Bernoulli **5**(5), 1999, 927–949

Nonparametric estimation of quadratic regression functionals

LI-SHAN HUANG[1] and JIANQING FAN[2]

[1]*Department of Statistics, Florida State University, Tallahassee FL 32306–4330, USA. E-mail: huang@stat.fsu.edu*
[2]*Department of Statistics, University of North Carolina, Chapel Hill NC 27599-3260, USA. E-mail: jfan@stat.unc.edu*

Quadratic regression functionals are important for bandwidth selection of nonparametric regression techniques and for nonparametric goodness-of-fit tests. Based on local polynomial regression, we propose estimators for weighted integrals of squared derivatives of regression functions. The rates of convergence in mean square error are calculated under various degrees of smoothness and appropriate values of the smoothing parameter. Asymptotic distributions of the proposed quadratic estimators are considered with the Gaussian noise assumption. It is shown that when the estimators are pseudo-quadratic (linear components dominate quadratic components), asymptotic normality with rate $n^{-1/2}$ can be achieved.

Keywords: asymptotic normality; equivalent kernel; local polynomial regression

1. Introduction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed observations with conditional mean $\mathrm{E}(y|x) = m(x)$ and variance $\mathrm{var}(y|x) = \sigma^2$. Consider the problem of estimating functionals of the form

$$\theta_\nu^* = \int [m^{(\nu)}(x)]^2 w(x)\,\mathrm{d}x, \qquad \nu = 0, 1, \ldots, \tag{1.1}$$

where $w(x)$ is a prescribed non-negative weight function. These functionals appear in expressions for the asymptotically optimal bandwidth for nonparametric function estimates; see, for example, Ruppert and Wand (1994) and Fan and Gijbels (1996). Doksum and Samarov (1995) consider particularly the applications of θ_0^* in measuring the explanatory power of covariates in regression. Another area of application is to examine the goodness of fit of a $(\nu - 1)$th-degree polynomial model by testing $H_0 : \theta_\nu^* = 0$, $\nu = 1, 2, \ldots$. When $w(\cdot)$ is the indicator function of an interval, H_0 tests the polynomial model at that given interval.

Problems similar to the estimation of θ_ν^* are considered in a number of papers. Hall and Marron (1987; 1991), Bickel and Ritov (1988) and Jones and Sheather (1991) discuss estimation of $\int [f^{(\nu)}(x)]^2\,\mathrm{d}x$ based on kernel density estimators, where $f(\cdot)$ is a probability density function. Birgé and Massart (1995) expand the study to the estimation of integrals of smoothed functionals of a density. Adaptive estimation of quadratic functionals can be found

1350–7265

in Efromovich and Low (1996). Laurent (1996; 1997) uses the orthonormal series method to estimate density functionals of the form $\int f^{(\nu)}(x) f^{(\nu')}(x) w(x)\,\mathrm{d}x$ and deal with this more general integral form via Taylor expansions. Optimality results of estimating $\int [m^{(\nu)}(x)]^2\,\mathrm{d}x$ under the Gaussian white noise model are obtained in Donoho and Nussbaum (1990) and Fan (1991). In the context of nonparametric regression, the corresponding problem is much less understood. Doksum and Samarov (1995) give estimators of $\int m^2(x) f(x) w(x)\,\mathrm{d}x$ ($f(\cdot)$ is the design density function) by local constant approximation, that is, the Nadaraya–Watson estimator. Here we develop estimates of θ_ν^* based on local polynomial regression estimators. Ruppert *et al*. (1995) consider functionals similar to θ_ν^* with weight $w(x) = f(x)$. However, no results as general as shown in this paper have been established, and our estimators for θ_ν^* are different from those of Doksum and Samarov (1995) and Ruppert *et al*. (1995).

Technically, θ_ν^*s are nonlinear functionals of $m(\cdot)$. While most theory focuses on estimation of linear functionals, it is also of theoretical interest to explore the difficulty of estimating nonlinear functionals. In the density estimation setting, the $n^{-1/2}$-consistency of estimating $\int [f^{(\nu)}(x)]^2\,\mathrm{d}x$ is established in Hall and Marron (1987; 1991), Bickel and Ritov (1988) and Birgé and Massart (1995). Laurent (1996; 1997) shows that similar results hold for estimated weighted integrals as well. Now a natural question is whether $n^{-1/2}$-consistent estimators can be constructed for regression functionals θ_ν^*. Note that the results for estimating density integral functionals are based on 'diagonal-out' type estimators. However, we do not have a 'diagonal-out' estimator for θ_ν^*. A similar 'bias-corrected' estimator (see (3.5) below) can be constructed for regression functionals θ_ν^*, but it requires $s > 2\nu + 1/2$ to achieve the $n^{-1/2}$ rate (see Theorem 4.4).

Our estimators are based on local polynomial regression estimators. Section 4 describes the asymptotic behaviour of the estimators, which are quadratic functionals of weighted regression estimates. Appropriate values of the smoothing parameter and smoothness conditions of $m(\cdot)$ are given for mean square rates of convergence. To study the asymptotic distributions of the proposed estimators, functionals of the form $\sum_{i=1}^n \sum_{j=1}^n a_{i,j} Y_i Y_j$ are considered in Section 5, where $Y_1, \ldots, Y_n$ are independent response variables. As a consequence, we provide further insights into the difficulty of estimating quadratic functionals: the $n^{-1/2}$ rate for estimating θ_ν^* can be achieved when estimators are pseudo-quadratic (linear component dominates). For genuinely quadratic estimators (quadratic component dominates), the $n^{-1/2}$ rate is not attainable.

This paper is organized as follows. Section 2 provides the background of local polynomial regression. The proposed estimators of θ_ν^* are given in Section 3. Section 4 contains rate-of-convergence results in mean square error, from both a theoretical and a practical point of view. Asymptotic distributions of the proposed estimators are studied in Sections 5 and 6. Proofs of lemmas and theorems are postponed to Section 7.

2. Local polynomial regression

Suppose that (X_i, Y_i), $i = 1, \ldots, n$, have been collected from the model

$$Y_i = m(X_i) + \varepsilon_i, \qquad \mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{var}(\varepsilon_i) = \sigma^2, \tag{2.1}$$

where $m(\cdot)$ is an unknown regression function, the ε_is are independent and identically distributed error terms, and X_is are independent of ε_is. The function of interest could be the regression curve $m(\cdot)$ or its derivatives.

The idea of local polynomial regression is to fit *locally* a low-order polynomial at grid points of interest, with observations receiving different weights. Assume that the $(p+1)$th derivative of $m(\cdot)$ exists, with p a non-negative integer. For a fixed point x, the regression function $m(\cdot)$ can be *locally* approximated by

$$m(z) \approx m(x) + m'(x)(z-x) + \cdots + m^{(p)}(x)(z-x)^p/p!,$$

for z in a neighbourhood of x. This leads to the following least-squares problem:

$$\min_{\beta} \sum_{i=1}^{n} \left(Y_i - \sum_{\nu=0}^{p} \beta_\nu (X_i - x)^\nu \right)^2 K\left(\frac{X_i - x}{h}\right), \tag{2.2}$$

where $\beta_\nu = m^{(\nu)}(x)/\nu!$, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$, and the dependence of β_ν on x is suppressed. The neighbourhood is controlled by a bandwidth h, and the weights are assigned via a kernel function K, a continuous, bounded and symmetric real function which integrates to one. Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ be the solution to the minimization problem (2.2). Then $\nu!\hat{\beta}_\nu$ is an estimator for $m^{(\nu)}(x)$, obtained by fitting a pth-degree weighted polynomial in a neighbourhood of x. See Fan and Gijbels (1996) for further details.

For convenience, some matrix notation is introduced here. Let

$$X = \begin{pmatrix} 1 & (X_1 - x) & \ldots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x) & \ldots & (X_n - x)^p \end{pmatrix}, \tag{2.3}$$

$$W = \operatorname{diag}\left(K\left(\frac{X_i - x}{h}\right)\right)_{n\times n} = \begin{pmatrix} K\left(\frac{X_1 - x}{h}\right) & & \\ & \ddots & \\ & & K\left(\frac{X_n - x}{h}\right) \end{pmatrix}, \tag{2.4}$$

$Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, and $\mathbf{m} = (m(X_1), \ldots, m(X_n))^{\mathrm{T}}$. By the standard least-squares theory, the estimator $\hat{\beta}$ can be written as

$$\hat{\beta} = (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W Y. \tag{2.5}$$

We further investigate $\hat{\beta}_\nu$, $\nu = 0, \ldots, p$, individually. Let e_ν, $\nu = 0, \ldots, p$, denote $(p+1) \times 1$ unit vectors having 1 as the $(\nu+1)$th component, 0 otherwise. Simple algebra shows that

$$\hat{\beta}_\nu(x) = e_\nu^{\mathrm{T}} \hat{\beta} = \sum_{i=1}^{n} W_\nu^n \left(\frac{X_i - x}{h}\right) Y_i, \tag{2.6}$$

where

$$W_\nu^n(t) = e_\nu^{\mathrm{T}}(X^{\mathrm{T}}WX)^{-1}(1,\ ht,\ \ldots,\ h^p t^p)^{\mathrm{T}}K(t), \qquad \nu = 0,\ \ldots,\ p. \tag{2.7}$$

The weight functions $W_\nu^n(\cdot)$, $\nu = 0, \ldots, p$, depend on both x and the design points $\{X_1, \ldots, X_n\}$, and satisfy

$$\sum_{i=1}^{n}(X_i - x)^k W_\nu^n\left(\frac{X_i - x}{h}\right) = \begin{cases} 1 & \text{if } k = \nu, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } 0 \leqslant \nu,\ k \leqslant p. \tag{2.8}$$

The dependence of $W_\nu^n(\cdot)$ on x and the design points $\{X_1, \ldots, X_n\}$ is the key to design adaptation and superior boundary behaviour of the local polynomial regression.

