general. Indeed,

$$v_{j,0}(x)\{f_0(x) - f(x)\} = \tfrac{1}{2}\sigma_K^2 h^2 \{v_{j,0}(f - f_0)\}''(x) + O(h^4) \tag{3.5}$$

under smoothness assumptions on f and the weight functions, writing $f_0(x) = f(x,\theta_0)$, $v_{j,0}(x) = v_j(x,\theta_0)$ and so on [and where $\theta_0 = \theta_0(x)$ also depends on x]. Furthermore $(v_{j,0}f_0)''(x)$, for example, means taking the second x-derivative of the $v_j(x,\theta)f(x,\theta)$ function, and then inserting the parameter value $\theta = \theta_0(x)$. Under mild regularity assumptions this also implies

$$\mathrm{E}\widehat{f}(x) = f(x) + \tfrac{1}{2}\sigma_K^2 h^2 b(x) + O(h^4 + (nh)^{-1}),$$

where the precise nature of the $b(x)$ function will be quite important and will be analyzed more later.

We need to assess the size of $J_h^{-1}M_h(J_h^{\mathrm{t}})^{-1}$ of (3.2) and of the variance appearing in (3.3), when h tends to zero. To this end, it proves to be convenient to reparametrize quantities in J_h and M_h. Rewrite $f(t,\theta_0)$ as $f(t-x,\psi_0)$, where the new parameters ψ are easily related to the old parameters θ and we note that the first element of $\widehat{\psi}$ is the only one directly specifying $\widehat{f}(x)$.

Also, replace $u(t, \theta_0)$ and $v(t, \theta_0)$ by $u_h(h^{-1}(t-x), \psi_0)$ and $v_h(h^{-1}(t-x), \psi_0)$ respectively, the subscript h referring to dependence of u_h and v_h on h to accommodate the h^{-1} attached to $t-x$. [For an example, reparametrize $\theta_1 + \theta_2 t + \theta_3 t^2$ to $\psi_1 + \psi_2 hz + \psi_3 h^2 z^2$, where $z = (t-x)/h$.] We then find that

$$
\begin{aligned}
J_h &= f_0(x) \int K(z) v_h(z, \psi_0) u_h(z, \psi_0)^{\mathrm{t}} \, \mathrm{d}z + O(h^2), \\
M_h &= f(x) \int K(z)^2 v_h(z, \psi_0) v_h(z, \psi_0)^{\mathrm{t}} \, \mathrm{d}z - h \xi_0 \xi_0^{\mathrm{t}} f(x)^2 + O(h^2),
\end{aligned} \tag{3.6}
$$

where $\xi_0 = \int K(z) v_h(z, \psi_0) \, \mathrm{d}z$.

3.3. *The one-parameter case.* Let $f(x, \theta)$ have just one parameter and let the weight function $v(t, \theta)$ be smooth and nonzero at x. From (3.5) and previous arguments one finds

$$
b(x) = f''(x) - f_0''(x) + 2\{v_0'(x)/v_0(x)\}\{f'(x) - f_0'(x)\}, \tag{3.7}
$$

differing from the kernel estimator's bias factor $f''(x)$ by a term depending on properties of $f(\cdot, \theta)$. If $f_0 = f$, that is, if we are working with the correct parametric class, then $b(x) = 0$. Otherwise, (3.7) should be small when f_0 is close to f and perhaps not too large in absolute value even when f and f_0 differ considerably. Notice that the expression for $b(x)$ simplifies when the weight function used is $v(t, \theta) = 1$. It also simplifies in the multiparameter case of the next section. An expression for the variance is found from (3.3) and (3.6). Assuming that $v_h(z)$ and $u_h(z)$ are of the form $c + O(hz)$ for small h, the weight function as well as other traces of the parametric model are seen to cancel out, for the leading terms, and the result is

$$
\operatorname{Var} \widehat{f}(x) = R(K)(nh)^{-1} f(x) - n^{-1} f(x)^2 + O(h/n). \tag{3.8}
$$

That is, the variance is the same, to the order of approximation used, as that of the ordinary kernel density estimator.

4. The multiparameter case. In this section, let the parametric model be $f_\theta(x) = f(x, \theta_1, \ldots, \theta_p)$ with $p \geq 2$. The results we shall obtain for approximate biases and variances again hold under suitable regularity assumptions, including permission to interchange limits and expectation. That these are met can be checked directly for the most important special cases, like those listed in Section 5.

4.1. *The bias.* We have $\mathrm{E}\widehat{f}(x) = f(x, \theta_0) + O((nh)^{-1})$ again, and the p equations $\int K_h(t-x) v_j(t, \theta_0)\{f(t) - f(t, \theta_0)\} \, \mathrm{d}t = 0$ can be used to see how far $f(x, \theta_0)$ is from $f(x)$. From (3.5) it is seen that

$$
\begin{aligned}
&f(x, \theta_0) - f(x) \\
&\quad = \tfrac{1}{2} \sigma_K^2 h^2 [f''(x) - f_0''(x) + 2\{v_{j,0}'(x)/v_{j,0}(x)\}\{f'(x) - f_0'(x)\}] + O(h^4)
\end{aligned}
$$

for each j, under smoothness assumptions. Since there are $p \geq 2$ equations giving the h^2 coefficient, this can hold only when $f'(x) - f_0'(x) = o(1)$ as $h \to 0$.

This is not, in general, true in the one-parameter case and is the cause of the extra term making up (3.7). For $p \geq 2$, however, we have

$$\mathrm{E}\widehat{f}(x) = f(x) + \tfrac{1}{2}\sigma_K^2 h^2\{f''(x) - f_0''(x)\} + O(h^3 + (nh)^{-1}). \tag{4.1}$$

Introduction of further local parameters has simplified the bias to depending solely on $f''(x) - f_0''(x)$. This is appealingly interpretable. The bias is of a familiar second derivative, local curvature type, and the way in which closeness of f_0 to f affects the bias is abundantly clear.

The foregoing remarks are really most relevant to the case of two parameters exactly. For $p \geq 3$, an extension of the above argument shows that $(f - f_0)''$ is also $o(1)$. To see this, write $g_r \equiv (f - f_0)^{(r)} \simeq \sum_{i=0}^{4-r} a(r, i)h^i$ for $r = 0, \ldots, 4$. Then look at the general equations governing asymptotic bias and equate terms in powers of h. These are

$$\begin{aligned} &g_0 + \tfrac{1}{2}k_2 h^2 (v_{j,0} g_0)''/v_{j,0} \\ &\quad + \tfrac{1}{24}k_4 h^4 (v_{j,0} g_0)^{(4)}/v_{j,0} + \tfrac{1}{720}k_6 h^6 (v_{j,0} g_0)^{(6)}/v_{j,0} = 0, \end{aligned}$$

for $j = 1, \ldots, p$, where we write $k_j = \int z^j K(z)\,\mathrm{d}z$; in particular, $k_2 = \sigma_K^2$. For instance, when $p = 3$, a little manipulation yields $a(0, i) = 0 = a(j, k)$ for $i = 0, 1, 2, 3$, $j = 1, 2$, $k = 1, 2$. Also of importance are $a(0, 4) + \frac{1}{2}k_2 a(2, 2) + \frac{1}{24}k_4 a(4, 0) = 0$ and $k_2 a(1, 2) + \frac{1}{6}k_4 a(3, 0) = 0$. To make further progress, we need to consider the h^6 term which involves three equations in four unknowns, where, in particular, two of these, say t and u, satisfy

$$t = \tfrac{1}{2}k_2 a(0, 4) + \tfrac{1}{4}k_4 a(2, 2) + \tfrac{1}{48}k_6 a(4, 0) \quad \text{and} \quad u = \tfrac{1}{6}k_4 a(1, 2) + \tfrac{1}{36}k_6 a(3, 0).$$

We can thus write $t = Au$ for appropriate A and hence find that

$$a(0, 4) = \frac{k_2 k_6 - k_4^2}{k_4 - k_2^2}\left\{\tfrac{1}{24}a(4, 0) - \tfrac{1}{18}Aa(3, 0)\right\}.$$

Reinterpreting this in bias terms results in

$$\begin{aligned} \mathrm{E}\widehat{f}(x) = f(x) - \frac{k_2 k_6 - k_4^2}{k_4 - k_2^2} h^4 \big[&\tfrac{1}{24}\{f^{(4)}(x) - f_0^{(4)}(x)\} \\ &- \tfrac{1}{18}A\{f^{(3)}(x) - f_0^{(3)}(x)\}\big] + o(h^4), \end{aligned} \tag{4.2}$$

where, being explicit about A, A is the solution to the system of equations $xv_{j,0} + yv_{j,0}' + Av_{j,0}'' = -v_{j,0}'''$ for $j = 1, 2, 3$.

