J. R. Statist. Soc. B (1999)
61, *Part* 4, *pp.* 793–815

Multivariate bandwidth selection for local linear regression

Lijian Yang

Michigan State University, East Lansing, USA

and Rolf Tschernig

Humboldt-Universität, Berlin, Germany

[Received December 1997. Final revision January 1999]

Summary. The existence and properties of optimal bandwidths for multivariate local linear regression are established, using either a scalar bandwidth for all regressors or a diagonal bandwidth vector that has a different bandwidth for each regressor. Both involve functionals of the derivatives of the unknown multivariate regression function. Estimating these functionals is difficult primarily because they contain multivariate derivatives. In this paper, an estimator of the multivariate second derivative is obtained via local cubic regression with most cross-terms left out. This estimator has the optimal rate of convergence but is simpler and uses much less computing time than the full local estimator. Using this as a pilot estimator, we obtain plug-in formulae for the optimal bandwidth, both scalar and diagonal, for multivariate local linear regression. As a simpler alternative, we also provide rule-of-thumb bandwidth selectors. All these bandwidths have satisfactory performance in our simulation study.

Keywords: Asymptotic optimality; Blocked quartic fit; Functional estimation; Partial local regression; Plug-in bandwidth; Rule-of-thumb bandwidth; Second derivatives

1. Introduction

Nonparametric estimation in general requires little *a priori* knowledge on the functions to be estimated. The estimation results, however, depend crucially on the choice of bandwidth. Whereas choosing too large a bandwidth may introduce a large bias, selecting too small a bandwidth may cause large estimation variance. An asymptotically optimal bandwidth usually exists and can be obtained through a bias–variance trade-off. Such an optimal bandwidth in general involves functionals of the unknown underlying functions, and the selection of the optimal bandwidth has always been a challenge. Bandwidth selection methods with good theoretical properties and practical performance exist for univariate density estimation, e.g. Jones *et al*. (1996a, b), Cheng (1997) and Grund and Polzehl (1998), and univariate local least squares regression, e.g. Fan and Gijbels (1995) and Ruppert *et al*. (1995). Wand and Jones (1993) provided an excellent analysis of bandwidth selection for bivariate density estimation, Sain *et al*. (1994) looked at multivariate cross-validation for density estimation bandwidth selection and Wand and Jones (1994) developed plug-in (PI) bandwidths for multivariate kernel density estimation. Herrmann *et al*. (1995) studied bandwidth selection for bivariate regression with fixed regular design, but it is unclear whether their method works for random

Address for correspondence: Lijian Yang, Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA.
E-mail: yang@stt.msu.edu

 1369–7412/99/61793

designs and higher dimensions. Ruppert (1997) proposed the empirical bias bandwidth selector (EBBS) which could be applied in a multivariate setting. However, the theoretical properties of the EBBS are unknown and its practical performance has only been studied in the univariate setting. As pointed out by Ruppert (1997), the difficulty in obtaining a reliable multivariate data-driven bandwidth is essentially due to the complexity of estimating higher order multivariate derivatives. To address this difficulty, we estimate multivariate second-order derivatives by a local cubic fit, where all unnecessary terms are left out. This both solves the theoretical problem and makes a practical implementation feasible. We are not aware of any other work that makes use of the desirable properties of a partial local polynomial fit.

Another new feature is the use of a different bandwidth for each of the regressors, in contrast with the EBBS method which uses the same bandwidth for all regressors. We have established some general assumptions that ensure the existence of asymptotic optimal bandwidths. This is not trivial because, when using a bandwidth vector, there is no simple closed formula for the solution of the bias–variance trade-off, and the existence of the solution needs to be implicitly proven. This use of a 'diagonal bandwidth' is a good compromise between flexibility and optimality on the one hand (which need more smoothing parameters, such as a bandwidth matrix) and interpretation and simplicity on the other hand (fewer parameters, such as one bandwidth for all explanatory variables). We have also studied the theoretical properties and practical performances of the latter option, which we call the 'scalar optimal bandwidth'.

The computation of our PI bandwidths requires the estimation of functionals that include derivatives up to the fourth order. The idea is to estimate unknown functionals by blocked polynomial fits, as in Härdle and Marron (1995) and Ruppert *et al.* (1995), but we had further to refine the idea. To do a quartic fit of a function of four variables could require up to 70 parameters in each block, which not only makes it impossible to have a sufficient number of blocks to adapt to local features of the function but also introduces much noise for the estimation within each block. Since the computation of one functional does not require all the derivatives included in a complete fourth-order Taylor expansion, we propose to use a partial quartic fit in the blocks, with only 23 parameters for the same function. This advantage in number grows drastically if the number of variables becomes even more.

The PI bandwidths are necessarily computation intensive, so we also investigated rule-of-thumb (ROT) bandwidths as simpler alternatives.

The paper is organized as follows. In the next section, we show the existence of an asymptotically optimal bandwidth vector. Section 3 defines an ROT and a PI bandwidth selector for local linear regression, and asymptotic optimality is established for the PI bandwidth. In Section 4, we describe in detail how functionals of second derivatives are estimated via partial local cubic regression, while parallel results using a scalar bandwidth selector are summarized in Appendix A. Implementation of the methods proposed is discussed in Section 5. In Section 6 we describe the results of a rather comprehensive Monte Carlo study. Section 7 concludes, while all assumptions and proofs are contained in Appendix B.

The programs that are used in the analysis can be obtained from

`http://www.blackwellpublishers.co.uk/rss/`

2. Background

To formulate the problem, consider a multivariate regression model

$$Y = f(\mathbf{X}) + g(\mathbf{X})\epsilon$$

where Y is a scalar dependent variable, $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ is a vector of explanatory variables, ϵ is independent of $\mathbf{X}$, $E(\epsilon) = 0$ and $\text{var}(\epsilon) = 1$. Let $(\mathbf{X}_i, Y_i)$, $i = 1, 2, \ldots, n$, be an independent and identically distributed sample. Then the local linear estimator of f at a given point $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ is obtained by doing a first-order Taylor expansion of the function f at point $\mathbf{x}$ for all the data points $\mathbf{X}_i$ and solving a least squares problem locally weighted by kernels: i.e. the estimator $\hat{f}(\mathbf{x})$ is the first element c of the vector

$$\{c, (c_\alpha)_{1 \leqslant \alpha \leqslant d}\}$$

that minimizes

$$\sum_{i=1}^{n} \left\{ Y_i - c - \sum_{\alpha=1}^{d} c_\alpha (X_{i\alpha} - x_\alpha) \right\}^2 K_{\mathbf{h}}(\mathbf{X}_i - \mathbf{x})$$

where K is a symmetric, compactly supported, univariate probability kernel (so that K is non-negative and $\int K(u) = 1$) and

$$K_{\mathbf{h}}(\mathbf{X}_i - \mathbf{x}) = \prod_{\alpha=1}^{d} \frac{1}{h_\alpha} K\left(\frac{X_{i\alpha} - x_\alpha}{h_\alpha} \right).$$

Denoting by $p(\mathbf{x})$ the density of $\mathbf{X}$, using a weight function $w(\mathbf{x})$, the weighted mean integrated squared error (MISE)

$$E\left[\int \{\hat{f}(\mathbf{x}) - f(\mathbf{x})\}^2 w(\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right]$$

is a function of the bandwidth vector $\mathbf{h} = (h_1, \ldots, h_d)^{\mathrm{T}}$, and the $\mathbf{h}$ that minimizes this error is called the optimal bandwidth vector. An asymptotic formula of the MISE in this setting is given by

$$\text{AMISE}(\mathbf{h}) = \frac{\sigma_K^4}{4} \sum_{\lambda,\mu=1}^{d} C_{\lambda\mu}(f) h_\lambda^2 h_\mu^2 + \|K\|_2^{2d} B(g) \frac{1}{n h_1 \ldots h_d}$$

in which $\sigma_K^2 = \int u^2 K(u) \, \mathrm{d}u$, $\|K\|_2^2 = \int K^2(u) \, \mathrm{d}u$ and

$$B(g) = \int g^2(\mathbf{x}) w(\mathbf{x}) \, \mathrm{d}\mathbf{x},$$

$$C_{\lambda\mu}(f) = \int f_{\lambda\lambda}(\mathbf{x}) f_{\mu\mu}(\mathbf{x}) p(\mathbf{x}) w(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \qquad \lambda, \mu = 1, \ldots, d,$$

where $f_{\lambda\lambda}(\mathbf{x})$ denotes the second derivative of $f(\mathbf{x})$ with respect to the λth variable x_λ. For the general formula of AMISE using a bandwidth matrix, see p. 140 of Wand and Jones (1995). We denote by $\mathbf{C}(f)$ the matrix whose (λ, μ)th entry is $C_{\lambda\mu}(f)$, which is always non-negative definite. Also, we always have $B(g) > 0$ by assumption 2 in Appendix B.

We introduce a function Q: $R_+^d \times M_+(d) \times R_+ \to R_+$

$$Q(\mathbf{v}, \mathbf{M}, a) = \frac{1}{4} \sum_{\lambda,\mu=1}^{d} M_{\lambda\mu} v_\lambda v_\mu + \frac{a}{\sqrt{(v_1 \ldots v_d)}} = \frac{1}{4} \mathbf{v}^{\mathrm{T}} \mathbf{M} \mathbf{v} + \frac{a}{\sqrt{(v_1 \ldots v_d)}} \tag{2.1}$$

where $M_+(d)$ is a set of non-negative definite $d \times d$ matrices whose definition is given by expression (B.1) in Appendix B, $R_+ = (0, \infty)$ and $M_{\lambda\mu}$ is the (λ, μ)th entry of a matrix $\mathbf{M}$. With $\mathbf{h}^2 = (h_1^2, \ldots, h_d^2)^{\mathrm{T}}$, we then have

$$\mathrm{AMISE}(\mathbf{h}) = Q\left\{\mathbf{h}^2,\ \sigma_K^4\,\mathbf{C}(f),\ \frac{\|K\|_2^{2d}\,B(g)}{n}\right\}.$$

Simple algebra gives

$$Q(\mathbf{v},\,\mathbf{M},\,a) = a^{4/(d+4)}c^{d/(d+4)}\,Q(a^{-2/(d+4)}c^{2/(d+4)}\mathbf{v},\ c^{-1}\mathbf{M},\ 1)$$

and therefore

$$\mathbf{v}(\mathbf{M},\,a) = a^{2/(d+4)}c^{-2/(d+4)}\,\mathbf{v}(c^{-1}\mathbf{M},\ 1) \tag{2.2}$$

where $\mathbf{v}(\mathbf{M},\,a)$ denotes the vector $\mathbf{v}$ that minimizes $Q(\mathbf{v},\,\mathbf{M},\,a)$. By theorem 6, $\mathbf{v}(\mathbf{M},\,a)$ is a well-defined and infinitely smooth function of $\mathbf{M}$ and a. Using equation (2.2), we have the following theorem.

Theorem 1. Under assumptions (1)–(4), $\mathrm{AMISE}(\mathbf{h})$ is minimized by a unique vector which depends on $B(g)$ and $\mathbf{C}(f)$ smoothly via

$$\mathbf{h}_{\mathrm{opt}} = \left\{\frac{\|K\|_2^{2d}\,B(g)}{n\sigma_K^4}\right\}^{1/(d+4)}\mathbf{v}^{1/2}\{\mathbf{C}(f),\ 1\}. \tag{2.3}$$

The optimal bandwidth $\mathbf{h}_{\mathrm{opt}}$ given in formula (2.3) contains unknown quantities $B(g)$ and $\mathbf{C}(f)$ and hence cannot be directly used in the estimation of $f(\mathbf{x})$. In Section 4, appropriate estimators of $f_{\lambda\lambda}(\mathbf{x})$ and $\mathbf{C}(f)$ are given based on a partial local cubic fit with many cross-terms left out. These estimators have very simple bias formulae and their biases and variances are of the same order as keeping the cross-terms in.

3. Bandwidth selection

A quick substitute of $\mathbf{h}_{\mathrm{opt}}$ is the simple ROT bandwidth defined as

$$\mathbf{h}_{\mathrm{ROT}} = \left\{\frac{\|K\|_2^{2d}\,B_{\mathrm{ROT}}(g)}{n\sigma_K^4}\right\}^{1/(d+4)}\mathbf{v}^{1/2}\{\mathbf{C}_{\mathrm{ROT}}(f),\ 1\} \tag{3.1}$$

in which $\mathbf{C}_{\mathrm{ROT}}(f)$ and $B_{\mathrm{ROT}}(g)$ are based on piecewise quartic estimation of $f(\mathbf{x})$, and whose formulae are given by expressions (5.1) and (5.2).

The bandwidth vector $\mathbf{h}_{\mathrm{ROT}}$ is of the same order as $\mathbf{h}_{\mathrm{opt}}$ but does not estimate $\mathbf{h}_{\mathrm{opt}}$ efficiently because f may not always be represented by a piecewise fourth-order polynomial. The idea of such an ROT bandwidth exists in the univariate case; see Fan and Gijbels (1995) and Ruppert *et al*. (1995). The latter additionally used a blocking idea that we have also adopted.

