Biometrika (1999), **86,** 2, *pp.* 417–428

Printed in Great Britain

High-derivative parametric enhancements of nonparametric curve estimators

BY MING-YEN CHENG
Department of Mathematics, National Taiwan University, Taipei 106, Taiwan
cheng@math.ntu.edu.tw

PETER HALL
Centre for Mathematics and its Applications, Australian National University, Canberra ACT 0200, Australia
peter.hall@anu.edu.au

AND BERWIN A. TURLACH
Department of Statistics, University of Adelaide, Adelaide SA 5005, Australia
bturlach@stats.adelaide.edu.au

SUMMARY

We suggest a method for using parametric information to modify a nonparametric estimator at the level of relatively high-order derivatives. The technique represents an alternative to methods that first fit a parametric model and then adjust it. In particular, relative to a 'nonparametric estimator with a parametric start', our estimator is not biased by the differences between parametric and nonparametric fits to low-order derivatives, since we effectively remove all the parametric information about low-order derivatives and replace it by nonparametric information. Thus, we employ parametric information only when the nonparametric information is unreliable, and do not use it elsewhere. The method has application to both nonparametric density estimation and nonparametric regression.

Some key words: Bias reduction; Curve estimation; Density estimation; Kernel regression; Local polynomial regression; Locally parametric methods; Log-polynomial model; Nonparametric regression.

1. INTRODUCTION

At least two different techniques have been proposed for combining parametric and nonparametric information in a curve estimator. One, based purely on locally parametric fitting, takes a potential parametric model and fits it in the neighbourhood of each point. Local polynomial regression, e.g. Fan & Gijbels (1996), is a relatively well-understood example; locally parametric methods in density estimation, e.g. Hjort (1994), Hjort & Jones (1996) and Loader (1996), are more recent. A second technique is first to fit a parametric model and then to attempt to adjust it, using nonparametric information.

The two approaches are related. For example, Hjort & Glad's (1995) 'nonparametric estimator with a parametric start' can be viewed as using locally parametric methods to modify an initial, global parametric approximation. This amounts to modifying the global

estimator simultaneously at several levels. While it has significant merits in some circumstances, it introduces a bias term for each level of correction.

We suggest an alternative way of combining local and global parametric models, making a nonparametric correction at only one level. We take the view that global parametric information is most valuable in the case of relatively high-order derivatives, where local information is either unreliable, as a result of high amounts of stochastic error, or not readily available. Throughout this paper, the terms 'high order' and 'low order' refer to relative orders of derivatives; thus, $f^{(r)}$ is a derivative of order r and is of lower order than $f^{(s)}$ if $r < s$.

Since local estimation of level and slope is often accurate even for relatively small sample sizes, but local estimation of curvature is reasonable only for large samples, information from a parametric model could be used to describe the second derivative explicitly. The zeroth and first derivatives could be approximated directly from the data.

For example, traditional local-linear methods in nonlinear regression involve fitting the model $\psi_1(u\,|\,\theta_1, \theta_2) \equiv \theta_1 + \theta_2(u - x)$ to data with design points in the neighbourhood of x. One could consider models of higher order, in particular the local quadratic

$$\psi_2(u\,|\,\theta_1, \theta_2, \theta_3) \equiv \theta_1 + \theta_2(u - x) + \tfrac{1}{2}\theta_3(u - x)^2.$$

However, numerical difficulties and the relatively high degree of inaccuracy associated with estimating θ_3, which corresponds to the second derivative of the regression mean, m say, at x, can make this unattractive. We suggest adopting a global parametric model, m_0, for the regression mean, and replacing θ_3 by $m_0''(x)$. Then we fit the two-parameter model $\psi_3(u\,|\,\theta_1, \theta_2) \equiv \psi_2\{u\,|\,\theta_1, \theta_2, m_0''(x)\}$ in place of $\psi_1(u\,|\,\theta_1, \theta_2)$.

Numerical difficulties associated with fitting ψ_3 are no greater than those when fitting ψ_1. Moreover, for a given bandwidth the stochastic error of the fit is no greater, and there can be significant reductions in bias. First-order properties of variance are identical to those under the original local model, ψ_1. Therefore, the high-order parametric enhancement provided by fitting second derivatives of m_0 is robust against even relatively serious misspecification of m_0. Additionally, there is the potential of reducing bias to zero. These results will be made more precise, in the cases of both nonparametric regression and nonparametric density estimation and for general local and global models, in § 5.

If the global parametric model were readily amenable to analysis then, in principle, one could take it as the local model and simply employ locally parametric methods. However, the global model is often not well suited to local use, for example because its complexity can lead to computational difficulties. In particular, if the global model is multimodal, then fitting it locally can be very difficult, but it is in just such complex cases that the extra qualitative information available about high-order factors, such as curvature, is of most value. Our method provides a way of using it, without being troubled by low-order inconsistencies between the model and the true curve.