Increasing p from 3 to 4 results in the considerable simplification that the term involving A in (4.2) disappears due to being able to set $t = u = 0$ so that we then have

$$\mathrm{E}\widehat{f}(x) = f(x) - \tfrac{1}{24}h^4 \frac{k_2 k_6 - k_4^2}{k_4 - k_2^2}\left\{f^{(4)}(x) - f_0^{(4)}(x)\right\} + o(h^4). \tag{4.3}$$

Therefore, one gets an exact parallel of properties of the local polynomial regression referred to in Section 1.2; see Ruppert and Wand (1994) and Fan

and Gijbels (1996). Fitting one or two parameters, using a second-order kernel K, corresponds to $O(h^2)$ bias, with two parameters exhibiting advantages in terms of simplicity. There is a parallel story for nonparametric regression, involving local models with one or two parameters; this is well known for the local constant and the local linear model. Three and four parameters yield $O(h^4)$ bias, as do local quadratic and cubic regressions, and four parameters afford a simple dependence on $(f - f_0)^{(4)}$. And, we conjecture, so on, with five and six parameters (sufficient smoothness of f and the parametric densities used permitting). An important point emerging here is that we have not had to impose any particular local parametric form to achieve this behavior. Rather it is a consequence of the number of local parameters fitted. See also Sections 4.2 and 4.3. Since the practical value of these asymptotic results is perhaps dubious, we prefer to concentrate on the two-parameter case and consequent improvements in leading constant rather than rate, allied with more obvious practical interpretation.

4.2. *The variance.* We use (3.3) with (3.6), assuming, as is reasonable, that the $v_h(z, \psi_0)$ and $u_h(z, \psi_0)$ functions are of the form $c_1 + c_2(hz) + c_3(hz)^2 + \cdots$ for small h, and that there is at least one nonzero c_i coefficient in each of the vectors v_h and u_h. It should be no surprise that v and u functions can be subjected to arbitrary linear transformations without effect on the resulting estimates, and it is easy to see by consideration of $u^{\mathrm{t}} J_h^{-1} M_h (J_h^{\mathrm{t}})^{-1} u$ and (3.6) that the variance is unaffected by this. As far as this asymptotic assessment is concerned, therefore, where $h \to 0$ and $nh \to \infty$, it follows that we can replace both v and u by the canonical function

$$V_h(z) = (1, hz, h^2 z^2, \ldots, h^{p-1} z^{p-1})^{\mathrm{t}}.$$

Thence, from (3.3) and (3.6), we see that

$$\operatorname{Var} \widehat{f}(x) = (nh)^{-1} f(x) \tau(K)^2 - f(x)^2/n + O(h + h/n), \tag{4.4}$$

where, letting $e_1 = (1, 0, \ldots, 0)^{\mathrm{t}}$,

$$\tau(K)^2 = e_1^{\mathrm{t}} \left(\int K V_h V_h^{\mathrm{t}} \, \mathrm{d}z \right)^{-1} \left(\int K^2 V_h V_h^{\mathrm{t}} \, \mathrm{d}z \right) \left(\int K V_h V_h^{\mathrm{t}} \, \mathrm{d}z \right)^{-1} e_1.$$

A particularly natural local parameterization takes $f(t, \theta)$ as $\exp(\sum_{j=0}^{p-1} \theta_j t^j)$, so that $u(t, \theta) = (1, t, \ldots, t^{p-1})^{\mathrm{t}}$. This is the special case—with $v = u$—explored by Loader (1996), who gives essentially the same variance expression as above, but we must emphasize that this variance result also holds for *any* (sensible) local parameterization and not just for Loader's: it is purely a consequence, as is the kernel-dependent part of the bias, of the *number* of local parameters fitted.

4.3. *Two, three, four parameters.* In the case of two parameters, (4.4) simply reduces to

$$\operatorname{Var} \widehat{f}(x) = (nh)^{-1} f(x) R(K) - n^{-1} f(x)^2 + O(h/n). \tag{4.5}$$

This nicely joins with the two-parameter bias to mean all the usual properties of the ordinary kernel density estimator with the single exception that the bias depends now on $(f - f_0)''$ rather than just f''.

For either three or four parameters, (4.4) yields

$$\text{(4.6)}\quad \operatorname{Var} \widehat{f}(x) = (nh)^{-1} f(x) \frac{\int (k_2 z^2 - k_4)^2 K(z)^2 \, dz}{(k_4 - k_2^2)^2} - n^{-1} f(x)^2 + O(h/n),$$

and this variance quantity associates appropriately with the kernel-dependent quantity given for the three parameters in (4.2) and for four parameters in (4.3). The two are the bias and variance of the fourth-order kernel $\{(k_2 z^2 - k_4)/(k_4 - k_2^2)\} K(z)$; see Jones and Foster (1993). This equivalence is familiar for local quadratic or cubic regression [Ruppert and Wand (1994)], but here we observe it for density estimation and, most importantly, for *any* local three- or four-parameter model.

As the pattern is that, for example, five and six parameters afford $O(h^6)$ bias, so $O((nh)^{-1})$ variance can be expected, and an equivalent kernel that is an appropriate quartic multiple of K.

We should note briefly that p parameters afford, again in parallel with $(p-1)$th degree polynomial fitting, natural estimators of the first $p-1$ derivatives of f. The usual rates for derivative estimation, which involves a variance contribution of order $n^{-1}h^{-(2r+1)}$ for the rth derivative, can be shown to obtain, and equivalent derivative kernels [Ruppert and Wand (1994)] will arise.

5. Special cases. This section exhibits various special cases of the general methodology.

5.1. *The classic kernel method.* The simplest special case is to set $f(x, \theta) = \theta$. Semiparametrically, this is not especially attractive since the limiting form of the estimator as $h \to \infty$ is uniform (albeit an improper uniform), but for small h, that is, locally to x, this makes perfect sense. Moreover, the resulting density estimator is given explicitly by

$$n^{-1} \sum_{i=1}^{n} K_h(x_i - x) \Big/ \int K_h(t - x) \, dt.$$

Since the integral is 1, the denominator may be ignored and the result is precisely the classical kernel density estimator $\widetilde{f}$. We mention the denominator, however, because it is not unity near any boundary of f's support, but rather effects a renormalization near the boundary as discussed further in Section 6.1.

Following on from this, a natural first two-parameter locally parametric estimator is provided by fitting a line $\theta_1 + \theta_2(t - x)$, say, locally to x. Provided we need not worry about boundaries, $\int K_h(t - x)(t - x) \, dt = 0$, and hence it turns out that $\widehat{f}(x) = \widetilde{f}(x)$ once more. Note that both local constant and linear models have $f_0''(x) = 0$, and the bias formula (4.1) gives the classic

answer $\frac{1}{2}\sigma_K^2 h^2 f''(x)$. (Near boundaries, local lines automatically adjust $\widehat{f}$ in a way that has good consequences which are described in Section 6.2.)

Local polynomials are the obvious further extension, higher degree polynomials corresponding to higher orders of bias in a way entirely analogous to local polynomial fitting in regression [(e.g., Ruppert and Wand (1994)]. Local polynomials are not so attractive (in density estimation) in semiparametric terms, however.