To improve on the ROT idea, we propose a PI bandwidth, so called as it approximates $\mathbf{h}_{\mathrm{opt}}$ by plugging in consistent estimates of the functionals $B(g)$ and $\mathbf{C}(f)$. Specifically, we define

$$\mathbf{h}_{\mathrm{PI}} = \left\{\frac{\|K\|_2^{2d}\,\hat{B}(g)}{n\sigma_K^4}\right\}^{1/(d+4)}\mathbf{v}^{1/2}\{\hat{\mathbf{C}}\,(f),\ 1\} \tag{3.2}$$

in which $\hat{\mathbf{C}}(f) = \{\hat{C}_{\lambda\mu}(f)\} = \{\hat{C}_{\lambda\mu}(f,\,h)\}$ where for each $\lambda,\ \mu = 1,\ \ldots,\ d$

$$\hat{C}_{\lambda\mu}(f) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{\lambda\lambda}(\mathbf{X}_i)\,\hat{f}_{\mu\mu}(\mathbf{X}_i)\,w(\mathbf{X}_i) = \int \hat{f}_{\lambda\lambda}(\mathbf{x})\,\hat{f}_{\mu\mu}(\mathbf{x})\,w(\mathbf{x})\,p(\mathbf{x})\,\mathrm{d}\mathbf{x} + O_p(n^{-1/2}) \tag{3.3}$$

and the $\hat{f}_{\lambda\lambda}(\mathbf{x})$ is a partial local cubic estimator of $f_{\lambda\lambda}(\mathbf{x})$ given in equation (4.2). The estimator $\hat{B}(g)$ is defined in equation (5.3).

Using standard results such as contained in Wand and Jones (1995) or Fan and Gijbels (1996), we have $\hat{B}(g)/B(g) = 1 + O_p(n^{-2/(d+4)})$ as $n \to \infty$. On the basis of this, the asymptotics of $\hat{\mathbf{C}}(f)$ as given in corollary 2 and the smooth dependence of $\mathbf{h}_{\text{opt}}$ on $\mathbf{C}(f)$ as in theorem 1, we have the following theorem.

Theorem 2. Under assumptions (1)–(4), for $n \to \infty$, the bandwidth defined in equation (3.2) is asymptotically optimal. In particular,

$$\mathbf{h}_{\text{PI}} = \mathbf{h}_{\text{opt}}\{I_d + O_p(n^{-2/(d+6)})\}.$$

Since $\mathbf{h}_{\text{PI}}$ is a consistent substitute for the optimal bandwidth $\mathbf{h}_{\text{opt}}$, it can be shown that $\mathbf{h}_{\text{PI}}$ performs asymptotically similarly to $\mathbf{h}_{\text{opt}}$ in terms of MISE. $\mathbf{h}_{\text{PI}}$ will on average work better than $\mathbf{h}_{\text{ROT}}$ since the latter is an inconsistent bandwidth estimator. Simulation results in Section 6 strongly support these expectations.

In Appendix A, we discuss the use of a scalar bandwidth h instead of $\mathbf{h}$, i.e. $\mathbf{h} = (h, \ldots, h)$. In that case, the corresponding ROT bandwidth h_{ROT} and PI bandwidth h_{PI} are given by equations (A.2) and (A.3).

4. Partial local estimation

In this section, we formulate an estimator $\hat{f}_{\lambda\lambda}(\mathbf{x})$ of $f_{\lambda\lambda}(\mathbf{x})$. In what follows, denote for any compact supported function L

$$\mu_r(L) = \int_{-\infty}^{\infty} u^r L(u)\,\mathrm{d}u;$$

hence in particular $\mu_2(K) = \sigma_K^2$. We write $\mu_4(K) = \kappa\sigma_K^4$ where κ denotes the kurtosis of the kernel K, which is always greater than 1. For the derivative functions, we denote

$$f_{\alpha\beta\gamma}(\mathbf{x}) = \frac{\partial^3}{\partial x_\alpha \partial x_\beta \partial x_\gamma} f(\mathbf{x}),$$

etc.

Denote

$$Z_\lambda = [1, \{(X_{i\alpha} - x_\alpha)^2\}, \{(X_{i\alpha} - x_\alpha)(X_{i\lambda} - x_\lambda)\}_{\alpha\neq\lambda}, (X_{i\alpha} - x_\alpha), \{(X_{i\alpha} - x_\alpha)(X_{i\lambda} - x_\lambda)^2\}_{\alpha\neq\lambda},$$
$$\{(X_{i\alpha} - x_\alpha)^2(X_{i\lambda} - x_\lambda)\}_{\alpha\neq\lambda}, (X_{i\lambda} - x_\lambda)^3]_{i=1}^n, \tag{4.1}$$

and then define

$$\hat{f}_{\lambda\lambda}(\mathbf{x}) = 2e_\lambda^{\mathrm{T}}(Z_\lambda^{\mathrm{T}} W Z_\lambda)^{-1} Z_\lambda^{\mathrm{T}} W Y \tag{4.2}$$

where e_λ is a $(5d-1)$-vector of 0s whose $(\lambda+1)$-element is 1, $Y = (Y_i)_{n\times 1}$ and with a scalar bandwidth $\mathbf{h} = (h, \ldots, h)$

$$W = \operatorname{diag}\left\{\frac{1}{n} K_h(\mathbf{X}_i - \mathbf{x})\right\}_{i=1}^n;$$

see Ruppert and Wand (1994) for a general definition of a local polynomial estimator. We then obtain the following theorem.

Theorem 3. Under assumptions (1)–(3), for $\lambda = 1, \ldots, d$, as $h \to 0$ and $nh^{d+4} \to \infty$, we have

$$\sqrt{(nh^{d+4})}\{\hat{f}_{\lambda\lambda}(\mathbf{x}) - f_{\lambda\lambda}(\mathbf{x}) - b_{\lambda\lambda}(\mathbf{x})h^2\} \to N\{0, \sigma^2_{\lambda\lambda}(\mathbf{x})\}$$

where

$$b_{\lambda\lambda}(\mathbf{x}) = \frac{\mu_6(K) - \kappa\sigma_K^6}{12(\kappa-1)\sigma_K^4} f_{\lambda\lambda\lambda\lambda}(\mathbf{x}) + \sum_{\alpha\neq\lambda} \frac{\sigma_K^2}{2} f_{\alpha\alpha\lambda\lambda}(\mathbf{x}), \tag{4.3}$$

$$\sigma^2_{\lambda\lambda}(\mathbf{x}) = \frac{4\{\mu_4(K^2) - 2\,\mu_2(K^2)\sigma_K^2 + \|K\|_2^2\sigma_K^4\}\,\|K\|_2^{2(d-1)}\,g^2(\mathbf{x})}{\sigma_K^8(\kappa-1)^2\,p(\mathbf{x})}. \tag{4.4}$$

The formation of the matrix Z_λ is much simpler than the corresponding matrix of a full local cubic expansion. This makes the explicit mathematical derivation of $(Z_\lambda^{\mathrm{T}} W Z_\lambda)^{-1}$ easier and computing expression (4.2) less costly.

On the basis of equation (4.2), the asymptotic properties of $\hat{C}_{\lambda\mu}(f, h)$ are presented in the next theorem. In the following, we denote $K^*(u) = K(u)(u^2 - \sigma_K^2)$, $K * K^*$ the convolution between K and K^*,

$$(K * K^*)(t) = \int K(t-u)\,K^*(u)\,\mathrm{d}u,$$

which is also $\int K(t+u)\,K^*(u)\,\mathrm{d}u$ by symmetry, and $K^{(2)} = K * K$, the self-convolution of K. Also, we denote by $\delta_{\lambda\mu}$ the Kronecker index, which equals 1 or 0 depending on whether $\lambda = \mu$ or not.

Theorem 4. Under assumptions (1)–(4), as $h \to 0$ and $nh^{d+4} \to \infty$, we have

$$\hat{C}_{\lambda\mu}(f, h) = C_{\lambda\mu}(f) + \{B_{\lambda\mu}(f) + B_{\mu\lambda}(f)\}h^2 + \frac{B(g)\,V_{\lambda\mu}(K)}{nh^{d+4}} + \zeta_{\lambda\mu} + O_p(h^4) + o_p\left(\frac{1}{n\sqrt{h^{d+8}}}\right)$$

in which

$$B_{\lambda\mu}(f) = \int f_{\mu\mu}(\mathbf{x})\,b_{\lambda\lambda}(\mathbf{x})\,w(\mathbf{x})\,p(\mathbf{x})\,\mathrm{d}\mathbf{x},$$

$$V_{\lambda\mu}(K) = \frac{4\|K\|_2^{2(d-2)}\{\delta_{\lambda\mu}\,\mu_4(K^2)\|K\|_2^2 + (1-\delta_{\lambda\mu})\,\mu_2^2(K^2) - 2\sigma_K^2\,\mu_2(K^2)\|K\|_2^2 + \sigma_K^4\|K\|_2^4\}}{\sigma_K^8(\kappa-1)^2},$$

$$n\sqrt{h^{d+8}}\,\zeta_{\lambda\mu} \xrightarrow{\mathrm{D}} N\{0, \sigma^2_{\lambda\mu}(g, K)\}$$

where

$$\sigma^2_{\lambda\mu}(g, K) = \frac{32\int g^4(\mathbf{x})\,w^2(\mathbf{x})\,\mathrm{d}\mathbf{x}}{\sigma_K^{16}(\kappa-1)^4}\int F_{\lambda\lambda}(\mathbf{t})\,F_{\mu\mu}(\mathbf{t})\,\mathrm{d}\mathbf{t}$$

and where we define for any $\lambda, \mu = 1, \ldots, d$

$$F_{\lambda\mu}(\mathbf{t}) = (1-\delta_{\lambda\mu})(K^* * K)(t_\lambda)(K^* * K)(t_\mu)\prod_{\gamma\neq\lambda,\mu} K^{(2)}(t_\gamma) + \delta_{\lambda\mu}\,K^{*(2)}(t_\lambda)\prod_{\gamma\neq\lambda} K^{(2)}(t_\gamma).$$

Note here that $\int g^4(\mathbf{x})\,w^2(\mathbf{x})\,\mathrm{d}\mathbf{x} > 0$ by assumption (2), and that therefore $\hat{C}_{\lambda\mu}(f, h)$ has a standard deviation of order $(n\sqrt{h^{d+8}})^{-1}$, which is smaller than one of the bias terms $(nh^{d+4})^{-1}$.

The two bias terms point to a trade-off if both are positive or cancellation if the h^2-term happens to be negative. Assuming that the h^2-term is never 0, we obtain the following results for estimating $C_{\lambda\mu}(f)$ optimally.

Corollary 1. Under the same assumptions as in theorem 4, the optimal bandwidth to estimate $C_{\lambda\mu}(f)$ by $\hat{C}_{\lambda\mu}(f, h)$ is

$$h_{C_{\lambda\mu}(f),\mathrm{opt}} = \begin{cases} \left[-\dfrac{B(g)\, V_{\lambda\mu}(K)}{\{B_{\lambda\mu}(f) + B_{\mu\lambda}(f)\}n} \right]^{1/(d+6)} & \text{if } B_{\lambda\mu}(f) + B_{\mu\lambda}(f) < 0, \\ \left[\dfrac{d+4}{2} \dfrac{B(g)\, V_{\lambda\mu}(K)}{\{B_{\lambda\mu}(f) + B_{\mu\lambda}(f)\}n} \right]^{1/(d+6)} & \text{if } B_{\lambda\mu}(f) + B_{\mu\lambda}(f) > 0, \end{cases} \tag{4.5}$$

which gives the standard error of order $n\sqrt{h^{d+8}_{C_{\lambda\mu}(f),\mathrm{opt}}} = O(n^{-(d+4)/2(d+6)})$ and the bias of order $h^4_{C_{\lambda\mu}(f),\mathrm{opt}} = O(n^{-4/(d+6)})$ when $B_{\lambda\mu}(f) + B_{\mu\lambda}(f) < 0$ and of order $h^2_{C_{\lambda\mu}(f),\mathrm{opt}} = O(n^{-2/(d+6)})$ when $B_{\lambda\mu}(f) + B_{\mu\lambda}(f) > 0$.

We can now complete the definition of $\hat{C}_{\lambda\mu}(f)$, and therefore of $\hat{\mathbf{C}}(f)$. Estimate all the derivatives $f_{\mu\mu}(\mathbf{x})$ etc. by blocked quartic fits and obtain the estimates of $B_{\lambda\mu}(f)$ as described in Section 5. Then plug them in formula (4.5) together with $\hat{B}(g)$ in place of $B(g)$. Use the resulting bandwidth in the estimator $\hat{C}_{\lambda\mu}(f)$. Definition (3.3) and theorem 4 yield the following corollary.

Corollary 2. Under assumptions (1)–(4), for $n \to \infty$, the $\hat{\mathbf{C}}(f)$ defined by equation (3.3) has the consistency property

$$\hat{\mathbf{C}}(f) = \mathbf{C}(f)\{I_d + O_p(n^{-2/(d+6)})\}.$$

5. Implementation

In this section we describe how to estimate $\mathbf{C}(f)$ and $B(g)$ to obtain the ROT and PI bandwidths (3.1) and (A.2), and (3.2) and (A.3) respectively.

