ACADEMIC PRESS

Available at WWW.MATHEMATICSWEB.ORG POWERED BY SCIENCE DIRECT®

Journal of Multivariate Analysis 86 (2003) 375–397

Journal of Multivariate Analysis

http://www.elsevier.com/locate/jmva

Reducing variance in nonparametric surface estimation

Ming-Yen Cheng[a,b] and Peter Hall[a,*]

[a] *Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia*

[b] *Department of Mathematics, National Taiwan University, Taipei 106, Taiwan*

Received 12 February 2001

Abstract

We suggest a method for reducing variance in nonparametric surface estimation. The technique is applicable to a wide range of inferential problems, including both density estimation and regression, and to a wide variety of estimator types. It is based on estimating the contours of a surface by minimising deviations of elementary surface estimates along a quadratic curve. Once a contour estimate has been obtained, the final surface estimate is computed by averaging conventional surface estimates along a portion of the contour. Theoretical and numerical properties of the technique are discussed.

AMS 1991 subject classifications: primary 62G07; secondary 62H12

Keywords: Bandwidth; Boundary effect; Kernel method; Nonparametric density estimation; Nonparametric regression; Variance reduction

1. Introduction

We suggest a variance reduction method for nonparametric surface estimators, based on approximating the projection of a contour into the design plane at the point x where we wish to construct the estimate. The contour estimator is then used as an axis along which a continuum of conventional surface estimates is averaged in order to achieve a final estimate at x. Since our technique does not alter asymptotic

*Corresponding author.
E-mail address: halpstat@pretty.anu.edu.au (P. Hall).

0047-259X/03/$ - see front matter
doi:10.1016/S0047-259X(02)00032-5

bias then the reduction in variance that it offers leads directly to a reduction in asymptotic mean squared error.

This method has several novel features. Firstly, it exploits the extra degree of freedom that is available in the problem of surface estimation. Secondly, it provides a new technique for estimating gradients and curvatures of contour lines, without passing explicitly to derivatives of surface estimates. Thirdly, when applied to a surface estimate that is always positive, in either density estimation or regression, our method produces a boundary-corrected estimate that is always positive. Our approach to estimating contours involves choosing either a line segment or a quadratic along which a conventional surface estimator is least variable, in the neighbourhood of the point x at which we wish to estimate the surface.

The technique is applicable to nonparametric methods in both density estimation and regression. Indeed, it is not tied to a particular estimator type in either of these settings; for example, in nonparametric regression it can be used in conjunction with spline, local linear or Nadaraya–Watson methods. In the case of density estimation, when a conventional kernel estimator is used as its basis, the technique can be viewed as a device for re-computing kernel shape.

As implied two paragraphs above, the technique also has potential application for overcoming edge effects. Modified boundary kernel methods have been proposed for addressing this problem (see e.g. [14,19,20]), but like their univariate counterparts they can produce negative estimates at boundaries. Local polynomial and local parametric methods are more successful in this regard, although the increase in variance of such techniques near the boundary means that good asymptotic performance is often not visible unless sample size is particularly large. Scott ([18], pp. 82–85) gives a particularly illuminating discussion of issues such as these.

Multivariate generalisations are of course possible. However, since the multivariate analogue of a contour is not so familiar, not as readily depicted, and not as easy to compute as in the bivariate case, then high-dimensional generalisations do not offer as convenient a vehicle for illustrating the potential of the method. If the distribution is d-variate then the contour corresponding to "height" H is the set of points y such that $g(y) = H$, and is a region with $d - 1$ degrees of freedom.

Our variance reduction method is related to the so-called balloon kernel techniques for density estimation. See [9,18, p. 149ff]. There is an extensive literature on approaches for remedying boundary effects in density estimation and regression, mainly in univariate cases. It includes methods based on special "boundary kernels", for example those considered by Gasser and Müller [6], Gasser et al. [7], Granovsky and Müller [8] and Müller [13]. Rice [15] suggested a dual-bandwidth approach. So-called "reflection methods" include those of [1a,10,17]. The projection method of Djojosugito and Speckman [2] is in the same spirit. Eubank and Speckman [3] proposed a method that involves combining a conventional curve estimator with a substantially undersmoothed estimator. Cheng, Fan and Marron [1] suggested methods that have optimal asymptotic performance at boundaries. The natural boundary-respecting properties of local polynomial methods have been discussed by

Fan [4], Hastie and Loader [11], Ruppert and Wand [16] and Fan and Gijbels [5], for example. See also [12].

Section 2 will introduce our method and discuss, in an heuristic and nontechnical way, its variance-reduction properties. Theoretical results, underpinning the informal arguments in Section 2, will be given in Section 3, and rigorous technical details will be outlined in Section 5. Section 4 will summarise a simulation study that complements the theory.

2. Methodology

2.1. The method

Let g denote a univariate function of a 2-vector; for example, g might be the density of a bivariate distribution, or the mean in a regression problem where the explanatory variable is bivariate and the response variable is scalar. We wish to estimate g nonparametrically, making only smoothness assumptions and exploiting the extra degree of freedom that is available in the context of surface estimation, relative to the conventional case where the argument of g is univariate.

To this end we first construct an elementary nonparametric estimator $\hat{g}$ of g. For example, when g is a probability density we might take

$$\hat{g}(x) = (nh^2)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.1}$$

where K is a radially symmetric bivariate kernel, h is a bandwidth, and $X_1, \ldots, X_n$ are independent and identically distributed random variables with density g.

Next we describe construction of a local quadratic estimator of the level set, or contour, of g in the neighbourhood of x; local linear estimators will be treated in Section 2.2. Let $\mathscr{C}(x|\theta, c)$ denote the parabola passing through $x = (x^{(1)}, x^{(2)})$, with its vertex at x and its tangent there in the direction of the unit vector $(\cos\theta, \sin\theta)$, and with curvature $2c$ at x. Thus, as a curve in the $(z^{(1)}, z^{(2)})$-plane, $\mathscr{C}(x|\theta, c)$ has equation

$$\begin{aligned}&(z^{(2)} - x^{(2)})\cos\theta - (z^{(1)} - x^{(1)})\sin\theta \\ &\quad = c\{(z^{(2)} - x^{(2)})\sin\theta + (z^{(1)} - x^{(1)})\cos\theta\}^2.\end{aligned}$$

We shall constrain θ and c by $-\pi/2 < \theta \leqslant \pi/2$ and $-\infty < c < \infty$, which ensures that each nondegenerate parabola in the plane is representable by $\mathscr{C}(x|\theta, c)$ for a unique triple (x, θ, c).

Given $\lambda > 0$, let $\mathscr{C}(x|\theta, c, \lambda)$ denote the set of points $z \in \mathscr{C}(x|\theta, c)$ that satisfy $||z - x|| \leqslant \lambda h$, where $||\cdot||$ denotes standard Euclidean distance. Let $|\mathscr{C}|$ denote the

length of a finite segment $\mathscr{C}$ of a rectifiable curve, and put $\xi(c,\lambda) = |\mathscr{C}(x|\theta,c,\lambda)|$,

$$\breve{g}(x|\theta,c,\lambda) = \xi(c,\lambda)^{-1} \int_{\mathscr{C}(x|\theta,c,\lambda)} \hat{g}(z)\, ds,$$

$$S(x|\theta,c,\lambda) = \xi(c,\lambda)^{-1} \int_{\mathscr{C}(x|\theta,c,\lambda)} \{\hat{g}(z) - \breve{g}(x|\theta,c,\lambda)\}^2\, ds, \tag{2.2}$$

$$(\hat{\theta}_x, \hat{c}_x) = \arg\min_{(\theta,c)} S(x|\theta,c,\lambda), \tag{2.3}$$

where ds is an infinitesimal element of arc length along $\mathscr{C} = \mathscr{C}(x|\theta,c,\lambda)$ at the point on $\mathscr{C}$ with coordinates z. Panel (a) of Fig. 1 depicts an example of the contour estimator $\mathscr{C}(x|\hat{\theta}_x,\hat{c}_x,\lambda)$. Our final estimator of $g(x)$ is

$$\tilde{g}(x|\lambda) = \breve{g}(x|\hat{\theta}_x,\hat{c}_x,\lambda). \tag{2.4}$$

In practice, one would not necessarily use the same value of λ when computing $(\hat{\theta}_x,\hat{c}_x)$ and when calculating $\tilde{g}$. That is, the λ's at (2.2) and (2.4) would not necessarily be identical. We shall argue in Section 3 that a relatively large value of λ (asymptotically, $\lambda\to\infty$) should be used to give accurate estimation of the "true" quadratic approximation $\mathscr{C}(x|\theta_x,c_x)$ to the contour line at x. On the other hand, a relatively small value of λ may be adequate for reducing variance and removing edge effects in the estimator $\tilde{g}$.

To give an intuitive explanation of this point, note that estimation of θ and c is closely related to estimation of second derivatives of g, for which a larger bandwidth is needed than when simply estimating g itself. This explains why λh, which is effectively a bandwidth for computation of $\hat{\theta}_x$ and $\hat{c}_x$, should be relatively large.

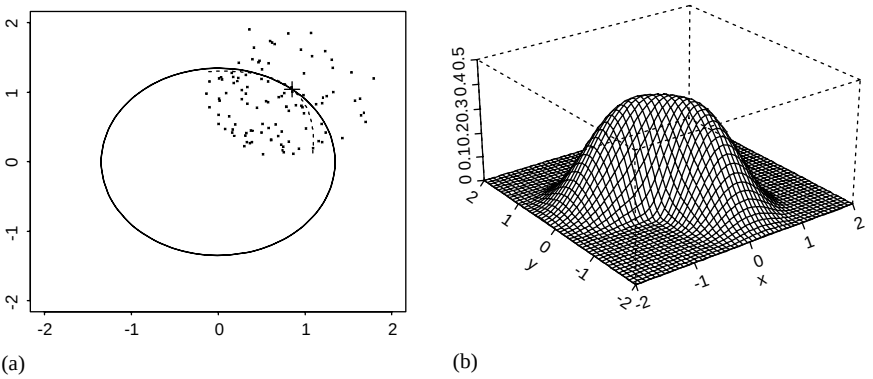

Fig. 1. *Sausage-shaped kernel.* In the context of density estimation, panel (a) depicts a portion of a point cloud, and the true contour line (solid line) that passes $x = (0.85, 1.04)$ (cross sign), when data are from the bivariate normal $N(0, I)$ distribution and $n = 500$. Dotted line is the contour line estimate $\mathscr{C}(x|\hat{\theta}_x,\hat{c}_x,\lambda)$, calculated at that point based on the spherical biweight kernel, $h = 0.8$, and $\lambda = 1.25$. Panel (b) shows a perspective plot of the corresponding "sausage-shaped" kernel K_x, defined at (2.5).