To understand the estimates $\hat{\beta}_\nu$, $\nu = 0, \ldots, p$, more intuitively, define

$$K_\nu^*(t) = e_\nu^{\mathrm{T}}S^{-1}(1,\ t,\ \ldots,\ t^p)^{\mathrm{T}}K(t), \qquad \nu = 0,\ \ldots,\ p, \tag{2.9}$$

where $S = (\mu_{i+j-2})_{(p+1)\times(p+1)}$ with $\mu_k = \int t^k K(t)\mathrm{d}t$. It is easy to see that the functions $K_\nu^*(\cdot)$, $\nu = 0, \ldots, p$, are independent of x and $\{X_1, \ldots, X_n\}$ and satisfy

$$\int t^k K_\nu^*(t)\mathrm{d}t = \begin{cases} 1 & \text{if } k = \nu, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } 0 \leqslant \nu,\ k \leqslant p. \tag{2.10}$$

The theoretical connection between $K_\nu^*(\cdot)$ and $W_\nu^n(\cdot)$ is shown in the following lemma, which extends the equivalent kernel results of Ruppert and Wand (1994).

Lemma 2.1. *If the marginal density of X, denoted by $f(\cdot)$, is Hölder continuous on an interval $[a, b]$ and $\min_{x\in[a,b]} f(x) > 0$, then*

$$\sup_{t\in[-1,1]}\ \sup_{x\in[a,b]} |nh^{\nu+1}W_\nu^n(t) - K_\nu^*(t)/f(x)| \xrightarrow{\text{a.s.}} 0,$$

provided that K has a bounded support $[-1, 1]$ and $h = h(n) \to 0$, $nh \to \infty$.

From (2.6) and Lemma 2.1, the local polynomial estimator of $\beta_\nu(x) = m^{(\nu)}(x)/\nu!$ has the asymptotic expansion

$$\hat{\beta}_\nu(x) = \frac{(1 + o_P(1))}{nh^{\nu+1}f(x)}\sum_{i=1}^{n} K_\nu^*\left(\frac{X_i - x}{h}\right)Y_i. \tag{2.11}$$

Indeed, (2.11) provides a simple technical device for studying the asymptotic behaviour of $\hat{\beta}$.

3. Estimators

By the definition of θ_ν^* in (1.1), a natural approach is to substitute an estimate $\hat{m}^{(\nu)}(x)$ for $m^{(\nu)}(x)$. One candidate is $(\nu!)^2\int\hat{\beta}_\nu^2(x)w(x)\,\mathrm{d}x$, where $\hat{\beta}_\nu(x)$ is obtained via fitting a local pth-degree polynomial, $p \geqslant \nu$, as in (2.6). Since $(\nu!)^2$ is a constant for any fixed ν, an estimator

$$\hat{\theta}_\nu = \int\hat{\beta}_\nu^2(x)w(x)\,\mathrm{d}x$$

is considered. Correspondingly, denote

$$\theta_\nu = \int \beta_\nu^2(x) w(x)\,\mathrm{d}x = \frac{1}{(\nu!)^2}\int [m^{(\nu)}(x)]^2 w(x)\,\mathrm{d}x.$$

We shall study the asymptotic properties of $\hat{\theta}_\nu$ later, but first an intuitive discussion may be helpful. In addition to the model assumption in (2.1), we further assume that $\mathrm{E}(\varepsilon_i^3) = 0$ and $\mathrm{E}(\varepsilon_i^4) < \infty$, $i = 1, \ldots, n$.

From (2.6),

$$\hat{\theta}_\nu = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}(\nu) Y_i Y_j, \tag{3.1}$$

where

$$a_{i,j}(\nu) = \int W_\nu^n\left(\frac{X_i - x}{h}\right) W_\nu^n\left(\frac{X_j - x}{h}\right) w(x)\,\mathrm{d}x. \tag{3.2}$$

Clearly $\hat{\theta}_\nu$ is a quadratic form in the Y_is and can be written as

$$\hat{\theta}_\nu = Y^{\mathrm{T}} A_\nu Y,$$

with $A_\nu = (a_{i,j}(\nu))_{n\times n}$. The conditional mean and variance of $\hat{\theta}_\nu$ follow at once from (3.1):

$$\mathrm{E}\{\hat{\theta}_\nu | X_1, \ldots, X_n\} = \mathbf{m}^{\mathrm{T}} A_\nu \mathbf{m} + \sigma^2 \mathrm{tr}(A_\nu), \tag{3.3}$$

$$\mathrm{var}\{\hat{\theta}_\nu | X_1, \ldots, X_n\} = 4\sigma^2 \mathbf{m}^{\mathrm{T}} A_\nu^2 \mathbf{m} + 2\sigma^4 \mathrm{tr}(A_\nu^2) + (\mathrm{E}(\varepsilon_1^4) - 3\sigma^4)\sum_{i=1}^{n} a_{i,i}^2(\nu), \tag{3.4}$$

where $\mathrm{tr}(A_\nu)$ denotes the trace of A_ν. Note that the second term on the right-hand side of (3.3) may be thought of as extra bias in the estimation. This motivates a 'bias-corrected' estimate:

$$\check{\theta}_\nu = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}(\nu) Y_i Y_j - \hat{\sigma}^2 \mathrm{tr}(A_\nu), \tag{3.5}$$

where $\hat{\sigma}^2$ is an estimator of σ^2. The non-negativity of $\check{\theta}_\nu$ is no longer guaranteed. This can be easily modified by taking the estimator $\check{\theta}_\nu^+ = \max(\check{\theta}_\nu, 0)$, whose performance is at least as good as $\check{\theta}_\nu$.

4. Rates of convergence

Asymptotic expressions for the conditional bias and variance of $\hat{\theta}_\nu$ and $\check{\theta}_\nu$ will be derived in this section. We shall use these to study how the mean square errors behave. It is convenient at this point to introduce some conditions.

Conditions A.

(A1) The kernel $K(\cdot)$ is a continuous, bounded and symmetric probability density function, having a support on $[-1, 1]$.

(A2) The design density function $f(\cdot)$ is positive and continuous for $x \in [-1, 1]$, and $f(\cdot)$ has derivatives of order $\max(\nu, (p-\nu+1)/2)$ *on* $[-1, 1]$.

(A3) The weight $w(\cdot)$ is a bounded and non-negative function on the interval $[-1, 1]$. *Further, $w(\cdot)$ has bounded derivatives of order* $\max(\nu, (p-\nu+1)/2)$ *and* $w^{(i)}(1) = w^{(i)}(-1) = 0,\ i = 0, \ldots, \max(\nu, (p-\nu+1)/2)$.

Condition A1 is assumed for simplicity of proofs, and Condition A2 is imposed in order to apply integration by parts in deriving the results (see (7.9) and (7.15)). With Condition A2, some smoothness condition on m is required. Condition A3 is mainly for eliminating the 'boundary effects' so that deeper and cleaner results can be obtained. Without loss of generality, the interval $[-1, 1]$ is used in Conditions A. The results will hold for any bounded intervals $[a_1, b_1]$, $[a_2, b_2]$ and $[a_2, b_2]$ in Conditions A1, A2 and A3, respectively. The regression function $m(\cdot)$ will be said to have smoothness of order s if there is a constant $M > 0$ such that, for all x and y in $[-1, 1]$,

$$|m^{(l)}(x) - m^{(l)}(y)| \leqslant M|x-y|^{\alpha}, \text{ where } 0 < \alpha \leqslant 1. \tag{4.1}$$

Let $s = \ell + \alpha$, and let $[s]$ denote the largest integer strictly less than s. Throughout we will use s to denote the smoothness of the regression function $m(\cdot)$ and p to denote the degree of local polynomial fitting. Note that p and s are two independent indices with $p \geqslant \nu$ and $s > \nu$. When estimating the νth derivative of $m(\cdot)$, it is argued in Ruppert and Wand (1994) and Fan and Gijbels (1996) that odd integers $(p-\nu)$ should be used. Thus, $p-\nu$ is assumed to be odd in this section so that $p \pm \nu + 1$ is a multiple of 2. For simplicity of notation, it is understood that *all expectations* in this section are conditioned on $\{X_1, \ldots, X_n\}$. Finally, let $(\phi_1 * \phi_2)(u) = \int \phi_1(t)\phi_2(u-t)\mathrm{d}t$ denote the convolution of two real-valued functions ϕ_1 and ϕ_2.

The asymptotic bias and variance of $\hat{\theta}_\nu$ are described in the following theorem. The proof is given in Section 7.