For the ROT estimation of $f(\mathbf{X}_i)$ and $f_{\lambda\lambda}(\mathbf{X}_i)$, Fan and Gijbels (1995) suggested the use of a quartic Taylor expansion. Ruppert *et al.* (1995) proposed to separate the data into equal-sized blocks and to use a quartic Taylor expansion on each block in case the function is very wiggly. Since the number of parameters per block increases dramatically with the increase in dimension and since the wiggliness of the function depends on each variable differently, our flexible blocking scheme sorts the data one variable at a time. Let N_λ denote the number of blocks in the λth direction and collect them in $\mathbf{N} = (N_1, \ldots, N_d)$. Then $N = \Pi_{\lambda=1}^{d} N_\lambda$ is the total number of blocks given the choices of N_λ, $\lambda = 1, 2, \ldots, d$. To choose the optimal blocks we follow Ruppert *et al.* (1995) and use Mallows's C_p,

$$C_\mathrm{p}(\mathbf{N}) = \frac{\mathrm{RSS}(\mathbf{N})\{n - k(d)[n/4k(d)]\}}{\min_{\mathbf{N}}\{\mathrm{RSS}(\mathbf{N})\}} - \{n - 2\,k(d)N\}$$

where $\mathrm{RSS}(\mathbf{N})$ denotes the residual sum of squares based on the quartic fit with blocking $\mathbf{N}$,

$$k(d) = 1 + \sum_{i=1}^{4} \binom{d+i-1}{i}$$

is the maximum number of parameters in one block and $[a]$ denotes the integer part of a.

Denoting the block vector chosen with $C_p(\mathbf{N})$ by $\mathbf{N}^*$, and by $\hat{f}_{\mathrm{ROT},\mathbf{N}^*}$, $\hat{f}_{\mathrm{ROT},\mathbf{N}^*,\lambda\lambda}$ the corresponding function estimators, we then estimate

$$B_{\mathrm{ROT}}(g) = \frac{1}{n - k(d)N^*} \sum_{i=1}^{n} \frac{\{Y_i - \hat{f}_{\mathrm{ROT},\mathbf{N}^*}(\mathbf{X}_i)\}^2 \, w(\mathbf{X}_i)}{p_{\mathrm{ROT}}(\mathbf{X}_i)}, \tag{5.1}$$

$$\mathbf{C}_{\mathrm{ROT}}(f) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{\mathrm{ROT},\mathbf{N}^*,\lambda\lambda}(\mathbf{X}_i) \, \hat{f}_{\mathrm{ROT},\mathbf{N}^*,\mu\mu}(\mathbf{X}_i) \, w(\mathbf{X}_i) \right\} \tag{5.2}$$

where $p_{\mathrm{ROT}}(\mathbf{x})$ is the uniform density on the range of the data. These estimates are necessarily inconsistent unless the true f is a piecewise quartic function.

This problem is avoided by PI estimation described in Section 4 and Appendix A. Now we estimate $B(g)$ by

$$\hat{B}(g) = \frac{1}{n} \sum_{i=1}^{n} \frac{w(\mathbf{X}_i)\{Y_i - \hat{f}_{\mathbf{h}_{\mathrm{ROT}}}(\mathbf{X}_i)\}^2}{\tilde{p}_{h_S}(\mathbf{X}_i)} \tag{5.3}$$

where $\mathbf{h}_{\mathrm{ROT}}$ is defined by equations (3.1), (5.1) and (5.2), h_S is the d-dimensional normal reference bandwidth (Silverman (1986), pages 86–87)) and $\tilde{p}_{h_S}(\mathbf{x})$ is a kernel density estimator.

To estimate $\mathbf{C}(f)$ we use partial local cubic regression as described in Section 4. This is better done with a bandwidth that is slightly larger than the optimal pilot bandwidth $h_{C_{\lambda\mu},\mathrm{opt}}$. The reason is that the bandwidth $h_{C_{\lambda\mu},\mathrm{opt}}$ only minimizes the absolute value of the bias, whereas it completely ignores the variance in $\mathbf{C}(f)$ estimation. Asymptotically this is justified since the standard deviation of $\hat{\mathbf{C}}(f, h)$ is of higher order than the bias. To use a bandwidth that is not equal to $h_{C_{\lambda\mu},\mathrm{opt}}$ would produce in the $\mathbf{C}(f)$ estimate a larger bias than the minimum bias, but not significantly so if the bandwidth is larger. To see this effect we take the bandwidth $\rho h_{C_{\lambda\mu},\mathrm{opt}}$ and plot in Fig. 1 the ratio of the bias against the minimum bias (achieved at $\rho = 1$) as a function $r(\rho)$ of $\rho \in [0.6, 2]$, for $d = 4$. Here, we assume that $B_{\lambda\mu} + B_{\mu\lambda} > 0$, and the ratio then is

$$r(\rho) = \frac{\{(d+4)/2)\}^{2/(d+6)}\rho^2 + \{(d+4)/2)\}^{-(d+4)/(d+6)}\rho^{-(d+4)}}{\{(d+4)/2)\}^{2/(d+6)} + \{(d+4)/2\}^{-(d+4)/(d+6)}}.$$

The broken line has the height of $r(2)$. We can see that $r(\rho)$ is about 3 or less for $\rho > 1$, whereas there is a dramatic increase for $\rho < 1$. Also it is easy to verify that $r(2) \leqslant 4$ for any d. Meanwhile, for $\rho = 2$, the ratio of decrease in the standard deviation is $\rho^{-(d+8)/2}$, which is $1/32$ if $d = 2$ and more for larger d. This decrease in standard deviation is drastic whereas the increase in bias is at most 4 times. We therefore use $2h_{C_{\lambda\mu},\mathrm{opt}}$, which should have very good finite sample performance.

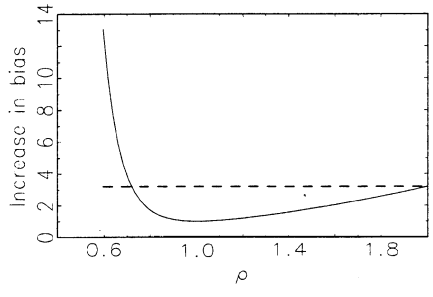

Fig. 1. Ratio of the increase in bias when estimating $\mathbf{C}(f)$ with $\rho h_{C_{\lambda\mu}(f),\mathrm{opt}}$, for $d = 4$

The estimation of the unknown functional $B_{\lambda\mu}$ is based on blocked quartic fits of f if $d \leqslant 3$. For larger d, the number of parameters can be so large that for medium sample sizes blocking or even global estimation becomes impossible. We therefore use a partial quartic fit for the estimation of each $B_{\lambda\mu}$, consisting of only those fourth-order derivatives $f_{\lambda\lambda\alpha\alpha}$, $\alpha = 1, \ldots, d$, in equation (4.3) plus all lower order derivatives that form a subset of those. As an example, such a partial quartic polynomial centred at zero reads

$$b_0 + \sum_{\alpha=1}^{d} b_\alpha X_\alpha + \sum_{\alpha=1}^{d} b_{\lambda\alpha} X_\lambda X_\alpha + \sum_{\alpha=1,\alpha\neq\lambda}^{d} b_{\alpha\alpha} X_\alpha^2 + \sum_{\alpha=1}^{d} b_{\lambda\alpha\alpha} X_\lambda X_\alpha^2 + \sum_{\alpha=1,\alpha\neq\lambda}^{d} b_{\lambda\lambda\alpha} X_\lambda^2 X_\alpha + \sum_{\alpha=1}^{d} b_{\lambda\lambda\alpha\alpha} X_\lambda^2 X_\alpha^2.$$

For this partial expansion the number of parameters $k(d) = 6d - 1$ grows linearly in d. It also includes all terms to estimate $f_{\mu\mu}$. For $d = 4$, the number of parameters drops to about a third of the complete expansion. The partial expansion allows not only more blocks but also the blocks to focus more on the variable directions that have more curvature. This is a new feature that is not an issue in the univariate case.

For efficient partial quartic fits we propose to use at least $4\,k(d)$ observations in each block. This amounts to requiring $n \geqslant 4\,k(d) = 4(6d - 1)$. For example, if $d = 4$, partial blocking needs $n \geqslant 92$ observations. If $n < 4(6d - 1)$, we cannot recommend any of our methods with confidence as there are just not enough data to estimate f accurately.

To use standard software for the numerical optimization problem in equations (3.1), (3.2) or (2.3), it may be convenient to reparameterize $\mathbf{v}$ as $\mathbf{v} = \exp(\mathbf{u})$, i.e. $v_1 = \exp(u_1), \ldots, v_d = \exp(u_d)$ where $\mathbf{u} = (u_1, \ldots, u_d)^{\mathrm{T}} \in R^d$ so there is no constraint on $\mathbf{u}$. The gradient and Hessian matrices are then

$$\frac{\partial Q\{\exp(\mathbf{u}), \mathbf{M}, a\}}{\partial \mathbf{u}} = \frac{1}{2} \begin{pmatrix} \exp(u_1) \sum_{\lambda=1}^{d} M_{1\lambda} \exp(u_\lambda) \\ \vdots \\ \exp(u_d) \sum_{\lambda=1}^{d} M_{d\lambda} \exp(u_\lambda) \end{pmatrix} - \frac{a}{2\sqrt{\{\exp(u_1) \ldots \exp(u_d)\}}} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$\begin{aligned} \frac{\partial^2 Q\{\exp(\mathbf{u}), \mathbf{M}, a\}}{\partial \mathbf{u} \partial \mathbf{u}^{\mathrm{T}}} &= \frac{1}{2} \operatorname{diag}\{\exp(u_1), \ldots, \exp(u_d)\} \mathbf{M} \operatorname{diag}\{\exp(u_1), \ldots, \exp(u_d)\} \\ &\quad + \frac{1}{2} \operatorname{diag}\{M_{11} \exp(2u_1), \ldots, M_{dd} \exp(2u_d)\} \\ &\quad + \frac{a}{4\sqrt{\{\exp(u_1) \ldots \exp(u_d)\}}} \mathbf{1}_{d\times d}, \end{aligned}$$

which make it easy to find $\mathbf{u}(\mathbf{M}, a) = \ln\{\mathbf{v}(\mathbf{M}, a)\}$, the $\mathbf{u}$-value that minimizes $Q\{\exp(\mathbf{u}), \mathbf{M}, a\}$. We then obtain the following substitute for equation (2.3):

$$\mathbf{h}_{\mathrm{opt}} = \left\{ \frac{\|K\|_2^{2d} B(g)}{n\sigma_K^4} \right\}^{1/(d+4)} \exp\left[\frac{\mathbf{u}\{\mathbf{C}(f), 1\}}{2} \right].$$

Similar substitutes are used to compute equations (3.1) and (3.2).

6. Simulation results

In this section we investigate the finite sample performance of the ROT bandwidths $\mathbf{h}_{\mathrm{ROT}}$

Table 1. Model 1: mean of the ISE for various bandwidth selectors

Bandwidth	*Kind*	*Means for the following sample sizes:*		
		100	*250*	*500*
Diagonal	Asymptotic optimal	0.0326	0.0165	0.00893
	PI	0.0409	0.0183	0.00931
	ROT	0.0656	0.0233	0.0112
Scalar	Asymptotic optimal	0.0338	0.0174	0.00934
	PI	0.0401	0.0180	0.00928
	ROT	0.0734	0.0244	0.0113

Table 2. Model 2: mean of the ISE for various bandwidth selectors

Bandwidth	*Kind*	*Means for the following sample sizes:*		
		100	*250*	*500*
Diagonal	Asymptotic optimal	0.0505	0.0246	0.0140
	PI	0.0523	0.0248	0.0139
	ROT	0.0752	0.0301	0.0168
Scalar	Asymptotic optimal	0.0505	0.0246	0.0140
	PI	0.0503	0.0240	0.0136
	ROT	0.0824	0.0314	0.0175

Table 3. Model 3: mean of the ISE for various bandwidth selectors

Bandwidth	*Kind*	*Means for the following sample sizes:*		
		100	*250*	*500*
Diagonal	Asymptotic optimal	0.0704	0.0327	0.0186
	PI	0.0711	0.0334	0.0185
	ROT	0.114	0.0469	0.0242
Scalar	Asymptotic optimal	0.0760	0.0361	0.0208
	PI	0.0747	0.0354	0.0202
	ROT	0.137	0.0560	0.0272

Table 4. Model 4: mean of the ISE for various bandwidth selectors

Bandwidth	*Kind*	*Means for the following sample sizes:*		
		100	*250*	*500*
Diagonal	Asymptotic optimal	0.0765	0.0344	0.0198
	PI	0.0873	0.0378	0.0219
	ROT	0.230	0.0525	0.0268
Scalar	Asymptotic optimal	0.147	0.0604	0.0349
	PI	0.105	0.0536	0.0337
	ROT	0.282	0.0823	0.0431

given by equation (3.1) and h_{ROT} given by equation (A.2), the PI bandwidths $\mathbf{h}_{\mathrm{PI}}$ given by equation (3.2) and h_{PI} given by equation (A.3) for the local linear estimation of $f(\mathbf{x})$. These results are compared with those obtained by using the asymptotic optimal bandwidths $\mathbf{h}_{\mathrm{opt}}$ as in equation (2.3) and h_{opt} as in equation (A.1), which we can compute since $f(\mathbf{x})$, $g(\mathbf{x})$ and $p(\mathbf{x})$ are explicitly given. For each pseudodata set generated, we compute the ISE