However, there is not the same pressing need for choosing λh large when estimating g itself.

2.2. Choice of contour estimator

To appreciate why minimising $S(x|\theta, c, \lambda)$ produces a parabola that approximates the contour $\mathscr{D}(x)$, note that we are in effect finding that choice of (θ, c) which renders $\hat{g}(z)$ least variable as we move z along the curve $\mathscr{C}(x|\theta, c)$. Indeed, if we were to replace $\hat{g}(z)$ by its true value, $g(z)$, when defining $\breve{g}(x|\theta, c, \lambda)$ and $S(x|\theta, c, \lambda)$, then the curve $\mathscr{C}(x|\hat{\theta}_x, \hat{c}_x)$ produced by minimising S would, if not constrained to have a quadratic equation, be exactly $\mathscr{D}(x)$. The curve $\mathscr{C}(x|\hat{\theta}_x, \hat{c}_x)$ represents an empirical, quadratic approximation to this contour.

An alternative technique is to take $\mathscr{C}$ to be a line segment, rather than a piece of a quadratic. The mechanics of implementing the approximation are virtually identical in this setting: we replace $\mathscr{C}(x|\theta, c, \lambda)$ by $\mathscr{C}_{\mathrm{lin}}(x|\theta, \lambda)$, denoting the line segment of length 2λ centred at x and inclined at angle θ; we replace $\xi(c, \lambda)$ at (2.2) by 2λ, and call the resulting integral $S(x|\theta, \lambda)$ instead of $S(x|\theta, c, \lambda)$; and we choose $\theta = \hat{\theta}_x$ to minimise $S(x|\theta, \lambda)$. This approach is adequate for the results described in Sections 3.1–3.3, but for the higher-order analogues described in Section 3.4 a local quadratic method, or something similar such as fitting local ellipses, is required.

A very different approach in estimating contour lines is to construct an appropriately oversmoothed estimator of the function g, and compute its contours. Oversmoothing is necessary in order to obtain sufficiently accurate estimates of derivatives of the surface; these are used explicitly or implicitly in constructing an estimate of the contour. We argue, however, that such a method is in general not as attractive as that proposed here, owing to the relative difficulty of drawing contours from differential-geometric properties of a surface.

Nevertheless, oversmoothing $\hat{g}$ is beneficial when it is necessary to construct $\tilde{g}$ at a place where the tangent plane to the surface is virtually horizontal. Minimising the function $S(x|\theta, c, \lambda)$ with respect to (θ, c) relies on detecting off-contour differences in g through variation of g; if the gradient of g is low then so too will be the variation. In such cases we rely on higher-order derivatives to provide "leverage" for detecting the contour—hence the need for more dramatic smoothing.

2.3. Removing edge effects

Let $\mathscr{R}$ denote the support of the distribution of the points X_i on which the estimator $\hat{g}$ is based. In the context of density estimation $\mathscr{R}$ would be the support of g, and in regression $\mathscr{R}$ would be the support of the density of X_i in the regression problem $Y_i = g(X_i) + \text{error}$. The basic estimator $\hat{g}(x)$ potentially suffers from edge effects whenever the support of the function $k_x(z) \equiv K\{(x - z)/h\}$ protrudes outside $\mathscr{R}$. However, assuming K is radially symmetric and vanishes outside a disc of unit radius, this problem is solved by the following trivial modification of the estimator suggested in Section 2.1: Re-define the parabola segment $\mathscr{C}(x|\theta, c, \lambda)$ to be the largest

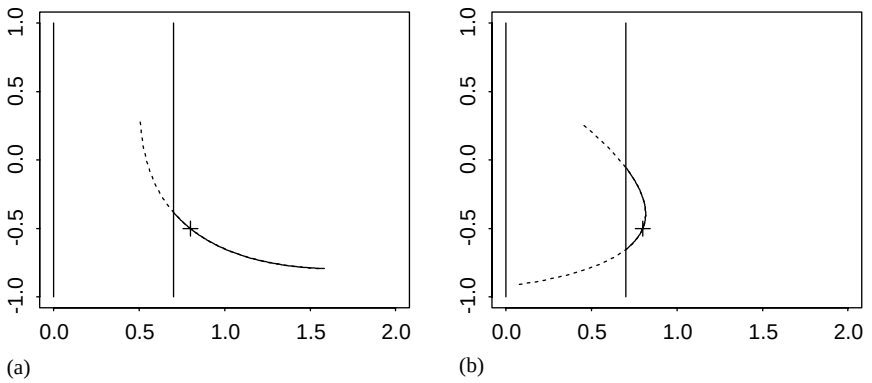

Fig. 2. *Removing edge effects*. In the presence of edge effects the subset of $\mathscr{C}(x|\theta,c)$ (dotted curve) that comprises $\mathscr{C}(x|\theta,c,\lambda)$ (solid curve) is reduced, to ensure that the resulting region $\mathscr{S}\{\theta,c,\mathscr{C}(x|\theta,c,\lambda)\}$, from which the estimator $\tilde{g}(\cdot|\lambda)$ is computed, lies wholly within the support $\mathscr{R}$ (right-hand side of the vertical line) of the design distribution. The point x is marked by a cross. Panels (a) and (b) illustrate cases where the contour is convex and concave, respectively, with respect to the boundary.

connected subset of $\mathscr{C}(x|\theta,c)$ inside the disc $\{z\colon ||z-x||\leqslant\lambda h\}$, subject to the set $\mathscr{S}\{\theta,c,\mathscr{C}(x|\theta,c,\lambda)\}$ being wholly contained within $\mathscr{R}$, where

$$\mathscr{S}(\theta,c,\mathscr{T})\equiv\{z_1\colon ||z_1-z_2||\leqslant h \text{ for some } z_2\in\mathscr{T}\}.$$

Fig. 2 illustrates the removal of edge effects in this context. Theoretical results, and their derivations, in the presence of edge effects are entirely analogous to their counterparts in the absence of those effects.

2.4. *Why the estimator $\tilde{g}$ has advantages*

The advantages stem from the property, established in Section 3, that $\tilde{g}$ is a good approximation to the average of $\hat{g}$ over a portion of the true contour of the surface represented by $y=g(x)$. Specifically, let $\mathscr{D}(x)$ denote the contour line that passes through x, and let $\mathscr{D}(x|\lambda)$ equal the largest connected subset of $\mathscr{D}(x)$ inside the disc $\{z\colon ||z-x||\leqslant\lambda h\}$, subject to $\mathscr{S}\{\theta,c,\mathscr{D}(x|\lambda)\}$ being wholly contained within $\mathscr{R}$. Write $\breve{g}_{\mathrm{cont}}(x|\lambda)$ for the integral average of $\hat{g}(z)$ over $z\in\mathscr{D}(x|\lambda)$. Then, as we shall show in Section 3.2, the difference between $\tilde{g}(x|\lambda)$ and $\breve{g}_{\mathrm{cont}}(x|\lambda)$ is of smaller order than the difference between the latter function and the true value of $g(x)$.

It is easy to see why $\breve{g}_{\mathrm{cont}}(x|\lambda)$ is likely to perform better than the conventional estimator $\hat{g}(x)$. Indeed, the averaging that is explicit in the definition of $\breve{g}_{\mathrm{cont}}$ will clearly tend to reduce variance, by an order of magnitude if λ is allowed to diverge with n. And the bias of $\breve{g}_{\mathrm{cont}}(x|\lambda)$ will equal the average value of the bias of $\hat{g}(z)$ over values of z for which $g(z)=g(x)$. Replacing bias by an average value is generally not deleterious, and in fact the asymptotic bias of $\breve{g}_{\mathrm{cont}}(x|\lambda)$ is identical to that of $\hat{g}(x)$.

2.5. Particular cases of $\tilde{g}$

In the case of density estimation the estimator $\tilde{g}$ may be thought of as having been computed using a kernel whose shape is symmetric about the parabolic axis represented by $\mathscr{C}(x|\hat{\theta}_x, \hat{c}_x)$. If $\hat{g}$ is given by (2.1) then this kernel is K_x, say, defined by

$$K_x(v) \equiv |\mathscr{C}(0|\hat{\theta}_x, \hat{c}_x h, \lambda/h)|^{-1} \int_{\mathscr{C}(0|\hat{\theta}_x, \hat{c}_x h, \lambda/h)} K(z+v)\, ds. \tag{2.5}$$

In this notation the estimator $\tilde{g}$ has the standard form at (2.1):

$$\tilde{g}(x|\lambda) = (nh^2)^{-1} \sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h}\right),$$

where the support of K_x is sausage-shaped with its axis represented by the quadratic $\mathscr{C}(0|\hat{\theta}_x, \hat{c}_x h)$.

Fig. 1 illustrates a typical local quadratic contour estimate, and the associated sausage-shaped kernel, in the case of nonparametric density estimation. There is an obvious analogue of the figure in the case of a local linear approximation to the contour.

In the context of kernel-based regression the estimator $\tilde{g}$ cannot be expressed simply as the result of replacing K in the definition of $\hat{g}(x)$ by K_x. An approach like this is still feasible, but it would generally involve at least two kernels like K_x, one ($K_{x,1}$ say) designed for estimating contours of fg, where f is the design density, and the other ($K_{x,2}$) designed to estimate contours of f. For example, in the case of local-constant or Nadaraya–Watson estimation of g one would use $K_{x,1}$ and $K_{x,2}$ in the numerator and denominator, respectively, of the estimator. The computational complexity of such an approach makes it unattractive, however.