Theorem 4.1. *Assume that as $n \to \infty$, $h = h(n) \to 0$, $nh^{2\nu+1} \to \infty$, and $nh^{2(s-[s])+1} \to \infty$. Then under Conditions A,*

(a) the asymptotic bias is

$$\mathrm{E}(\hat{\theta}_\nu) - \theta_\nu \equiv B_{n,\nu}$$

$$= \begin{cases} O_P(h^{s+[s]-2\nu}) + (C_2 + o_P(1))n^{-1}h^{-2\nu-1} & \text{if } s \leqslant (p+\nu+1)/2, \\ (C_1 + o_P(1))h^{p+1-\nu} + (C_2 + o_P(1))n^{-1}h^{-2\nu-1} & \text{if } s > (p+\nu+1)/2, \end{cases} \tag{4.2}$$

where

$$C_1 = 2\left(\int t^{p+1} K_\nu^*(t)\mathrm{d}t\right)\left(\int \beta_{p+1}(x)\beta_\nu(x)w(x)\,\mathrm{d}x\right),$$

$$C_2 = \sigma^2\left(\int [K_\nu^*(t)]^2\,\mathrm{d}t\right)\left(\int f(x)^{-1}w(x)\,\mathrm{d}x\right); \tag{4.3}$$

(b) the asymptotic variance is

$$\begin{aligned}\mathrm{var}(\hat\theta_\nu) &\equiv V_{n,\nu}\\ &= \begin{cases}(D_1 + o_P(1))n^{-2}h^{-4\nu-1} + O_P(n^{-1}h^{-2\nu+s}) & \text{if } s \leqslant 2\nu,\\ (D_1 + o_P(1))n^{-2}h^{-4\nu-1} + (D_2 + o_P(1))n^{-1} & \text{if } s > 2\nu,\end{cases}\end{aligned} \tag{4.4}$$

where

$$D_1 = 2\sigma^2\left(\left(\sigma^2\int [(K_\nu^* * K_\nu^*)(z)]^2\,\mathrm{d}z\right)\left(\int f(x)^{-2}w^2(x)\,\mathrm{d}x\right)\right.$$
$$\left. + 2\iint m^2(x)f^{-1}(x)f^{-1}(y)w(x)w(y)\,\mathrm{d}x\,\mathrm{d}y\right),$$

$$D_2 = \frac{4\sigma^2}{(\nu!)^2}\int [G_\nu^{(\nu)}(y)]^2 f(y)\mathrm{d}y,$$

with $G_\nu(y) = \beta_\nu(y)w(y)f^{-1}(y)$.

Remark 1. Note that in Theorem 4.1(a), when $s \leqslant p+1$, the integral $\int \beta_{p+1}(x)\beta_\nu(x)w(x)\,\mathrm{d}x$, which involves $m^{(p+1)}(x)$, is to be understood as

$$\int \beta_{\frac{p+\nu+1}{2}}(x)\frac{\mathrm{d}^r}{\mathrm{d}x^r}(\beta_\nu(x)w(x))\,\mathrm{d}x \qquad \text{with } r = (p-\nu+1)/2,$$

via integration by parts.

The asymptotic minimum mean square error of $\hat\theta_\nu$ can be obtained from Theorem 4.1. We state the results in two versions. The theoretical version (part (a)) on the best possible rate obtained by choosing a sufficiently large P when the degree of smoothness is given. Practically (part (b)), given the order of local polynomial fitting p, one wants to know what the best rate is if $m(\cdot)$ is sufficiently smooth.

Theorem 4.2. *Under the assumptions of Theorem 4.1, the asymptotic minimum mean square errors of* $\hat\theta_\nu$ *are achieved as follows.*

(a) Given the degree of smoothness $s > \nu$, *taking* $h = O(n^{-1/(s+[s]+1)})$ *and* $p \geqslant (2s-\nu-1)$, *then*

$$\mathrm{E}(\hat\theta_\nu - \theta_\nu)^2 = \begin{cases} O_P(n^{-2(s+[s]-2\nu)/(\mathrm{s}+[\mathrm{s}]+1)}) & \text{if } s \leqslant 2\nu,\\ (D_2 + o_P(1))n^{-1} + O_P(n^{-2(s+[s]-2\nu)/(\mathrm{s}+[\mathrm{s}]+1)}) & \text{if } s > 2\nu.\end{cases}$$

In particular, when $s > 2\nu+1$, $\mathrm{E}(\hat\theta_\nu - \theta_\nu)^2 = D_2 n^{-1} + o_P(n^{-1})$.

(b) Given the order of local polynomial fit $p \geqslant \nu$, if $s > (p+\nu+1)/2$, then

$$\mathrm{E}(\hat{\theta}_\nu - \theta_\nu)^2 = \begin{cases} O_P(n^{-(2p-2\nu+3)/(p+\nu+2)}) & \text{if } C_1 < 0 \text{ and } s \leqslant 2\nu, \\ O_P(n^{-2(p-\nu+1)/(p+\nu+2)}) & \text{if } C_1 > 0 \text{ and } s \leqslant 2\nu, \\ (D_2 + o_P(1))n^{-1} + O_P(n^{-(2p-2\nu+3)/(p+\nu+2)}) & \text{if } C_1 < 0 \text{ and } s > 2\nu, \\ (D_2 + o_P(1))n^{-1} + O_P(n^{-2(p-\nu+1)/(p+\nu+2)}) & \text{if } C_1 > 0 \text{ and } s > 2\nu, \end{cases} \tag{4.5}$$

by taking

$$h_{\mathrm{OPT}} = \begin{cases} \left(\dfrac{C_2}{-C_1 n}\right)^{1/(p+\nu+2)} & \text{if } C_1 < 0, \\ \left(\dfrac{C_2(2\nu+1)}{C_1(p+1-\nu)n}\right)^{1/(p+\nu+2)} & \text{if } C_1 > 0. \end{cases} \tag{4.6}$$

In particular, $\mathrm{E}(\hat{\theta}_\nu - \theta_\nu)^2 = D_2 n^{-1} + o_P(n^{-1})$, if $p > 3\nu - 1$ for the case of $C_1 < 0$, and $p > 3\nu$ for $C_1 > 0$.

Remark 2. The bandwidth h_{OPT} in (4.6) minimizes the mean square error of $\hat{\theta}_\nu$. In particular, when $D_2 n^{-1}$ is the leading term in (4.5), h_{OPT} minimizes the second-order terms of the mean square error of $\hat{\theta}_\nu$. In this case, any h satisfying

$$h = o(n^{-1/(2(p+1-\nu))}) \quad \text{and} \quad n^{-1/(4\nu+2)} = o(h)$$

can be the optimal smoothing parameter. Thus, the choice of bandwidth is not sensitive in this case.

Bickel and Ritov (1988) and Laurent (1996; 1997) give the efficient information bound for estimating respectively unweighted and weighted integrals of squared density derivatives. For the current regression setting, we conjecture that $D_2 n^{-1}$ is the semi-parametric information bound. In other words, $\hat{\theta}_\nu$ is efficient when the degree of smoothness is sufficiently large. But the construction of the estimator $\hat{\theta}_\nu$ is conceptually simpler than that of Bickel and Ritov (1988).

Note that $C_2 n^{-1} h^{-2\nu-1}$ in the asymptotic bias (4.2) converges to zero more slowly than the square root of $D_1 n^{-2} h^{-4\nu-1}$ in the variance expression (4.4). In kernel density estimation, Jones and Sheather (1991) argue in favour of estimators of the type $\hat{\theta}_\nu$, since one can choose an ideal bandwidth such that the leading bias terms (analogous to the second expression in (4.2)) cancel altogether. However, this bias reduction technique is not applicable here, since C_1 in (4.3) is not necessarily negative and C_2 is always positive. To correct the extra bias term, we use the estimator $\check{\theta}_\nu$ defined in (3.5).

Suppose that σ^2 can be estimated at rate $O(\epsilon_n)$, that is, $\mathrm{E}\{(\hat{\sigma}^2 - \sigma^2)^2\} = O(\epsilon_n^2)$. It will be seen in Theorem 4.4 that the rate $\epsilon_n = o(h^{1/2})$ is fast enough for the purpose of bias correction. The following theorem gives the asymptotic bias and variance of $\check{\theta}_\nu$.

Theorem 4.3. *Under the assumptions of Theorem 4.1,*

(a) the asymptotic bias of $\check{\theta}_\nu$ *is*

$$\mathrm{E}(\check{\theta}_\nu) - \theta_\nu \equiv b_{n,\nu} = \begin{cases} O_P(h^{s+[s]-2\nu} + \epsilon_n n^{-1} h^{-2\nu-1}) & \text{if } s \leqslant (p+\nu+1)/2, \\ (C_1 + o_P(1))h^{p+1-\nu} + O_P(\epsilon_n n^{-1} h^{-2\nu-1}) & \text{if } s > (p+\nu+1)/2; \end{cases} \tag{4.7}$$

(b) the asymptotic variance of $\check{\theta}_\nu$ *is*

$$\mathrm{var}(\check{\theta}_\nu) \equiv \upsilon_{n,\nu} = \begin{cases} (D_1 + o_P(1))n^{-2}h^{-4\nu-1} + O_P(n^{-1}h^{-2\nu+s} + \epsilon_n^2 n^{-2} h^{-4\nu-2}) & \text{if } s \leqslant 2\nu, \\ (D_1 + o_P(1))n^{-2}h^{-4\nu-1} + (D_2 + o_P(1))n^{-1} + O_P(\epsilon_n^2 n^{-2} h^{-4\nu-2}) & \text{if } s > 2\nu. \end{cases}$$

We now summarize some results from Theorem 4.3. Again they are stated in both theoretical (part (a)) and practical (part (b)) versions.

Theorem 4.4. *Under the assumptions of Theorem 4.1, the asymptotic minimum mean square errors of* $\check{\theta}_\nu$ *are achieved as follows.*

(a) Given s, *taking* $h = O(n^{-2/(2s+2[s]+1)})$ *and* $p \geqslant (2s - \nu - 1)$, *if* $\epsilon_n = o(h^{1/2})$, *then*

$$\mathrm{E}(\check{\theta}_\nu - \theta_\nu)^2 = \begin{cases} O_P(n^{-4(s+[s]-2\nu)/(2s+2[s]+1)}) & \text{if } s \leqslant 2\nu, \\ (D_2 + o_P(1))n^{-1} + O_P(n^{-4(s+[s]-2\nu)/(2s+2[s]+1)}) & \text{if } s > 2\nu. \end{cases} \tag{4.8}$$

In particular, $\mathrm{E}(\check{\theta}_\nu - \theta_\nu)^2 = D_2 n^{-1} + o_P(n^{-1})$ *when* $s > 2\nu + \frac{1}{2}$.