Table 5. Model 5: mean of the ISE for various bandwidth selectors

Bandwidth	*Kind*	*Means for the following sample sizes:*		
		100	*250*	*500*
Diagonal	Asymptotic optimal	0.0662	0.0324	0.0193
	PI	0.142	0.0573	0.0297
	ROT	0.306	0.100	0.0466
Scalar	Asymptotic optimal	0.150	0.0716	0.0425
	PI	0.145	0.0692	0.0410
	ROT	0.338	0.116	0.0606

Table 6. Model 6: mean of the ISE for various bandwidth selectors

Bandwidth	*Kind*	*Means for the following sample sizes:*	
		250	*500*
Diagonal	Asymptotic optimal	0.459	0.291
	PI	0.363	0.225
	ROT	0.641	0.502
Scalar	Asymptotic optimal	0.791	0.502
	PI	0.447	0.308
	ROT	0.793	0.747

$$\frac{1}{n}\sum_{i=1}^{s}\{\hat{f}_h(\mathbf{U}_i) - f(\mathbf{U}_i)\}^2$$

for the diagonal bandwidths $\mathbf{h} = \mathbf{h}_{\text{opt}}$, $\mathbf{h}_{\text{PI}}$, $\mathbf{h}_{\text{ROT}}$ and scalar bandwidths $h = h_{\text{opt}}$, h_{PI}, h_{ROT}, where the $\mathbf{U}_i$ are taken from an equally spaced grid of roughly 2500 points. In total we consider six models of varying complexity and in all cases 100 replications are conducted. For all the models the random design matrix $\mathbf{X}$ is an independent sample from the uniform distribution on $[0, 1]^d$ with sample size n, and for the estimation of $\mathbf{C}(f)$ we screen off observations $\mathbf{X}_i$ that are outside $[a, 1-a]^d$ where $a = (1 - 0.9^{1/d})/2$. The sample size n is 100, 250 or 500 except for model 6 where $n = 250$ and $n = 500$. The models are

$$Y = f(\mathbf{X}) + \epsilon, \qquad \epsilon \sim N(0, 0.25),$$

where the regression function is

$$f(\mathbf{x}) = x_1^2 + x_2^4$$

for model 1,

$$f(\mathbf{x}) = \sin\{\pi(x_1 + x_2)\}$$

for model 2,

$$f(\mathbf{x}) = \sin(\pi x_1)\,\sin(2\pi x_2)$$

for model 3,

$$f(\mathbf{x}) = \{\sin(\pi x_1) + \sin(4\pi x_2)\}/2$$

for model 4,

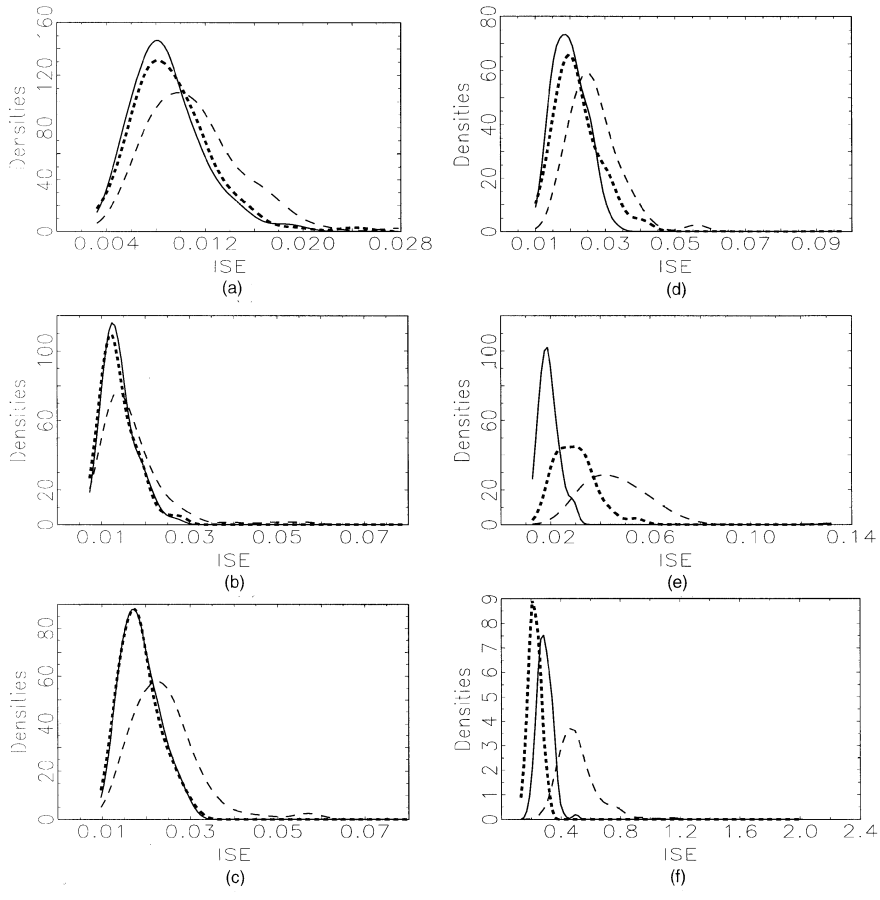

Fig. 2. Density of the ISE for (a) model 1, (b) model 2, (c) model 3, (d) model 4, (e) model 5 and (f) model 6 (———, $\mathbf{h}_{\text{opt}}$; - - - - -, $\mathbf{h}_{\text{PI}}$; — — —, $\mathbf{h}_{\text{ROT}}$; $n = 500$)

$$f(\mathbf{x}) = \{(x_1 - 0.5)^2 + x_2^2\} \sin(2\pi x_3)$$

for model 5 and

$$f(\mathbf{x}) = \sin\{\pi(x_1 + 0.5x_2 + 2x_3 + 0.5x_4)\}$$

for model 6.

Table 1 reports in its first three rows the mean of the ISE based on $\mathbf{h}_{\text{opt}}$, $\mathbf{h}_{\text{PI}}$, $\mathbf{h}_{\text{ROT}}$ for each sample size of model 1. The last three rows present the corresponding results for the scalar bandwidth h_{opt}, h_{PI} and h_{ROT}. Similarly, Tables 2–6 display the results for models 2–6. In addition, Fig. 2 presents plots of the densities of the ISE based on $\mathbf{h}_{\text{opt}}$ (full curve), $\mathbf{h}_{\text{PI}}$ (short broken curve) and $\mathbf{h}_{\text{ROT}}$ (long broken curve) for all models, where the sample size is 500. Inspecting Tables 1–6 and Fig. 2, we find that the PI bandwidth is always superior to the ROT bandwidth as it delivers function estimates that have smaller ISEs. Furthermore, the mean of the ISE declines with sample size as expected.

If we use $C(f)$ to measure complexity, the two-dimensional models 1–4 are of increasing complexity with $C(f) = 48.8$, 194.8, 608.8, 3129.3 respectively. Since the ROT bandwidth is based on piecewise quartic fits, we might expect the ROT function estimates to perform worse

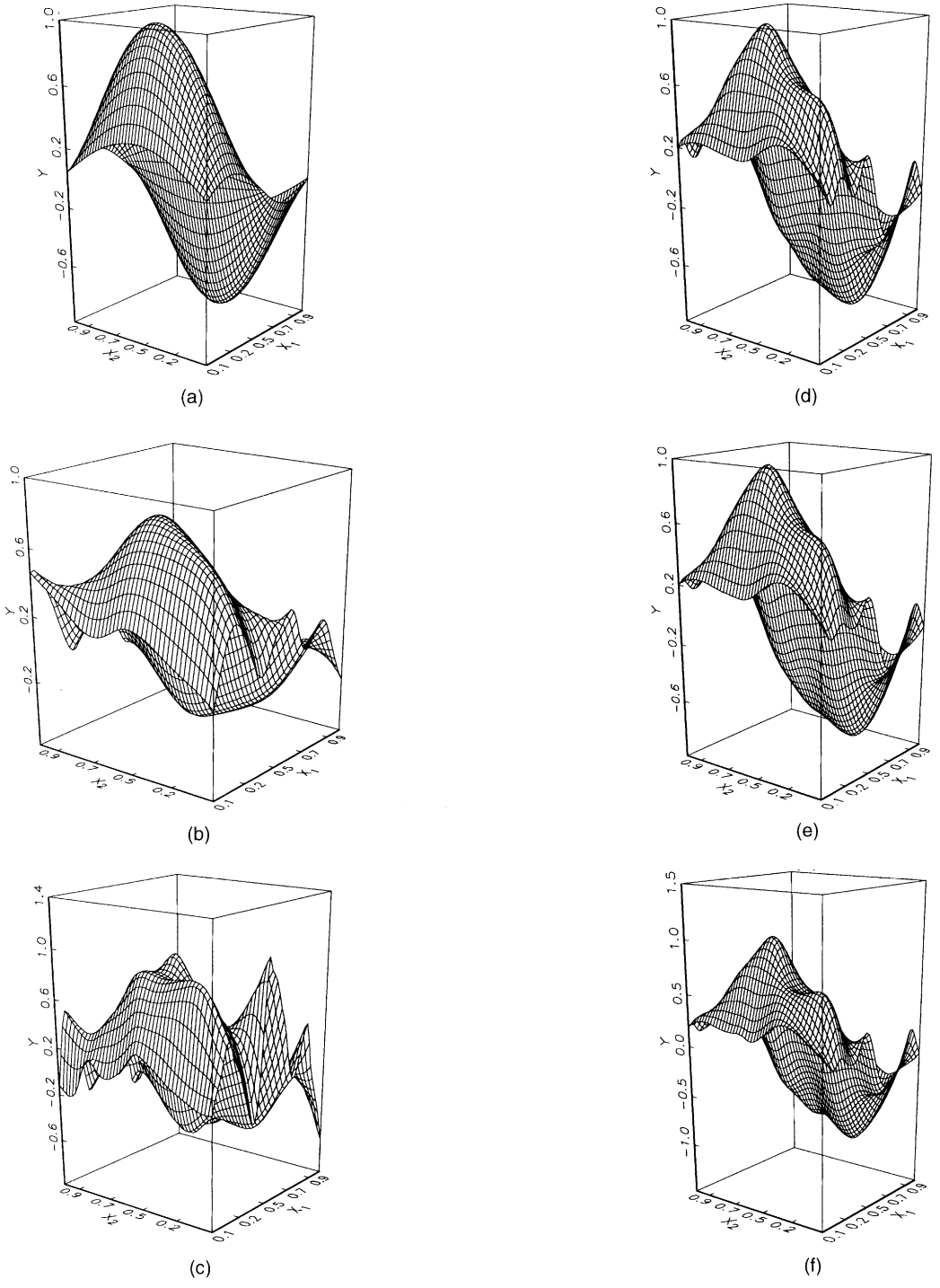

Fig. 3. Function $f(\cdot)$ of model 3: (a) true function; (b) estimate with $\mathbf{h}_{\mathrm{PI}}$, $n = 100$; (c) estimate with $\mathbf{h}_{\mathrm{ROT}}$, $n = 100$; (d) estimate with $\mathbf{h}_{\mathrm{opt}}$, $n = 500$; (e) estimate with $\mathbf{h}_{\mathrm{PI}}$, $n = 500$; (f) estimate with $\mathbf{h}_{\mathrm{ROT}}$, $n = 500$