3. Theoretical properties

3.1. Contour approximation

Our aim in this section is to describe the accuracy with which the empirical contour line $\mathscr{C}(x|\hat{\theta}_x, \hat{c}_x)$ estimates a nonrandom, quadratic approximation $\mathscr{C}(x|\theta_x, c_x)$ to $\mathscr{D}(x)$. For brevity we confine our detailed treatment to the case of nonparametric density estimation, noting in Section 3.6 the similarities to nonparametric regression. We deal initially only with situations where edge effects do not arise; Section 3.5 discusses how our results change in the presence of edge effects.

Let $\mathscr{S}$ denote a bounded, open set in the plane. We assume of the kernel that

K is a compactly supported, radially symmetric, probability density with Hölder-continuous first derivatives; (C_K)

of h and λ that

$$h \asymp n^{-1/6}, \quad \lambda^2 h/(\log n)^{5/4} \to \infty, \quad \text{and} \quad \lambda h = O(n^{-\varepsilon}) \text{ for some } \varepsilon > 0, \text{ as } n \to \infty; \tag{$C_{h,\lambda}$}$$

and of the density g that it is differentiable on $\mathscr{S}$ and satisfies

$$\text{the gradient of the steepest vector in the tangent plane at } x \text{ to the surface represented by } y = g(x) \text{ does not vanish for } x \in \mathscr{S} \tag{C_{1g}}$$

and

$$g \text{ has two Hölder-continuous derivatives, of all types, in } \mathscr{S}. \tag{C_{2g}}$$

In respect of $(C_{h,\lambda})$, note that $h \asymp n^{-1/6}$ is the optimal size of bandwidth for estimating a density g with two derivatives.

Conditions (C_{1g}) and (C_{2g}) imply that, for each $x \in \mathscr{S}$, the contour line $\mathscr{D}(x)$ that passes through x may be represented locally as a quadratic, in the sense that there exist a real number c_x, and $\theta_x \in (-\pi/2, \pi/2]$, both uniquely determined, such that the distance from any given point z on $\mathscr{D}(x)$ to the nearest point on $\mathscr{C}(x|\theta_x, c_x)$ converges to 0 at rate $o(r^2)$, uniformly in z satisfying $||z - x|| \leqslant r$, as $r \to 0$.

From a sample $X_1, \ldots, X_n$ of independent and identically distributed random variables drawn from the distribution with density g, compute first the density estimator $\hat{g}$ given at (2.1), and then $(\hat{\theta}_x, \hat{c}_x)$ defined at (2.3). Our first result describes rates of convergence of the estimators $\hat{\theta}_x$ and $\hat{c}_x$ to θ_x and c_x, respectively. Immediately below the theorem we discuss its analogue when contours are estimated using local linear methods.

Given $\varepsilon > 0$ let $\mathscr{S}_\varepsilon \subseteq \mathscr{S}$ equal the set of all points $x \in \mathscr{S}$ such that the closed disc of radius ε, centred at x, is contained in $\mathscr{S}$. Let $\langle \theta_1 - \theta_2 \rangle$ denote the distance between arbitrary real numbers θ_1 and θ_2, modulo π.

Theorem 3.1. *Assume conditions* (C_K), $(C_{h,\lambda})$, (C_{1g}) *and* (C_{2g}). *Constrain* c *to satisfy* $|c| \leqslant C/(\lambda h)$, *where* $C > 0$ *is fixed, when choosing* (θ, c) *to minimise* $S(x|\theta, c, \lambda)$, *defined at* (2.2). *Then for each* $\varepsilon > 0$, *and with probability* 1,

$$(\log n)^{1/2} \sup_{x \in \mathscr{S}_\varepsilon} (\langle \hat{\theta}_x - \theta_x \rangle + \lambda h \, |\hat{c}_x - c_x|) \to 0. \tag{3.1}$$

The theorem holds with only minor modifications if we use local linear, rather than local quadratic, approximations to contour lines. Indeed, consistent estimation of c_x is not required for our method to produce asymptotic improvements on the conventional estimator $\hat{g}$. If we take $S(x|\theta, \lambda)$ to be the "linear" analogue of $S(x|\theta, c, \lambda)$ defined in Section 2.2, and $\hat{\theta}_x$ to be its minimiser; and if we assume (C_K), $(C_{h,\lambda})$, (C_{1g}) and (C_{2g}); then (3.1) continues to hold in the sense that with probability 1,

$$(\log n)^{1/2} \sup_{x \in \mathscr{S}_\varepsilon} \langle \hat{\theta}_x - \theta_x \rangle \to 0. \tag{3.2}$$

In practical terms, the assumption "$\lambda^2 h/(\log n)^{5/4} \to \infty$" in $(C_{h,\lambda})$ asks that the square of the radius, λh, of the parabola fragment $\mathscr{C}(x|\theta, c, \lambda)$ be of larger order than the bandwidth, h.

3.2. Density estimation

In this section we show that any sufficiently accurate empirical, quadratic approximation $\mathscr{C}(x|\tilde{\theta}_x, \tilde{c}_x)$ to $\mathscr{C}(x|\theta_x, c_x)$ leads to an estimator $\breve{g}(x|\tilde{\theta}_x, \tilde{c}_x, \lambda)$ that is a uniformly good approximation to $\breve{g}_{\text{cont}}(x|\lambda)$.

Let $\tilde{\theta}_x, \tilde{c}_x$ denote general estimators of θ_x, c_x respectively. Write λ_0 for a new version of λ, which for the sake of simplicity we shall keep fixed. Our next result describes properties of the estimator $\breve{g}(x|\tilde{\theta}_x, \tilde{c}_x, \lambda_0)$. The version of (3.1) for $\tilde{\theta}_x$ and $\tilde{c}_x$, and fixed λ, is: with probability 1,

$$(\log n)^{1/2} \sup_{x \in \mathscr{S}_\varepsilon} (\langle \tilde{\theta}_x - \theta_x \rangle + h|\tilde{c}_x - c_x|) \to 0. \tag{3.3}$$

Recall the definition of

$$\breve{g}_{\text{cont}}(x|\lambda) = |\mathscr{D}(x|\lambda)|^{-1} \int_{\mathscr{D}(x|\lambda)} \hat{g}(z)\, ds.$$

Theorem 3.2. *Assume conditions* (C_K), (C_{2g}) *and* (3.3). *Suppose too that* $h \asymp n^{-1/6}$ *and* $\lambda_0 > 0$ *is fixed. Then with probability* 1,

$$\breve{g}(x|\tilde{\theta}_x, \tilde{c}_x, \lambda_0) = \breve{g}_{\text{cont}}(x|\lambda_0) + o_p(h^2) \tag{3.4}$$

uniformly in $x \in \mathscr{S}_\varepsilon$, *for each* $\varepsilon > 0$.

The estimators $\hat{\theta}_x, \hat{c}_x$ described in Theorem 3.1 are examples of $\tilde{\theta}_x, \tilde{c}_x$, and then (3.1) immediately implies (3.3). However, taking $\hat{\theta}_x$ to be a local linear estimator is also adequate; there we should take $\hat{c}_x = 0$, and (3.3) follows from (3.2).

We should stress that in Theorem 3.2 the value λ_0 of λ is taken fixed, while in Theorem 3.1 it diverges slowly with n. The latter requirement is symptomatic of the degree of oversmoothing that is necessary when estimating quantities that are linked to density derivatives, such as the tangent angle θ_x or the curvature c_x, rather than the density itself.

3.3. Performance advantages

To appreciate the variance reduction properties of the estimator $\breve{g}_{\text{cont}}$ (and hence of $\tilde{g}$), relative to its standard kernel counterpart $\hat{g}$, let $L(v)$ denote the integral average of $K(v+z)$ over $z \in \mathscr{L}$ where $\mathscr{L}$ is any line segment of length $2\lambda_0$, and put $\kappa_M = \int M^2$ for $M = K$ or L. We shall show shortly that the variances of $\hat{g}(x)$ and

$\check{g}_{\rm cont}(x|\lambda_0)$ are asymptotic to $(nh^2)^{-1}g(x)\,\kappa_M$ as $n\to\infty$, where $M=K$ and L in the respective cases. Moreover, $\kappa_L<\kappa_K$, and so our method reduces variance; and also, $\kappa_L/\kappa_K\sim C\lambda_0^{-1}$, for a constant $C>0$, as $\lambda_0\to\infty$. The latter result shows that as the fixed value of λ_0 becomes larger, the extent of variance reduction increases in proportion to λ_0. (Note that κ_L does not depend on the particular choice of $\mathscr{L}$.)

The asymptotic bias of $\check{g}_{\rm cont}(x|\lambda_0)$ is readily seen to be identical to that of $\hat{g}(x)$, and in fact the expected value of either estimator equals $g(x)+\frac{1}{2}\kappa_2\nabla^2 g(x)+o(h^2)$, where $\kappa_2=\int(v^{(1)})^2\,K(v)\,dv$, $v^{(1)}$ denotes the first component of the vector v, and $\nabla^2 g$ equals the Laplacian. Combining this result with that in the previous paragraph, and with (3.4), we see that the empirical contour estimator $\check{g}(x|\tilde{\theta}_x,\tilde{c}_x,\lambda_0)$ has the same asymptotic bias as the conventional kernel estimator $\hat{g}(x)$, but has reduced asymptotic variance.

Therefore the estimator $\check{g}(x|\tilde{\theta}_x,\tilde{c}_x,\lambda_0)$ has less minimum asymptotic mean squared error (AMSE) than $\hat{g}(x)$. In particular, if $h=Hn^{-1/6}$ then the AMSE equals $n^{-2/3}\,A_L(H)$, where

$$A_L(H)=\tfrac{1}{4}H^4\kappa_2^2\{\nabla^2 g(x)\}^2+H^{-2}\,g(x)\kappa_L.$$

Through the fact that $\kappa_L<\kappa_K$ this is always (unless $g(x)=0$) strictly less than the AMSE of $\hat{g}(x)$; in the obvious notation the AMSE of $\hat{g}(x)$ equals $n^{-2/3}\,A_K(H)$. Likewise the asymptotic mean integrated squared error of $\check{g}(x|\tilde{\theta}_x,\tilde{c}_x,\lambda_0)$, computed for example over $x\in\mathscr{S}_\varepsilon$, is less than that for $\hat{g}(x)$.