5.2. *Local log-linear density.* Consider the local model $a\exp(b(t-x))$ for f around x [as does Loader (1996)]. The score function is $(1/a, t-x)'$, and the two equations to solve, in order to maximize the local likelihood, are

$$n^{-1}\sum_{i=1}^{n} K_h(x_i - x)\begin{pmatrix} 1/a \\ x_i - x \end{pmatrix} = \int K_h(t-x)\begin{pmatrix} 1/a \\ t-x \end{pmatrix} a\exp(b(t-x))\,\mathrm{d}t.$$

The components on the right-hand side can be written $\psi(bh)$ and $ah\psi'(bh)$, where $\psi(u) = \int \exp(uz)K(z)\,\mathrm{d}z$ is the moment-generating function for K. The two equations therefore become $\widetilde{f}(x) = a\psi(bh)$ and $\widetilde{g}(x) = ah\psi'(bh)$, where $\widetilde{g}(x)$ is the average of $K_h(x_i - x)(x_i - x)$. Note that the general recipe says $\widehat{f}(x) = f(x, \widehat{a}(x), \widehat{b}(x)) = \widehat{a}(x)$, so the $\widehat{b}(x)$ is only somewhat silently present when using this local reparametrization. Here one solves $\widetilde{g}(x)/\widetilde{f}(x) = h\psi'(bh)/\psi(bh)$ for b and in the end uses $\widehat{f}(x) = \widetilde{f}(x)/\psi(\widehat{b}h)$.

This apparatus can be used in particular when K is the standard normal. Some mild caution is called for since K then has unbounded support, to the effect that the local model is only trusted when $t \in x \pm 2.5\,h$, say. In this case, $\widetilde{g}(x)$ above is directly related to the derivative $\widetilde{f}'(x)$ of the ordinary kernel estimator; indeed, $\widetilde{g}(x) = h^2\widetilde{f}'(x)$. [(In fact, $\widetilde{g}(x)/(\sigma_K^2 h^2)$ is quite generally an estimator of f', usually a different one from $(\widetilde{f})'$. For comparisons, see Jones (1994).] This fact, combined with $\psi(u) = \exp(\frac{1}{2}u^2)$ and $\psi'(u) = \psi(u)u$, gives $\widehat{b} = \widetilde{f}'(x)/\widetilde{f}(x)$ and

$$\widehat{f}(x) = \widetilde{f}(x)\exp\bigl(-\tfrac{1}{2}h^2\widehat{b}^2\bigr) = \widetilde{f}(x)\exp\bigl[-\tfrac{1}{2}h^2\{\widetilde{f}'(x)/\widetilde{f}(x)\}^2\bigr]. \tag{5.1}$$

This particular version of our general local likelihood method performs accordingly an explicit correction to the traditional estimator, attempting to get the local slope right. Its bias in general is $\frac{1}{2}h^2\{f'' - (f')^2/f\} + O(h^4)$, which will be only $O(h^4)$ if the true model agrees with $a\exp(bt)$ on $|t-x| \le 2.5\,h$.

As mentioned in Section 4.3, $\widehat{b}$ will be more variable than $\widehat{a}$, and might require a larger window parameter for its estimation. The correction factor $\widehat{b}(x) = \widetilde{f}'(x)/\widetilde{f}(x)$ in (5.1) could therefore either be computed separately, for a somewhat larger h than that used for $\widetilde{f}$, or the values of $\widehat{b}(x)$ could be post-smoothed before being plugged into (5.1).

5.3. *Local level, slope, and curvature.* As a continuation of the previous special case, as well as of the theory of Section 4.3 and of Loader (1996), one

can try out $f(t) = a\exp\{b(t-x)+\frac{1}{2}c(t-x)^2\}$ for t in a neighborhood of x. This local model is meant to be able to capture local level, slope and curvature of the true density, in the neighborhood $t \in x \pm kh$, as above. For each given x there are now three equations to solve:

$$n^{-1}\sum_{i=1}^{n} K_h(x_i - x)\begin{pmatrix} 1/a \\ x_i - x \\ (x_i - x)^2 \end{pmatrix} = \int K_h(t-x)\begin{pmatrix} 1/a \\ t - x \\ (t-x)^2 \end{pmatrix} a\exp(b(t-x)+\tfrac{1}{2}c(t-x)^2)\,\mathrm{d}t.$$

The right-hand side gives three functions in (a, b, c) to equate to $\widetilde{f}(x)$, $\widetilde{g}(x)$ (given above) and $\widetilde{g}_2(x) = n^{-1}\sum_{i=1}^{n} K_h(x_i - x)(x_i - x)^2$. In the end the local likelihood estimator is $\widehat{f}(x) = f(x, \widehat{a}, \widehat{b}, \widehat{c}) = \widehat{a}$.

Finite-support kernels are perhaps advisable here, to secure finiteness of the integrals on the right-hand side. The equations must, in general, be solved numerically for each x; existence and uniqueness of a solution is guaranteed by concavity in $(\log a, b, c)$ of the local likelihood. Let us give the fairly explicit solution that is possible for the case of the standard normal ϕ being used for K, interpreting the local model to be an approximation on $t \in x \pm 2.5h$. In this case, $\widetilde{g} = h^2\widetilde{f}'$ and $\widetilde{g}_2 = h^2\widetilde{f} + h^4\widetilde{f}''$, bringing in information about the first and second derivative of the standard estimator. The three equations become

$$\begin{aligned} \widetilde{f}(x) &= (a/R)\exp\bigl(\tfrac{1}{2}h^2b^2/R^2\bigr), \\ \widetilde{f}'(x) &= (ab/R^3)\exp\bigl(\tfrac{1}{2}h^2b^2/R^2\bigr), \\ h^2\widetilde{f}(x) + h^4\widetilde{f}''(x) &= (ah^2/R^3)(1 + h^2b^2/R^2)\exp\bigl(\tfrac{1}{2}h^2b^2/R^2\bigr), \end{aligned}$$

where $R = (1 - ch^2)^{1/2}$. There is a unique solution if the $\widehat{c}$ found in a minute obeys $1 > \widehat{c}h^2$. Some manipulations show that $\widehat{R} = (1 - \widehat{c}h^2)^{1/2}$ can be found from

$$(1/\widehat{R}^2) - 1 = h^2\bigl[\widetilde{f}''(x)/\widetilde{f}(x) - \{\widetilde{f}'(x)/\widetilde{f}(x)\}^2\bigr] = h^2\widehat{D}.$$

This gives $\widehat{c} = \widehat{D}/(1 + h^2\widehat{D})$ and $\widehat{R} = (1 + h^2\widehat{D})^{-1/2}$, and, in the end,

$$\widehat{f}(x) = \widetilde{f}(x)\widehat{R}\exp\bigl[-\tfrac{1}{2}h^2\widehat{R}^2\{\widetilde{f}'(x)/\widetilde{f}(x)\}^2\bigr]. \tag{5.2}$$

Note that $\widehat{f}$ can be computed quite explicitly in cases (5.1) and (5.2). This is quite fortunate, of course, in view of the general complexity of our scheme.

Again, Loader (1996) has also, independently of the present authors, worked with local likelihood estimation of densities that are log-linear in polynomials. Formulae (5.1) and (5.2) are not in Loader (1996), but he comments further on the general implementation issues involved. The manipulations that led to (5.1) and (5.2) do not, unfortunately, extend so neatly to the log-cubic case.

5.4. *A running normal density estimate.* Let us fit the normal density locally using 1 and $t-x$ as weight functions in (1.4), that is,

$$n^{-1}\sum_{i=1}^{n} K_h(x_i-x)\begin{pmatrix}1\\ x_i-x\end{pmatrix} = \int K_h(t-x)\begin{pmatrix}1\\ t-x\end{pmatrix}\frac{1}{\sigma}\phi\Big(\frac{t-\mu}{\sigma}\Big)\,\mathrm{d}t$$

are solved to get hold of the local $\widehat{\mu}(x)$ and $\widehat{\sigma}(x)$. This should essentially take care of the local level and slope. If $K=\phi$ is used, then these equations after some calculations become

$$\begin{aligned}\widetilde{f}(x) &= \phi\Big(\frac{x-\mu}{(\sigma^2+h^2)^{1/2}}\Big)\frac{1}{(\sigma^2+h^2)^{1/2}},\\ \widetilde{f}'(x) &= -\frac{x-\mu}{(\sigma^2+h^2)^{3/2}}\phi\Big(\frac{x-\mu}{(\sigma^2+h^2)^{1/2}}\Big),\end{aligned}\tag{5.3}$$

essentially matching traditional estimates of f and f' with quantities predicted by the model. It follows that $\widetilde{q}(x)=\widetilde{f}'(x)/\widetilde{f}(x)=-(x-\mu)/(\sigma^2+h^2)$ and, when inserted in the first equation, this gives a single equation to solve for the local $\sigma=\widehat{\sigma}(x)$:

$$\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\sigma^2+h^2}}\exp\Big\{-\frac{1}{2}\widetilde{q}(x)^2(\sigma^2+h^2)\Big\}=\widetilde{f}(x).$$

There is a unique solution provided only $\phi(h\widetilde{q}(x))>h\widetilde{f}(x)$. Then the local $\mu=\widehat{\mu}(x)$ is found from $\widehat{\mu}(x)=x+\{\widehat{\sigma}(x)^2+h^2\}\widetilde{q}(x)$.