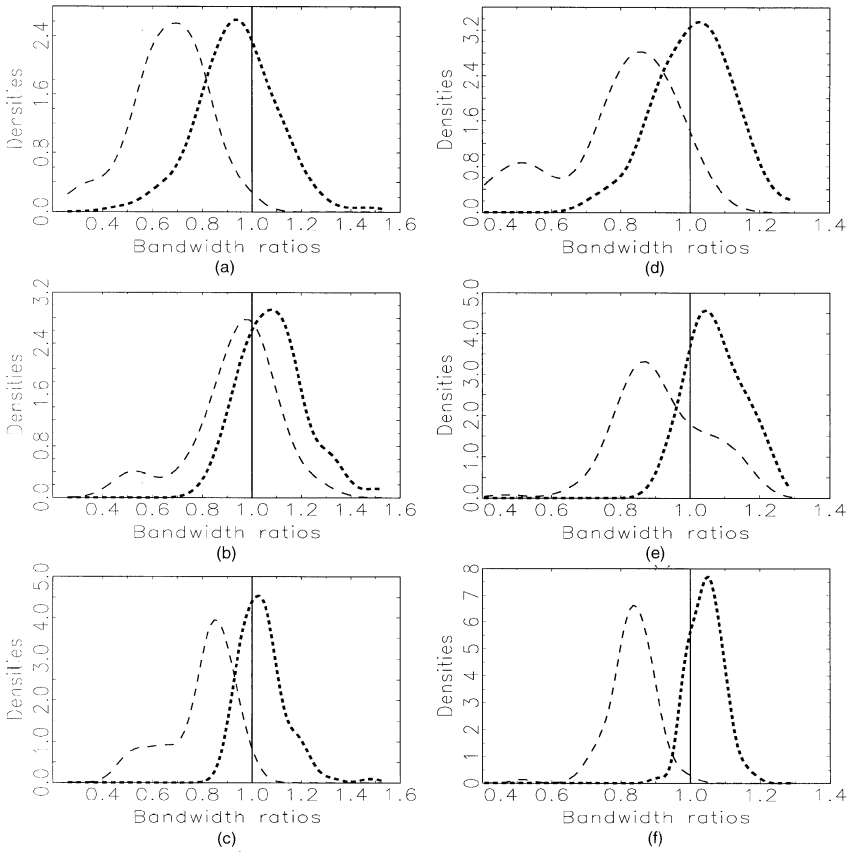

Fig. 4. Densities of bandwidth ratios for model 3 (- - - - -, PI; — — —, ROT): (a) first elements of $\mathbf{h}_{PI}/\mathbf{h}_{opt}$ and $\mathbf{h}_{ROT}/\mathbf{h}_{opt}$, $n = 100$; (b) second elements of $\mathbf{h}_{PI}/\mathbf{h}_{opt}$ and $\mathbf{h}_{ROT}/\mathbf{h}_{opt}$, $n = 100$; (c) h_{PI}/h_{opt} and h_{ROT}/h_{opt}, $n = 100$; (d) first elements of $\mathbf{h}_{PI}/\mathbf{h}_{opt}$ and $\mathbf{h}_{ROT}/\mathbf{h}_{opt}$, $n = 500$; (e) second elements of $\mathbf{h}_{PI}/\mathbf{h}_{opt}$ and $\mathbf{h}_{ROT}/\mathbf{h}_{opt}$, $n = 500$; (f) h_{PI}/h_{opt} and h_{ROT}/h_{opt}, $n = 500$

than the PI estimates for complex models that are far from being piecewise quartic. This is confirmed by Tables 1–4 and the ISE densities in Figs 2(a)–2(d). Fig. 3 shows the true function of model 3 and its estimates for one replication based on $\mathbf{h}_{PI}$ and $\mathbf{h}_{ROT}$, $n = 100$, or $\mathbf{h}_{opt}$, $\mathbf{h}_{PI}$ and $\mathbf{h}_{ROT}$, $n = 500$.

For higher dimension, the advantage of PI over ROT estimates becomes more significant. For the three-dimensional model 5 and four-dimensional model 6, Tables 5 and 6 show a reduction in MISE up to 60%. The ISE densities are plotted in Figs 2(e) and 2(f).

For models 1 and 2, there is practically no difference between the diagonal and scalar bandwidth, whereas for model 3 the diagonal has only a slight advantage. This is due to the form of f, which does not require different amounts of smoothing in each direction. For models 4–6, the improvement in diagonal bandwidth becomes significant as the f-function requires very different amounts of smoothing for various directions. Overall, we clearly prefer the diagonal to the scalar bandwidth.

Although the function estimate is the ultimate goal, it is also informative to analyse how well the PI and ROT bandwidths approximate the asymptotically optimal bandwidth $\mathbf{h}_{opt}$. When comparing the densities of ratios, we find that $\mathbf{h}_{PI}/\mathbf{h}_{opt}$ is always to the right of $\mathbf{h}_{ROT}/\mathbf{h}_{opt}$. The

ratio for the PI bandwidth is always closer to 1 except for a few cases where the ROT ratio is also close to 1. The PI ratio $\mathbf{h}_{\mathrm{PI}}/\mathbf{h}_{\mathrm{opt}}$ also approaches 1 faster than the ROT ratio $\mathbf{h}_{\mathrm{ROT}}/\mathbf{h}_{\mathrm{opt}}$ as the sample size n increases. All these conclusions can be drawn from inspecting the density plots. Fig. 4 presents the density plots for the complicated model 3 based on 100 (Figs 4(a)–4(c)) and 500 (Figs 4(d)–4(f)) observations. The densities of the first and second elements of $\mathbf{h}_{\mathrm{PI}}/\mathbf{h}_{\mathrm{opt}}$ and $\mathbf{h}_{\mathrm{ROT}}/\mathbf{h}_{\mathrm{opt}}$ are depicted in Figs 4(a), 4(b), 4(d) and 4(e). Figs 4(c) and 4(f) show the densities of $h_{\mathrm{PI}}/h_{\mathrm{opt}}$ and $h_{\mathrm{ROT}}/h_{\mathrm{opt}}$. Increasing complexity or dimension makes the bandwidth estimation more difficult.

7. Conclusion

We have shown the existence of both a diagonal and a scalar optimal bandwidth for multivariate regression and have proposed both an ROT and a PI bandwidth selector. The ROT method is simple to use but may perform poorly because of its inconsistency. The PI method is achieved by using partial local cubic regression to estimate the unknown functionals of second derivatives. The partial local cubic expansion does not need many terms and has simple bias and variance formulae. The pilot bandwidths for the functional estimation are determined by blocked partial quartic fits. In Monte Carlo simulations we investigated the performance of these automatic bandwidth selectors for models up to four dimensions and various sample sizes. The diagonal PI method is found to be the best in terms of the ISE, although the ROT method can be useful if the model is not too complicated. Furthermore, a bandwidth vector is always preferable to a scalar bandwidth. These conclusions are observed with a sample size as small as 100.

The results of this paper can be generalized to autoregressive time series under conditions that lead to geometric mixing; see, for instance, Härdle *et al*. (1998) for such conditions. The scalar plug-in bandwidth was used to obtain a data-driven asymptotic final prediction error for the non-linear time series lag selection developed by Tschernig and Yang (1999).

Acknowledgements

The authors thank Rong Chen and Michael Neumann for their many helpful discussions. The very insightful comments from the referees and the Joint Editor are also gratefully acknowledged. These comments stimulated a substantial extension of the method. This work was supported by Sonderforschungsbereich 373 'Quantifikation und Simulation Ökonomischer Prozesse' of the Deutsche Forschungsgemeinschaft, at Humboldt-Universität zu Berlin.

Appendix A

We demonstrate that a scalar bandwidth h can be used instead of a bandwidth vector $\mathbf{h} = (h_1, \ldots, h_d)$ for estimating $f(\mathbf{x})$. Estimation based on a single bandwidth h achieves the same convergence rate as multiple bandwidths but may be less efficient when in each variable direction a different amount of smoothing is appropriate. However, a single bandwidth is more easily computed.

When using a single bandwidth h to estimate the function $f(\mathbf{x})$, the AMISE is the following function of h:

$$\mathrm{AMISE}(h) = \frac{\sigma_K^4}{4} C(f) h^4 + \|K\|_2^{2d} B(g) \frac{1}{nh^d}$$

where $C(f) = \Sigma_{\lambda,\mu=1}^{d} C_{\lambda\mu}(f)$. The scalar bandwidth h that minimizes AMISE(h) is

$$h_{\text{opt}} = \left\{ \frac{d\|K\|_2^{2d} B(g)}{4n\sigma_K^4 C(f)} \right\}^{1/(d+4)}. \tag{A.1}$$

Denote $\hat{C}(f, h) = \Sigma_{\lambda,\mu=1}^{d} \hat{C}_{\lambda\mu}(f, h)$; then we have the following theorem.

Theorem 5. Under assumptions (1)–(4) in Appendix B, as $h \to 0$ and $nh^{d+4} \to \infty$, we have

$$\hat{C}(f, h) = C(f) + 2h^2 \sum_{\lambda,\mu=1}^{d} B_{\lambda\mu}(f) + \frac{B(g) \sum_{\lambda,\mu=1}^{d} V_{\lambda\mu}(K)}{nh^{d+4}} + \zeta + O_p(h^4) + o_p\left(\frac{1}{n\sqrt{h^{d+8}}}\right)$$

in which

$$n\sqrt{h^{d+8}}\zeta \xrightarrow{\text{D}} N\{0, \sigma^2(g, K)\}$$

where

$$\sigma^2(g, K) = \frac{32 \int g^4(\mathbf{x}) w^2(\mathbf{x}) \, d\mathbf{x}}{\sigma_K^{16}(\kappa - 1)^4} \int \left\{ \sum_{\lambda,\mu=1}^{d} F_{\lambda\mu}(\mathbf{t}) \right\}^2 d\mathbf{t}.$$

The optimal bandwidth to estimate $C(f)$ by $\hat{C}(f, h)$ is

$$h_{C(f),\text{opt}} = \begin{cases} \left\{ -\dfrac{B(g) \sum_{\lambda,\mu=1}^{d} V_{\lambda\mu}(K)}{2 \sum_{\lambda,\mu=1}^{d} B_{\lambda\mu}(f) n} \right\}^{1/(d+6)} & \text{if } \sum_{\lambda,\mu=1}^{d} B_{\lambda\mu}(f) < 0, \\ \left\{ (d+4) \dfrac{B(g) \sum_{\lambda,\mu=1}^{d} V_{\lambda\mu}(K)}{4 \sum_{\lambda,\mu=1}^{d} B_{\lambda\mu}(f) n} \right\}^{1/(d+6)} & \text{if } \sum_{\lambda,\mu=1}^{d} B_{\lambda\mu}(f) > 0. \end{cases}$$

By using the formulae in theorem 5, we can obtain both the ROT and the PI bandwidth

$$h_{\text{ROT}} = \left\{ \frac{d\|K\|_2^{2d} B_{\text{ROT}}(g)}{4n\sigma_K^4 \sum_{\lambda,\mu=1}^{d} C_{\text{ROT},\lambda\mu}(f)} \right\}^{1/(d+4)}, \tag{A.2}$$

$$h_{\text{PI}} = \left\{ \frac{d\|K\|_2^{2d} \hat{B}(g)}{4n\sigma_K^4 \hat{C}(f)} \right\}^{1/(d+4)} \tag{A.3}$$

in which $B_{\text{ROT}}(g)$ and $C_{\text{ROT},\lambda\mu}(f)$ are the same as used in equation (3.1). It is easy to verify that $h_{\text{PI}} = h_{\text{opt}}\{1 + O_p(n^{-2/(d+6)})\}$.

Appendix B

We denote by S the support of $w(\mathbf{x})$, and all integrals in variable $\mathbf{x}$ are over S. The following assumptions are needed to prove theorems 1, 3 and 4.

(a) The density $p(\mathbf{x})$ of $\mathbf{X}$ exists, is continuously differentiable up to order 2 and is bounded below away from 0 on S (assumption 1).
(b) The function $f(\mathbf{x})$ is continuously differentiable on S up to order 4 whereas $g(\mathbf{x})$ is continuous on S and positive on an open subset of S (assumption 2).
(c) The bandwidth $h = h_n$ is a positive number depending on n such that $h \to 0$ and $nh^{d+4} \to \infty$ (assumption 3).

Denote by $M_+(d)$ the set of all non-negative definite matrices $\mathbf{M}$ that are positive definite for non-negative vectors, i.e.

$$M_+(d) = \left\{ \mathbf{M}\text{: } \mathbf{M} \text{ non-negative definite and } \inf_{\mathbf{h} \in S_+^{d-1}} (\mathbf{h}^{\mathrm{T}} \mathbf{M} \mathbf{h}) = c(\mathbf{M}) > 0 \right\} \tag{B.1}$$

where $S_+^{d-1} = S^{d-1} \cap \bar{R}_+^d$ denotes the portion of the $(d-1)$-dimensional unit sphere S^{d-1} with non-negative co-ordinates. For the diagonal bandwidth to exist, we assume that the matrix $\mathbf{C}(f) \in M_+(d)$ (assumption 4).

Assumption 4 is not trivial. Take the function $f(x_1, x_2) = \exp(x_1)\ \sin(\pi x_2)$; together with $p(\mathbf{x})$ equal to the uniform density on $[0, 1]^2$ and $w(\mathbf{x}) = p(\mathbf{x})$, we have

$$\mathbf{C}(f) = \int_{[0,1]^2} \exp(2x_1)\ \sin^2(\pi x_2)\, \mathrm{d}\mathbf{x} \begin{pmatrix} 1 & -\pi^2 \\ -\pi^2 & \pi^4 \end{pmatrix} = \frac{1}{4} \{\exp(2) - 1\} \begin{pmatrix} 1 & -\pi^2 \\ -\pi^2 & \pi^4 \end{pmatrix}$$

and therefore $(\pi^2, 1)\, \mathbf{C}(f) (\pi^2, 1)^{\mathrm{T}} = 0$ even though $(\pi^2, 1)^{\mathrm{T}} \in R_+^2$.

For the proof of theorem 1, we also need lemmas 1 and 2 and theorem 6.