The estimator $\check{g}(x|\tilde{\theta}_x,\tilde{c}_x,\lambda_0)$ is also asymptotically normally distributed, in the sense that

$$\begin{aligned}\check{g}(x|\tilde{\theta}_x,\tilde{c}_x,\lambda_0)&=g(x)+\tfrac{1}{2}h^4\kappa_2\nabla^2 g(x)+(nh)^{-1/2}\{g(x)\kappa_L\}^{1/2}N_n\\&\quad+o_p(n^{-1/3}),\end{aligned}\tag{3.5}$$

where N_n is asymptotically distributed as normal $N(0,1)$.

To rigorously establish the claims made above, note that $\check{g}_{\rm cont}(x|\lambda_0)$ may be written in a form similar to that at (2.1):

$$\check{g}_{\rm cont}(x|\lambda_0)=(nh^2)^{-1}\sum_{i=1}^n K_{{\rm cont},x}\left(\frac{x-X_i}{h}\right),\tag{3.6}$$

where

$$K_{{\rm cont},x}(v)\equiv|\mathscr{D}_0|^{-1}\int_{\mathscr{D}_0}K(v+z)\,ds,\tag{3.7}$$

with $\mathscr{D}_0$ denoting the image of the contour line segment $\mathscr{D}(x)$ after rescaling by h^{-1} in each coordinate and translating x to the origin. As $h\to0$ the kernel $K_{{\rm cont},x}$ converges to L, if the line segment $\mathscr{L}$ is chosen to have its centre at the origin and to be parallel to the contour tangent at x. (This does not affect the value of κ_L, however.) Therefore, the claim that the variances of $\hat{g}(x)$ and $\check{g}_{\rm cont}(x|\lambda_0)$ are

asymptotic to $(nh^2)^{-1}g(x)\kappa_M$ follows from standard arguments for the variance of a kernel density estimator; see for example [22, pp. 19–23]. The result $\kappa_L < \kappa_K$ follows from the Cauchy–Schwarz inequality; equality cannot arise unless $\lambda_0 = 0$. It may be shown too that

$$\lambda_0\kappa_L/\kappa_K \to C \equiv \int_{-\infty}^{\infty} ds \int K(v)K\{v + (s,0)^{\mathrm{T}}\}\, dv,$$

whence it follows that $\kappa_L \sim C\kappa_K/\lambda_0$. Result (3.5) follows from (3.4), (3.6) and the bias properties of $\breve{g}_{\mathrm{cont}}(x|\lambda_0)$ noted in the paragraph containing (3.5). Asymptotic normality of the variable N_n in (3.5) is an immediate consequence of the fact that $\breve{g}_{\mathrm{cont}}(x|\lambda_0)$ is a sum of n independent and identically distributed random variables.

3.4. High-order generalisations, and optimality

In Section 3.2 we simplified our theory by considering only the case where h is of the size appropriate for optimal construction of $\hat{g}$, and λ_0 is fixed. In the present section we discuss improvements in the overall convergence rate that are available using other choices of bandwidth, and taking λ_0 to diverge with n. Our first result is a version of Theorems 3.1 and 3.2 in this setting.

Theorem 3.3. *Assume* (C_K), $h = c_1 n^{-2/11}$ *and* $\lambda_0 = c_2 n^{1/11}$ *where* $c_1, c_2 > 0$ *are fixed, and that*

$$\begin{aligned} &g \text{ has two Hölder continuous derivatives,} \\ &\text{where the Hölder coefficient exceeds } \tfrac{2}{3}. \end{aligned} \tag{3.8}$$

Then estimators $\tilde{\theta}_x$ *and* $\tilde{c}_x$ *of* θ_x *and* c_x, *respectively, can be constructed such that with probability* 1,

$$h^{-1}(\log n)^{1/2} \sup_{x \in \mathscr{S}_\varepsilon} (\langle \tilde{\theta}_x - \theta_x \rangle + \lambda_0 h |\tilde{c}_x - c_x|) \to 0. \tag{3.9}$$

Furthermore, if $(\tilde{\theta}_x, \tilde{c}_x)$ *satisfies* (3.9), *then* (3.4) *continues to hold with probability* 1, *uniformly in* $x \in \mathscr{S}_\varepsilon$ *for each* $\varepsilon > 0$.

If the Hölder coefficient mentioned in (3.8) equals $1 - \xi_1 \in (0, \frac{1}{3})$; if, when constructing $\hat{g}$ for use with the local quadratic contour estimation method outlined in Section 2.1, we take $h = n^{-(3-\xi_2)/22}$ where $0 < \xi_2 < 9\xi_1/(2 + 3\xi_1)$; and if we take $(\tilde{\theta}_x, \tilde{c}_x)$ to be $(\hat{\theta}_x, \hat{c}_x)$, defined in Section 2.1; then (3.9) holds. (An outline proof will be given in Section 5.3.) Thus, as noted in the last paragraph of Section 3.2, it is necessary to use a larger order of bandwidth when estimating quantities such as θ_x and c_x, which depend on derivatives of g, than when estimating g itself.

The estimator $\breve{g}_{\mathrm{cont}}$ in (3.4) again admits representation (3.6), with kernel $K_{\mathrm{cont},x}$ given by (3.7). It may be shown from those formulae that $\breve{g}_{\mathrm{cont}}(x|\lambda_0)$ has standard

deviation $O\{(nh^2\lambda_0)^{-1/2}\}$ and bias $O(h^2)$. Both are of order $n^{-4/11}$, and so $\check{g}(x|\tilde{\theta}_x, \tilde{c}_x, \lambda_0) = g(x) + O_p(n^{-4/11})$. This represents an improvement by an order of magnitude on the rate of convergence, $O_p(n^{-1/3})$, of the estimator discussed in Theorem 3.2. Faster rates of convergence, up to $O_p(n^{-(1/2)+\xi})$ for any given $\xi > 0$, can be obtained for sufficiently smooth densities by using more accurate contour estimators.

As is well known (see e.g. [21]), the optimal rate of convergence of estimators of bivariate densities with two bounded derivatives equals $O(n^{-1/3})$. The results discussed in Section 3.3 might seem to contradict this result, since they signal the possibility of achieving a convergence rate of $o(n^{-1/3})$ by choosing $n^{1/6}h$ to decrease to zero sufficiently slowly, and λ_0 to diverge to infinity sufficiently slowly, as $n \to \infty$. However, there is in fact no violation, since we need a little more than just two bounded derivatives, specifically the Hölder continuity assumption in condition (C_{2g}), in order to achieve the rate. Likewise, the assumption in (3.8) that the Hölder coefficient exceeds (rather than equals) $\frac{2}{3}$ is slightly more than necessary for optimal performance under minimal conditions. In each case, however, a biproduct of the additional assumption is the uniform strong approximation of the empirical contour-based estimator $\check{g}(x|\tilde{\theta}_x, \tilde{c}_x, \lambda_0)$ by its generalised kernel form $\check{g}_{\rm cont}(x|\lambda_0)$, as evidenced by (3.4).

3.5. The case of edge effects

In Section 2.3 we showed that, in the context of density estimation, there is a natural version of $\tilde{g}$ that addresses edge effects. As a prelude to stating the results of Section 3.2 for estimators of this type, redefine the contour segment $\mathscr{D}_0$ at (3.7) by taking it to be that subset of the original $\mathscr{D}_0$ which is as large as possible subject to the support of $\int_{\mathscr{D}_0} K\{(\cdot + z)/h\}\, ds$ not protruding outside the support set $\mathscr{R}$ (introduced in Section 2.3). With this modification, continue to define $K_{{\rm cont},x}$ by (3.7) and $\check{g}_{\rm cont}$ by (3.6).

Take $\mathscr{R}$ to be a compact set whose boundary has two Hölder-continuous derivatives and is such that at no point on the boundary is the tangent to the boundary equal to the corresponding contour line; and assume the conditions of Theorems 3.1 and 3.2 on $\mathscr{R}$ rather than $\mathscr{S}$. Then (3.1) and (3.4) hold uniformly in $x \in \mathscr{R}$ (rather than $x \in \mathscr{S}_\varepsilon$). Moreover, an argument identical to that in Section 3.2 shows that the asymptotic variance of $\check{g}_{\rm cont}(x|\lambda_0)$, and hence of $\tilde{g}(x|\lambda_0)$, decreases to 0 at rate λ_0^{-1} as $\lambda_0 \to \infty$.

3.6. Nonparametric regression

A model for nonparametric regression is that where data pairs (X_i, Y_i) are generated by the formula $Y_i = g(X_i) + \varepsilon_i$, the errors ε_i having zero mean. Theory in this case is similar to that for density estimation, although regularity conditions are required on the design variables X_i and the error distribution. For the latter it is

sufficient to suppose that the ε_is are independent and identically distributed with all moments finite and zero mean. In this case, terms in $\log n$ in (3.1) and (3.3) should be replaced by terms in n^δ, for $\delta > 0$ fixed but arbitrarily small. In this vein, the assumption "$\lambda^2 h/(\log n)^{5/4} \to \infty$" in $(C_{h,\lambda})$ should be replaced by "$\lambda^2 h/n^\delta \to \infty$ for some $\delta > 0$". Let $(C'_{h,\lambda})$ denote the corresponding version of $(C_{h,\lambda})$.

Of the design variables it is adequate to suppose that they are independent and identically distributed with density f, which is bounded away from 0 on $\mathscr{S}$ and has two Hölder-continuous derivatives there. With this assumption, (C_K), $(C'_{h,\lambda})$, (C_{1g}) and (C_{2g}); using λ when estimating $(\hat\theta_x, \hat c_x)$, and employing a fixed λ_0 when constructing $\tilde g$; and taking the basic estimator $\hat g$ to be of either Nadaraya–Watson or local-linear type; results described in Sections 3.2 and 3.3 hold in the case of nonparametric regression.

4. Numerical examples

Three estimation methods, local quadratic approximation to contour lines (giving $\tilde g_B(x|\lambda)$), local linear approximation (giving $\tilde g_L(x|\lambda)$), and the standard kernel estimator $\hat g(x)$, were used to estimate the probability density functions of two distributions. We generated 200 random samples of size $n = 500$ from each. For each sample, integrated squared error (ISE) values for the three estimators were approximated by numerical integration. Values of MISE were approximated by averaging 200 of the respective ISE values. The spherical biweight kernel, $K(z) = \frac{15}{8\pi}(1 - ||z||^2)_+^2$, was employed throughout.