The log-linear approach to density estimation is a good example of a locally parametric method that is chosen significantly for its good computational properties, and has problems capturing certain important qualitative features of a density. In particular, estimators based on locally fitted log-linear models are even less able than standard kernel methods to reach into the mode of a density, since they are more negatively biased there than second-order kernel approaches. However, this drawback may be largely overcome by combining local log-linear fitting with a moderately accurate global model which has modes in approximately the same places as the true density.

The global model might be viewed as a Bayesian prior, which we modify by incorporat-

ing local information from the data. In the main examples considered in § 2, in particular (2·3) and (2·4), the prior is used to supply information about curvature, and is modified at a local level by using the data to estimate level and slope. A case of particular interest is that where the global and local models are polynomials, or log-polynomials in the case of density estimation, of degrees d and $d' < d$, respectively. Once the parameters of the global model have been estimated, only d' parameters in the local model remain to be chosen. Our exposition will address $(d', d) = (1, 2)$, and other cases may be treated similarly.

Section 2 introduces low-order nonparametric approximations, derived by combining high-order properties of the global model with low-order, locally parametric methods. In the context of density estimation we consider both quadratic and Kullback–Leibler quantification of loss in locally parametric fits, but for regression we treat only quadratic loss. Methods of assessing loss are described in § 3, and numerical properties of the resulting estimators are discussed in § 4. Theoretical properties are sketched in § 5.

2. Low-order nonparametric corrections

2·1. *Corrections in the case of density estimation*

Two models for f are involved: a local model, $\psi(.\,|\,\theta)$, which might for example be log-polynomial, and would be chosen at least partly on grounds of computational expediency; and a global model, f_0, which would typically be determined by a relatively small number of parameters estimated from the data. The local model is not intended to capture the shape of f in any qualitative sense; that is done by f_0.

Next we describe construction of ψ, beginning by introducing a locally parametric precursor, ϕ. Given integers $1 \leqslant r < s$, let $\phi(.\,|\,\alpha_1, \ldots, \alpha_s)$ be an s-variate model for the true density, f, in a neighbourhood of x. For the sake of simplicity, assume that ϕ has been parameterised so that $\alpha_i = f^{(i-1)}(x)$, for $1 \leqslant i \leqslant r$, and that $\alpha_{r+1}, \ldots, \alpha_s$ denote known functions of $f(x), f^{(1)}(x), \ldots, f^{(s-1)}(x)$; see (2·2) for an example. In formulating the local model we shall estimate $\alpha_1, \ldots, \alpha_r$ wholly nonparametrically, and employ information in the global model to assist approximation to $\alpha_{r+1}, \ldots, \alpha_s$.

There are at least two ways in which this can be done. First, we can simply replace each $f^{(i-1)}(x)$ by $f_0^{(i-1)}(x)$, for $1 \leqslant i \leqslant s$, whenever it appears in the known formula for α_j $(r+1 \leqslant j \leqslant s)$. Call this Case I. Secondly, we can make this substitution for $f^{(i-1)}(x)$ when $r+1 \leqslant i \leqslant s$, but replace $f^{(i-1)}(x)$ by θ_i when $1 \leqslant i \leqslant r$, whenever either of these quantities appears in the formula for α_j for $r+1 \leqslant j \leqslant s$. Call this case II. In either case the local model is

$$\psi(.\,|\,\theta) = \phi(.\,|\,\theta_1, \ldots, \theta_r, \alpha^*_{r+1}, \ldots, \alpha^*_s),$$

where, for $r+1 \leqslant i \leqslant s$, we take

$$\alpha_i^* = \alpha_i\{f_0(x), \ldots, f_0^{(s-1)}(x)\}$$

in Case I and

$$\alpha_i^* = \alpha_i\{\theta_1, \ldots, \theta_r, f_0^{(r)}(x), \ldots, f_0^{(s-1)}(x)\}$$

in Case II. We propose fitting $\psi(.\,|\,\theta)$ using locally parametric methods, with either quadratic or Kullback–Leibler loss describing performance.

The model ϕ would be constructed so that $\psi^{(i)}(x\,|\,\theta) \doteq f^{(i)}(x)$ for $0 \leqslant i \leqslant s-1$ if both (a) $\theta_i \doteq f^{(i-1)}(x)$ for $1 \leqslant i \leqslant r$, and (b) $f_0^{(i-1)}(x) \doteq f^{(i-1)}(x)$ for $r+1 \leqslant i \leqslant s$. The approxi-

mation for $1 \leqslant i \leqslant r$ is guaranteed by the flexibility of locally parametric methods, and the approximation for $r+1 \leqslant i \leqslant s$ is valid if the global model is moderately accurate at high orders. We do not need the global model to be particularly accurate, only approximate; after all, it corrects the local model only in high-order terms.