(b) Given p, *if* $s > (p+\nu+1)/2$ *and* $\epsilon_n = o(h^{1/2})$, *then*

$$\mathrm{E}(\check{\theta}_\nu - \theta_\nu)^2 = \begin{cases} O_P(n^{-4(p+1-\nu)/(2p+2\nu+3)}) & \text{if } s \leqslant 2\nu, \\ (D_2 + o_P(1))n^{-1} + O_P(n^{-4(p+1-\nu)/(2p+2\nu+3)}) & \text{if } s > 2\nu, \end{cases} \tag{4.9}$$

by taking

$$h_{\mathrm{opt}} = \left(\frac{D_1(4\nu+1)}{2C_1^2(p+1-\nu)}\right)^{1/(2p+2\nu+3)} n^{-2/(2p+2\nu+3)}. \tag{4.10}$$

In particular, $\mathrm{E}(\check{\theta}_\nu - \theta_\nu)^2 = D_2 n^{-1} + o_P(n^{-1})$ *when* $p > 3\nu - \frac{1}{2}$.

Remark 3. As shown in Theorem 4.4, the $n^{-1/2}$ rate is achievable and in this case the smoothing parameter affects only the second-order terms of the mean square error. In fact, when $D_2 n^{-1}$ is the dominant term in (4.9), the $n^{-1/2}$ rate of convergence is attained for any bandwidth satisfying

$$h = o(n^{-1/(2(p+1-\nu))}) \quad \text{and} \quad n^{-1/(4\nu+1)} = o(h).$$

Remark 4. Theorem 4.4(a) shows that if $s > 2\nu + \frac{1}{2}$, θ_ν can be estimated at the $n^{-1/2}$ rate. This minimal smoothness condition is slightly stronger than $s > 2\nu + \frac{1}{4}$ for estimating

functionals of a density, as shown in Bickel and Ritov (1988), Laurent (1996; 1997) and Birgé and Massart (1995). One can possibly follow a similar idea to Laurent (1996) to construct better bias-corrected estimators, at the expense of obtaining much more complex estimators in our setting. We have not chosen this option because we are primarily interested in understanding the performance of the intuitively natural estimators.

In practical implementation one should attempt to estimate σ^2 and h_{OPT} or h_{opt}. The parameter σ^2 can be estimated at rate $n^{-1/2}$ using the estimators of Rice (1984) and Hall *et al*. (1990), for example. The problem of estimating the ideal bandwidth for $\hat{\theta}_\nu$ or $\check{\theta}_\nu$ is beyond the scope of this paper, but (4.6) and (4.10) provide a guideline. A simple rule of thumb is the following. Fit a global polynomial of order $(p+\nu+5)/2$ to obtain an estimate for C_1 (see Remark 1) by regarding the regression function as a polynomial function; estimate $f(x)$ and plug the resulting estimate into C_2 or D_1. Then an estimate of h_{OPT} or h_{opt} is formed. This 'plug-in rule' is expected to work reasonably well in many situations, since $\hat{\theta}_\nu$ and $\check{\theta}_\nu$ are robust against the choice of bandwidth (see Remarks 2 and 3).

Ruppert *et al.*, (1995) consider estimation of

$$\theta_{rs} = \int m^{(r)}(x) m^{(s)}(x) f(x)\,\mathrm{d}x,\ r,\ s \geqslant 0 \quad \text{and} \quad r+s \text{ even}. \tag{4.11}$$

They propose an estimator

$$\hat{\theta}_{rs} = n^{-1} \sum_{i=1}^{n} \hat{m}_r(X_i,\ g)\hat{m}_s(X_i,\ g),$$

where $\hat{m}_r(\cdot,\ g)$ and $\hat{m}_s(\cdot,\ \mathrm{g})$ are obtained via fitting a pth-degree local polynomial with a bandwidth g. Also $p-r$ and $p-s$ are both odd. Comparing with their estimator, our error criterion (weighted mean integrated square error) is more general and the estimators $\hat{\theta}_\nu$ and $\check{\theta}_\nu$ are different from $\hat{\theta}_{rs}$.

5. Asymptotic distributions of quadratic functionals

The asymptotic distribution of

$$\hat{\theta} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j} Y_i Y_j \tag{5.1}$$

will be considered in this section. To make the technical arguments simple, we restrict our attention to the Gaussian model,

$$Y_i = m(X_i) + \varepsilon_i, \qquad \varepsilon_i \text{ i.i.d.} \sim N(0,\ \sigma^2), \qquad i = 1,\ \ldots,\ n. \tag{5.2}$$

This Gaussian noise assumption can be removed with additional work on proofs, for example via applying the martingale central limit theorem. The $a_{i,j}$s in (5.1) possibly depend on n and the design points $\{X_1,\ \ldots,\ X_n\}$. Let $\mathbf{m}$ be an $n \times 1$ column vector with entries $m(X_i)$, $i = 1,\ \ldots,\ n$, $\varepsilon = (\varepsilon_1,\ \ldots,\ \varepsilon_n)^{\mathrm{T}}$, and $A = (a_{i,j})_{n\times n}$. Then $\hat{\theta}$ can be written as

$$\hat{\theta} = Y^{\mathrm{T}}AY = \mathbf{m}^{\mathrm{T}}A\mathbf{m} + 2\mathbf{m}^{\mathrm{T}}A\varepsilon + \varepsilon^{\mathrm{T}}A\varepsilon. \tag{5.3}$$

The mean and variance of $\hat{\theta}$ can be calculated directly from (5.3):

$$\mathrm{E}\{\hat{\theta}|X_1, \ldots, X_n\} = \mathbf{m}^{\mathrm{T}}A\mathbf{m} + \sigma^2\mathrm{tr}(A), \tag{5.4}$$

$$\mathrm{var}\{\hat{\theta}|X_1, \ldots, X_n\} = 2\sigma^4\mathrm{tr}(A^2) + 4\sigma^2\mathbf{m}^{\mathrm{T}}A^2\mathbf{m}. \tag{5.5}$$

Conditioned on $\{X_1, \ldots, X_n\}$, $2\mathbf{m}^{\mathrm{T}}A\varepsilon$ and $\varepsilon^{\mathrm{T}}A\varepsilon$ are the stochastic terms in (5.3). It is easy to see that $2\mathbf{m}^{\mathrm{T}}A\varepsilon$ is linear in ε_is and contributes $4\sigma^2\mathbf{m}^{\mathrm{T}}A^2\mathbf{m}$ to the variance in (5.5), while $\varepsilon^{\mathrm{T}}A\varepsilon$ is quadratic in ε_is with variance $2\sigma^4\mathrm{tr}(A^2)$.

The objective is to show that $\hat{\theta}$ is asymptotically normally distributed under some conditions. Two cases will be considered, depending on whether the linear term or the quadratic term dominates (see (5.3) and (5.5)):

(i) $\mathrm{tr}(A^2)/\mathbf{m}^{\mathrm{T}}A^2\mathbf{m} \xrightarrow{P} 0$ ($\mathbf{m}^{\mathrm{T}}A\varepsilon$ dominates),
(ii) $\mathrm{tr}(A^2)/\mathbf{m}^{\mathrm{T}}A^2\mathbf{m} \xrightarrow{P} \infty$ ($\varepsilon^{\mathrm{T}}A\varepsilon$ dominates).

If the linear term dominates (case (i)), that is, $\hat{\theta}$ is 'pseudo-quadratic', then the normality of $\hat{\theta}$ follows directly under the Gaussian noise assumption. The distribution theory of quadratic forms in normal variables will be used to prove the asymptotic normality for case (ii), where $\hat{\theta}$ is 'genuinely quadratic'. As will be seen in Section 6, the separate treatments of these two cases are natural, corresponding respectively to root-n and non-root-n rates of convergence. There is a 'boundary case' when $\mathrm{tr}(A^2)$ and $\mathbf{m}^{\mathrm{T}}A^2\mathbf{m}$ are of the same order; this may be handled with additional work. The general theory of quadratic functionals in Whittle (1964) and Khatri (1980) is useful, but not directly applicable to the situation described here. We provide the following result.

