NONPARAMETRIC ESTIMATION OF QUADRATIC REGRESSION FUNCTIONALS

BY

LI - SHAN HUANG*, & JIANQING FAN+

FSU STATISTICS REPORT M905
Nonparametric Estimation of Quadratic Regression Functionals

Li-Shan Huang,* and Jianqing Fan+

October 3, 1995

FSU Technical Report No. M-905

*Department of Statistics
Florida State University
Tallahassee, FL 32306-3033

+Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260
Nonparametric Estimation of Quadratic Regression Functionals

Li-Shan Huang
Department of Statistics
Florida State University
Tallahassee, F.L. 32306-3033

Jianqing Fan
Department of Statistics
University of North Carolina
Chapel Hill, N.C. 27599-3260

Abstract: Quadratic regression functionals are important for the bandwidth selection problem of nonparametric regression techniques. 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 with appropriate values of the smoothing parameter and the amount of smoothness assumed. Asymptotic distributions of the 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 the $n^{-1/2}$ rate can be achieved.

Key words and phrases. Asymptotic normality, estimation of quadratic functionals, local polynomial regression.

Abbreviated title: Estimation of Quadratic Regression Functionals.

AMS 1991 subject classification. Primary 62G07; Secondary 60F05.

1 Introduction

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

$$\theta^*_\nu = \int [m^{(\nu)}(x)]^2 w(x) dx, \quad \nu \geq 0,$$

where $w(x)$ is a prescribed nonnegative weight function. These functionals appear in expressions for the asymptotically optimal bandwidth for nonparametric function estimates; see for example Ruppert and Wand (1994). Doksum and Samarov (1995) consider particularly the applications of $\theta^*_0$. Another area of application is to examine the goodness of fit of a $(\nu - 1)$-th degree polynomial model by testing $H_0 : \theta^*_\nu = 0$. When $w(\cdot)$ is the indicator function of an interval, $H_0$ tests the polynomial modeling at that given interval. Such tests are useful diagnostic tools when one wishes to extrapolate the model fitting beyond the boundaries of design points.

Problems similar to estimation of $\theta^*_\nu$ are considered in a number of papers. Hall and Marron (1987, 1991), Bickel and Ritov (1988), and Jones and Sheather (1991) discuss estimation of $\int [f^{(\nu)}(x)]^2 dx$, where $f(\cdot)$ is a probability density function. Optimality results of estimating $\int [m^{(\nu)}(x)]^2 dx$ under the Gaussian white noise model are obtained in Donoho and Nussbaum (1990) and Fan (1991). In the context of nonparametric regression, Doksum and Samarov (1995) give estimators of $\int m^2(x)f(x)w(x) dx$ ($f(\cdot)$ is the design density function) by local constant approximation,
i.e. the Nadaraya-Watson estimator. Here we develop estimates of $\theta^*_v$ based on local polynomial regression estimators. Ruppert, Sheather, and Wand (1995) consider functionals similar to $\theta^*_v$ with weight $w(x) = f(x)$. However, no results as general as shown in this paper have been established and our estimators for $\theta^*_v$ are different from that of Doksum and Samarov (1995) and Ruppert, Sheather, and Wand (1995).

Technically, $\theta^*_v$'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 $\int [f^{(v)}(x)]^2 dx$ is established in Hall and Marron (1987, 1991) and Bickel and Ritov (1988), using integration by parts: $\int [f^{(v)}(x)]^2 dx = (-1)^v \int f^{(2v)}(x) f(x) dx$. But this identity does not hold when a weight function is introduced. Note that the weight scheme in (1.1) has to be introduced for regression problems due to lack of data at the tails. Hence, a natural question is whether $n^{-1/2}$-consistent estimators can be constructed for $\theta^*_v$. Section 4 describes the asymptotic behavior of quadratic functionals of weighted regression estimates that are used to estimate $\theta^*_v$. 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 $\sum_{i=1}^n \sum_{j=1}^n a_{ij} Y_i Y_j$ 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 $\theta^*_v$ can be achieved when estimators are pseudo-quadratic (linear component dominates). For genuine-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 $\theta^*_v$ are given in Section 3. Section 4 contains rate-of-convergence results in mean square error, from both a theoretical and a practical points of view. Asymptotic distributions of the proposed estimators are studied in Sections 5 and 6. Proofs of lemmas and theorems are in Section 7.