If r is even, as in the example at (2·3), then the most important of the high-order approximations is the first, with $i = r+1$:

$$\psi^{(r)}(x \mid \theta) \doteq f^{(r)}(x). \tag{2·1}$$

When this condition is satisfied, properties of locally parametric estimators ensure that bias is low; see § 5.

The example to which we shall devote most attention is that where ϕ is a log-quadratic model and ψ is log-linear. Thus, we rely on the global model for qualitative information about curvature. To this end, define $\alpha = (\alpha_1, \alpha_2, \alpha_3)^{\mathrm{T}}$ and $\theta = (\theta_1, \theta_2)^{\mathrm{T}}$, and in a neighbourhood of x consider the models

$$\phi(u \mid \alpha) = \alpha_1 \exp\{(\alpha_2/\alpha_1)(u-x) + \tfrac{1}{2}\alpha_3(u-x)^2\}, \tag{2·2}$$

$$\psi(u \mid \theta) = \theta_1 \exp\{(\theta_2/\theta_1)(u-x) + \tfrac{1}{2}\alpha_3(u-x)^2\}. \tag{2·3}$$

Noting that $\phi(x \mid \alpha) = \alpha_1$, $\phi'(x \mid \alpha) = \alpha_2$ and $\phi''(x \mid \alpha) = \alpha_1^{-1}\alpha_2^2 + \alpha_1\alpha_3$, we see that we should choose α_3 so that $f''(x) \doteq f(x)^{-1}f'(x)^2 + f(x)\alpha_3$. Therefore, Case I of our procedure would use

$$\alpha_3^* = \{f_0''(x)/f_0(x)\} - \{f_0'(x)/f_0(x)\}^2$$

in (2·3), while, in Case II,

$$\alpha_3^* = \alpha_3^*(\theta) = \{f_0''(x)/\theta_1\} - (\theta_2/\theta_1)^2.$$

2·2. *Corrections in the case of regression*

The same methods apply, although the class of potential models is different. For brevity we pass directly to regression versions of the polynomial models at (2·2) and (2·3). Let m denote the true regression mean, and let m_0 be a global parametric model for m. Local parametric models are given by

$$\begin{aligned} \phi(u \mid \alpha) &= \alpha_1 + \alpha_2(u-x) + \tfrac{1}{2}\alpha_3(u-x)^2, \\ \psi(u \mid \theta) &= \theta_1 + \theta_2(u-x) + \tfrac{1}{2}m_0''(x)(u-x)^2. \end{aligned} \tag{2·4}$$

In this setting, Cases I and II of our approach are identical, since $\phi''(x \mid \alpha) = \alpha_3$ alone. Again we suggest fitting $\psi(. \mid \theta)$ using locally parametric methods, employing quadratic loss to select θ.

In the case of regression there do not exist general, canonical models to compare with such as the normal density and its mixtures in density estimation. However, in some special cases there are natural counterparts. These include parametric models that are traditional in settings where nonparametric methods are being considered. See, for example, models for the analysis of growth curve data (Gasser et al., 1985) and household income data (Hildenbrand & Hildenbrand, 1986); and logistic models and, more generally, five-parameter models of Richards type, e.g. Ratkowsky (1983, pp. 61–8), for the analysis of data on biological population size. Härdle & Mammen (1993) have addressed the problem of comparing parametric and nonparametric regression fits in a wide class of contexts.

3. Locally parametric methods

3·1. *Empirical measures of loss in density estimation*

Suppose we observe data $X_1, \ldots, X_n$ from a distribution with density f on a closed interval $\mathscr{I}$, such as the whole real line or a compact subset of it. Let $\psi(.|\theta)$ be the model for f, governed by the parameter vector $\theta = (\theta_1, \ldots, \theta_r)^{\mathrm{T}}$, and put $w(u) = h^{-1}K\{(u-x)/h\}$, where K is a bounded, symmetric, compactly supported probability density. An example of $\psi(.|\theta)$, in the case $r = 2$, is given at (2·3). Note, however, that, as explained two sentences below that display, α_3 depends on θ in Case II but not in Case I. Define $\hat{\theta} = \hat{\theta}(x, h)$ to be the minimiser of either

$$Q(\theta) = \int w(u)\psi(u|\theta)^2\, du - 2n^{-1}\sum_{i=1}^{n} w(X_i)\psi(X_i|\theta)$$

or

$$L(\theta) = \int w(u)\psi(u|\theta)\, du - n^{-1}\sum_{i=1}^{n} w(X_i)\log\psi(X_i|\theta),$$

denoting quadratic loss and Kullback–Leibler loss respectively. A locally parametric estimator of $f(x)$ is $\hat{f}(x) = \psi(x|\hat{\theta})$.