Lemma 1.

$$\frac{\partial Q(\mathbf{v}, \mathbf{M}, a)}{\partial \mathbf{v}} = \frac{1}{2} \begin{pmatrix} \sum_{\lambda=1}^{d} M_{1\lambda} v_\lambda \\ \vdots \\ \sum_{\lambda=1}^{d} M_{d\lambda} v_\lambda \end{pmatrix} - \frac{a}{2\sqrt{(v_1 \ldots v_d)}} \begin{pmatrix} v_1^{-1} \\ \vdots \\ v_d^{-1} \end{pmatrix}, \tag{B.2}$$

$$\frac{\partial^2 Q(\mathbf{v}, \mathbf{M}, a)}{\partial \mathbf{v} \partial \mathbf{v}^{\mathrm{T}}} = \frac{1}{2} \mathbf{M} + \frac{a}{4\sqrt{(v_1 \ldots v_d)}} \begin{pmatrix} 3v_1^{-2} & v_1^{-1} v_2^{-1} & \ldots & v_1^{-1} v_d^{-1} \\ v_2^{-1} v_1^{-1} & 3v_2^{-2} & \ldots & v_2^{-1} v_d^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ v_d^{-1} v_1^{-1} & v_d^{-1} v_2^{-1} & \ldots & 3v_d^{-2} \end{pmatrix}. \tag{B.3}$$

Proof. Direct matrix calculation plus definition (2.1) yield the results.

Lemma 2. For any fixed $\mathbf{M} \in M_+(d)$ and $a > 0$, $Q(\mathbf{v}, \mathbf{M}, a)$ is a strictly convex function of $\mathbf{v} \in R_+^d$ and therefore can be minimized by at most one vector $\mathbf{v}$ in R_+^d.

Proof. Both terms in equation (B.3) are non-negative definite, whereas the second term is positive definite, so the Hessian of $Q(\mathbf{v}, \mathbf{M}, a)$ with respect to $\mathbf{v}$ is positive definite.

Theorem 6. For any fixed $\mathbf{M} \in M_+(d)$ and $a > 0$, $Q(\mathbf{v}, \mathbf{M}, a)$ has a unique minimum vector $\mathbf{v}_{\mathrm{opt}} = \mathbf{v}(\mathbf{M}, a)$, which is infinitely continuously differentiable in both $\mathbf{M}$ and a.

Proof. $\mathbf{M} \in M_+(d)$ implies that

$$Q(\mathbf{v}, \mathbf{M}, a) \geqslant \frac{1}{4} \sum_{\lambda,\mu=1}^{d} M_{\lambda\mu} v_\lambda v_\mu \geqslant \frac{1}{4}\, c(\mathbf{M}) \sum_{\lambda=1}^{d} v_\lambda^2 = \frac{1}{4}\, c(\mathbf{M}) \|\mathbf{v}\|^2$$

for all $\mathbf{v} \in R_+^d$ and Euclidean norm $\|\mathbf{v}\| = (\Sigma_{\lambda=1}^d v_\lambda^2)^{1/2}$, so

$$\lim_{\|\mathbf{v}\| \to \infty} \{Q(\mathbf{v}, \mathbf{M}, a)\} \geqslant \frac{1}{4} \lim_{\|\mathbf{v}\| \to \infty} \{c(\mathbf{M}) \|\mathbf{v}\|^2\} = \infty$$

while it is also clear that we always have

$$\lim_{\|\mathbf{v}\| \to 0} \{Q(\mathbf{v}, \mathbf{M}, a)\} \geqslant \frac{1}{4} \lim_{\|\mathbf{v}\| \to 0} \left\{ \frac{a}{\sqrt{(v_1 \ldots v_d)}} \right\} = \infty.$$

Thus the infimum of $Q(\mathbf{v}, \mathbf{M}, a)$ is reached at a $\mathbf{v}$ whose norm is bounded away from ∞ and 0, and by the previous lemma this $\mathbf{v}$ is unique. Equation (B.2) and the implicit function theorem show that $\mathbf{v}(\mathbf{M}, a)$ is an infinitely continuously differentiable function of $\mathbf{M}$ and a.

To prove theorem 3, we need the dispersion matrix of the partial local cubic estimator.

Lemma 3. Let Z_λ be as in equation (4.1) and

$$H = \operatorname{diag}(1, h^{-2}I_{2d-1}, h^{-1}I_d, h^{-3}I_{2d-2}, h^{-3}).$$

As $n \to \infty$,

$$H^{\mathrm{T}} Z_\lambda^{\mathrm{T}} W Z_\lambda H = \operatorname{diag}\left[\begin{pmatrix} 1 & \sigma_K^2 \mathbf{1}_{1\times d} \\ \sigma_K^2 \mathbf{1}_{d\times 1} & \sigma_K^4\{(\kappa-1)I_d + \mathbf{1}_{d\times d}\} \end{pmatrix}, \sigma_K^4 I_{d-1}, D\right]\{p(\mathbf{x})I_{5d-1} + o_p(1)\} \tag{B.4}$$

and

$$(H^{\mathrm{T}} Z_\lambda^{\mathrm{T}} W Z_\lambda H)^{-1} = \operatorname{diag}\left\{\begin{pmatrix} (\kappa-1+d)(\kappa-1)^{-1} & -\sigma_K^{-2}(\kappa-1)^{-1}\mathbf{1}_{1\times d} \\ -\sigma_K^{-2}(\kappa-1)^{-1}\mathbf{1}_{d\times 1} & \sigma_K^{-4}(\kappa-1)^{-1}I_d \end{pmatrix}, \sigma_K^{-4} I_{d-1}, D^{-1}\right\}\left\{\frac{I_{5d-1}}{p(\mathbf{x})} + o_p(1)\right\} \tag{B.5}$$

uniformly in a compact neighbourhood of x. Here

$$D = \begin{pmatrix} \sigma_K^2 I_d & A & B & C \\ A^{\mathrm{T}} & \kappa\sigma_K^6 I_{d-1} & \mathbf{0}_{(d-1)\times(d-1)} & \mathbf{0}_{(d-1)\times 1} \\ B^{\mathrm{T}} & \mathbf{0}_{(d-1)\times(d-1)} & \sigma_K^6\{(\kappa-1)I_{d-1} + \mathbf{1}_{(d-1)\times(d-1)}\} & \kappa\sigma_K^6 \mathbf{1}_{(d-1)\times 1} \\ C^{\mathrm{T}} & \mathbf{0}_{1\times(d-1)} & \kappa\sigma_K^6 \mathbf{1}_{1\times(d-1)} & \mu_6(K) \end{pmatrix}$$

in which

$$A = \sigma_K^4 \begin{pmatrix} I_{\lambda-1} & \mathbf{0}_{(\lambda-1)\times(d-\lambda)} \\ \mathbf{0}_{1\times(\lambda-1)} & \mathbf{0}_{1\times(d-\lambda)} \\ \mathbf{0}_{(d-\lambda)\times(\lambda-1)} & I_{d-\lambda} \end{pmatrix}, \qquad B = \sigma_K^4 \begin{pmatrix} \mathbf{0}_{(\lambda-1)\times(d-1)} \\ \mathbf{1}_{1\times(d-1)} \\ \mathbf{0}_{(d-\lambda)\times(d-1)} \end{pmatrix}, \qquad C = \kappa\sigma_K^4 \begin{pmatrix} \mathbf{0}_{(\lambda-1)\times 1} \\ 1 \\ \mathbf{0}_{(d-\lambda)\times 1} \end{pmatrix}.$$

Proof. Equation (B.4) follows by the usual array-type central limit theorem, K being a symmetric compact product kernel; see, for example, Wand and Jones (1995). Equation (B.5) follows by direct verification. The exact inverse of D can be obtained by using the tools in Lütkepohl (1996).

Proof of theorem 3. Given any $\lambda = 1, \ldots, d$ and denoting Z_λ by Z, we first note that

$$e_\lambda^{\mathrm{T}} H(H^{\mathrm{T}} Z^{\mathrm{T}} W Z H)^{-1} H^{\mathrm{T}} Z^{\mathrm{T}} W Z e_0 f(\mathbf{x}) = e_\lambda^{\mathrm{T}}(Z^{\mathrm{T}} W Z)^{-1} Z^{\mathrm{T}} W Z e_0 f(\mathbf{x}) = 0,$$
$$e_\lambda^{\mathrm{T}} H(H^{\mathrm{T}} Z^{\mathrm{T}} W Z H)^{-1} H^{\mathrm{T}} Z^{\mathrm{T}} W Z e_\lambda f_{\lambda\lambda}(\mathbf{x}) = f_{\lambda\lambda}(\mathbf{x}),$$

whereas for any $\lambda' = 1, \ldots, 5d-1$, $\lambda' \neq \lambda$,

$$e_\lambda^{\mathrm{T}} H(H^{\mathrm{T}} Z^{\mathrm{T}} W Z H)^{-1} H^{\mathrm{T}} Z^{\mathrm{T}} W Z e_{\lambda'} = e_\lambda^{\mathrm{T}}(Z^{\mathrm{T}} W Z)^{-1} Z^{\mathrm{T}} W Z e_{\lambda'} = 0.$$

Combining these with equation (B.5) of lemma 3, we have

$$\begin{aligned} \hat{f}_{\lambda\lambda}(\mathbf{x}) - f_{\lambda\lambda}(\mathbf{x}) &= 2e_\lambda^{\mathrm{T}} H(H^{\mathrm{T}} Z^{\mathrm{T}} W Z H)^{-1} H^{\mathrm{T}} Z^{\mathrm{T}} W Y - f_{\lambda\lambda}(\mathbf{x}) \\ &= T_1 + T_2 + T_3 \end{aligned}$$

where

$$\begin{aligned} T_1 = \frac{2 + o_p(1)}{\sigma_K^4(\kappa-1)\,p(\mathbf{x}) n h^2} \sum_{i=1}^n K_h(\mathbf{X}_i - \mathbf{x}) \left\{\frac{(X_{i\lambda} - x_\lambda)^2}{h^2} - \sigma_K^2\right\} \Big\{ & f(\mathbf{X}_i) - f(\mathbf{x}) - (\mathbf{X}_i - \mathbf{x})^{\mathrm{T}} \nabla f(\mathbf{x}) \\ & - \frac{1}{2}(\mathbf{X}_i - \mathbf{x})^{\mathrm{T}} \nabla^2 f(\mathbf{x})(\mathbf{X}_i - \mathbf{x}) - \frac{1}{3!} f_{\lambda\lambda\lambda}(\mathbf{x})(X_{i\lambda} - x_\lambda)^3 - \sum_{\alpha\neq\lambda} \frac{3}{3!} f_{\alpha\lambda\lambda}(\mathbf{x})(X_{i\lambda} - x_\lambda)^2 (X_{i\alpha} - x_\alpha) \\ & - \sum_{\alpha\neq\lambda} \frac{3}{3!} f_{\alpha\alpha\lambda}(\mathbf{x})(X_{i\lambda} - x_\lambda)(X_{i\alpha} - x_\alpha)^2 \Big\}, \end{aligned}$$

$$T_2 = -\frac{2+o_p(1)}{\sigma_K^4(\kappa-1)\,p(\mathbf{x})nh^2}\sum_{i=1}^{n} K_h(\mathbf{X}_i-\mathbf{x})\left\{\frac{(X_{i\lambda}-x_\lambda)^2}{h^2}-\sigma_K^2\right\}\left\{\sum_{\alpha<\beta,\alpha\neq\lambda\neq\beta}(X_{i\alpha}-x_\alpha)(X_{i\beta}-x_\beta)\,f_{\alpha\beta}(\mathbf{x})\right\},$$

$$T_3 = \frac{2+o_p(1)}{\sigma_K^4(\kappa-1)\,p(\mathbf{x})nh^2}\sum_{i=1}^{n} K_h(\mathbf{X}_i-\mathbf{x})\left\{\frac{(X_{i\lambda}-x_\lambda)^2}{h^2}-\sigma_K^2\right\} g(\mathbf{X}_i)\,\epsilon_i.$$

The asymptotic variance of T_3 is calculated as

$$\frac{4+o(1)}{\sigma_K^8(\kappa-1)^2\,p(\mathbf{x})^2nh^4}\int K_h(\mathbf{u}-\mathbf{x})^2\left\{\frac{(u_\lambda-x_\lambda)^2}{h^2}-\sigma_K^2\right\}^2 g^2(\mathbf{u})\,p(\mathbf{u})\,\mathrm{d}\mathbf{u}$$

which (using $\mathbf{u}=\mathbf{x}+h\mathbf{v}$)

$$\begin{aligned}
&= \frac{4+o(1)}{\sigma_K^8(\kappa-1)^2\,p(\mathbf{x})^2nh^{d+4}}\int K(\mathbf{v})^2(v_\lambda^2-\sigma_K^2)^2\,g^2(\mathbf{x}+h\mathbf{v})\,p(\mathbf{x}+h\mathbf{v})\,\mathrm{d}\mathbf{v}\\
&= \frac{4\{\mu_4(K^2)-2\,\mu_2(K^2)\sigma_K^2+\|K\|_2^2\sigma_K^4\}\,\|K\|_2^{2(d-1)}\,g^2(\mathbf{x})}{\sigma_K^8(\kappa-1)^2\,p(\mathbf{x})nh^{d+4}}\{1+o(1)\}=\sigma_{\lambda\lambda}^2(\mathbf{x})+o(h^2) \qquad \text{(B.6)}
\end{aligned}$$

with $\sigma_{\lambda\lambda}^2(\mathbf{x})$ as given in equation (4.4).