One may alternatively use the local likelihood function (1.3), that is, minimize

$$n^{-1}\sum_{i=1}^{n} K_h(x_i-x)\Big\{\log\sigma+\frac{\frac{1}{2}(x_i-\mu)^2}{\sigma^2}\Big\}+\phi\Big(\frac{x-\mu}{(\sigma^2+h^2)^{1/2}}\Big)\frac{1}{(\sigma^2+h^2)^{1/2}}\tag{5.4}$$

to produce $\widehat{\mu}(x)$ and $\widehat{\sigma}(x)$. This can be thrown to an optimizer or one could use, say, Newton–Raphson to solve the two equations that use the score functions $\sigma^{-2}(t-\mu)$ and $\sigma^{-1}\{(t-\mu)^2/\sigma^2-1\}$ as weight functions. These equations can be worked out to be

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} K_h(x_i-x)\frac{x_i-\mu}{\sigma} &= \frac{\sigma(x-\mu)}{(\sigma^2+h^2)^{3/2}}\phi\Big(\frac{x-\mu}{\sqrt{\sigma^2+h^2}}\Big),\\ \frac{1}{n}\sum_{i=1}^{n} K_h(x_i-x)\Big(\frac{(x_i-\mu)^2}{\sigma^2}-1\Big) &= \frac{\sigma^2}{(\sigma^2+h^2)^{3/2}}\phi\Big(\frac{x-\mu}{\sqrt{\sigma^2+h^2}}\Big)\Big(\frac{(x-\mu)^2}{\sigma^2+h^2}-1\Big).\end{aligned}$$

The running parameter estimates for both versions (5.3) and (5.4) would now have to be computed over a grid of x values. A practical suggestion would be to start optimizing or equation solving at a new x at the optimized values for the previous x.

The local log-likelihood $L_n(x,\theta)$ is not necessarily concave, but it should be so with high probability since the matrix of second derivatives goes to the $-J_h$ matrix, defined in Section 3, and the J_h matrix is symmetric and positive definite in this $v=u$ case. We are hopeful that simplistic computational schemes

should work well, a problem currently under investigation by J. Fosen, a student of the first author.

5.5. *Correcting a parametric start.* An alternative approach to semiparametric estimation might be to start with a known, or globally estimated parametric, model $f_{\text{init}}(t)$ and to multiply it with a local correction factor. Estimation of the local correction factor can conveniently take place within our local likelihood framework as follows. First, let $f(t, \theta) = f_{\text{init}}(t)\theta$. We think of $\theta = \theta(x)$ as the local correction factor for t near x. The local log-likelihood is $\widetilde{f}(x) \log \theta - \theta \int K_h(t - x) f_{\text{init}}(t)\, \mathrm{d}t$. The resulting estimator is

$$\widehat{f}(x) = f_{\text{init}}(x) \frac{\widetilde{f}(x)}{(K_h * f_{\text{init}})(x)} = \widetilde{f}(x) \frac{f_{\text{init}}(x)}{(K_h * f_{\text{init}})(x)}, \tag{5.5}$$

where $*$ denotes convolution. Note the simplicity and explicitness of this solution. The two expressions are meant to make clear two useful viewpoints: the estimator is a (typically parametric) start estimator times a nonparametric correction, and also the nonparametric kernel estimator times a parametric update.

We could also try $f(t) = f_{\text{init}}(t)\, a \exp(b(t - x))$ for t near x. The local log-likelihood becomes

$$\log a\, \widetilde{f}(x) + b\widetilde{g}(x) - a \int K(z) f_{\text{init}}(x + hz) \exp(bhz)\, \mathrm{d}z,$$

where $\widetilde{g}$ is as before. Note that the log-likelihood is concave in $(\log a, b)$. Maximizing the local likelihood gives two equations which will not be solvable explicitly in general. However, for the normal case and with a normal kernel we find

$$\widehat{f}(x) = \widetilde{f}(x)(1 + h^2/\sigma^2)^{1/2} \exp\bigl[-\tfrac{1}{2}h^2(1 + h^2/\sigma^2)\{\widetilde{f}'(x)/\widetilde{f}(x)\}^2\bigr]. \tag{5.6}$$

This is a (simpler) close relation of (5.2). In a way, however, the normal case is misleading in its potential: formulas like (5.2) and (5.6) are utilizing special properties of the normal to approximate the obvious bias correction $\widetilde{f}(x) - \frac{1}{2}h^2\widetilde{f''}(x)$, where $\widetilde{f''}$ is an appropriate estimator of f''.

Asymptotic bias properties of the above cases are interesting. Both have $b(x)$ of the form $(f - f_0)''(x)$, since when the local correction is a constant, u_0 is a constant also. In the local constant correction case, b can be written $f'' - ff''_{\text{init}}/f_{\text{init}}$; in the local exponential-of-linear correction case, some further analysis shows that the b function can be written $f'' - (f')^2/f + f(f'_{\text{init}})^2/f^2_{\text{init}} - ff''_{\text{init}}/f_{\text{init}}$. Each is zero if $f_{\text{init}} = f$. Another appealing b function which can be reached within this correction factor framework is $b = f_{\text{init}}(f/f_{\text{init}})''$, which is the bias factor function for Hjort and Glad's (1995) estimator. These authors' nonparametric correction to a parametric start arises if the kernel $K(z)$ is replaced by the modified (local) kernel $K(z)f_{\text{init}}(x)/f_{\text{init}}(x + hz)$ in the local constant correction described above. See Hjort (1996a) for further analysis and comparisons.

An interesting feature of semiparametric estimators of the form $\widetilde{f}$ times a parametric correction, as the second expression in (5.5), is that if taking a likelihood approach, one need not localize the likelihood, but may use a global likelihood to estimate the parameters in the parametric part, the localization already being attended to by $\widetilde{f}$. Efron and Tibshirani (1996) develop such an approach; in this connection, see also Hjort (1996a). Finally we point out that these local nonparametric multiplicative correction methods also work well when the initial estimator is itself nonparametric. When f_{init} is the kernel method, for example, (5.5) gives $\widetilde{f}_K^2/\widetilde{f}_{K*K}$, where the subscripts indicate the kernel functions used. Such estimators have bias of order h^4 and performance generally similar to that of an estimator investigated in Jones, Linton and Nielsen (1995); see Hjort (1996b).

5.6. *Local L_2-fitting.* Consider the local distance measure

$$\int K_h(t-x)\{f(t)-f(t,\theta)\}^2\,\mathrm{d}t,$$

an alternative to the local Kullback–Leibler distance (2.1). Multiplying out and disregarding the one term which does not depend on the parameter, we arrive at the following natural proposal: minimize, for each local x, the criterion function

$$Q_n(x,\theta)=\int K_h(t-x)f(t,\theta)^2\,\mathrm{d}t-2n^{-1}\sum_{i=1}^n K_h(x_i-x)f(x_i,\theta),$$

and use the accompanying version of $f(x,\widehat{\theta}(x))$. This would constitute a third possible avenue for computing a running normal estimate, for example. Taking the derivative it is seen that this local L_2-method is a special case of the general (1.4) method, with weight function $v(t,\theta)=f(t,\theta)u(t,\theta)$. Thus the theory developed applies to this case and suggests, in particular, that the behavior would be quite comparable to that of the other methods for small bandwidths. We would prefer the local likelihood to the local integrated quadratic for large and moderate h, that is, in situations where the parametric model used is not entirely inadequate, since the likelihood method is more efficient then. In the normal case, if h is large, the variance of the μ estimator is about 1.54 times higher with the L_2 method and the variance of the σ estimator about 1.85 times higher. However, the corresponding parameter estimates are more robust than the maximum likelihood ones. Further results and discussion are in Hjort (1994).