3·2. *Deterministic loss functions associated with Q and L*

Except for terms that do not depend on θ, we may regard $Q(\theta)$ and $L(\theta)$ as empirical versions of the deterministic loss functions $\omega(\theta) \equiv \omega\{f, \psi(.|\theta)\}$ and $\lambda(\theta) \equiv \lambda\{f, \psi(.|\theta)\}$, respectively, where

$$\omega(f, g) = \int w(g-f)^2\, du, \quad \lambda(f, g) = -\int w\{f\log(g/f) - (g-f)\}\, du$$

are distance functions on the set of all probability densities. The relationships between ω and Q and between λ and L may be seen by noting that $Q(\theta) \to \omega(\theta) - D_1$ and $L(\theta) \to \lambda(\theta) - D_2$ as $n \to \infty$, where $D_1 = \int \omega f^2$ and $D_2 = \int \omega(f\log f - f)$ do not depend on θ; and that $\omega(\theta) = E\{Q(\theta)\} + D_1$ and $\lambda(\theta) = E\{L(\theta)\} + D_2$.

3·3. *Quadratic loss in nonparametric regression*

Assume data pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are generated by the model $Y_i = m(X_i) + \varepsilon_i$, where the X_i's are independent and identically distributed with density f, and, conditional on the X_i's, the ε_i's are independent with density f. Let $\psi(.|\theta)$ be a model for m, such as that at (2·4). In this setting we define

$$Q(\theta) = n^{-1}\sum_{i=1}^{n} w(X_i)\{Y_i - \psi(X_i|\theta)\}^2,$$

choose $\hat{\theta}$ to minimise Q, and take $\hat{m}(x) = \psi(x|\hat{\theta})$ as our estimator of m. The analogue of ω here is

$$\omega(\theta) = n^{-1}\sum_{i} w(X_i)\{\psi(X_i|\theta) - m(X_i)\}^2.$$

Note that ω depends on $X_1, \ldots, X_n$, although not on $Y_1, \ldots, Y_n$.

4. NUMERICAL PROPERTIES

We conducted a comprehensive simulation study to compare the different locally parametric methods proposed. As target densities we used a skewed distribution, #2, two symmetric bimodal distributions, #6 and #7, and two asymmetric bimodal distributions, #8 and $0{\cdot}25\mathcal{N}(-1, 0{\cdot}5^2) + 0{\cdot}75\mathcal{N}(1, 0{\cdot}5^2)$. The numbers refer to the normal mixtures introduced by Marron & Wand (1992).

The methods used were Case I and Case II of the technique proposed at (2·2) and (2·3), hereafter referred to as version I and version II, respectively. As global models for version II of our method we used a normal distribution in unimodal cases and a normal mixture, $0{\cdot}5\mathcal{N}(\mu - c, \sigma^2) + 0{\cdot}5\mathcal{N}(\mu + c, \sigma^2)$, in bimodal cases. For the latter we estimated μ, c and σ using method of moments estimators. The estimator was calculated by an iterative Newton–Raphson algorithm. Version I of our method was implemented using an explicit formula based on a normal distribution as global model. We also employed the local log-linear and local log-quadratic estimator using the explicit formulae given by Hjort & Jones (1996).

Since we were mostly interested in the performance of the methods for small to moderate sample sizes, we took $n = 30$, 50, 75 and 100 in our simulation study. The bandwidth parameter h was varied between 0·2 and 0·9. For each setting we generated 200 replications. For each replication, the density estimates were evaluated at 201 equispaced points in $[-3, 3]$. Using these evaluations we calculated integrated squared error using the trapezoidal rule, and mean integrated squared error by averaging all 200 integrated squared error values. We also recorded at each point the mean and variance of the 200 estimates, as well as the 10 and 90 percentiles.

To obtain an impression of the visual behaviour of the estimates we calculated several roughness measures:

$$R_1 = \int \hat{f}'(u)^2\, du, \quad R_2 = \int \hat{f}''(u)^2\, du, \quad R_3 = \int \frac{|\hat{f}''(u)|}{\{1 + \hat{f}'(u)^2\}^{3/2}}\, du.$$

The first two are well-known descriptions of roughness (Scott, 1992, p. 53). The third measures the average curvature of the estimate, with increased weight being given to places where $\hat{f}$ is flat and where, consequently, traditional kernel methods tend to suffer visibly from noisy estimates of curvature. Derivatives were calculated by first- and second-order central difference formulae, and integrated using the trapezoidal rule.