A similar procedure applied to T_1 yields

$$\begin{aligned}
T_1 = {}& \frac{2+o_p(1)}{\sigma_K^4(\kappa-1)\,p(\mathbf{x})h^2}\int K_h(\mathbf{u}-\mathbf{x})\left\{\frac{(u_\lambda-x_\lambda)^2}{h^2}-\sigma_K^2\right\}\Big\{f(\mathbf{u})-f(\mathbf{x})-(\mathbf{u}-\mathbf{x})^{\mathrm{T}}\,\nabla f(\mathbf{x})\\
&-\frac{1}{2}(\mathbf{u}-\mathbf{x})^{\mathrm{T}}\,\nabla^2 f(\mathbf{x})(\mathbf{u}-\mathbf{x})-\frac{1}{3!}f_{\lambda\lambda\lambda}(\mathbf{x})(u_\lambda-x_\lambda)^3-\sum_{\alpha\neq\lambda}\frac{3}{3!}f_{\alpha\lambda\lambda}(\mathbf{x})(u_\alpha-x_\alpha)(u_\lambda-x_\lambda)^2\\
&-\sum_{\alpha\neq\lambda}\frac{3}{3!}f_{\alpha\alpha\lambda}(\mathbf{x})(u_\alpha-x_\alpha)^2(u_\lambda-x_\lambda)\Big\}\,p(\mathbf{u})\,\mathrm{d}\mathbf{u}\\
= {}& \frac{2+o_p(1)}{\sigma_K^4(\kappa-1)\,p(\mathbf{x})h^2}\int K(\mathbf{v})(v_\lambda^2-\sigma_K^2)\Big\{f(\mathbf{x}+h\mathbf{v})-f(\mathbf{x})-h\mathbf{v}^{\mathrm{T}}\,\nabla f(\mathbf{x})-\frac{h^2}{2}\mathbf{v}^{\mathrm{T}}\,\nabla^2 f(\mathbf{x})\mathbf{v}-\frac{h^3}{3!}f_{\lambda\lambda\lambda}(\mathbf{x})v_\lambda^3\\
&-\sum_{\alpha\neq\lambda}\frac{h^3}{2}f_{\alpha\lambda\lambda}(\mathbf{x})v_\alpha v_\lambda^2-\sum_{\alpha\neq\lambda}\frac{h^3}{2}f_{\alpha\alpha\lambda}(\mathbf{x})v_\alpha^2 v_\lambda\Big\}\,p(\mathbf{x}+h\mathbf{v})\,\mathrm{d}\mathbf{v}\\
= {}& \frac{h^2\int K(\mathbf{v})(v_\lambda^2-\sigma_K^2)v_\alpha^2v_\beta^2\,\mathrm{d}\mathbf{v}}{\sigma_K^4(\kappa-1)\,p(\mathbf{x})}\left[\sum_{\alpha<\beta,\alpha\neq\lambda\neq\beta}\{f_{\alpha\alpha\beta}(\mathbf{x})\,p_\beta(\mathbf{x})+f_{\alpha\beta\beta}(\mathbf{x})\,p_\alpha(\mathbf{x})\}+\sum_{\alpha<\beta}\frac{f_{\alpha\alpha\beta\beta}(\mathbf{x})\,p(\mathbf{x})}{2}\right]\\
&+\frac{h^2\int K(\mathbf{v})(v_\lambda^2-\sigma_K^2)v_\gamma^4\,\mathrm{d}\mathbf{v}}{\sigma_K^4(\kappa-1)\,p(\mathbf{x})}\left\{\sum_{\gamma\neq\lambda}\frac{f_{\gamma\gamma\gamma}(\mathbf{x})\,p_\gamma(\mathbf{x})}{3}+\sum_{\gamma=1}^{d}\frac{f_{\gamma\gamma\gamma\gamma}(\mathbf{x})\,p(\mathbf{x})}{12}\right\}+o(h^2).
\end{aligned}$$

We note then that

$$\int K(\mathbf{v})(v_\lambda^2-\sigma_K^2)v_\alpha^2v_\beta^2\,\mathrm{d}\mathbf{v}=\begin{cases}0 & a<\beta,\ \alpha\neq\lambda\neq\beta,\\ (\kappa-1)\sigma_K^6 & \beta\neq\lambda=\alpha,\end{cases} \qquad \text{(B.7)}$$

$$\int K(\mathbf{v})(v_\lambda^2-\sigma_K^2)v_\gamma^4\,\mathrm{d}\mathbf{v}=\begin{cases}0 & \gamma\neq\lambda,\\ \mu_6(K)-\kappa\sigma_K^6 & \gamma=\lambda,\end{cases}$$

so

$$T_1 = \frac{h^2(\kappa-1)\,\sigma_K^6}{\sigma_K^4(\kappa-1)\,p(\mathbf{x})} \sum_{\alpha\neq\lambda} \frac{f_{\alpha\alpha\lambda\lambda}(\mathbf{x})\,p(\mathbf{x})}{2} + \frac{h^2\{\mu_6(K)-\kappa\sigma_K^6\}}{\sigma_K^4(\kappa-1)\,p(\mathbf{x})} \frac{f_{\lambda\lambda\lambda\lambda}(\mathbf{x})\,p(\mathbf{x})}{12} + o(h^2)$$

or

$$T_1 = b_{\lambda\lambda}(\mathbf{x})h^2 + o(h^2), \tag{B.8}$$

where $b_{\lambda\lambda}(\mathbf{x})$ is as given in equation (4.3).

The term T_2 is treated as

$$\begin{aligned} T_2 &= -\frac{2+o_p(1)}{\sigma_K^4(\kappa-1)\,p(\mathbf{x})n} \sum_{i=1}^{n} K_h(\mathbf{X}_i-\mathbf{x})\left\{\frac{(X_{i\lambda}-x_\lambda)^2}{h^2}-\sigma_K^2\right\} \sum_{\alpha<\beta,\alpha\neq\lambda\neq\beta} \frac{X_{i\alpha}-x_\alpha}{h}\,\frac{X_{i\beta}-x_\beta}{h}\,f_{\alpha\beta}(\mathbf{x}) \\ &= -\sum_{\alpha<\beta,\alpha\neq\lambda\neq\beta} \frac{2f_{\alpha\beta}(\mathbf{x})\{1+o_p(1)\}}{\sigma_K^4(\kappa-1)\,p(\mathbf{x})} \int K_h(\mathbf{u}-\mathbf{x})\left\{\frac{(u_\lambda-x_\lambda)^2}{h^2}-\sigma_K^2\right\}\frac{u_\alpha-x_\alpha}{h}\,\frac{u_\beta-x_\beta}{h}\,p(\mathbf{u})\,\mathrm{d}\mathbf{u} \\ &= -\sum_{\alpha<\beta,\alpha\neq\lambda\neq\beta} \frac{2f_{\alpha\beta}(\mathbf{x})}{\sigma_K^4(\kappa-1)\,p(\mathbf{x})} \int K(\mathbf{v})(v_\lambda^2-\sigma_K^2)v_\alpha v_\beta\,p(\mathbf{x}+h\mathbf{v})\,\mathrm{d}\mathbf{v}\{1+o_p(1)\} \\ &= -2h^2 \sum_{\alpha<\beta,\alpha\neq\lambda\neq\beta} \frac{\int K(\mathbf{v})(v_\lambda^2-\sigma_K^2)v_\alpha^2v_\beta^2\,\mathrm{d}\mathbf{v}\,f_{\alpha\beta}(\mathbf{x})\,p_{\alpha\beta}(\mathbf{x})}{\sigma_K^4(\kappa-1)\,p(\mathbf{x})}\{1+o_p(1)\} = o(h^2) \end{aligned} \tag{B.9}$$

since $\int K(\mathbf{v})(v_\lambda^2-\sigma_K^2)v_\alpha^2v_\beta^2\,\mathrm{d}\mathbf{v}=0$ for $\alpha<\beta$, $\alpha\neq\lambda\neq\beta$, by equation (B.7). Now equations (B.8), (B.9) and (B.6) together prove theorem 3.

Proof of theorem 4. Since

$$\hat{f}_{\lambda\lambda}(\mathbf{x}) = f_{\lambda\lambda}(\mathbf{x}) + \{b_{\lambda\lambda}(\mathbf{x})h^2 + T_{3,\lambda\lambda}\}\left[1+o_p\left\{h^2+\frac{1}{\sqrt{(nh^{d+4})}}\right\}\right]$$

where

$$T_{3,\lambda\lambda} = \frac{2}{\sigma_K^4(\kappa-1)\,p(\mathbf{x})nh^2} \sum_{i=1}^{n} K_h(\mathbf{X}_i-\mathbf{x})\left\{\frac{(X_{i\lambda}-x_\lambda)^2}{h^2}-\sigma_K^2\right\} g(\mathbf{X}_i)\epsilon_i$$

we have

$$\begin{aligned} \hat{C}_{\lambda\mu}(f,h) &= \int \hat{f}_{\lambda\lambda}(\mathbf{x})\,\hat{f}_{\mu\mu}(\mathbf{x})\,w(\mathbf{x})\,p(\mathbf{x})\,\mathrm{d}\mathbf{x} + O_p(n^{-1/2}) \\ &= C_{\lambda\mu}(f) + h^2\{B_{\lambda\mu}(f)+B_{\mu\lambda}(f)\} + \int T_{3,\lambda\lambda}T_{3,\mu\mu}\,w(\mathbf{x})\,p(\mathbf{x})\,\mathrm{d}\mathbf{x} + o_p\left\{h^4+\frac{h^2}{\sqrt{(nh^{d+4})}}\right\}. \end{aligned} \tag{B.10}$$

Thus it remains to compute the term $\int T_{3,\lambda\lambda}T_{3,\mu\mu}\,p(\mathbf{x})\,\mathrm{d}\mathbf{x}$, which has a simple decomposition as

$$\sum_{1\leqslant i<j\leqslant n} 2\,H_{\lambda\mu}(\mathbf{X}_i,\mathbf{X}_j)\epsilon_i\epsilon_j + \sum_{1\leqslant i\leqslant n} H_{\lambda\mu}(\mathbf{X}_i,\mathbf{X}_i)\epsilon_i^2$$

where

$$H_{\lambda\mu}(\mathbf{X}_i,\mathbf{X}_j) = \int \frac{4\,K_h(\mathbf{X}_i-\mathbf{x})\,K_h(\mathbf{X}_j-\mathbf{x})\,w(\mathbf{x})}{\sigma_K^8(\kappa-1)^2\,p(\mathbf{x})n^2h^4} \left\{\frac{(X_{i\lambda}-x_\lambda)^2}{h^2}-\sigma_K^2\right\}\left\{\frac{(X_{j\mu}-x_\mu)^2}{h^2}-\sigma_K^2\right\} g(\mathbf{X}_i)\,g(\mathbf{X}_j)\,\mathrm{d}\mathbf{x}.$$

Note first that

$$E\{H_{\lambda\mu}(\mathbf{X}_1, \mathbf{X}_1)\epsilon_1^2\} = E\left[\int \frac{4\, K_h^2(\mathbf{X}_1 - \mathbf{x})\, w(\mathbf{x})\, d\mathbf{x}}{\sigma_K^8(\kappa-1)^2\, p(\mathbf{x}) n^2 h^4} \left\{ \frac{(X_{1\lambda} - x_\lambda)^2}{h^2} - \sigma_K^2 \right\} \left\{ \frac{(X_{1\mu} - x_\mu)^2}{h^2} - \sigma_K^2 \right\} g^2(\mathbf{X}_1) \right]$$

$$= \int \int \frac{4\, K_h^2(\mathbf{y} - \mathbf{x})\, w(\mathbf{x})\, d\mathbf{x}}{\sigma_K^8(\kappa-1)^2\, p(\mathbf{x}) n^2 h^4} \left\{ \frac{(y_\lambda - x_\lambda)^2}{h^2} - \sigma_K^2 \right\} \left\{ \frac{(y_\mu - x_\mu)^2}{h^2} - \sigma_K^2 \right\} g^2(\mathbf{y})\, p(\mathbf{y})\, d\mathbf{y}$$

$$\overset{\mathbf{y}-\mathbf{x}=h\mathbf{u}}{=} \int \int \frac{4\, w(\mathbf{x})}{\sigma_K^8(\kappa-1)^2\, p(\mathbf{x}) n^2 h^{d+4}} K^2(\mathbf{u})(u_\lambda^2 - \sigma_K^2)(u_\mu^2 - \sigma_K^2)\, g^2(\mathbf{x} + h\mathbf{u})\, p(\mathbf{x} + h\mathbf{u})\, d\mathbf{x}\, d\mathbf{u}$$

$$= \frac{4 \int g^2(\mathbf{x})\, w(\mathbf{x})\, d\mathbf{x} \int K^2(\mathbf{u})(u_\lambda^2 u_\mu^2 - \sigma_K^2 u_\mu^2 - \sigma_K^2 u_\lambda^2 + \sigma_K^4)\, d\mathbf{u}\{1 + o(1)\}}{\sigma_K^8(\kappa-1)^2 n^2 h^{d+4}}$$

$$= \frac{B(g)\, V_{\lambda\mu}(K)}{n^2 h^{d+4}}\{1 + o(1)\}$$

and similarly that

$$E\{H_{\lambda\mu}(\mathbf{X}_1, \mathbf{X}_1)^2 \epsilon_1^4\} = O\left(\frac{1}{n^4 h^{8+3d}} \right).$$