Our first example is the unimodal bivariate normal $N(0, I)$ distribution. We took the bandwidth to equal 0.8. To construct $\tilde g_B(x|\lambda)$ and $\tilde g_L(x|\lambda)$, λ in (2.2) was taken as $\min(0.1 + \frac{2}{3}d(x), 1.1)$, where $d(x)$ was the distance from x to the location of the mode of $\hat g$. Three-quarters of this value was used for λ in (2.4). See the second-last paragraph of Section 2.1 and the last paragraph of Section 3.2. Notice that "radii" of contour lines of the density surface degenerate near the mode, and that linearly increasing the value of λ ensures appropriate approximation of the contour lines.

Among the 200 random samples, the three samples that give median ISE values for the three estimators are plotted in Fig. 3, which also shows the corresponding values of $\hat g(x)$, $\tilde g_L(x|\lambda)$ and $\tilde g_B(x|\lambda)$. In multivariate cases, often a density surface estimate fluctuates significantly due to data sparseness difficulties. The averaging step of our contour approximation methods remedies this problem. This effect is clearly demonstrated by the middle and bottom rows of Fig. 3. There, for each of the three samples, the surfaces corresponding to $\tilde g_L(x|\lambda)$ and $\tilde g_B(x|\lambda)$ have less wiggly contour lines than $\hat g$ at places away from the mode.

Table 1 gives ISE values for the nine estimates. Table 2 provides average ISE values, these being approximation to MISEs, for the three estimators. These results demonstrate clear gains of $\tilde g_L(x|\lambda)$ over $\hat g$.

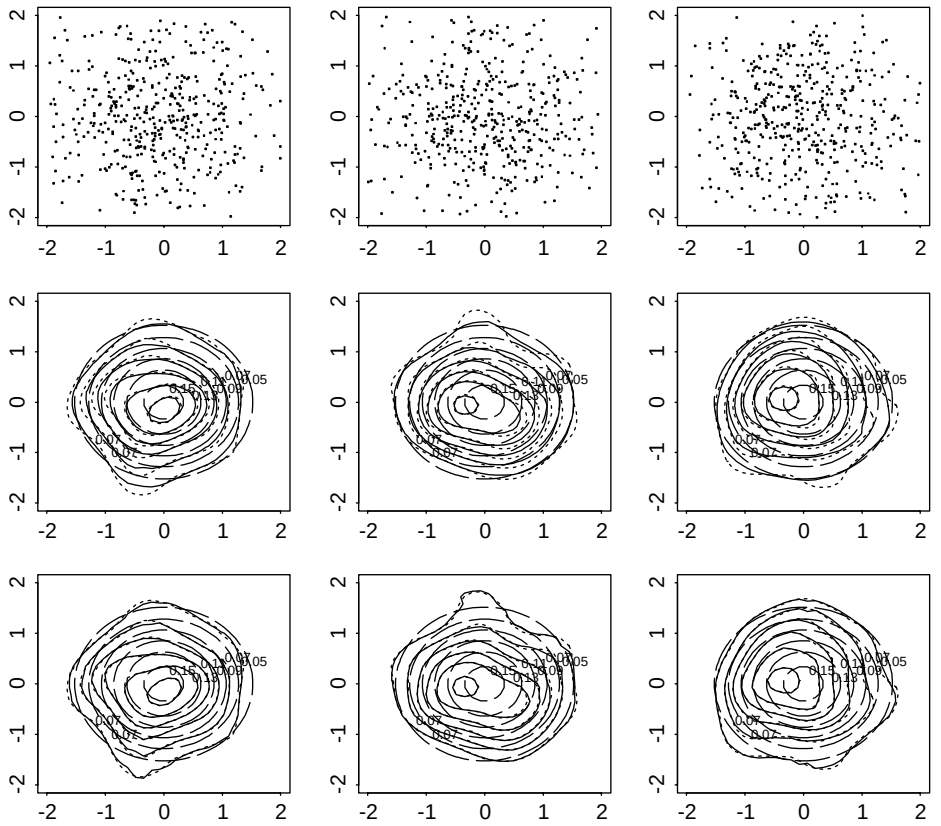

Fig. 3. *Unimodal density estimates*. The top row depicts three samples of size 500 drawn from the bivariate normal $N(0, I)$ distribution. Middle and bottom rows show contour lines of the true density surface (dashed lines), $\hat{g}$ (dotted lines) and contour approximation estimators, for the three respective samples. The middle row compares the local linear contour approximation method (solid lines) with $\hat{g}$. The bottom row compares the local quadratic contour approximation method (solid lines) with $\hat{g}$.

Table 1
ISE values of the density estimates shown in Figs. 3 and 4

	Unimodal normal		
$\hat{g}(x)$	0.001550	0.001660	0.001602
$\tilde{g}_L(x\|\lambda)$	0.001398	0.001332	0.001450
$\tilde{g}_B(x\|\lambda)$	0.001437	0.001635	0.001520
	Bimodal normal mixture		
$\hat{g}(x)$	0.003616	0.003717	0.003610
$\tilde{g}_L(x\|\lambda)$	0.003311	0.003261	0.003310
$\tilde{g}_B(x\|\lambda)$	0.003548	0.003661	0.003553

Table 2
Average ISE values of $\hat{g}(x)$, $\tilde{g}_L(x|\lambda)$, and $\tilde{g}_B(x|\lambda)$ when applied to 200 random samples of size 500, drawn from the unimodal $N(0, I)$ or the bimodal normal mixture distribution at (4.1)

	Unimodal normal	Bimodal normal mixture
$\hat{g}(x)$	0.001650	0.003724
$\tilde{g}_L(x|\lambda)$	0.001413	0.003377
$\tilde{g}_B(x|\lambda)$	0.001619	0.003678

Our next example illustrates performance of our estimators in a more complex, bimodal setting, a mixture of two bivariate normal distributions:

$$0.7N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right) + 0.3N\left(\begin{pmatrix}1.5\\-1\end{pmatrix}, \begin{pmatrix}0.26 & 0.1\\0.1 & 0.26\end{pmatrix}\right). \tag{4.1}$$

Bandwidth was $h = 0.6$. To construct $\tilde{g}(x|\lambda)$ we took λ in (2.2) to be $\min(0.1 + \frac{2}{3}d(x), 1.1)$, and three-quarters of its value to be λ in (2.4), where $d(x)$ was the distance from x to the location of the mode of $\hat{g}$ nearest to x. This prevents our using too-large values of λ at places between the modes, where the contour lines are curved, and hence helps preserve the bimodal feature of the density surface estimate. (In practice, there may not be prior information about the number of modes of the true distribution. In this case one can make a judgment from plots of preliminary estimates, such as $\hat{g}$.) For this example our approach again reduces fluctuations in the density surface estimates caused by stochastic variability, particularly in regions away from either of the modes; see the panels in the middle and bottom rows of Fig. 4. The ISE and average ISE values are given in Tables 1 and 2.

In summary, our simulation results demonstrate advantages of the contour approximation methods: the density surface estimates are more regularly shaped and the MISE values are reduced, compared to the usual kernel density estimate. Notably, the local linear contour approximation estimator enjoys good numerical performance. The local quadratic approximation method performs less well; it involves fitting two, rather than one, parameter, and thus will outperform $\hat{g}$ in MISE terms only when sample size is relatively large.

5. Proofs

5.1. Proof of Theorem 3.1

Put $\gamma = E(\hat{g})$, $\bar{\gamma} = E(\check{g})$, $\Delta = \hat{g} - \gamma$ and $\bar{\Delta} = \check{g} - \bar{\gamma}$. Define $A_1(x|\theta, c, \lambda)$, $A_2(x|\theta, c, \lambda)$ and $A_3(x|\theta, c, \lambda)$ to equal the integrals of $\{\gamma(z) - \bar{\gamma}(x|\theta, c)\}^2$, $\{\Delta(z) - \bar{\Delta}(x|\theta, c)\}^2$ and $\{\gamma(z) - \bar{\gamma}(x|\theta, c)\}\{\Delta(z) - \bar{\Delta}(x|\theta, c)\}$, respectively, over $z \in \mathscr{C}(x|\theta, c, \lambda)$. Then,

$$A_2(x|\theta, c, \lambda) = \int_{\mathscr{C}(x|\theta,c,\lambda)} \Delta(z)^2 \, ds - \xi(c, \lambda)\bar{\Delta}(x|\theta, c)^2, \tag{5.1}$$

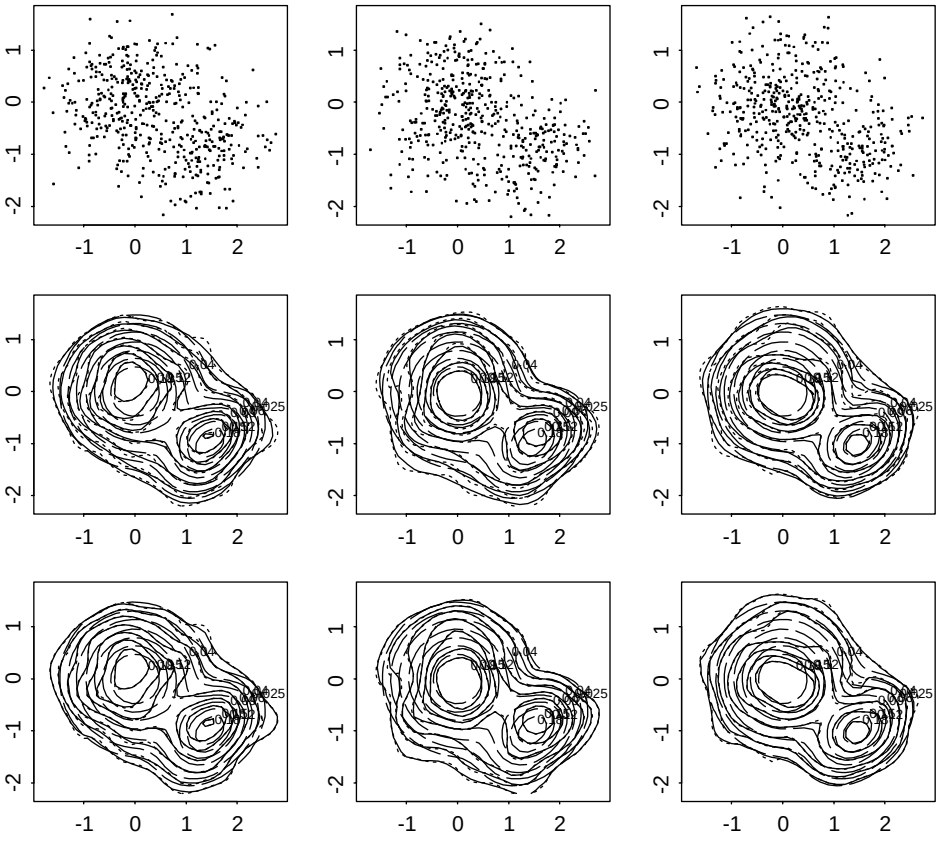

Fig. 4. *Bimodal density estimates*. Same as Fig. 3, except that the samples are from the bivariate normal mixture distribution at (4.1).