The simulation study showed that our methods generally have lower variance than do log-quadratic estimators. Thus, the global-model approach is achieving the purpose for which it was designed, i.e. reducing stochastic variability by replacing error-prone nonparametric estimates of high-order terms by less variable ones resulting from a simpler, parametric global model. If the global model is significantly in error then of course, at least in large samples, our approach can be more biased than log-quadratic estimators, but this bias occurs only in high-order terms.

Visual inspection of plots with either pointwise variances or percentiles showed that both versions of our estimator had approximately the same amount of variability. In terms of mean integrated squared error, however, version I is superior to version II. Table 1 gives the achieved minimal mean integrated squared error for all estimators in each setting. With one exception discussed below, the two versions of our method and local-linear estimator achieved minimal mean integrated squared error at similar bandwidths, while

Table 1. *Minimum mean integrated squared error, multiplied by* 10^3, *achieved by all estimators for each target density and sample size. The numbers refer to the normal mixtures introduced by Marron & Wand* (1992). *The mixture for the distribution 'bimodal' is given in the text.*

		Estimator			
Target density	n	Version I	Version II	Log-linear	Log-quadratic
Bimodal #6	30	17·393	21·037	17·052	20·359
	50	12·636	14·677	12·860	13·715
	75	9·415	10·618	9·683	9·777
	100	7·381	8·231	7·674	7·400
Bimodal #7	30	25·709	32·419	28·161	32·634
	50	18·446	22·031	20·477	21·019
	75	13·471	15·467	15·047	14·196
	100	10·509	11·973	11·853	10·461
Bimodal #8	30	22·129	24·515	21·214	24·776
	50	16·126	17·067	15·825	17·167
	75	12·410	13·092	12·232	12·610
	100	9·973	10·424	9·904	9·917
Bimodal	30	28·257	32·298	29·732	29·982
	50	18·450	20·956	19·756	18·857
	75	14·335	15·961	15·565	14·085
	100	11·138	12·230	12·307	10·626
Skewed #2	30	13·987	16·351	17·089	15·073
	50	9·958	10·525	12·349	9·067
	75	8·026	7·650	9·545	6·309
	100	6·874	6·093	7·795	4·922

the local-quadratic estimator achieved its minimal mean integrated squared error typically at a much larger bandwidth.

These mean integrated squared error figures show that, for most of the bimodal target densities and for $n = 50$ or 75, version I of our method performed best in terms of mean integrated squared error. It is followed by the log-linear estimator, the log-quadratic estimator and version II of our method. Plots of pointwise biases showed that the log-quadratic estimator typically has smallest pointwise bias but it needs larger sample sizes before this reduction in bias offsets the increase in variance with respect to mean integrated squared error.

For the skewed target density, version I of our method and the log-quadratic estimator both achieved minimal mean integrated squared error at the largest bandwidth used. The mean integrated squared error curve for version I of our method had, for each sample size, a local minimum close to the bandwidths at which the log-linear and version II of our method attained their minimum. The behaviour of these estimators for this target density is not surprising, as the skewed distribution is unimodal and therefore well approximated by a quadratic function on a logarithmic scale. Hence, the bias component of mean integrated squared error is practically negligible and a large bandwidth may be used to reduce the variance component.

The superiority of our methods in terms of variance and visual appearance is supported by the roughness criteria calculated in the simulation study. For all three criteria, version I of our method and the log-linear method achieved comparable values, although our

method was usually better, especially at bandwidths where minimal mean integrated squared error was obtained. The log-quadratic method was markedly rougher with respect to all three criteria. In terms of R_1, version II of our method behaved much like version I and the log-linear method. In terms of R_2 and R_3, version II displayed behaviour similar to the log-quadratic method. Closer analysis showed that this is because this estimator is calculated by an iterative procedure and not a closed formula. Consequently, version II is more susceptible to second-order differences. The behaviour of these estimators is depicted in Fig. 1.