Theorem 5.1. *Under the model assumption in (5.2), if $\hat{\theta}$ is pseudo-quadratic (case (i)), then the conditional distribution of $\hat{\theta}$ given $\{X_1, \ldots, X_n\}$ is asymptotically normal:*

$$Y^{\mathrm{T}}AY \xrightarrow{C} N\big(\mathbf{m}^{\mathrm{T}}A\mathbf{m} + \sigma^2\mathrm{tr}(A),\, 2\sigma^4\mathrm{tr}(A^2) + 4\sigma^2\mathbf{m}^{\mathrm{T}}A^2\mathbf{m}\big), \tag{5.6}$$

where the symbol $\xrightarrow{C}$ denotes convergence conditioned on $\{X_1, \ldots, X_n\}$. Similarly, (5.6) holds for case (ii) if

$$\frac{\mathrm{tr}(A^4)}{(\mathrm{tr}(A^2))^2} \xrightarrow{P} 0, \quad \text{as } n \to \infty.$$

More precisely, the conditional asymptotic normality in (5.6) means that

$$P\left\{\frac{Y^{\mathrm{T}}AY - \mathbf{m}^{\mathrm{T}}A\mathbf{m} - \sigma^2\mathrm{tr}(A)}{(2\sigma^4\mathrm{tr}(A^2) + 4\sigma^2\mathbf{m}^{\mathrm{T}}A^2\mathbf{m})^{1/2}} \leqslant t|X_1, \ldots, X_n\right\} \xrightarrow{P} \Phi(t) \quad \text{for all } t, \tag{5.7}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable. Expression (5.7) also implies the unconditional asymptotic normality by the dominated convergence theorem:

$$P\left\{\frac{Y^{\mathrm{T}}AY - \mathbf{m}^{\mathrm{T}}A\mathbf{m} - \sigma^2\mathrm{tr}(A)}{(2\sigma^4\mathrm{tr}(A^2) + 4\sigma^2\mathbf{m}^{\mathrm{T}}A^2\mathbf{m})^{1/2}} \leqslant t\right\} \to \Phi(t) \qquad \textit{for all } t.$$

6. Asymptotic normality

We now establish the asymptotic normality of $\hat{\theta}_\nu$ and $\check{\theta}_\nu$. In addition to Conditions A, the bandwidth is assumed to satisfy the following:

Condition B. *$h = h(n) \to 0$ and $nh + \log(h) \to \infty$ as $n \to \infty$.*

This requirement is minor in the sense that if a local smoothing neighbourhood contains at least $(\log n)$ data points ($h > n^{-1}(\log n)^k$, $k > 1$), then Condition B is satisfied. We need the following lemmas to show the asymptotic normality of $\hat{\theta}_\nu$ and $\check{\theta}_\nu$.

Lemma 6.1. *Suppose $Z_{n,1}, Z_{n,2}, \ldots$ is a sequence of random variables having a multinomial distribution $(n; p_{n,1}, p_{n,2}, \ldots)$ with parameters satisfying $p_{n,j} \geqslant 0$, $j = 1, 2, \ldots,$ and $\max_j p_{n,j} < ch$, where c is a constant. Under Condition B,*

$$P\left\{\max_j Z_{n,j} > nhb_n\right\} \to 0, \qquad \textit{as } n \to \infty,$$

for any sequence b_n such that $b_n \to \infty$.

Lemma 6.2. *Let $\lambda_{\max}(\nu)$ be the maximum eigenvalue of $A_\nu = (a_{i,j}(\nu))_{n\times n}$, where $a_{i,j}(\nu)$ is defined in (3.2). Under Conditions A and B,*

$$\frac{\lambda^2_{\max}(\nu)}{\mathrm{tr}(A_\nu^2)} \xrightarrow{P} 0. \tag{6.1}$$

Since the design matrix X (see (2.3)) and the matrix A_ν are the same for $(X_1, \psi(\mathrm{Y}_1))$, $\ldots$, $(X_n, \psi(\mathrm{Y}_n))$ with $\psi(\cdot)$ a known function, Lemmas 6.1 and 6.2 are applicable to the case of estimating functionals of the form

$$\int [m_\psi^{(\nu)}(x)]^2 w(x)\,\mathrm{d}x,$$

where $m_\psi(x) = \mathrm{E}(\psi(y)|x)$.

From Lemmas 6.1 and 6.2, the asymptotic normality of $\hat{\theta}_\nu$ and $\check{\theta}_\nu$ follows.

Theorem 6.1. *Under the Gaussian model assumption in (5.2), if Conditions A and B are satisfied and $nh^{2\nu+1} \to \infty$, $nh^{2(s-[s])+1} \to \infty$, as $n \to \infty$, then the conditional distributions of $\hat{\theta}_\nu$ and $\check{\theta}_\nu$ are asymptotically normal:*

$$(\hat{\theta}_\nu - \theta_\nu) \xrightarrow{C} N(B_{n,\nu}, V_{n,\nu}),$$

$$(\check{\theta}_\nu - \theta_\nu) \xrightarrow{C} N(b_{n,\nu}, \upsilon_{n,\nu}),$$

where $B_{n,\nu}$, $V_{n,\nu}$, $b_{n,\nu}$ and $v_{n,\nu}$ are given in Theorems 4.1 and 4.3.

In particular, the $n^{-1/2}$-consistency is achieved with appropriate smoothing parameters (see Theorems 4.2 and 4.4):

(a) Given $s > 2\nu + 1$, taking $h = O(n^{-1/(s+[s]+1)})$ and $p \geqslant (2s - \nu - 1)$,

$$\sqrt{n}(\hat{\theta}_\nu - \theta_\nu) \xrightarrow{C} N(0,\ D_2).$$

(b) Given p such that

$$\begin{cases} p > 3\nu - 1 & \text{if } C_1 < 0, \\ p > 3\nu & \text{if } C_1 > 0, \end{cases}$$

if $s > (p + \nu + 1)/2$ and $h = h_{\mathrm{OPT}}$ given in (4.6), then

$$\sqrt{n}(\hat{\theta}_\nu - \theta_\nu) \xrightarrow{C} N(0,\ D_2).$$

(c) Given $s > 2\nu + \frac{1}{2}$, taking $h = O(n^{-2/(2s+2[s]+1)})$ and $p \geqslant (2s - \nu - 1)$, if $\epsilon_n = o(h^{1/2})$, then

$$\sqrt{n}(\check{\theta}_\nu - \theta_\nu) \xrightarrow{C} N(0,\ D_2).$$

(d) Given $p > 3\nu - \frac{1}{2}$, taking $h = h_{\mathrm{opt}}$ in (4.10), if $s > (p + \nu + 1)/2$ and $\epsilon_n = o(h^{1/2})$, then

$$\sqrt{n}(\check{\theta}_\nu - \theta_\nu) \xrightarrow{C} N(0,\ D_2).$$

Note that the above $n^{-1/2}$ rate of convergence is achieved when estimators are pseudo-quadratic, since $D_2 n^{-1}$ is one of the terms in the asymptotic variance of $2\mathbf{m}^{\mathrm{T}} A_\nu \varepsilon$.

Nonparametric estimation of θ_ν has important implications for the choice of the smoothing parameter for local regression. Based on the theory in this paper, Fan and Huang (1996) show that the relative rate of convergence of the bandwidth selector proposed by Fan and Gijbels (1995) for fitting local linear regression is of order $n^{-2/7}$ if local cubic fitting is used at the pilot stage, and the rate can be improved to $n^{-2/5}$ when a local polynomial of degree 5 is used in the pilot fitting. The theoretical results appear to be new in the nonparametric regression literature.

7. Proofs

Proof of Lemma 2.1. Stone (1980) shows that

$$\sup_{x \in [c,d]} \left| n^{-1} h^{-1} \sum_{i=1}^{n} (X_i - x)^j h^{-j} K\left(\frac{X_i - x}{h}\right) - f(x)\mu_j \right| \xrightarrow{\text{a.s.}} 0,\ j = 0, \ldots, 2p.$$

Substituting $f(x)\mu_j$, $j = 0, \ldots, 2p$, for the corresponding moments in the matrix $(X^{\mathrm{T}}WX)$, we obtain

$$\|n^{-1}h^{-1}H^{-1}(X^{\mathrm{T}}WX)H^{-1} - f(x)S\|_{\infty} \xrightarrow{\text{a.s.}} 0, \tag{7.1}$$

where $H = \mathrm{diag}(1, h, h^2, \ldots, h^p)_{(p+1)\times(p+1)}$. Lemma 2.1 follows from (2.7) and (7.1). □

Proof of Theorem 4.1. We begin by estimating the conditional bias using (3.3). From Lemma 2.1 and Condition A1, it can be shown that

$$\begin{aligned}
\mathrm{tr}(A_\nu) &= \sum_{i=1}^{n} \int \left(W_\nu^n \left(\frac{X_i - x}{h} \right) \right)^2 w(x)\,\mathrm{d}x \\
&= \frac{1}{(nh^{\nu+1})^2} \sum_{i=1}^{n} \int \left(\frac{1}{f(x)} K_\nu^* \left(\frac{X_i - x}{h} \right) + o_P(1) I_{[X_i - h, X_i + h]}(x) \right)^2 w(x)\,\mathrm{d}x \\
&= \frac{1 + o_P(1)}{(nh^{\nu+1})^2} \sum_{i=1}^{n} \int \frac{1}{f(x)^2} K_\nu^{*2} \left(\frac{X_i - x}{h} \right) w(x)\,\mathrm{d}x.
\end{aligned} \tag{7.2}$$

Since

$$\mathrm{var}\left(\sum_{i=1}^{n} K_\nu^{*2} \left(\frac{X_i - x}{h} \right) \right) = O(nh) = o\left(\left(n\mathrm{E}\left\{ K_\nu^{*^2} \left(\frac{X_i - x}{h} \right) \right\} \right)^2 \right),$$

the summation in (7.2) can be replaced by its expectation:

$$\mathrm{tr}(A_\nu) = \frac{1}{nh^{2\nu+1}} \left(\int K_\nu^{*2}(t)\mathrm{d}t \right) \left(\int f(x)^{-1} w(x)\,\mathrm{d}x \right) + o_P(n^{-1}h^{-2\nu-1}). \tag{7.3}$$

We now evaluate the first term of (3.3). A Taylor series expansion gives

$$\begin{aligned}
\sum_{i=1}^{n} W_\nu^n \left(\frac{X_i - x}{h} \right) m(X_i) &= \sum_{i=1}^{n} W_\nu^n \left(\frac{X_i - x}{h} \right) [m(x) + m'(x)(X_i - x) + \ldots] \\
&= \beta_\nu(x) + r(x),
\end{aligned} \tag{7.4}$$

where $r(x)$ denotes the remainder terms. If $s > (p + \nu + 1)/2$, assume initially that $m(\cdot)$ has $p + 1$ bounded derivatives. It is shown in Ruppert and Wand (1994) that

$$r(x) = h^{p+1-\nu} \left(\int t^{p+1} K_\nu^*(t)\mathrm{d}t \right) \beta_{p+1}(x) + o_P(h^{p+1-\nu}). \tag{7.5}$$

Hence

$$\mathbf{m}^{\mathrm{T}} A_\nu \mathbf{m} = \int \beta_\nu^2(x) w(x)\,\mathrm{d}x + 2h^{p+1-\nu}\left(\int t^{p+1} K_\nu^*(t)\mathrm{d}t\right) \times \left(\int \beta_{p+1}(x)\beta_\nu(x)w(x)\,\mathrm{d}x\right) + o_P(h^{p+1-\nu}). \tag{7.6}$$

As noted in Remark 1, $\int \beta_{p+1}(x)\beta_\nu(x)w(x)\,\mathrm{d}x$ depends only on the first $(p+\nu+1)/2$ derivatives of $m(\cdot)$. Since any function with $(p+\nu+1)/2$ derivatives can be approximated arbitrarily close by a function with $p+1$ derivatives, (7.6) must hold for $s > (p+\nu+1)/2$. Similar arguments can be seen in Hall and Marron (1991, p. 170).