$$A_3(x|\theta,c,\lambda) = \int_{\mathscr{C}(x|\theta,c,\lambda)} \gamma(z)\Delta(z)\,ds - \xi(c,\lambda)\bar{\gamma}(x|\theta,c)\bar{\Delta}(x|\theta,c), \tag{5.2}$$

$$\xi S = A_1 + A_2 + 2A_3. \tag{5.3}$$

Without loss of generality, $\lambda \geqslant 1$ and $\lambda h \leqslant 1$. Let ψ denote a differentiable function defined in the plane, write $|D\psi|(z)$ for the supremum of the absolute value of the directional derivative of ψ (at z) over all directions, let $C>0$, and put $\eta = \langle \theta - \theta_x \rangle$ and $\zeta = |c - c_x|$. There exists $C_1 > 0$ with the property that

$$\left| \left(\int_{\mathscr{C}(x|\theta,c,\lambda)} - \int_{\mathscr{C}(x|\theta_x,c_x,\lambda)} \right) \psi(z)\,ds \right|$$
$$\leqslant C_1(\eta + \lambda h\zeta)(\lambda h)^2 \sup_{z:\ ||z-x|| \leqslant \lambda h} \{|D\psi|(z) + |\psi(z)|\} \tag{5.4}$$

uniformly in (θ, c) such that $|\theta|, |\theta_x| \leqslant \pi$, $|c|, |c_x| \leqslant C/(\lambda h)$ and $\eta, \lambda h\zeta \leqslant C$. (Below we shall refer to this uniform sense as "uniform*". At (5.4) and below the constants $C_1, \ldots, C_4$ depend only on C.) To derive (5.4), note that the distance between a given point on $\mathscr{C}(x|\theta, c)$ and its counterpart on $\mathscr{C}(x|\theta_x, c_x, \lambda)$, to which the former may be rotated about x, is dominated by a constant multiple of $(\eta + \lambda h\zeta)\,\lambda h$. Therefore, the difference of function values at the two points is dominated by a constant multiple of $(\eta + \lambda h\zeta)\,\lambda h$ times $\sup|D\psi|(z)$. To obtain the bound at (5.4) this should be multiplied by a constant times the lengths of the curves, i.e. by a constant times λh. There is an additional contribution to the right-hand side, coming from the difference between 1 and the Jacobian of the transformation, based on a rotation, which takes $\mathscr{C}(x|\theta, c)$ to $\mathscr{C}(x|\theta_x, c_x, \lambda)$; but it too is dominated by a constant multiple of the right-hand side of (5.4).

The quantity $\xi(c, \lambda)$, being the length of $\mathscr{C}(x|\theta, c, \lambda)$, is asymptotic to $2\lambda h$ uniformly in $|c| \leqslant C/(\lambda h)$; and $|\xi(c, \lambda) - \xi(c_x, \lambda)| \leqslant C_2(\lambda h)^3\zeta$ uniformly in $|c| \leqslant C$ such that $\zeta \leqslant C/(\lambda h)$. Therefore,

$$|\xi(c, \lambda)^{-1} - \xi(c_x, \lambda)^{-1}| \leqslant C_3\lambda h\zeta, \tag{5.5}$$

in the same uniform sense. Combining (5.1) with the results in this paragraph, and defining $B_j(x|\theta, c, \lambda) = \xi(c, \lambda)^{-1}A_j(x|\theta, c, \lambda)$, we conclude that in the uniform* sense,

$$\begin{aligned} &|B_2(x|\theta, c, \lambda) - B_2(x|\theta_x, c_x, \lambda)| \\ &\quad \leqslant C_4(\eta + \lambda h\zeta)\lambda h \sup_{z:\, \|z-x\| \leqslant \lambda h} \{|\Delta(z)|\,|D\Delta|(z) + \Delta(z)^2\}. \end{aligned} \tag{5.6}$$

Given $j = 0, 1$, let $A_{4j}(x|\theta, c, \lambda)$ denote the integral of $\gamma(z)^j\Delta(z)$ over $z \in \mathscr{C}(x|\theta, c, \lambda)$. An argument similar to that leading to (5.6) implies that in the uniform* sense,

$$|\bar{\gamma}(\theta, c) - \bar{\gamma}(\theta_x, c_x)| \leqslant C_5(\eta + \lambda h\zeta)\lambda h, \tag{5.7}$$

where the constants C_5, C_6, C_7 here and below depend only on C, g and K. From (5.2), (5.7) and the properties of $\xi(c, \lambda)$ discussed in the previous paragraph, we deduce that in the uniform* sense,

$$\begin{aligned} &|B_3(x|\theta, c, \lambda) - B_3(x|\theta_x, c_x, \lambda)| \\ &\quad \leqslant C_6\Big[(\lambda h)^{-1} \max_{j=1,2}\{|A_{4j}(x|\theta, c, \lambda) - A_{4j}(x|\theta_x, c_x, \lambda)|\} \\ &\qquad + (\eta + \lambda h\zeta)\lambda h \max\{|A_{40}(x|\theta, c, \lambda)|, |A_{40}(x|\theta_x, c_x, \lambda)|\}\Big]. \end{aligned} \tag{5.8}$$

Combining (5.3), (5.6) and (5.8) we conclude that in the uniform* sense,

$$
\begin{aligned}
&|S(x|\theta,c,\lambda) - S(x|\theta_x,c_x,\lambda) - \{B_1(x|\theta,c,\lambda) - B_1(x|\theta_x,c_x,\lambda)\}| \\
&\quad \leqslant C_7 \Bigg((\lambda h)^{-1} \max_{j=1,2} \{|A_{4j}(x|\theta,c,\lambda) - A_{4j}(x|\theta_x,c_x,\lambda)|\} \\
&\qquad + (\eta + \lambda h\zeta)\lambda h \Bigg[\sup_{z:\ \|z-x\| \leqslant \lambda h} \{|\Delta(z)|\,|D\Delta|(z) + \Delta(z)^2\} \\
&\qquad + \max\{|A_{40}(x|\theta,c,\lambda)|, |A_{40}(x|\theta_x,c_x,\lambda)|\} \Bigg] \Bigg).
\end{aligned} \tag{5.9}
$$

The quantities $T_1 \equiv \Delta(z)$, $T_2 \equiv A_{40}(x|\theta,c,\lambda)$ and $T_3 \equiv A_{4j}(x|\theta,c,\lambda) - A_{4j}(x|\theta_x,c_x,\lambda)$ all have zero mean, and have variances equal to $O(s_1^2)$, $O(s_2^2)$ and $O(s_3^2)$, respectively, where $s_1^2 = h^4$, $s_2^2 = \lambda^2 h^6$ and $s_3^2 = (\eta + \lambda h\zeta)(\lambda h^2)^2$. Also, $T_4 \equiv |D\Delta|(z)$ has mean square equal to $O(s_4^2)$, where $s_4^2 = h^2$. For example, to obtain the order of the variance of T_3, note that the area between the curves $\mathscr{C}(x|\theta,c,\lambda)$ and $\mathscr{C}(x|\theta_x,c_x,\lambda)$ equals $O(\alpha)$, where $\alpha = (\eta + \lambda h\zeta)(\lambda h)^2$. The variance of nh^2T_3 is essentially the variance of a Poisson variable with mean $O(n\alpha)$, and so the variance of T_3 equals $O\{(nh^2)^{-2}n\alpha\}$, which, since $h \asymp n^{-1/6}$, equals $O(\alpha h^2) = O(s_3^2)$.

Using Bennett's inequality we may prove that, provided

$$
n^{1-\varepsilon}h^2 s_i \to \infty, \text{ for some } \varepsilon > 0 \text{ and } i = 1, \ldots, 4, \tag{5.10}
$$

the probability that $U_1 \equiv |T_1T_4|$, $U_2 \equiv |T_2|$ or $U_3 \equiv |T_3|$ exceeds $u_1 \equiv C_8 s_1 s_4 \log n$, $u_2 \equiv C_8 s_2 (\log n)^{1/2}$ or $u_3 \equiv C_8 s_3 (\log n)^{1/2}$, respectively, equals $O(n^{-C_9})$ in each case, where C_9 may be made arbitrarily large by choosing C_8 sufficiently large; and these probabilities are of the stated orders uniformly in $x, z \in \mathscr{S}_\varepsilon$, and in c, c_x, θ, θ_x complying with the "uniform*" sense. From this result, using standard methods of approximation (see below), we may deduce that with probability 1 the right-hand side of (5.9), denoted below by RHS, satisfies

$$
\begin{aligned}
&\text{RHS} = O(\delta_n) \\
&\text{where } \delta_n = (\eta + \lambda h\zeta)^{1/2} h (\log n)^{1/2} + (\eta + \lambda h\zeta)\lambda^2 h^4 (\log n)^{1/2},
\end{aligned} \tag{5.11}
$$

the former identity holding uniformly in $x \in \mathscr{S}_\varepsilon$ and in c, c_x, θ, θ_x complying with the "uniform*" sense. (Below we shall refer to this alternative uniform sense as "uniform†".)