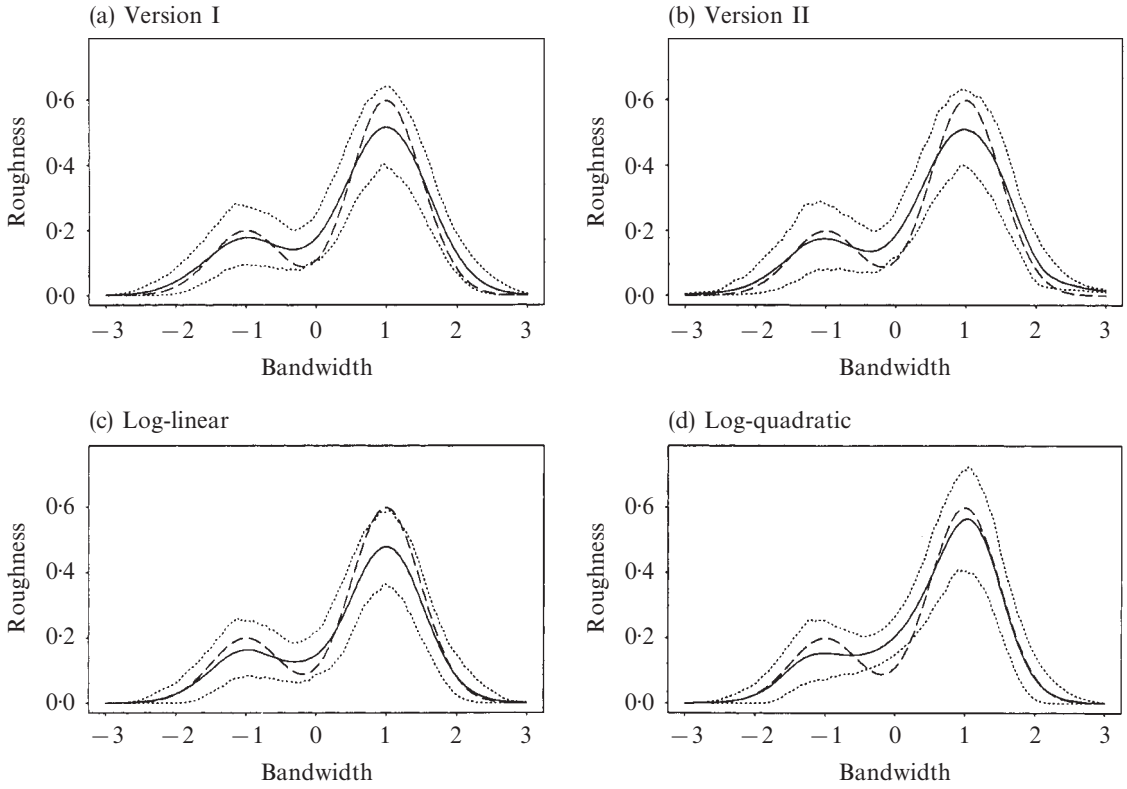

Fig. 1. Behaviour of the estimators for the bimodal distribution given in the text and sample size $n = 30$. Dashed line, the true density; solid line, the mean over 200 replications; dotted lines, the 10 and 90 percentiles of the replications.

In summary, our simulation study showed that our method is robust against misspecification of the global model. It is not necessary that the class from which the global model stems contain the actual distribution, as long as one gets the general 'shape' correct. This is especially true for version I of our method. Furthermore, the reduction in bias for this version is sufficient to make it preferable, in terms of mean integrated squared error, to log-linear methods. For small samples, at least, it is superior to log-quadratic methods on account of its lower variance.

5. THEORETICAL PROPERTIES

5·1. *Preliminary remarks*

Usually, the global parametric model f_0, in the case of density estimation, or m_0, for regression, would suffer stochastic fluctuations of order only $n^{-\frac{1}{2}}$, where n denotes sample

size. This is not to say that the accuracy of the parametric model would be $O(n^{-\frac{1}{2}})$; indeed, we do not need even the global model to be consistent. Since errors of this size are of smaller order than the errors introduced by local parametric modelling, then in developing first-order theory we may assume without loss of generality that the fitted global model f_0 is nonstochastic. That we do below.

Formulae for bias alter if we use a global model to enhance the local model. In the case of local linear regression the effects on bias are particularly transparent, so we address them here. If, in place of the local-linear model $\theta_1 + \theta_2(u-x)$, we fit the 'enhanced' local quadratic

$$\theta_1 + \theta_2(u-x) + \tfrac{1}{2}m_0''(x)(u-x)^2,$$

then in the calculations that lead to $\hat{m}$ we are simply replacing the response variable Y_i by $Y_i - \frac{1}{2}m_0''(x)(Y_i - x)^2$. Therefore the asymptotic bias, which was previously a constant multiple of $(d/du)^2 m(u)|_{u=x} = m''(x)$, is now the same constant multiple of

$$(d/du)^2 \{m(u) - \tfrac{1}{2}m_0(x)(u-x)^2\}|_{u=x} = m''(x) - m_0''(x).$$

That is, the effect of fitting a global model is simply to 'correct' in the obvious way, by subtraction, for the dominant term in the expression for bias. Section 5·4 will describe more general results for bias and variance in the local polynomial case, while §§ 5·2 and 5·3 will address bias and variance in density estimation.