5.7. *Uniform kernel.* Let K be uniform on $[-\frac{1}{2},\frac{1}{2}]$. In this case the local log-likelihood function is $L_n(x,\theta)=n^{-1}\sum_W \log f(x_i,\theta)-\{F(x+\frac{1}{2}h,\theta)-F(x-\frac{1}{2}h,\theta)\}$, where the sum is over the window where $x_i\in x\pm\frac{1}{2}h$. Maximizing this essentially aims to match empirical facts from the local window $x\pm\frac{1}{2}h$ to behavior predicted by the parametric $f(\cdot,\theta)$ on this window. If $v_1(x,t,\theta)=1$ is one of the weights used in (1.4), then that equation simply matches the empirical and theoretical probabilities of falling inside this window.

5.8. *Relationship with moment estimation.* Note that as h becomes large, the (1.4) recipe ends up choosing as estimate the parameter value that solves $n^{-1}\sum_{i=1}^{n} v(x_i, \theta) = \mathrm{E}_\theta v(X_i, \theta)$, which is ordinary moment estimation with the $v_j(X_i, \theta)$ functions. This also indicates that having $v_1(t, \theta) = 1$ as first weight function, which we partly used in special cases above, does not work well with large hs. We would expect the two methods of obtaining a running normal density estimate, based on (5.3) and (5.4), respectively, to perform similarly for small hs, but the second method would perhaps be the best one for moderate and large hs.

6. Estimating the density at a boundary. Throughout the theoretical exposition so far, we have assumed that f has as its support the whole real line. In this section, we consider the presence of known boundaries to f's support. It will be general enough to consider positive data, and hence one boundary at zero. Consider estimation points x at and near the boundary in the sense that $x = ph$ for $0 \le p < 1$ and suppose K has support $[-1, 1]$. (This setup can easily be extended to infinite support kernels. However, finite support is a standard assumption, delineating boundary and interior regions. Results proved earlier continue to hold for x in the interior.) Define $a_l(p) = \int_{-1}^{p} u^l K(u)\,\mathrm{d}u$ and $b(p) = \int_{-1}^{p} K^2(u)\,\mathrm{d}u$. [Note that for $p \ge 1$, $a_0(p) = 1$, $a_1(p) = 0$, $a_2(p) = \sigma_K^2$ and $b(p) = R(K)$.]

6.1. *The one-parameter case.* For x near the boundary, formula (3.4) changes to

$$\int K_h(t-x)g(t)\,\mathrm{d}t = a_0(p)g(x) - a_1(p)hg'(x) + \tfrac{1}{2}a_2(p)h^2 g''(x) + O(h^3). \tag{6.1}$$

From this, it immediately follows that

$$\mathrm{E}\widehat{f}(x) \simeq f(x) - \{a_1(p)/a_0(p)\}h(f - f_0)'(x).$$

With a single locally fit parameter, therefore, boundary bias is of the undesirable $O(h)$ type unless one has been fortunate enough to choose one's parametric class equal to the true f near the boundary. The boundary variance can be found using the arguments of Section 4.2, with $V_h(z) = 1$:

$$J_h = a_0(p)f_0(x) + O(h) \quad \text{and} \quad M_h = b_0(p)f(x) + O(h). \tag{6.2}$$

These give a variance of

$$\mathrm{Var}\,\widehat{f}(x) \simeq (nh)^{-1}\{b(p)/a_0^2(p)\}f(x). \tag{6.3}$$

Bias and variance in (6.1) and (6.3) exactly match those of a standard kernel estimator divided by $a_0(p)$ save the replacement of f' by $f' - f_0'$ [e.g., Jones (1993b)]. That is, the one parameter local likelihood estimator behaves much like a renormalized kernel estimator with respect to boundaries. In Section 5.1 we noted that if the single parameter were a constant, such a renormalization explicitly and exactly takes place; however, the current asymptotic observations apply more generally to any one-parameter fitting.

6.2. *The two-parameter case.* Just as the local linear regression fit has an appealing $O(h^2)$ boundary bias (Fan and Gijbels, 1992), so too does the two-parameter locally parametric density estimator, as we shall now demonstrate.

Write $v_{0,j}$ for the derivative of v with respect to θ_j, $j = 1, 2$, evaluated at θ_0. To obtain the bias, we need to study the expansions of

$$\int_0^\infty K_h(t-x)v_{0,1}(t)\{f(t)-f(t_0)\}\,\mathrm{d}t = 0 = \int_0^\infty K_h(t-x)v_{0,2}(t)\{f(t)-f(t_0)\}\,\mathrm{d}t.$$

Expanding each to order h^3 and writing $(f - f_0) \simeq Ah^2$, $(f - f_0)' \simeq Bh$ and $(f - f_0)'' \simeq C$, we find that the $O(h^2)$ term in either side of the above expression involves $A - \{a_1(p)/a_0(p)\}B + \frac{1}{2}\{a_2(p)/a_0(p)\}C$ and that the difference between left- and right-hand sides yields an $O(h^3)$ term involving $-a_1(p)A + a_2(p)B - \frac{1}{2}a_3(p)C$. Setting these two quantities to zero and solving for A yields

$$\mathrm{E}\widehat{f}(x) = f(x) + \tfrac{1}{2}Q(p)h^2\{f''(x) - f_0''(x)\}, \tag{6.4}$$

where

$$Q(p) = \frac{a_2^2(p) - a_1(p)a_3(p)}{a_2(p)a_0(p) - a_1^2(p)}.$$

For the variance in the two-parameter case, simply use $V_h(z) = (1, hz)'$ in formula (4.4). We get

$$\mathrm{Var}\,\widehat{f}(x) = (nh)^{-1}f(x)\frac{\int\{a_2(p) - a_1(p)z\}^2K(z)^2\,\mathrm{d}z}{\{a_0(p)a_2(p) - a_1(p)^2\}^2} + O(n^{-1}). \tag{6.5}$$

The kernel-dependent asymptotic bias and variance terms are precisely those of the popular boundary kernel

$$\frac{a_2(p) - a_1(p)z}{a_0(p)a_2(p) - a_1(p)^2}K(z)$$

[see, e.g., Jones (1993b)]. That is, with two parameters we achieve $O(h^2)$ boundary bias (regardless of choice of local model) in an appealing way, and there is also the potential of further decrease in bias due to a good choice of model.

Three parameters can be expected to achieve $O(h^3)$ boundary bias, four parameters $O(h^4)$, but this is not pursued here.

7. Multidimensional data. The local likelihood method based on (1.3) generalizes easily to the case of d-dimensional data vectors, using d-dimensional kernel functions. The general weight function version of the method, through solving (1.4) for as many equations as there are parameters in the model used, is also operable in the vector case. A bivariate example could be to smooth the product-normal model, where the final estimator is of

the form

$$f(x, y, \widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}_1, \widehat{\sigma}_2) = \widehat{\sigma}_1^{-1}\phi(\widehat{\sigma}_1^{-1}(x - \widehat{\mu}_1))\,\widehat{\sigma}_2^{-1}\phi(\widehat{\sigma}_2^{-1}(x - \widehat{\mu}_2)).$$

This would smooth toward normal marginals, but also smooth somewhat toward independence.

Defining such estimators is, therefore, easy in principle, although computational matters become more complicated with the increasing number of running parameters to solve for. The local minimum Kullback–Leibler distance result of Section 2.1 is also seen to hold, giving support to the idea. Another question is to what extent the theory of the previous sections can be generalized, to establish properties of the resulting density estimators. We shall briefly go through the two-dimensional case to illustrate that the theory indeed goes through with appropriate extensions of previous techniques. Again it will be seen that the new method has scope for reduction of bias in a large neighborhood of densities around the parametric model employed. Our machinery could perhaps turn out to be of particular value in the multidimensional case, where there is much to lose and appalling convergence rates to meet by not imposing any structure at all.