The "standard methods of approximation" alluded to above may be summarised as follows. Since $\mathscr{S}$ is bounded then, for any $c > 0$, a square lattice with edge width n^{-c} has only $O(n^{2c})$ of its vertices in $\mathscr{S}$. Since the derivatives of K are Hölder continuous then we may choose c so large that the difference between the value of U_j at a general point u (say) within $\mathscr{S}_\varepsilon$, and the value of U_j at the point of the lattice (within $\mathscr{S}$) that is nearest to u, equals $O(n^{-1})$ uniformly in u and in $j = 1, 2, 3$, with probability 1. Call this result (R_1). By choosing C_8 (introduced in the previous

paragraph) so large that we may take $C_9 \geqslant 2c+2$, and applying the Borel–Cantelli lemma, we may show that the supremum of $U_j(u)$, over all u in the lattice, equals $O(t_j)$ for each j, with probability 1. Call this result (R_2). Since $n^{-1} = O(t_j)$ then, combining (R_1) and (R_2), we have shown that the supremum of $U_j(u)$, over all $u \in \mathscr{S}_\varepsilon$, equals $O(t_j)$ for each j. This implies (5.11).

Define $\bar{g}(x|\theta, c, \lambda)$ to equal the integral of $\xi(c,\lambda)^{-1} g(z)$ over $z \in \mathscr{C}(x|\theta, c, \lambda)$. Given two bounded functions a and b defined in the plane, and a smooth, rectifiable, planar curve $\mathscr{C}$ of finite length $|\mathscr{C}|$, put

$$||a-b||_{\mathscr{C}} = \left[|\mathscr{C}|^{-1} \int_{\mathscr{C}} \{a(z)-b(z)\}^2 \, ds \right]^{1/2}.$$

The conditions assumed of g imply that $\gamma = g + O(h^2)$, whence it follows that

$$B_1(x|\theta, c, \lambda)^{1/2} = ||g - \bar{g}(x|\theta, c, \lambda)||_{\mathscr{C}(x|\theta,c,\lambda)} + O(h^2), \tag{5.12}$$

in the uniform† sense. Moreover, writing $\beta_n = \beta_n(x)$ for a sequence of positive functions satisfying $\beta_n(x) \asymp 1$ uniformly in $x \in \mathscr{S}_\varepsilon$, we claim that

$$||g - \bar{g}(x|\theta, c, \lambda)||_{\mathscr{C}(x|\theta,c,\lambda)} = \beta_n^{1/2}(\eta + \lambda h\zeta)\lambda h \tag{5.13}$$

in the uniform† sense.

To derive (5.13), note that each point on the curve segment $\mathscr{C}(x|\theta, c, \lambda)$ (the length of which is asymptotic to $2\lambda h$) is distant $O\{(\eta + \lambda h\zeta)\,\lambda h\}$ from the nearest point on the true contour line that passes through x. Moreover, along a portion of the curve segment, the portion having length equal to at least constant multiple of λh for all sufficiently large n, the nearest distance is at least a constant multiple of $(\eta + \lambda h\zeta)\lambda h$. Let $\bar{g}_{\mathrm{cont}}(x|\lambda)$ denote the average of $g(z)$ for z in the contour segment $\mathscr{D}(x|\lambda)$. In view of (C_{1g}) and the results just noted,

$$||g - \bar{g}(x|\theta, c, \lambda)||_{\mathscr{C}(x|\theta,c,\lambda)} - ||g - \bar{g}_{\mathrm{cont}}(x|\lambda)||_{\mathscr{D}(x|\lambda)} = \beta_n^{1/2}(\eta + \lambda h\zeta)\lambda h, \tag{5.14}$$

where β_n has the properties claimed of the quantity at (5.13). A similar argument shows that, since g has two Hölder-continuous derivatives in $\mathscr{S}$,

$$||g - \bar{g}(x|\theta_x, c_x, \lambda)||_{\mathscr{C}(x|\theta_x,c_x,\lambda)} - ||g - \bar{g}_{\mathrm{cont}}(x|\lambda)||_{\mathscr{D}(x|\lambda)} = O\{(\lambda h)^{2+t}\}, \tag{5.15}$$

where $t > 0$ depends on the Hölder exponent. But by definition of the contour line $\mathscr{D}(x)$, $\bar{g}_{\mathrm{cont}}(x|\lambda) = g(x)$ and $g(z) = g(x)$ for all $z \in \mathscr{D}(x)$, and so (5.14) and (5.15) are respectively identical to (5.13) and

$$||g - \bar{g}(x|\theta_x, c_x, \lambda)||_{\mathscr{C}(x|\theta_x,c_x,\lambda)} = O\{(\lambda h)^{2+t}\}. \tag{5.16}$$

Combining (5.12) and (5.13) we deduce that

$$B_1(x|\theta, c, \lambda) = \beta_n\{(\eta + \lambda h\zeta)\lambda h\}^2 + O\{(\eta + \lambda h\zeta)\lambda h^3 + h^4\}. \tag{5.17}$$

Likewise, (5.16) and the version of (5.12) for $(\theta, c) = (\theta_x, c_x)$ gives

$$B_1(x|\theta_x, c_x, \lambda) = O\{(\lambda h)^{4+2t} + h^4\}. \tag{5.18}$$

Combining (5.9), (5.11), (5.17) and (5.18), and noting that the quantity $(\eta+\lambda h\zeta)\lambda h^3$ at (5.17) is of smaller order than the term $(\eta+\lambda h\zeta)\lambda h^3(\log n)^{1/2}$ at (5.11), we see that with probability 1, uniformly in $x\in\mathscr{S}_\varepsilon$ for each $\varepsilon>0$,

$$S(x|\theta,c,\lambda)-S(x|\theta_x,c_x,\lambda)=\beta_n\{(\eta+\lambda h\zeta)\lambda h\}^2+O\{\delta_n+(\lambda h)^{4+2t}+h^4\}. \tag{5.19}$$

Therefore, in the operation of minimising $S(x|\theta,c,\lambda)$ over θ and c, $\varepsilon_n\equiv\eta+\lambda h\zeta$ can be made as small as a sufficiently large constant multiple of $\varepsilon_n'\equiv(\lambda h)^{-1}\{\delta_n^{1/2}+(\lambda h)^{2+t}+h^2\}$. Now, the relation $\varepsilon_n\asymp\varepsilon_n'$ is equivalent to

$$\varepsilon_n\asymp(\lambda^2 h)^{-2/3}(\log n)^{1/3}+(\lambda h)^{1+t}. \tag{5.20}$$

Note too that the property $(\lambda^2 h)^{-2/3}(\log n)^{1/3}+(\lambda h)^{1+t}=O(\varepsilon_n)$ implies (5.10). Therefore, with probability 1,

$$\langle\hat\theta_x-\theta_x\rangle+\lambda h|\hat c_x-c_x|=O\{(\lambda^2 h)^{-2/3}(\log n)^{1/3}+(\lambda h)^{1+t}\}.$$

The theorem follows directly from this result.

5.2. Proof of Theorem 3.2

In a slight abuse of notation, write $\breve g$, $\bar\gamma$, $\mathscr{C}$ and ξ for $\breve g(x|\tilde\theta_x,\tilde c_x,\lambda_0)$, $\bar\gamma(x|\tilde\theta_x,\tilde c_x,\lambda_0)$, $\mathscr{C}(x|\tilde\theta_x,\tilde c_x,\lambda_0)$ and $\xi(\tilde c_x,\lambda_0)$, and let $\breve g_0$, $\bar\gamma_0$, $\mathscr{C}_0$ and ξ_0 denote the respective versions of those quantities when $(\tilde\theta_x,\tilde c_x)$ is replaced by (θ_x,c_x). In a slight change of notation from the previous proof, put $\eta=\eta(x)=\langle\tilde\theta_x-\theta_x\rangle$ and $\zeta=\zeta(x)=|\tilde c_x-c_x|$. Standard methods of strong approximation, similar to those used to derive (5.11), may be used to show that under the conditions of the theorem, $|\hat g(z)-\gamma(z)|=O\{h^2(\log n)^{1/2}\}$ and $|D(\hat g-\gamma)|(z)=O\{h(\log n)^{1/2}\}$ uniformly in $z\in\mathscr{S}_\varepsilon$, for each $\varepsilon>0$, with probability 1. Using this result, (5.4), (5.5) and the representations

$$\breve g-\bar\gamma=\xi^{-1}\int_{\mathscr{C}}(\hat g-\gamma),\quad \breve g_0-\bar\gamma_0=\xi_0^{-1}\int_{\mathscr{C}_0}(\hat g_0-\gamma),$$

we may prove that with probability 1,

$$\begin{aligned}&|\breve g(x|\tilde\theta_x,\tilde c_x,\lambda_0)-\bar\gamma(x|\tilde\theta_x,\tilde c_x,\lambda_0)-\{\breve g(x|\theta_x,c_x,\lambda_0)-\bar\gamma(x|\theta_x,c_x,\lambda_0)\}|\\&\quad=O\{(\eta+h\zeta)h^2(\log n)^{1/2}\}.\end{aligned} \tag{5.21}$$

Similarly, using the fact that $|D(\gamma-g)|(z)=O(h)$ uniformly in $z\in\mathscr{S}_\varepsilon$, and applying (5.4), (5.5) and the relation

$$\bar\gamma(x|\tilde\theta_x,\tilde c_x,\lambda_0)-\bar\gamma(x|\theta_x,c_x,\lambda_0)=\xi^{-1}\int_{\mathscr{C}}(\gamma-g)-\xi_0^{-1}\int_{\mathscr{C}_0}(\gamma-g),$$

we may prove that

$$|\bar\gamma(x|\tilde\theta_x,\tilde c_x,\lambda_0)-\bar\gamma(x|\theta_x,c_x,\lambda_0)|=O\{(\eta+h\zeta)h^2\}. \tag{5.22}$$

Both (5.21) and (5.22) are valid uniformly in $x\in\mathscr{S}_\varepsilon$, for each $\varepsilon>0$.