5·2. *Representation for* $\hat{f}$

Here and in § 5·3 we treat the case of density estimation, where both quadratic and Kullback–Leibler loss are candidates for measuring performance. The relationship between empirical and deterministic loss functions translates into the following simple representation for $\hat{f}$, available under general conditions: if $\hat{\theta}$ and θ^0 are defined as the minimisers of $Q(\theta)$ and $\omega(\theta)$ respectively, or as the minimisers of $L(\theta)$ and $\lambda(\theta)$ respectively, then

$$\hat{f}(x) = \psi(x|\theta^0) + \{(nh)^{-1} f(x)\rho\}^{\frac{1}{2}} Z, \tag{5·1}$$

where Z is asymptotically normal $\mathcal{N}(0, 1)$ and ρ is a constant, given by (5·2) and identical for all choices of loss and all choices of the model ψ. Regularity conditions are discussed in the Appendix. Formula (5·1) describes the error-about-the-mean term, i.e. the stochastic term, in the difference $\hat{f} - f$. Formula (5·5) will give the bias term.

Since ρ has this invariance property, it may be written down directly from a formula given by Loader (1996) in the special case where loss is interpreted in the Kullback–Leibler sense and $\psi(.|\theta)$ is an exponential polynomial. When x is an interior point of $\mathcal{I}$, $\rho = \kappa \equiv \int K^2$ for all $r \geqslant 1$; and when x is a finite endpoint of $\mathcal{I}$ and $r = 1$, $\rho = 2\kappa$. More generally, suppose the support of K equals the interval $[-c, c]$, let $\mathcal{S}$ be the set of limit points of $\{z \in [-c, c] : x + hz \in \mathcal{I}\}$, let e_1 be the column vector of length r with 1 in the first position and 0's elsewhere, and let K_l be the $r \times r$ matrix with $k_{i+j-2,l}$ in position (i, j), where $k_{il} = k_{il}(\mathcal{S}) = \int_{\mathcal{S}} u^i K(u)^l \, du$. Then,

$$\rho = e_1^{\mathrm{T}} K_1^{-1} K_2 K_1^{-1} e_1, \tag{5·2}$$

implying in particular that

$$\rho = \begin{cases} k_{02}/k_{01}^2 & \text{if } r = 1, \\ \{k_{21}(k_{02}k_{21} - k_{12}k_{11}) - k_{11}(k_{12}k_{21} - k_{11}k_{22})\}/(k_{01}k_{21} - k_{11}^2)^2 & \text{if } r = 2. \end{cases}$$

5·3. *Approximate equivalence of quadratic and Kullback–Leibler loss*

A close connection between ω and λ may be seen directly from Taylor expansion: if f is continuous and $f(x) > 0$ then, as $g \to f$,

$$\lambda(f, g) = \frac{1}{2}\int w(f-g)^2 f^{-1} + \frac{1}{3}\int w(f-g)^3 f^{-2} + \ldots$$

$$\sim \frac{1}{2} f(x)^{-1}\omega(f, g).$$

Therefore, it comes as no surprise to learn that the asymptotic bias, $\psi(x\,|\,\theta^0) - f(x)$, often does not depend, to first order, on the type of loss function. Since variance components are first-order equivalent, see § 5·2, then optimising quadratic or Kullback–Leibler loss often produces first-order equivalent estimators. This contrasts markedly with properties of these loss functions in a global setting; see Hall (1987), for example.

To describe bias, we define $\psi^{(i)}(u\,|\,\theta)$ to equal the ith partial derivative of $\psi(u\,|\,\theta)$ with respect to u. Assume there exists a unique r-vector $\theta^1 = \theta^1(x)$ such that $\psi^{(i)}(x\,|\,\theta^1) = f^{(i)}(x)$ for $0 \leqslant i \leqslant r-1$. Then, for θ^0 defined under either quadratic or Kullback–Leibler loss,

$$\psi(x\,|\,\theta^0) - f(x) = h^r\{f^{(r)}(x) - \psi^{(r)}(x\,|\,\theta^1)\}(r!)^{-1}k_{r1}e_1^{\mathrm{T}}K_1^{-1}e_1 + O(h^{r+1}). \tag{5·3}$$

Note particularly that the difference between the true density f and the fitted model $\psi(.\,|\,\theta^1)$ appears only at the first order, here the rth order, at which nonparametric information is not being used. By way of comparison, when we employ a nonparametric estimator with a parametric start the difference between the true f and the global model f_0, at each level up to and including the $(r-1)$th, makes a contribution to the term of order h^r.

Higher-order expansions are series in the differences $f^{(i)}(x) - \psi^{(i)}(x\,|\,\theta^1)$. Comparing (2·1) and (5·3) we see that, if the initial global model is reasonably accurate in terms of approximating rth derivatives, then bias will be reduced relative to what it would be if we used ordinary kernel methods. For the latter, the term $f^{(r)}(x) - \psi^{(r)}(x\,|\,\theta^1)$ in (5·3) would be replaced by $f^{(r)}(x)$, and so would not admit correction by the global model.