Hence by an array-type central limit theorem

$$\sum_{1 \leqslant i \leqslant n} H_{\lambda\mu}(\mathbf{X}_i, \mathbf{X}_i)\epsilon_i^2 = \frac{B(g)\, V_{\lambda\mu}(K)}{n h^{d+4}} + O_p\left\{ \frac{1}{\sqrt{(n^3 h^{8+3d})}} \right\}. \tag{B.11}$$

Meanwhile, using a central limit theorem for non-degenerate *U-statistics* as contained in Hall (1984), one can verify that

$$\sum_{1 \leqslant i < j \leqslant n} 2\, H_{\lambda\mu}(\mathbf{X}_i, \mathbf{X}_j)\epsilon_i \epsilon_j$$

is asymptotically normal with mean 0 and asymptotic variance

$$\frac{n^2}{2} E\{4 H_{\lambda\mu}(\mathbf{X}_1, \mathbf{X}_2)^2 \epsilon_1^2 \epsilon_2^2\} = 2n^2\, E\{H_{\lambda\mu}(\mathbf{X}_1, \mathbf{X}_2)^2\}$$

$$= 2n^2\, E\left[\int \frac{4\, K_h(\mathbf{X}_1 - \mathbf{x})\, K_h(\mathbf{X}_2 - \mathbf{x})\, w(\mathbf{x})}{\sigma_K^8(\kappa-1)^2\, p(\mathbf{x}) n^2 h^4} \left\{ \frac{(X_{1\lambda} - x_\lambda)^2}{h^2} - \sigma_K^2 \right\} g(\mathbf{X}_1) \left\{ \frac{(X_{2\mu} - x_\mu)^2}{h^2} - \sigma_K^2 \right\} g(\mathbf{X}_2)\, d\mathbf{x} \right]^2$$

which is

$$2n^2\, E\left[\int \frac{4\, K_h(\mathbf{X}_1 - \mathbf{x})\, K_h(\mathbf{X}_2 - \mathbf{x})\, w(\mathbf{x})}{\sigma_K^8(\kappa-1)^2 n^2 h^4\, p(\mathbf{x})} \left\{ \frac{(X_{1\lambda} - x_\lambda)^2}{h^2} - \sigma_K^2 \right\} \left\{ \frac{(X_{2\mu} - x_\mu)^2}{h^2} - \sigma_K^2 \right\} g(\mathbf{X}_1)\, g(\mathbf{X}_2)\, d\mathbf{x} \right]^2$$

$$= 2n^2 \int \left[\int \frac{4\, K_h(\mathbf{y}_1 - \mathbf{x})\, K_h(\mathbf{y}_2 - \mathbf{x})\, w(\mathbf{x})}{\sigma_K^8(\kappa-1)^2 n^2 h^4\, p(\mathbf{x})} \left\{ \frac{(y_{1\lambda} - x_\lambda)^2}{h^2} - \sigma_K^2 \right\} \left\{ \frac{(y_{2\mu} - x_\mu)^2}{h^2} - \sigma_K^2 \right\} g(\mathbf{y}_1)\, g(\mathbf{y}_2)\, d\mathbf{x} \right]^2$$

$$\times\, p(\mathbf{y}_1)\, p(\mathbf{y}_2)\, d\mathbf{y}_1\, d\mathbf{y}_2$$

$$= \int \frac{32\, K_h(\mathbf{y}_1 - \mathbf{x}_1)\, K_h(\mathbf{y}_2 - \mathbf{x}_1)\, K_h(\mathbf{y}_1 - \mathbf{x}_2)\, K_h(\mathbf{y}_2 - \mathbf{x}_2)\, w(\mathbf{x}_1)\, w(\mathbf{x}_2)}{\sigma_K^{16}(\kappa-1)^4 n^2 h^8\, p(\mathbf{x}_1)\, p(\mathbf{x}_2)} \left\{ \frac{(y_{1\lambda} - x_{1\lambda})^2}{h^2} - \sigma_K^2 \right\}$$

$$\times \left\{ \frac{(y_{2\mu} - x_{1\mu})^2}{h^2} - \sigma_K^2 \right\} \left\{ \frac{(y_{1\lambda} - x_{2\lambda})^2}{h^2} - \sigma_K^2 \right\} \left\{ \frac{(y_{2\mu} - x_{2\mu})^2}{h^2} - \sigma_K^2 \right\}$$

$$\times\, g^2(\mathbf{y}_1)\, g^2(\mathbf{y}_2)\, p(\mathbf{y}_1)\, p(\mathbf{y}_2)\, d\mathbf{x}_1\, d\mathbf{x}_2\, d\mathbf{y}_1\, d\mathbf{y}_2.$$

This, after doing a change of variables $\mathbf{y}_1 - \mathbf{x}_1 = h\mathbf{u}_1$ and $\mathbf{y}_2 - \mathbf{x}_2 = h\mathbf{u}_2$, becomes

$$\frac{32}{\sigma_K^{16}(\kappa-1)^4 n^2 h^8} \int\int \frac{K(\mathbf{u}_1)\,K_h(\mathbf{x}_2-\mathbf{x}_1+h\mathbf{u}_2)\,K_h(\mathbf{x}_1-\mathbf{x}_2+h\mathbf{u}_1)\,K(\mathbf{u}_2)\,w(\mathbf{x}_1)\,w(\mathbf{x}_2)}{p(\mathbf{x}_1)\,p(\mathbf{x}_2)} (u_{1\lambda}^2 - \sigma_K^2)$$

$$\times \left\{ \frac{(x_{2\mu} - x_{1\mu} + hu_{2\mu})^2}{h^2} - \sigma_K^2 \right\} \left\{ \frac{(x_{1\lambda} - x_{2\lambda} + hu_{1\lambda})^2}{h^2} - \sigma_K^2 \right\} (u_{2\mu}^2 - \sigma_K^2)\, g^2(\mathbf{x}_1 + h\mathbf{u}_1)\, g^2(\mathbf{x}_2 + h\mathbf{u}_2)$$

$$\times p(\mathbf{x}_1 + h\mathbf{u}_1)\, p(\mathbf{x}_2 + h\mathbf{u}_2)\, \mathrm{d}\mathbf{x}_1\, \mathrm{d}\mathbf{x}_2\, \mathrm{d}\mathbf{u}_1\, \mathrm{d}\mathbf{u}_2.$$

With another change of variables $\mathbf{x}_2 - \mathbf{x}_1 = h\mathbf{t}_1$, this becomes

$$\frac{32}{\sigma_K^{16}(\kappa-1)^4 n^2 h^{d+8}} \int\int \frac{K(\mathbf{u}_1)\,K(\mathbf{t}_1+\mathbf{u}_2)\,K(-\mathbf{t}_1+\mathbf{u}_1)\,K(\mathbf{u}_2)\,w(\mathbf{x}_1)\,w(\mathbf{x}_1+h\mathbf{t}_1)}{p(\mathbf{x}_1)\,p(\mathbf{x}_1+h\mathbf{t}_1)} (u_{1\lambda}^2 - \sigma_K^2)$$

$$\times \{(t_{1\mu} + u_{2\mu})^2 - \sigma_K^2\} \{(-t_{1\lambda} + u_{1\lambda})^2 - \sigma_K^2\} (u_{2\mu}^2 - \sigma_K^2)\, g^2(\mathbf{x}_1 + h\mathbf{u}_1)\, g^2(\mathbf{x}_1 + h\mathbf{t}_1 + h\mathbf{u}_2)$$

$$\times p(\mathbf{x}_1 + h\mathbf{u}_1)\, p(\mathbf{x}_1 + h\mathbf{t}_1 + h\mathbf{u}_2)\, \mathrm{d}\mathbf{x}_1\, \mathrm{d}\mathbf{t}_1\, \mathrm{d}\mathbf{u}_1\, \mathrm{d}\mathbf{u}_2$$

$$= \frac{32 \int g^4(\mathbf{x}_1)\, w^2(\mathbf{x}_1)\, \mathrm{d}\mathbf{x}_1}{\sigma_K^{16}(\kappa-1)^4 n^2 h^{d+8}} \int K(\mathbf{u}_1)(u_{1\lambda}^2 - \sigma_K^2)\, K(-\mathbf{t}_1 + \mathbf{u}_1)\{(-t_{1\lambda} + u_{1\lambda})^2 - \sigma_K^2\}\, K(\mathbf{u}_2)$$

$$\times (u_{2\mu}^2 - \sigma_K^2)\, K(\mathbf{t}_1 + \mathbf{u}_2)\{(t_{1\mu} + u_{2\mu})^2 - \sigma_K^2\}\, \mathrm{d}\mathbf{t}_1\, \mathrm{d}\mathbf{u}_1\, \mathrm{d}\mathbf{u}_2\, \{1 + o(1)\}$$

$$= \frac{32 \int g^4(\mathbf{x}_1)\, w^2(\mathbf{x}_1)\, \mathrm{d}\mathbf{x}_1}{\sigma_K^{16}(\kappa-1)^4 n^2 h^{d+8}} \int \mathrm{d}\mathbf{t}_1 \left[\int K(\mathbf{u}_1)\, K(\mathbf{u}_1 - \mathbf{t}_1)(u_{1\lambda}^2 - \sigma_K^2)\{(u_{1\lambda} - t_{1\lambda})^2 - \sigma_K^2\}\, \mathrm{d}\mathbf{u}_1 \right]$$

$$\times \left[\int K(\mathbf{u}_2)\, K(\mathbf{u}_2 + \mathbf{t}_1)(u_{2\mu}^2 - \sigma_K^2)\{(u_{2\mu} + t_{2\mu})^2 - \sigma_K^2\}\, \mathrm{d}\mathbf{u}_2 \right] \{1 + o(1)\}$$

$$= \frac{32 \int g^4(\mathbf{x})\, w^2(\mathbf{x})\, \mathrm{d}\mathbf{x}}{\sigma_K^{16}(\kappa-1)^4 n^2 h^{d+8}} \int F_{\lambda\lambda}(\mathbf{t})\, F_{\mu\mu}(\mathbf{t})\, \mathrm{d}\mathbf{t}\, \{1 + o(1)\}$$

where $F_{\lambda\mu}(\mathbf{t})$ is as given in theorem 4. Hence

$$\zeta_{\lambda\mu} = \sum_{1 \leqslant i < j \leqslant n} 2\, H_{\lambda\mu}(\mathbf{X}_i, \mathbf{X}_j)\epsilon_i\epsilon_j$$

has mean 0 and asymptotic variance

$$\frac{32 \int g^4(\mathbf{x})\, w^2(\mathbf{x})\, \mathrm{d}\mathbf{x}}{\sigma_K^{16}(\kappa-1)^4 n^2 h^{d+8}} \int F_{\lambda\lambda}(\mathbf{t})\, F_{\mu\mu}(\mathbf{t})\, \mathrm{d}\mathbf{t}.$$

This, plus equations (B.10) and (B.11), completes the proof of theorem 4.

References

Cheng, M. Y. (1997) A bandwidth selector for local linear density estimators. *Ann. Statist.*, **25**, 1001–1013.

Fan, J. and Gijbels, I. (1995) Adaptive order polynomial fitting: bandwidth robustification and bias reduction. *J. Comput. Graph. Statist.*, **4**, 213–227.

———(1996) *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Grund, B. and Polzehl, J. (1998) Bias corrected bootstrap bandwidth selection. *J. Nonparam. Statist.*, **8**, 97–126.

Hall, P. (1984) Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multiv. Anal.*, **14**, 1–16.

Härdle, W. and Marron, J. S. (1995) Fast and simple smoothing parameter selection. *Comput. Statist. Data Anal.*, **20**, 1–17.

Härdle, W., Tsybakov, A. B. and Yang, L. (1998) Nonparametric vector autoregression. *J. Statist. Planng Inf.*, **68**, 221–245.

Herrmann, E., Wand, M. P., Engel, J. and Gasser, Th. (1995) A bandwidth selector for bivariate kernel regression. *J. R. Statist. Soc.* B, **57**, 171–180.

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996a) A brief survey of bandwidth selection for density estimation. *J. Am. Statist. Ass.*, **91**, 401–407.

———(1996b) Progress in data-based bandwidth selection for kernel density estimation. *Comput. Statist.*, **11**, 337–381.

Lütkepohl, H. (1996) *Handbook of Matrices*. Chichester: Wiley.

Ruppert, D. (1997) Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Am. Statist. Ass.*, **92**, 1049–1062.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995) An effective bandwidth selector for local least squares regression. *J. Am. Statist. Ass.*, **90**, 1257–1270.

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

Sain, S. R., Baggerly, K. A. and Scott, D. W. (1994) Cross-validation of multivariate densities. *J. Am. Statist. Ass.*, **89**, 807–817.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Tschernig, R. and Yang, L. (1999) Nonparametric lag selection for time series. *J. Time Ser. Anal.*, to be published.

Wand, M. P. and Jones, M. C. (1993) Comparison of smoothing parametrizations in bivariate kernel density estimation. *J. Am. Statist. Ass.*, **88**, 520–528.

———(1994) Multivariate plug-in bandwidth selection. *Comput. Statist.*, **9**, 97–117.

———(1995) *Kernel Smoothing*. London: Chapman and Hall.