If $s \leqslant (p+\nu+1)/2$, set $l = [s] + [s-\nu]$. Assume initially that $m^{(l)}$ exists and is bounded (note that $l < (p+1)$). From a Taylor expansion with an integral remainder,

$$m(X_i) = m(x) + \sum_{k=1}^{l-1}(X_i - x)^k m^{(k)}(x)/k! + \frac{(X_i - x)^l}{(l-1)!}\int_0^1 (1-t)^{l-1} m^{(l)}(x + t(X_i - x))\mathrm{d}t, \tag{7.7}$$

for $i = 1, \ldots, n$. It follows from (7.4) and (7.7) that

$$\mathbf{m}^{\mathrm{T}} A_\nu \mathbf{m} = \int \beta_\nu^2(x) w(x)\,\mathrm{d}x + \int \beta_\nu(x) r(x) w(x)\,\mathrm{d}x + O_P(h^{2(s-\nu)}). \tag{7.8}$$

The second term on the right-hand side of (7.8) can be written as

$$\begin{aligned}
&\sum_{i=1}^n \int_0^1 \frac{(1-t)^{l-1}}{(l-1)!}\int \frac{\mathrm{d}^{[s-\nu]}}{\mathrm{d}x^{[s-\nu]}}\left(\beta_\nu(x) W_\nu^n\left(\frac{X_i - x}{h}\right)(X_i - x)^l w(x)\right) \\
&\quad \times (m^{([s])}(x + t(X_i - x)) - m^{([s])}(x))\,\mathrm{d}x\,\mathrm{d}t \\
&= h^{-\nu-1}\int_0^1 \frac{(1-t)^{l-1}}{(l-1)!}\mathrm{E}\left\{\int \frac{\mathrm{d}^{[s-\nu]}}{\mathrm{d}x^{[s-\nu]}}\left(\beta_\nu(x) K_\nu^*\left(\frac{X_i - x}{h}\right)(X_i - x)^l w(x) f^{-1}(x)\right)\right. \\
&\quad \left. \times (m^{([s])}(x + t(X_i - x)) - m^{([s])}(x))\mathrm{d}x\right\}\mathrm{d}t + O_p(h^{l-\nu}a_n) \\
&= h^{l-\nu}\int_0^1 \frac{(1-t)^{l-1}}{(l-1)!}\int\int \frac{\mathrm{d}^{[s-\nu]}}{\mathrm{d}x^{[s-\nu]}}(\beta_\nu(x) K_\nu^*(u) u^l w(x) f^{-1}(x)) \\
&\quad \times (m^{([s])}(x + htu) - m^{([s])}(x)) f(x + hu)\mathrm{d}u\,\mathrm{d}x\,\mathrm{d}t + O_P(h^{l-\nu}a_n),
\end{aligned} \tag{7.9}$$

where $a_n = h + (nh)^{-1/2}$ is obtained by more careful evaluation of the $o_P(1)$ term in Lemma 2.1. It follows from (7.8), (7.9) and (4.1) that if $s \leqslant (p+\nu+1)/2$, then

$$\mathbf{m}^{\mathrm{T}} A_\nu \mathbf{m} = \int \beta_\nu^2(x) w(x)\,\mathrm{d}x + O_P(h^{s+[s]-2\nu}). \tag{7.10}$$

The combination of (7.3), (7.6) and (7.10) gives the bias expression in Theorem 4.1(a).

Next, we compute the asymptotic conditional variance of $\hat{\theta}_\nu$. In what follows, we treat

explicitly the three terms given in the right-hand side of (3.4). The first term (omitting the factor $4\sigma^2$) is

$$\begin{aligned}\mathbf{m}^{\mathrm{T}} A_\nu^2 \mathbf{m} &= \sum_{j=1}^{n} \left(\sum_{i=1}^{n} a_{i,j}(\nu) m(X_i) \right)^2 \\ &= \sum_{j=1}^{n} \sum_{i=1}^{n} \sum_{k=1}^{n} a_{i,j}(\nu) m(X_i) a_{k,j}(\nu) m(X_k) \\ &= \sum_{i,j,k \text{ all different}} + \sum_{i=k\neq j} + 2 \sum_{i=j\neq k} + \sum_{i=j=k} . \end{aligned} \tag{7.11}$$

Now

$$\begin{aligned}\sum_{i=k\neq j} &= \sum_{i\neq j} m^2(X_i) \left(\int W_\nu^n \left(\frac{X_i - x}{h} \right) W_\nu^n \left(\frac{X_j - x}{h} \right) w(x)\, \mathrm{d}x \right)^2 \\ &= \sum_{i\neq j} \frac{m^2(X_i)}{n^4 h^{4\nu+4}} \left\{ \int K_\nu^* \left(\frac{X_i - x}{h} \right) K_\nu^* \left(\frac{X_j - x}{h} \right) w(x) f^{-2}(x) \right. \\ &\quad \left. \times (1 + o_P(1)) I_{[-h+X_i, h+X_i]}(x) I_{[-h+X_j, h+X_j]}(x))\, \mathrm{d}x \right\}^2 \\ &= n^{-2} h^{-4\nu-2} \iint \left\{ \int K_\nu^*(u_1) K_\nu^* \left(u_1 + \frac{x-y}{h} \right) m^2(x + hu_1) f(x + hu_1) \right. \\ &\quad \left. \times (1 + o_P(1)) I_{[-1,1]}(u_1) I_{[-1-(x-y)/h, 1-(x-y)/h]}(u_1))\mathrm{d}u_1 \right\} \\ &\quad \times \left\{ \int K_\nu^*(u_2) K_\nu^* \left(u_2 + \frac{x-y}{h} \right) f(y + hu_2)(1 + o_P(1) I_{[-1,1]}(u_2) \right. \\ &\quad \left. \times I_{[-1-(x-y)/h, 1-(x-y)/h]}(u_2))\mathrm{d}u_2 \right\} w(x) w(y) f^{-2}(x) f^{-2}(y)\, \mathrm{d}x\, \mathrm{d}y \\ &\quad + O_P(n^{-3} h^{-4\nu-3/2}) \\ &= n^{-2} h^{-4\nu-1} (1 + o_P(1)) \left(\int (K_\nu^* * K_\nu^*(z))^2\, \mathrm{d}z \right) \\ &\quad \times \iint m^2(x) f^{-1}(x) f^{-1}(y) w(x) w(y)\, \mathrm{d}x\, \mathrm{d}y. \end{aligned} \tag{7.12}$$

By (7.4) and Lemma 2.1,

$$\sum_{i,j,k \text{ all different}} = n^{-1}h^{-2\nu}\iiint(\beta_\nu(y)+r(y))(\beta_\nu(y+hz)+r(y+hz))K_\nu^*(t)$$

$$\times K_\nu^*(t+z)f(y+h(z+t))w(y)w(y+hz)f^{-1}(y+hz)$$

$$\times f^{-1}(y)\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}y + O_P(\sqrt{nh}n^{-2}h^{-2\nu-2})$$

$$= n^{-1}h^{-2\nu}\iiint g_\nu(y)g_\nu(y+hz)K_\nu^*(t)K_\nu^*(t+z)$$

$$\times f(y+h(z+t))\mathrm{d}t\,\mathrm{d}z\,\mathrm{d}y + O_P(\sqrt{nh}n^{-2}h^{-2\nu-2}), \tag{7.13}$$

with $g_\nu(y) = (\beta_\nu(y)+r(y))w(y)f^{-1}(y)$. If $s>2\nu$, it is assumed initially that $s>3\nu$. Then write

$$g_\nu(y+hz) = \sum_{l=0}^{2\nu} g_\nu^{(l)}(y)h^l z^l/l! + o(h^{2\nu}) \tag{7.14}$$

and

$$f(y+h(z+t)) = \sum_{k=0}^{\nu} f^{(k)}(y)h^k(z+t)^k/k! + o(h^\nu). \tag{7.15}$$

Note that from (2.10), for non-negative integers l, k, and $l+k \leqslant 2\nu$,

$$\iint z^l(z+t)^k K_\nu^*(t)K_\nu^*(t+z)\mathrm{d}t\,\mathrm{d}z = \begin{cases} (-1)^\nu \dfrac{l!}{\nu!(l-\nu)!} & \text{if } l \geqslant \nu,\ k \leqslant \nu \\ & \text{and } l+k=2\nu, \\ 0 & \text{elsewhere.} \end{cases} \tag{7.16}$$