Let $K(z_1, z_2) = K_1(z_1)K_2(z_2)$ be a product kernel. A good version of the traditional estimator is

$$\widetilde{f}(x_1, x_2) = n^{-1}\sum_{i=1}^{n} K_{h_1, h_2}(x_{i,1} - x_1, x_{i,2} - x_2) = n^{-1}\sum_{i=1}^{n} K_{h_1, h_2}(\mathbf{x}_i - \mathbf{x}),$$

where $K_{h_1, h_2}(\mathbf{t}) = h_1^{-1}K_1(h_1^{-1}t_1)h_2^{-1}K_2(h_2^{-1}t_2)$ and where we write $\mathbf{x} = (x_1, x_2)$ and so on; see Wand and Jones (1993). It has

$$\begin{aligned} \text{bias} &\simeq \sum_{i=1}^{2} \tfrac{1}{2}\sigma(K_i)^2 h_i^2 f''_{ii}(\mathbf{x}) \quad \text{and} \\ \text{variance} &\simeq \frac{R(K_1)R(K_2)f(\mathbf{x})}{nh_1h_2} - \frac{f(\mathbf{x})^2}{n}, \end{aligned} \tag{7.1}$$

where $\sigma(K_i)^2 = \int z^2 K_i(z)\,\mathrm{d}z$ and $R(K_i) = \int K_i(z)^2\,\mathrm{d}z$. We also use $f''_{ii}(\mathbf{x})$ for $\partial^2 f(\mathbf{x})/\partial x_i^2$ and so on.

The new locally parametric estimator is defined as $\widehat{f}(\mathbf{x}) = f(\mathbf{x}, \widehat{\theta}(\mathbf{x}))$, where the local parameter estimate solves

$$n^{-1}\sum_{i=1}^{n} K_{h_1, h_2}(\mathbf{x}_i - \mathbf{x})v(\mathbf{x}_i, \theta) - \int K_{h_1, h_2}(\mathbf{t} - \mathbf{x})v(\mathbf{t}, \theta)f(\mathbf{t}, \theta)\,\mathrm{d}\mathbf{t} = \mathbf{0}$$

around each given $\mathbf{x}$ point. Four independent equations are needed to handle the product-normal model above, for example. The expected value of $\widehat{f}(\mathbf{x})$ is $f(\mathbf{x}, \theta_0) + O((nh)^{-1})$, where θ_0 is locally least false and solves

$$V_j(\mathbf{x}, \theta) = \int K_{h_1, h_2}(\mathbf{t} - \mathbf{x})v_j(\mathbf{t}, \theta)\{f(\mathbf{t}) - f(\mathbf{t}, \theta)\}\,\mathrm{d}\mathbf{t} = 0 \quad \text{for } j = 1, \ldots, p.$$

Using

$$\int K_{h_1,h_2}(\mathbf{t}-\mathbf{x})g(\mathbf{t})\,\mathrm{d}\mathbf{t} = g(\mathbf{x}) + \sum_{i=1}^{2} \tfrac{1}{2}\sigma(K_i)^2 h_i^2 g''_{ii}(\mathbf{x}) + O((h_1^2+h_2^2)^2),$$

which is proved by Taylor expansions and properly generalizes (3.4), one finds that

$$\text{bias} \simeq \sum_{i=1}^{2} \frac{1}{2}\sigma(K_i)^2 h_i^2 \left[f''_{ii}(\mathbf{x}) - f''_{0,ii}(\mathbf{x}) + 2\frac{v'_{j,0,i}(\mathbf{x})}{v_{j,0}(\mathbf{x})}\{f'_i(\mathbf{x}) - f'_{0,i}(\mathbf{x})\}\right],$$

where f_0 and v_0 indicate the $f(\mathbf{t},\theta)$ and $v(\mathbf{t},\theta)$ functions with $\theta_0 = \theta_0(\mathbf{x})$ inserted. If there is more than one v_j function in direction x_i, then $f'_i(\mathbf{x}) - f'_{0,i}(\mathbf{x})$ is necessarily $o(1)$ and

$$\text{bias} \simeq \sum_{i=1}^{2} \tfrac{1}{2}\sigma(K_i)^2 h_i^2 \{f''_{ii}(\mathbf{x}) - f''_{0,ii}(\mathbf{x})\}. \tag{7.2}$$

Further, even $f''_{ii}(\mathbf{x}) - f''_{0,ii}(\mathbf{x})$ is $o(1)$ in directions involving three or more v_j functions and bias order can then be reduced, and so on. Turning next to the variance, one needs to consider

$$M = h_1 h_2 \,\mathrm{Var}_f\{K_{h_1,h_2}(\mathbf{X}_i - \mathbf{x})v_0(\mathbf{X}_i)\}$$

and

$$J = \int K_{h_1,h_2}(\mathbf{t}-\mathbf{x})[v_0(\mathbf{t})u_0(\mathbf{t})' f_0(\mathbf{t}) + v_0^*(\mathbf{t})\{f_0(\mathbf{t}) - f(\mathbf{t})\}]\mathrm{d}\mathbf{t}.$$

Using the same type of method as that used in Sections 4.2 and 4.3 in this more laborious situation, one ends up with exactly the same variance as in (7.1), to the order of approximation used, provided there are no more than two local parameters in each direction. [Extensions to higher numbers of parameters can be carried out, as with the case that led to (4.4).]

An interesting special case of the general method is that of a local model $f(t_1,t_2) = a\exp(b_1(t_1-x_1)+b_2(t_2-x_2))$, for $\mathbf{t}$ around $\mathbf{x}$, modelling local level and local slopes. The score function is $(1/a, t_1-x_1, t_2-x_2)'$ and gives three equations to solve for the three parameters. If the product normal kernel is used, calculations generalizing those of Section 5.2 yield

$$\widehat{f}(\mathbf{x}) = \widetilde{f}(\mathbf{x})\exp\left[-\tfrac{1}{2}\sum_{i=1}^{2} h_i^2\{\widetilde{f}'_i(\mathbf{x})/\widetilde{f}(\mathbf{x})\}^2\right]. \tag{7.3}$$

A more involved version can be given where the local curvatures $\exp\{\frac{1}{2}c_i(t_i - x_i)^2\}$ and/or the local covariance factor $\exp\{d(t_1-x_1)(t_2-x_2)\}$ are taken into account, thus generalizing the one-dimensional (5.2). Yet another estimator of interest evolves by modelling $f(\mathbf{t})$ as a global $f_{\text{init}}(\mathbf{t})$ times a local log-linear correction factor, in the spirit of Section 5.5. Explicit estimators can be written

out, similar to formula (5.6), for the case of a binormal start and Gaussian kernels.

8. Supplementing results and remarks.

8.1. *MSE and MISE analysis.* The approximate mean squared error (AMSE) for the new estimator is

$$\mathrm{AMSE}\{\widehat{f}(x)\} = \tfrac{1}{4}\sigma_K^4 h^4 b(x)^2 + R(K)(nh)^{-1} f(x),$$

with $b(x) = f''(x) - f_0''(x)$ in the typical case, and ignoring terms of order $n^{-1} + h^6 + h/n$ or smaller. For estimation consistency we need $h \to 0$ (forcing the bias to zero) while $nh \to \infty$ (forcing variance to zero). The theoretically best choice of h at x is therefore of the form $\{R(K)/\sigma_K^4\}^{1/5}\{f(x)/b(x)^2\}^{1/5}\, n^{-1/5}$, and the theoretically best AMSE is $\frac{5}{4}\{R(K)\sigma_K\}^{4/5} f(x)^{4/5} b(x)^{2/5}\, n^{-4/5}$. Choosing the best h for every x is generally too ambitious, and it is convenient to study the approximate or asymptotic mean integrated squared error $\mathrm{AMISE}(\widehat{f}) = \frac{1}{4}\sigma_K^4 h^4 R_{\mathrm{new}}(f) + R(K)(nh)^{-1}$, where $R_{\mathrm{new}}(f) = \int b(x)^2\,\mathrm{d}x$. The theoretically best global h-value is

$$h_0 = \{R(K)/\sigma_K^4\}^{1/5} R_{\mathrm{new}}(f)^{-1/5}\, n^{-1/5}, \tag{8.1}$$

leading to the theoretically best AMISE $\frac{5}{4}\{R(K)\sigma_K\}^{4/5} R_{\mathrm{new}}^{1/5}(f)\, n^{-4/5}$. We note that the Yepanechnikov kernel $K_0(z) = \frac{3}{2}(1 - 4z^2)_+$ (and scaled versions thereof) is optimal in that it manages to minimize $R(K)\sigma_K$; see, for example, Wand and Jones [(1995), Section 2.7].