If r is odd and x is an interior point of $\mathscr{I}$ then $k_{r1} = 0$, and so the first term on the right-hand side of (5·3) vanishes. Then the bias is of order h^{r+1}, and the term of that order depends on choice of loss function as well as on the model. Details in the case of Kullback–Leibler loss and log-polynomial models for f are given by Loader (1996). For the case of general loss functions and models with $r = 1, 2, 3$, see Hjort & Jones (1996).

5·4. *Quadratic loss in nonparametric regression*

Let $\hat{\theta}$ and θ^0 denote the minimisers of $Q(\theta)$ and $\omega(\theta)$, respectively, and assume the errors ε_i in the regression model have variance σ^2 in a neighbourhood of x. See § 3·3 for definitions of $Q(\theta)$, $\omega(\theta)$ and the regression model. Then, analogously to (5·1), $\hat{m}$ admits the representation

$$\hat{m}(x) = \psi(x\,|\,\theta^0) + \{(nh)^{-1}f(x)^{-1}\rho\}^{\frac{1}{2}}\sigma Z,$$

where again Z is asymptotically normal $\mathcal{N}(0, 1)$ and ρ is given by (5·2). Moreover, (5·3) continues to hold if f there is replaced by m.

Acknowledgement

The authors are grateful to two reviewers for helpful comments.

Appendix

Background to theoretical results

For (5·1) to hold we require f to have $r-1 \geqslant 0$ continuous derivatives in a neighbourhood of x, $f(x) > 0$, $h \to 0$ and $nh \to \infty$ as $n \to \infty$, and $\psi(.|\theta)$ to be able to represent uniquely all values of $(f, \ldots, f^{(r-1)})^{\mathrm{T}}$ that may potentially arise in a neighbourhood of $(f(x), \ldots, f^{(r-1)}(x))^{\mathrm{T}}$. Recall that θ is of length r; this is the 'derivative-matching condition'. It is concisely described by the assumption that, after a reparameterisation, $\psi(.|\theta)$ is expressible as

$$\psi(u|\theta) = \sum_{i=1}^{r} \frac{1}{(i-1)!}\theta_i(u-x)^{i-1} + o(|u-x|^{r-1}) \tag{A·1}$$

as $u \to x$, and that the inverse reparameterisation is continuous and one-to-one in a neighbourhood of $(f(x), \ldots, f^{(r-1)}(x))^{\mathrm{T}}$. We ask too that $\psi(.|\theta)$ be continuously differentiable as a function of θ, and that the remainder term in (A·1) be of the stated order after a differentiation with respect to θ.

Sufficient conditions for (5·5) are that f have $r+1$ bounded derivatives in a neighbourhood of x, that $f(x) > 0$, if loss is measured in Kullback–Leibler terms, that the derivative-matching condition hold, that $\psi(u|\theta)$ have $r+1$ bounded derivatives as a function of u, and that these derivatives have themselves two bounded, continuous derivatives as functions of θ. Thus, in the expansion at (A·1), representing $\psi(.|\theta)$ after reparameterisation, the remainder term is $O(|u-x|^r)$, and is also of this order after two differentiations with respect to θ.

Details of assumptions and technical arguments behind results in § 5·4 are given by Ruppert & Wand (1994).

References

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

Gasser, T., Köhler, W., Müller, H. G., Lago, R., Molinari, L. & Prader, A. (1985). Human height growth: correlation and multivariate structure of velocity and acceleration. *Ann. Human Biol.* **12**, 501–15.

Hall, P. (1987). On Kullback–Leibler loss and density estimation. *Ann. Statist.* **15**, 1491–519.

Härdle, W. & Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21**, 1926–47.

Hildenbrand, K. & Hildenbrand, W. (1986). On the mean income effect: a data analysis of the U.K. family expenditure survey. In *Contributions to Mathematical Economics*, Ed. W. Hildenbrand and A. Mas-Colell, pp. 247–68. Amsterdam: North-Holland.

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*, Ed. P. Lachout and J. Á. Víšek, pp. 102–5. Prague: Academy of Sciences of the Czech Republic.

Hjort, N. L. & Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.* **23**, 882–904.

Hjort, N. L. & Jones, M. C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24**, 1619–47.

Loader, C. R. (1996). Local likelihood density estimation. *Ann. Statist.* **24**, 1602–18.

Marron, J. S. & Wand, M. P. (1992). Exact mean integrated squared errors. *Ann. Statist.* **20**, 712–36.

RATKOWSKY, D. A. (1983). *Nonlinear Regression Modelling*. New York: Marcel Dekker.
RUPPERT, D. & WAND, M. P. (1994). Multivariate weighted least squares regression. *Ann. Statist*. **22**, 1346–70.
SCOTT, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley.

[*Received December* 1997. *Revised December* 1998]