Likewise, recalling that $\check{g}_{\rm cont}(x|\lambda_0)$ is the average value of $\hat{g}$ along the contour segment $\mathscr{D}(x|\lambda_0)$; and noting that, in view of the Hölder continuity of second derivatives of g, $\mathscr{D}(x|\lambda_0)$ and the parabola segment $\mathscr{C}(x|\theta_x, c_x, \lambda_0)$ are uniformly distant $h^2 n^{-\varepsilon}$ apart, for some $\varepsilon > 0$; we may show that with probability 1,

$$|\check{g}(x|\theta_x, c_x, \lambda_0) - \check{g}_{\rm cont}(x|\lambda_0)| = o(h^2). \tag{5.23}$$

Combining (5.21)–(5.23) we deduce that with probability 1, and uniformly in $x \in \mathscr{S}_\varepsilon$,

$$\check{g}(x|\tilde{\theta}_x, \tilde{c}_x, \lambda_0) - \check{g}_{\rm cont}(x|\lambda_0) = O\{(\eta + h\zeta)h^2(\log n)^{1/2}\} + o(h^2). \tag{5.24}$$

The theorem follows from this property and (3.4).

5.3. Proof of Theorem 3.3

Here we show that (C_K), (3.8) and (3.9) are sufficient for (3.4) when $h = c_1 n^{-2/11}$ and $\lambda_0 = c_2 n^{1/11}$, and that estimators $(\tilde{\theta}_x, \tilde{c}_x)$ satisfying (3.9) are readily constructed when (3.8) holds.

The arguments leading to (5.21) and (5.22) apply as before, although the terms ζh and h^2 on the right-hand sides of those formulae should be replaced by $\zeta\lambda_0 h$ and $(\lambda_0 h)^2$, respectively. Therefore, in view of (3.9), for the present choices of h and λ_0, the right-hand sides of (5.21) and (5.22) equal $o(h^2)$ with probability 1, uniformly in $x \in \mathscr{S}_\varepsilon$.

For some $\xi > 0$,

$$\begin{aligned}
&|E\hat{g}(y_1) - g(y_1) - \{E\hat{g}(y_2) - g(y_2)\}| = O(h^2\,|y_1 - y_2|^\xi),\\
&|\hat{g}(y_1) - E\hat{g}(y_1) - \{\hat{g}(y_2) - E\hat{g}(y_2)\}|\\
&\quad = O\{(nh^2\lambda_0)^{-1/2}(||y_1 - y_2||/h)^\xi(\log n)^{1/2}\}
\end{aligned}$$

uniformly in points $y_1 \in \mathscr{D}(x|\lambda_0)$ and $y_2 \in \mathscr{C}(x|\theta_x, c_x, \lambda_0)$ that are both distant s from x and are on the same side of x, and are in $x \in \mathscr{S}_\varepsilon$. (In the case of the second identity the result holds with probability 1.) For some $\eta > 0$, $||y_2 - y_1|| = O\{(\lambda_0 h)^{2+2\eta}\} = O(h^{1+\eta})$, uniformly in pairs (y_1, y_2). Therefore, with probability 1 the difference between the integral averages of $\hat{g} - g$ over $\mathscr{D}(x|\lambda_0)$ and $\mathscr{C}(x|\theta_x, c_x, \lambda_0)$ equals $o(h^2)$, uniformly in $x \in \mathscr{S}_\varepsilon$.

Given y_1 and y_2 as before,

$$g(y_2) = g(y_1) + \sum_{i=1}^{2} (y_2 - y_1)^{(i)}\, g_i(y_1) + O(||y_2 - y_1||^2),$$

where g_1 and g_2 represent first partial derivatives, and bracketed superscripts denote vector components. Recall that $||y_2 - y_1|| = o(h)$, uniformly in (y_1, y_2), and observe that the integral average of $(y_2 - y_1)^{(i)}\, g_i(y_1)$ over $y_2 \in \mathscr{C}(x|\theta_x, c_x, \lambda_0)$ is bounded by a constant multiple of the integral average of $||y_2 - y_1||^2$, and hence equals $o(h^2)$.

From this property and the fact that $g(y_1) = g(x)$ for each y_1 we deduce that the integral average of $g(y_2)$ over $\mathscr{C}(x|\theta_x, c_x, \lambda_0)$ equals the integral average of $g(y_1)$ over $\mathscr{D}(x|\lambda_0)$, plus a term equal to $o(h^2)$.

Combining these results we see that the difference between the integral averages of $\hat{g}$ over $\mathscr{D}(x|\lambda_0)$ and $\mathscr{C}(x|\theta_x, c_x, \lambda_0)$ equals $o(h^2)$, uniformly in $x \in \mathscr{S}_\varepsilon$. This is the analogue of (5.23) in the present setting. Combining this property and the versions of (5.21) and (5.22) we obtain the following version of (5.24): $\breve{g}(x|\tilde{\theta}_x, \tilde{c}_x, \lambda_0) - \breve{g}_{\text{cont}}(x|\lambda_0) = o(h^2)$ uniformly in $x \in \mathscr{S}_\varepsilon$, with probability 1. This is equivalent to (3.4).

Next we show that, if (3.8) holds, estimators $\tilde{\theta}_x$ and $\tilde{c}_x$ can be constructed such that (3.9) is true. Note that, in view of the present choice of h and λ_0, (3.9) is equivalent to

$$n^{2/11}(\log n)^{1/2} \sup_{x \in \mathscr{S}_\varepsilon} \langle \tilde{\theta}_x - \theta_x \rangle \to 0, \quad n^{1/11}(\log n)^{1/2} \sup_{x \in \mathscr{S}_\varepsilon} |\tilde{c}_x - c_x| \to 0 \qquad (5.25)$$

with probability 1. Now, (3.8) implies that, simply by forming the respective derivatives of $\hat{g}$, one may estimate first and second derivatives of g with respective rates $n^{-(5/22)-\eta}$ and $n^{-(1/11)-\eta}$, for uniform convergence in $\mathscr{S}_\varepsilon$ with probability 1. Therefore we may estimate contour tangent angle and contour curvature with the same respective rates. (In fact we may achieve this end by fitting a local quadratic to contours, as suggested in Section 2.1.) Result (5.25) follows from this property.

If, when using the local quadratic contour estimation method outlined in Section 2.1, we choose the bandwidth for $\hat{g}$ to be $h = n^{-(3-\xi_2)/22}$ where $0 < \xi_2 < 9\xi_1/(2 + 3\xi_1)$ and $\xi_1 \in (0, \frac{1}{3})$ is the Hölder coefficient mentioned in (3.8), then for some $\eta > 0$ the convergence rates $n^{-(5/22)-\eta}$ and $n^{-(1/11)-\eta}$ (for, respectively, first and second derivatives of g) mentioned in the previous paragraph are obtained. It follows that the local quadratic contour estimators also enjoy these rates.

Acknowledgments

Helpful comments of an editor and two reviewers have helped improve the paper.

References

[1] M.-Y. Cheng, J. Fan, J.S. Marron, On automatic boundary corrections, Ann. Statist. 25 (1997) 1691–1708.

[1a] A. Cowling, P. Hall, On pseudodata methods for removing boundary effects in kernel density estimation, J. Roy. Statist. Soc. Ser. B 58 (1996) 551–563.

[2] R.A. Djojosugito, P.L. Speckman, Boundary bias correction in nonparametric density estimation, Comm. Statist. Theory Methods 21 (1992) 69–88.

[3] R.L. Eubank, P. Speckman, A bias reduction theorem with applications in nonparametric regression, Scand. J. Statist. 18 (1991) 211–222.

[4] J. Fan, Local linear regression smoothers and their minimax efficiencies, Ann. Statist. 21 (1993) 196–216.

[5] J. Fan, I. Gijbels, Local Polynomial Modelling and its Applications, Chapman & Hall, London, 1996.

[6] T. Gasser, H.G. Müller, Kernel estimation of regression functions, in: T. Gasser, M. Rosenblatt (Eds.), Smoothing Techniques for Curve Estimation, Lecture Notes in Mathematics, Vol. 757, Springer, New York, 1979, pp. 23–68.
[7] T. Gasser, H.G. Müller, V. Mammitzsch, Kernels for nonparametric curve estimation, J. Roy. Statist. Soc. Ser. B 47 (1985) 238–252.
[8] B.L. Granovsky, H.G. Müller, Optimizing kernel methods: a unifying variational principle, Internat. Statist. Rev. 59 (1991) 373–388.
[9] P. Hall, C. Huber, A. Owen, A. Coventry, Asymptotically optimal balloon density estimates, J. Multivariate Anal. 51 (1994) 352–371.
[10] P. Hall, T.E. Wehrly, A geometrical method for removing edge effects from kernel-type nonparametric regression estimators, J. Amer. Statist. Assoc. 86 (1991) 665–672.
[11] T. Hastie, C. Loader, Local regression: automatic kernel carpentry, Statist. Sci. 8 (1993) 120–143.
[12] M.C. Jones, Simple boundary correction for kernel density estimation, Statist. Comput. 3 (1993) 135–146.
[13] H.G. Müller, Smooth optimal kernel estimators near endpoints, Biometrika 78 (1991) 521–530.
[14] H.G. Müller, U. Stadtmüller, Multivariate boundary kernels and a continuous least squares principle, J. Roy. Statist. Soc. Ser. B 61 (1999) 439–458.
[15] J. Rice, Boundary modification for kernel estimators, Commun. Statist. Theory Methods 13 (1984) 893–900.
[16] D. Ruppert, M.P. Wand, Multivariate locally weighted least squares regression, Ann. Statist. 22 (1994) 1346–1370.
[17] E.F. Schuster, Incorporating support constraints into nonparametric estimators of densities, Commun. Statist. Theory Methods 14 (1985) 1123–1136.
[18] D.W. Scott, Multivariate Density Estimation—Theory, Practice and Visualization, Wiley, New York, 1992.
[19] J.G. Staniswalis, K. Messer, D.R. Finston, Kernel estimators for multivariate regression, J. Nonparamet. Statist. 3 (1993) 103–121.
[20] J.G. Staniswalis, K. Messer, Addendum to Staniswalis, Messer and Finston (1993), J. Nonparamet. Statist. 7 (1996) 67–68.
[21] C.J. Stone, Optimal rates of convergence for nonparametric estimators, Ann. Statist. 8 (1980) 1348–1360.
[22] M.P. Wand, M.C. Jones, Kernel Smoothing, Chapman & Hall, London, 1995.