Combining (7.13)–(7.16),

$$\sum_{i,j,k \text{ all different}} = \frac{(-1)^\nu}{n(\nu!)^2}\int g_\nu(y)\left(\sum_{i=0}^{\nu} g_\nu^{(\nu+i)}(y)f^{(\nu-i)}(y)\frac{\nu!}{i!(\nu-i)!}\right)\mathrm{d}y + o_P(n^{-1})$$

$$= \frac{(-1)^\nu}{n(\nu!)^2}\int g_\nu(y)\frac{\mathrm{d}^\nu}{\mathrm{d}y^\nu}[g_\nu^{(\nu)}(y)f(y)]\,\mathrm{d}y + o_P(n^{-1}). \tag{7.17}$$

Applying integration by parts and Condition A3 to (7.17), it follows that

$$\sum_{i,j,k \text{ all different}} = \frac{1}{n(\nu!)^2}\int [g_\nu^{(\nu)}(y)]^2 f(y)\mathrm{d}y + o_P(n^{-1})$$

$$= \frac{1}{n(\nu!)^2}\int [G_\nu^{(\nu)}(y)]^2 f(y)\mathrm{d}y + o_P(n^{-1}), \tag{7.18}$$

with $G_\nu(y) = \beta_\nu(y)w(y)f^{-1}(y)$. Expression (7.18) only involves derivatives of $m(\cdot)$ up to the 2νth, and hence must hold for $s > 2\nu$. To modify arguments (7.13)–(7.18) for the case of $s \leqslant 2\nu$, the change required is in (7.14) and (7.15). The initial assumption is that $m(\cdot)$ has $(2[s] - \nu)$ bounded derivatives, and Taylor expansions of orders $2([s] - \nu)$ and $(2\nu - [s])$ for $(g_\nu(y + hz) - g_\nu(y))$ and $f(y + h(z + t))$, respectively. Then, applying integration by parts $[s] - \nu$ times, together with the smoothness definition (4.1), we obtain

$$\sum_{i,j,k \text{ all different}} = O_P(n^{-1}h^{-2\nu+s}).$$

The other two terms in (7.11), $2\sum_{i=j\neq k}$ and $\sum_{i=j=k}$, are of smaller order than $\sum_{i,j,k \text{ all different}}$ and $\sum_{i=k\neq j}$.

The second term in the right-hand side of (3.4) (omitting the factor $2\sigma^4$) is

$$\begin{aligned}\operatorname{tr}(A_\nu^2) &= \sum_{i=1}^n a_{i,i}^2(\nu) + \sum_{i\neq j} a_{i,j}^2(\nu) \\ &\equiv I_1 + I_2,\end{aligned}$$

where I_1 and I_2 denote the diagonal and non-diagonal terms, respectively. By means of arguments similar to those used in establishing the term $\mathbf{m}^{\mathrm{T}}A_\nu^2\mathbf{m}$, we obtain that $I_1 = O_P(n^{-3}h^{-4\nu-2})$ and

$$\begin{aligned}I_2 &= \frac{n(n-1)(1+o_P(1))}{n^4h^{4\nu+4}} \sum_{i\neq j} \mathrm{E}\left\{\int K_\nu^*\left(\frac{X_i - x}{h}\right) K_\nu^*\left(\frac{X_j - x}{h}\right) f(x)^{-2}w(x)\,\mathrm{d}x\right\}^2 \\ &= n^{-2}h^{-4\nu-1}\iint \{(K_\nu^* * K_\nu^*)(z)\}^2\,\mathrm{d}z f(x)^{-2}w^2(x)\,\mathrm{d}x + o_P(n^{-2}h^{-4\nu-1}).\end{aligned}$$

Consequently,

$$\operatorname{tr}(A_\nu^2) = n^{-2}h^{-4\nu-1}(1 + o_P(1))\left(\int [(K_\nu^* * K_\nu^*)(z)]^2\,\mathrm{d}z\right)\left(\int f(x)^{-2}w^2(x)\,\mathrm{d}x\right). \tag{7.19}$$

Observe that the last term (omitting the factor $\mathrm{E}(\varepsilon_1^4) - 3\sigma^4$) in the right-hand side of (3.4) is

$$\sum_{i=1}^n a_{i,i}^2(\nu) = I_1 = o_P(\operatorname{tr}(A_\nu^2)). \tag{7.20}$$

The asymptotic variance follows from (7.12), (7.18), (7.19) and (7.20). □

Proof of Theorem 4.3. This theorem follows directly by the following two expressions:

$$\begin{aligned}
\mathrm{E}(\breve{\theta}_\nu) &= \mathrm{E}(\hat{\theta}_\nu) - \mathrm{E}(\hat{\sigma}^2 \mathrm{tr}(A_\nu)) \\
&= \mathrm{E}(\hat{\theta}_\nu) - \sigma^2 \mathrm{tr}(A_\nu) + \mathrm{E}((\hat{\sigma}^2 - \sigma^2)\mathrm{tr}(A_\nu)) \\
&= \mathrm{E}(\hat{\theta}_\nu) - \sigma^2 \mathrm{tr}(A_\nu) + O_P(\epsilon_n n^{-1} h^{-2\nu-1}). \\
\mathrm{var}(\breve{\theta}_\nu) &= \mathrm{var}(\hat{\theta}_\nu - (\hat{\sigma}^2 - \sigma^2)\mathrm{tr}(A_\nu)) \\
&= \mathrm{var}(\hat{\theta}_\nu) + O_P(\epsilon_n^2 n^{-2} h^{-4\nu-2}).
\end{aligned}$$

Proof of Theorem 5.1. If the linear term dominates, i.e. $\mathrm{tr}(A^2)/\mathbf{m}^{\mathrm{T}} A^2 \mathbf{m} \xrightarrow{P} 0$ (case (i)), then $\varepsilon^{\mathrm{T}} A \varepsilon - \mathrm{E}(\varepsilon^{\mathrm{T}} A \varepsilon) = o_P\big((\mathbf{m}^{\mathrm{T}} A^2 \mathbf{m})^{1/2}\big)$. Conditioning on $\{X_1, \ldots, X_n\}$, it is trivial that $\mathbf{m}^{\mathrm{T}} A \varepsilon$ is normally distributed with mean 0 and variance $\sigma^2 \mathbf{m}^{\mathrm{T}} A^2 \mathbf{m}$. This implies

$$\begin{aligned}
(4\sigma^2 \mathbf{m}^{\mathrm{T}} A^2 \mathbf{m})^{1/2} (\hat{\theta} - \mathrm{E}(\hat{\theta})) &= (4\sigma^2 \mathbf{m}^{\mathrm{T}} A^2 \mathbf{m})^{1/2} \{2\mathbf{m}^{\mathrm{T}} A \varepsilon + (\varepsilon^{\mathrm{T}} A \varepsilon - \mathrm{E}(\varepsilon^{\mathrm{T}} A \varepsilon))\} \\
&\xrightarrow{C} N(0, 1).
\end{aligned}$$

For case (ii), $\mathrm{tr}(A^2)/\mathbf{m}^{\mathrm{T}} A^2 \mathbf{m} \xrightarrow{P} \infty$, we need only consider $\varepsilon^{\mathrm{T}} A \varepsilon$. Factor A into a product $\Gamma^{\mathrm{T}} \Lambda \Gamma$, where Γ is an orthonormal matrix and Λ is a diagonal matrix with the eigenvalues of A, λ_i, as its entries. Now

$$\varepsilon^{\mathrm{T}} A \varepsilon = \delta^{\mathrm{T}} \Lambda \delta = \sum_{i=1}^{n} \lambda_i \delta_i^2,$$

with $\delta = \Gamma \varepsilon = (\delta_1, \ldots, \delta_n)$. Note that by the assumption that $\mathrm{tr}(A^4)/\mathrm{tr}(A^2))^2 \xrightarrow{P} 0$, as $n \to \infty$,

$$\sum_{i=1}^{n} \mathrm{E}^c |\lambda_i \delta_i^2 - \lambda_i|^4 = \sum_{i=1}^{n} \lambda_i^4 \mathrm{E}^c (\delta_i^2 - 1)^4 = 60\sigma^4 \sum_{i=1}^{n} \lambda_i^4 = o_P([\mathrm{tr}(A^2)]^2),$$

where E^c denotes the conditional expectation given $\{X_1, \ldots, X_n\}$. Hence the Lindeberg condition for the asymptotic normality of $\sum_{i=1}^{n} \lambda_i \delta_i^2$ holds:

$$\frac{\sum_{i=1}^{n} \mathrm{E}^c |\lambda_i \delta_i^2 - \lambda_i|^4}{\left[\mathrm{var}\left(\sum_{i=1}^{n} \lambda_i \delta_i^2 | X_1, \ldots, X_n\right)\right]^2} = \frac{15\sigma^4 \sum_{i=1}^{n} \lambda_i^4}{[\mathrm{tr}(A^2)]^2} \xrightarrow{P} 0,$$

and so $\varepsilon^{\mathrm{T}} A \varepsilon$ is asymptotic normally distributed. Consequently, the asymptotic normality of $\hat{\theta}$ for cases (i) and (ii) is proved. □

Proof of Lemma 6.1. Note that

$$P\Big(\max_j Z_{n,j} > nhb_n\Big) \leqslant \sum_j P(Z_{n,j} > nhb_n). \tag{7.21}$$

Clearly, it is sufficient to show that $[\sum_{j=1}^n P(Z_{n,j} > nhb_n)] \to 0$.