8.2. *Comparison with the traditional method.* The calculations above are quite analogous to well known ones for the ordinary kernel method (which, in any case, are a special case). This also makes it easy to compare the two methods. Using the global (approximate) MISE criterion, we see that the new method is better provided $R_{\mathrm{new}}(f) < R_{\mathrm{trad}}(f)$, where the latter roughness quantity is $\int (f'')^2\,\mathrm{d}x$. This statement refers to the situation where both methods use the same kernel and the same bandwidth. If R_{new} really is smaller, then $\widehat{f}$ can be made even better by selecting a better h. This also defines a relatively broad nonparametric neighborhood of densities around the parametric model at which the new method is better. That R_{new} really offers a significant improvement on R_{trad} in many practical situations, for some of the new estimators displayed in Section 5, will be substantiated and reported in future work.

At a pointwise level, several points made by Hjort (1997) in the analogous locally parametric hazard estimation case are worth repeating, in modified form, here. First, it is easy to show that the locally parametric estimator is (asymptotically) better than the classical estimator whenever $0 \le f_0''(x)/f''(x) \le 2$. As long as f_0'' and f'' have the same sign, $|f_0''(x)|$ can afford to range over $[0, 2|f''(x)|]$. Note that this observation holds for small h, that is, at "the nonparametric end" of our semiparametric estimator. We should also note,

however, that differences in the constant involved in the bias may not be all that important, since the squared bias makes up only 1/5 of optimized mean squared error, the remainder being due to variance.

The locally parametric estimator is also designed to have special advantages over the kernel estimator when f is, in fact, close to f_0. Regardless of this, the kernel estimator has mean squared error of order $h^4 + (nh)^{-1}$ which is minimized by taking $h \sim n^{-1/5}$ and hence optimal mean squared error of $O(n^{-4/5})$. On the other hand, one might quantify closeness of f_0 and f by setting $(f_0 - f)'' \sim n^{-\epsilon}$ for some $0 < \epsilon < \frac{1}{2}$. The mean squared error of the locally parametric estimator is thus $h^4 n^{-2\epsilon} + (nh)^{-1}$ which is optimized by taking $h \sim n^{-(1-2\epsilon)/5}$. The optimized mean squared error is then of order $n^{-(4+2\epsilon)/5}$. For instance, if f_0'' and f'' are $O(n^{-1/4})$ apart, the mean squared error is improved to $O(n^{-9/10})$, and as the difference tends to $n^{-1/2}$, the mean squared error tends to n^{-1}.

8.3. *Choosing the bandwidth.* Methods for automatic bandwidth selection for the traditional kernel density estimator are reviewed by Jones, Marron and Sheather (1995). They might be utilized unaltered for locally parametric estimates, at least as a first attempt. However, if we are using an estimator that does indeed improve on the basic one, we will be oversmoothing relative to the new optimal choice. An argument in Section 8.2 suggests that the degree of oversmoothing may not often be very great, however.

The best of the bandwidth selectors in the ordinary case are founded on good estimates of unknown quantities in MISE expressions. The key is usually in the estimation of $R_{\text{trad}} = R(f'')$, and this transfers to the need to estimate $R_{\text{new}} = R((f - f_0)'')$ (one might think of adapting traditional selectors by multiplying them by an estimate of $(R_{\text{trad}}/R_{\text{new}})^{1/5}$). The estimation of R_{new} is not straightforward since it involves the second derivatives of both the true f and its best possible approximant of the form $f(x, \theta(x))$. One *possibility* might be to estimate f'' by $\widetilde{f}''$ using a different bandwidth g which is optimal for estimation of $R(f'')$ [this is what happens in a good bandwidth selector for the traditional estimator; see e.g., Sheather and Jones (1991)] and f_0'' by $\widetilde{f}''$ using the same h as for estimation of f. This type of difficulty extends to rule-of-thumb approaches too. We should also mention that it could be worthwhile to employ more than one bandwidth when forming an estimator based on several equations, as for the methods of Sections 5.2 and 5.3. This is because local slope equations typically would benefit from larger bandwidths than for local level equations.

Least squares cross-validation, which for the traditional estimator is less reliable than the best methods [Jones, Marron and Sheather (1996)], has the advantage that it does not involve f_θ explicitly. One can just follow the usual idea of estimating $\mathrm{E}\{\int \widehat{f}(x)^2 \,\mathrm{d}x - 2\int f(x)\widehat{f}(x)\,\mathrm{d}x\}$ by $\int f(x, \widehat{\theta}(x))^2 \,\mathrm{d}x - 2n^{-1}\sum_{i=1}^n f(x_i, \widehat{\theta}_{(i)}(x_i))$, where numerical integration is used for the first term and $\widehat{\theta}_{(i)}(x)$ is the leave-one-out version of $\widehat{\theta}(x)$.

Alternative methods are also worth considering, particularly since one sometimes would be interested in using moderate or large hs, namely, in situations where the data fit the local model well. A changing and adaptively defined h could be advantageous in some cases. Hjort (1997) considers a local goodness-of-fit approach in the hazard case: increase the bandwidth until the local model fails to pass a goodness-of-fit criterion. Extension of this methodology to the density case is an interesting topic for further research, one possibility being to exploit results of Section 8.5.

8.4. *Large-sample normality.* The basic bias and variance results for our estimator $f(x, \widehat{\theta}(x))$ were derived in Sections 3 and 4. Our arguments were based on claims (3.2) and (3.3) about limiting normality for $\widehat{\theta}(x)$, and in fact also on variants of these that work in the framework where the smoothing parameter h is not fixed but goes to zero with n. Here we outline proofs of precise versions of these claims.

The $\widehat{\theta}(x)$ we consider is the solution to (1.4). For convenience we partly suppress the fixed x in the notation now. Taylor expansion analysis for $V_n(\widehat{\theta}) = 0$ gives

$$(nh)^{1/2}(\widehat{\theta} - \theta_0) \simeq -V_n^*(\theta_0)^{-1}(nh)^{1/2}V_n(\theta_0), \tag{8.2}$$

where V_n^* is the $p \times p$ matrix of partial derivatives of the $V_{n,j}(\theta)$ functions, and this leads to a $J_h^{-1}\mathcal{N}_p\{0, M_h\}$ limit by well known arguments. A more formal proof starts out by observing that $\widehat{\theta}$ can be seen as the functional $T(F_n)$, where $T(F)$ is the solution to $w(F, \theta) = \int K_h(t-x)v(t, \theta)\{\mathrm{d}F(t) - f(t, \theta)\,\mathrm{d}t\} = 0$; see (3.1). Under regularity assumptions this is a second-order smooth functional in the sense of Shao (1991), with influence function

$$I(F, t) = J_h^{-1}\left\{K_h(t-x)v(t, \theta_0) - \int K_h(t-x)v(t, \theta_0)f(t, \theta_0)\,\mathrm{d}t\right\},$$

in which $\theta_0 = T(F)$. This is seen from a Taylor expansion of $v((1-\varepsilon)F + \varepsilon\delta_t, \theta)$ around θ_0, where δ_t is unit point mass at t. This is sufficient for consistency and a normal $\{0, J_h^{-1}M_h(J_h')^{-1}\}$ limit for $(nh)^{1/2}(\widehat{\theta} - \theta_0)$; see Shao (1991). These arguments, in conjunction with the theory and tools developed in Sections 4.2 and 4.3, can also be used to prove

$$(nh)^{1/2}\{f(x, \widehat{\theta}(x)) - f(x) - b_n(x)\} \to_d \mathcal{N}\{0, \tau(K)^2 f(x)\} \tag{8.3}$$

when $h \to 0$ and $nh \to \infty$. Here $b_n(x)$ is the bias of $f(x, \widehat{\theta}(x))$ and is of the form $\frac{1}{2}\sigma_K^2 h^2 b(x) + o(h^2)$ for appropriate $b(x)$ functions in the case of one- and two-parameter local families and of the form $c(K)h^4 b(x) + o(h^4)$ for certain other $b(x)$ functions in the case of three- and four-parameter local families; see equations (3.7), (4.1), (4.2) or (4.3). Also, $\tau(K)^2$ is the general variance factor appearing in formula (4.4).