Given an arbitrary positive constant d,

$$\begin{aligned} P(Z_{n,j} > nhb_n) &= P(\exp(Z_{n,j}d) > \exp(nhb_n d)) \\ &\leqslant \exp(-nhb_n d)\mathrm{E}\{\exp(Z_{n,j}d)\} \\ &= \exp(-nhb_n d)(1 - p_{n,j} + \mathrm{e}^d p_{n,j})^n \\ &\leqslant \exp(-nhb_n d + np_{n,j}\mathrm{e}^d). \end{aligned} \tag{7.22}$$

Take $d = \log(hb_n/p_{n,j})$ in (7.22). If n is sufficiently large, then

$$P(Z_{n,j} > nhb_n) \leqslant \left(\frac{p_{n,j}\mathrm{e}}{hbn_n}\right)^{nhb_n}. \tag{7.23}$$

Since $\max_j p_{n,j} \leqslant ch$, the right-hand side of (7.23) is bounded by

$$(ch)^{nhb_n-1}\left(\frac{\mathrm{e}}{hb_n}\right)^{nhb_n} p_{n,j} = \frac{1}{ch}\left(\frac{c\mathrm{e}}{b_n}\right)^{nhb_n} p_{n,j}.$$

Now $b_n(\log(c\mathrm{e}) - \log(b_n)) \leqslant -1$ for sufficiently large n, which, together with Condition B, gives

$$\begin{aligned} \sum_j P(Z_{n,j} > nhb_n) &\leqslant \sum_j \frac{1}{ch}\left(\frac{1}{\mathrm{e}}\right)^{nh} p_{n,j} \\ &= \frac{1}{ch}\left(\frac{1}{\mathrm{e}}\right)^{nh} \to 0. \end{aligned} \tag{7.24}$$

The conclusion follows from (7.21) and (7.24). □

Proof of Lemma 6.2. Observe that from (7.19)

$$\mathrm{tr}(A_\nu^2) \geqslant c_1 n^{-2}h^{-4\nu-1} + o_P(n^{-2}h^{-4\nu-1}) \tag{7.25}$$

for some non-negative constant c_1. It suffices to consider the rate of $\lambda_{\max}(\nu)$. From basic theory of linear algebra,

$$\lambda_{\max}(\nu) = \sup_{\|b\|=1} \sum_{i=1}^n \sum_{j=1}^n a_{i,j}(\nu) b_i b_j. \tag{7.26}$$

From Lemma 2.1, for n sufficiently large,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}(\nu) b_i b_j \leqslant \frac{2}{(nh^{\nu+1})^2} \int f(x)^{-2} \left(\sum_{i=1}^{n} K_\nu^* \left(\frac{X_i - x}{h} \right) b_i \right)^2 w(x)\, \mathrm{d}x. \tag{7.27}$$

Under Conditions A, the right-hand side of (7.27) is further bounded by

$$\begin{aligned} & c_2 n^{-2} h^{-2\nu-2} \int \left(\sum_{i=1}^{n} |b_i| I_{[X_i - h, h + X_i]}(x) \right)^2 \mathrm{d}x \\ & = c_2 n^{-2} h^{-2\nu-2} \left(\sum_{i=1}^{n} b_i^2 \int I_{[X_i - h, X_i + h]}(x)\, \mathrm{d}x \right. \\ & \quad \left. + \sum_{i \neq j} |b_i|\, |b_j| \int I_{[X_i - h, X_i + h]}(x) I_{[X_j - h, X_j + h]}(x)\, \mathrm{d}x \right) \\ & \leqslant c_2 n^{-2} h^{-2\nu-2} \left(2h \sum_{i=1}^{n} b_i^2 + 2h \sum_{i \neq j} I_{[|X_i - X_j| \leqslant 2h]} |b_i|\, |b_j| \right) \end{aligned} \tag{7.28}$$

for some non-negative constant c_2.

We use the following argument to evaluate the double summation in (7.28). Let $I_k = (4(k-1)h, 4kh]$, $k \in \mathbb{Z}$ be a partition of $(-\infty, \infty)$, and n_k be the number of design points X_i that fall in the interval I_k. Since $I_{[|X_i - X_j| \leqslant 2h]} = 0$ for $X_i \in I_{k_1}$, $X_j \in I_{k_2}$, and $|k_1 - k_2| > 1$,

$$\sum_{i,j} I_{[|X_i - X_j| \leqslant 2h]} |b_i|\, |b_j| \leqslant \sum_{k=-\infty}^{\infty} \left(\sum_{X_i, X_j \in I_k \cup I_{(k+1)}} |b_i|\, |b_j| \right) = \sum_{k=-\infty}^{\infty} \left(\sum_{X_i \in I_k \cup I_{(k+1)}} |b_i| \right)^2. \tag{7.29}$$

Put $N = \max_{k \in \mathbb{Z}} (n_k + n_{k+1}) \equiv (n_{k_0} + n_{k_0+1})$ for some $k_0 \in \mathbb{Z}$. The maximum of the right-hand side of (7.29) over $\|b\| = 1$ is attained at $|b_i| = 1/\sqrt{N}$ for those indices such that $X_i \in I_{k_0} \cup I_{k_0+1}$. Therefore,

$$\max_{\|b\|=1} \sum_{i,j} I_{[|X_i - X_j| \leqslant 2h]} |b_i| |b_j| \leqslant N.$$

Note that the $\{n_k\}$ have a multinomial distribution and hence, by Lemma 6.1, $N = O_P(nh\alpha_n)$, for any sequence $\alpha_n \to \infty$. Taking $\alpha_n = h^{-1/4}$, it follows from (7.28) that

$$\lambda_{\max}(\nu) \leqslant c_2 n^{-2} h^{-2\nu-2} O_P(h + nh^{7/4}). \tag{7.30}$$

Combine (7.25) and (7.30) to conclude that

$$\frac{\lambda_{\max}^2(\nu)}{\mathrm{tr}(A_\nu^2)} = O_P(n^{-2}h^{-1} + h^{1/2}) \to 0. \qquad \square$$

Proof of Theorem 6.1. From Theorem 5.1, one only needs to verify

$$\frac{\mathrm{tr}(A_\nu^4)}{(\mathrm{tr}(A_\nu^2))^2} \xrightarrow{P} 0, \quad \text{as } n \to \infty.$$

Let $\lambda_i(\nu)$, $i = 1, \ldots, n$, be eigenvalues of A_ν and $\lambda_{\max}(\nu)$ be the maximum eigenvalue among all $\lambda_i(\nu)$s. Then,

$$\frac{\mathrm{tr}(A_\nu^4)}{(\mathrm{tr}(A_\nu^2))^2} = \frac{\sum_{i=1}^n \lambda_i^4(\nu)}{\left(\sum_{i=1}^n \lambda_i^2(\nu)\right)^2} \leqslant \frac{\lambda_{\max}^2(\nu)}{\sum_{i=1}^n \lambda_i^2(\nu)} = \frac{\lambda_{\max}^2(\nu)}{\mathrm{tr}(A_\nu^2)} \xrightarrow{P} 0, \quad \text{as } n \to \infty. \tag{7.31}$$

Hence the asymptotic normality of $\hat{\theta}_\nu$ and $\check{\theta}_\nu$ follows. □

Acknowledgements

This work was part of Li-Shan Huang's doctoral dissertation at the University of North Carolina. She would like to thank the Department of Statistics, University of North Carolina, National Institute of Statistical Sciences, Professors M.R. Leadbetter and R.L. Smith for the financial support during her graduate studies.

References

Birgé, L. and Massart, P. (1995) Estimation of integral functionals of a density. *Ann. Statist.,* **23**, 11–29.

Bickel, P.J. and Ritov, Y. (1988) Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A*, **50**, 381–393.

Doksum, K. and Samarov, A. (1995) Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *Ann. Statist.,* **23**, 1443–1473.

Donoho, D.L. and Nussbaum, M. (1990) Minimax quadratic estimation of a quadratic functional. *J. Complexity,* **6**, 290–323.

Efromovich, S. and Low, M. (1996) On Bickel and Ritov's conjecture about adaptive estimation of the integral of the square of density derivative. *Ann. Statist.*, **24**, 682–686.

Fan, J. (1991) On the estimation of quadratic functionals. *Ann. Statist.,* **19**, 1273–1294.

Fan, J. and Gijbels, I. (1995) Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B,* **57**, 371–394.

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

Fan, J. and Huang, L.-S. (1996) Rates of convergence for the pre-asymptotic substitution bandwidth selector. Technical report no. 912, Department of Statistics, Florida State University.

Hall, P. and Marron, J.S. (1987) Estimation of integrated squared density derivatives. *Statist. Probab. Lett.,* **6**, 109–115.

Hall, P. and Marron, J.S. (1991) Lower bounds for bandwidth selection in density estimation. *Probab.*

Theory Related Fields, **90**, 149–173.

Hall, P., Kay, J.W. and Titterington, D.M. (1990) Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika,* **77**, 521–528.

Jones, M.C. and Sheather, S.J. (1991) Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statist. Probab. Lett.,* **11**, 511–514.

Khatri, C.G. (1980) Quadratic forms in normal variables. In P.R. Krishnaiah (ed.), *Handbook of Statistics*, Vol. I, pp. 443–469. New York: North-Holland.

Laurent, B. (1996) Efficient estimation of integral functionals of a density. *Ann. Statist.*, **24**, 659–681.

Laurent, B. (1997) Estimation of integral functionals of a density and its derivatives. *Bernoulli*, **3**, 181–211.

Rice, J. (1984) Bandwidth choice for nonparametric regression. *Ann. Statist.,* **12**, 1215–1230.

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

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

Stone, C.J. (1980) Optimal rates of convergence for nonparametric estimators. *Ann. Statist.,* **8**, 1348–1360.

Whittle, P. (1964) On the convergence to normality of quadratic forms in independent random variables. *Theory Probab. Appl.,* **9**, 103–108.

Received March 1997 and revised April 1998