Another useful version of such a precise result, valid in the general log-linear case (cf. the special cases treated in Sections 4.2, 4.3, 5.2 and 5.3) is as

follows. Let the model be of the form $f(t, \theta) = \exp\{\theta' w(t)\}$, where $w(t)$ is a vector of p functionally independent and twice differentiable weight functions. We assume that $\theta' w(t)$ spans the full real line as θ varies. The local log-likelihood

$$L_n(x, \theta) = n^{-1} \sum_{i=1}^{n} K_h(x_i - x)\theta' w(x_i) - \int K_h(t - x) \exp\{\theta' w(t)\}\, \mathrm{d}t$$

is concave in θ. Let $\theta_{0,h}$ be the unique maximizer of the limit function or, equivalently, the unique solution to $\int K_h(t-x)w(t)[f(t)-\exp\{\theta' w(t)\}]\, \mathrm{d}t = 0$. Next study the function

$$A_n(s) = nh\{L_n(x, \theta_{0,h} + s/(nh)^{1/2}) - L_n(x, \theta_{0,h})\}.$$

It is concave in s and inspection shows that it can be expressed as $s'U_n - \frac{1}{2}s'J_n s + O(\|s\|^3/(nh)^{1/2})$. Here

$$U_n = n^{-1/2} \sum_{i=1}^{n} h^{1/2}\{K_h(x_i - x)w(x_i) - \xi_n\}$$

with $\xi_n = \int K_h(t - x)w(t)f(t)\, \mathrm{d}t$, and $J_n = \int K_h(t - x)f_0(t)w(t)w(t)'\, \mathrm{d}t$. The point is now that the maximizer of $A_n(s)$, which is $(nh)^{1/2}(\widehat{\theta} - \theta_{0,h})$, must be close to the maximizer of the quadratic approximation $s'U_n - \frac{1}{2}s'J_n s$, which is $J_n^{-1}U_n$. Precise general concavity-based arguments are in Hjort and Pollard (1996). Now $J_n^{-1}U_n$ has a covariance matrix which stabilizes as n grows, and using the Lindeberg theorem it is not difficult to show that it is asymptotically normal. The delta method, combined with the arguments that led to (3.6) and (4.4), then gives the appropriate version of (8.3) again.

8.5. *Parameter inspection.* Plotting the estimated running parameter $\widehat{\theta}(x)$ against x is a natural idea. This could be used for model exploration purposes and for goodness-of-fit testing. Monitoring $\widehat{\theta}(x)$ based on a pilot value of h can also be used for choosing the final bandwidth or for post-smoothing before being used in the final $f(x, \widehat{\theta}(x))$.

From the discussion of Section 3.1, it is clear that $\widehat{\theta}(x)$ aims at the locally least false parameter value $\theta_0(x)$, which is a constant value θ_0 independent of x if and only if the parametric model used is perfect. The approximate precision of $\widehat{\theta}(x)$ can be worked out from $J_h^{-1}M_h(J_h')^{-1}$ of (3.2) using methods developed in connection with (3.6) and (4.4). To a first order approximation the variances for the components of $\widehat{\theta}(x)$ are inversely proportional to $nhf(x)$ and hence their plots cannot normally be trusted in regions of small density. We note that both weight functions $v(t, \theta)$ as well as characteristics of the model used show up in explicit calculations for the variance matrix for $\widehat{\theta}(x)$, in contrast with the analogous calculation for the variance of $f(x, \widehat{\theta}(x))$, ending with (3.8) and (4.4).

9. Conclusions. We believe we have been studying *the* most attractive way of doing semiparametric density estimation. The estimators run the gamut from a fully parametric fit to almost fully nonparametric (except with some small change in performance which may well be beneficial) with only a single smoothing parameter to be chosen. The number of parameters in the "local model" crucially affects performance: one and two fitted parameters are most readily comparable with ordinary kernel density estimation, three and four fitted parameters with fourth-order kernel estimation and more parameters with higher order estimates. Even numbers of fitted parameters have advantages in terms of simplicity and interpretability of bias. These comments parallel the fitting of local polynomials in regression, but we note that they are driven by numbers of parameters only (which are effectively automatically reparametrized into intercept, slope, curvature, etc., parameters) and not by the specific functional form. Together with and in generalization of Loader (1996), we believe we have laid firm theoretical foundations for locally parametric nonparametric density estimation. Much still remains to be done in terms of exploring practical issues and applications.

Acknowledgments. We have had fruitful discussions with John Copas, Oliver Linton and Jens Perch Nielsen. Comments from referees inspired improvements over an earlier version. Part of this work was carried out while the second author visited the University of Oslo with partial support from its Department of Mathematics.

REFERENCES

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

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836.

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

EFRON, B. and TIBSHIRANI, R. (1996). Using specially designed exponential families for density estimation. *J. Amer. Statist. Assoc.* To appear.

FAN, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87** 998–1004.

FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21** 196–216.

FAN, J. and GIJBELS, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20** 2008–2036.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman and Hall, London.

FAN, J., HECKMAN, N. E. and WAND, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90** 141–150.

FENSTAD, G. U. and HJORT, N. L. (1996). Two Hermite expansion density estimators, and a comparison with the kernel method. Unpublished manuscript.

HASTIE, T. and LOADER, C. R. (1993). Local regression: Automatic kernel carpentry (with discussion). *Statist. Sci.* **8** 120–143.

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

HJORT, N. L. (1986). *Theory of Statistical Symbol Recognition.* Norwegian Computing Centre, Oslo.

HJORT, N. L. (1991). Semiparametric estimation of parametric hazard rates. In *Survival Analysis: State of the Art* (P. S. Goel and J. P. Klein, eds.) 211–236. Kluwer, Dordrecht.

HJORT, N. L. (1994). Minimum L2 and robust Kullback–Leibler estimation. In *Proceedings of the 12th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes* (P. Lachout and J. Á. Víšek, eds.) 102–105. Academy of Sciences of the Czech Republic, Prague.

HJORT, N. L. (1995). Bayesian approaches to semiparametric density estimation (with discussion). In *Bayesian Statistics* V (J. Bernardo, J. Berger, P. Dawid and A. F. M. Smith, eds.) 223–253. Oxford Univ. Press.

HJORT, N. L. (1996a). Performance of Efron and Tibshirani's semiparametric density estimator. Statistical research report, Dept. Mathematics, Univ. Oslo.

HJORT, N. L. (1996b). Multiplicative higher order bias kernel density estimators. Statistical research report, Dept. Mathematics, Univ. Oslo.

HJORT, N. L. (1997). Dynamic likelihood hazard rate estimation. *Biometrika* **84**. To appear.

HJORT, N. L. and GLAD, I. K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.* **23** 882–904.

HJORT, N. L. and POLLARD, D. B. (1996). Asymptotics for minimizers of convex processes. Unpublished manuscript.

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

JONES, M. C. (1993b). Simple boundary correction for kernel density estimation. *Statistics and Computing* **3** 135–146.

JONES, M. C. (1994). On kernel density derivative estimation. *Comm. Statist. Theory Methods* **23** 2133–2139.

JONES, M. C. (1995). On close relations of local likelihood density estimation. Unpublished manuscript.

JONES, M. C., DAVIES, S. J. and PARK, B. U. (1994). Versions of kernel-type regression estimators. *J. Amer. Statist. Assoc.* **89** 825–832.

JONES, M. C. and FOSTER, P. J. (1993). Generalized jackknifing and higher order kernels. *J. Nonparametr. Statist.* **3** 81–94.

JONES, M. C. and HJORT, N. L. (1994). Local fitting of regression models by likelihood: what's important? Statistical research report, Dept. Mathematics, Univ. Oslo.

JONES, M. C., LINTON, O. and NIELSEN, J. P. (1995). A simple and effective bias reduction method for density and regression estimation. *Biometrika* **82** 327–338.

JONES, M. C., MARRON, J. S. and SHEATHER, S. J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.* **91** 401–407.

LINDSEY, J. K. (1974). Comparison of probability distributions. *J. Roy. Statist. Soc. Ser. B* **36** 38–47.

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

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.

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

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.

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. Assoc. B* **53** 683–690.

STANISWALIS, J. (1989). The kernel estimate of a regression function in likelihood-based models. *J. Amer. Statist. Assoc.* **84** 276–283.

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–620.

TIBSHIRANI, R. and HASTIE, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82** 559–567.

WAND, M. P. and JONES, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.* **88** 520–528.

WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF OSLO
P.O. BOX 1053 BLINDERN
N-0316 OSLO
NORWAY
E-MAIL: nils@math.uio.no

DEPARTMENT OF STATISTICS
THE OPEN UNIVERSITY
WALTON HALL
MILTON KEYNES, MK7 6AA
ENGLAND