2 Local polynomial regression

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

$$Y_i = m(X_i) + \varepsilon_i, \quad E(\varepsilon_i) = 0, \text{Var}(\varepsilon_i) = \sigma^2,$$

(2.1)

where $m(\cdot)$ is an unknown regression function, $\varepsilon_i$'s are independent and identically distributed error terms, and $X_i$'s are independent of $\varepsilon_i$'s. 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 nonnegative integer. For a fixed point \(x\), the regression function \(m(\cdot)\) can be locally approximated by

\[
m(z) \approx m(x) + m'(x)(z-x) + \ldots + m^{(p)}(x)(z-x)^p / p!,
\]

for \(z\) in a neighborhood of \(x\). This leads to the following least squares problem:

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

where \(\beta_\nu = m^{(\nu)}(x)/\nu!\), \(\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T\), and the dependence of \(\beta_\nu\) on \(x\) is suppressed. The neighborhood 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 \(p\)-th degree weighted polynomial in a neighborhood of \(x\). See also Fan (1992), Fan and Gijbels (1995), Hastie and Loader (1993), and Ruppert and Wand (1994) for further details.

For convenience, some matrix notation is introduced here. Let

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

\[
W = \text{diag}(K(\frac{X_i - x}{h}))_{n \times n} = \begin{pmatrix}
K(\frac{X_1 - x}{h}) & \ldots \\
\vdots & \ddots \\
K(\frac{X_n - x}{h})
\end{pmatrix}, \tag{2.4}
\]

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

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

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

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

3
where

\[ W_\nu^n(t) = e_\nu^T (X^T W X)^{-1} (1, ht, \ldots, h^p t^p)^T K(t), \quad \nu = 0, \ldots, p. \]  \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

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

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 behavior of local polynomial regression.

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

\[ K_\nu^*(t) = e_\nu^T S^{-1} (1, t, \ldots, t^p)^T K(t), \quad \nu = 0, \ldots, p, \]  \tag{2.9}

where \( S = (\mu_{i+j-2})_{(p+1) \times (p+1)} \) with \( \mu_k = \int t^k K(t) dt. \) 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

\[ \int t^k K_\nu^*(t) dt = \begin{cases} 1 & \text{if } k = \nu, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } 0 \leq \nu, k \leq p. \tag{2.10} \]

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 for \([c, d] \subset [a, b],\)

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

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 behavior of \( \hat{\beta}. \)
3 Estimators

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

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

is considered. Correspondingly, denote

$$\theta_\nu = \int \beta^2_\nu(x) w(x) dx = \frac{1}{(\nu!)^2} \int [m^{(\nu)}(x)]^2 w(x) dx.$$

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 $E(\varepsilon_i^3) = 0$ and $E(\varepsilon_i^4) < \infty$, $i = 1, \ldots, n$.

From (2.6),

$$\hat{\theta}_\nu = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j}(\nu)Y_iY_j,$$  \hspace{1cm} (3.1)

where

$$a_{i,j}(\nu) = \int W^*_\nu \left( \frac{X_i - x}{h} \right) W^*_\nu \left( \frac{X_j - x}{h} \right) w(x) dx.$$  \hspace{1cm} (3.2)

Clearly $\hat{\theta}_\nu$ is in a quadratic form of $Y_i$'s and can be written as

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

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):

$$E\{\hat{\theta}_\nu | X_1, \ldots, X_n \} = \mathbf{m}^T A_\nu \mathbf{m} + \sigma^2 \text{tr}(A_\nu),$$  \hspace{1cm} (3.3)

$$\text{Var}\{\hat{\theta}_\nu | X_1, \ldots, X_n \} = 4\sigma^2 \mathbf{m}^T A^2_\nu \mathbf{m} + 2\sigma^4 \text{tr}(A^2_\nu)$$

$$+ (E(\varepsilon_1^4) - 3\sigma^4) \sum_{i=1}^{n} a^2_{i,i}(\nu),$$  \hspace{1cm} (3.4)

where $\text{tr}(A_\nu)$ denotes the trace of $A_\nu$. Note that the second term in the RHS of (3.3) may be thought of as extra bias in the estimation. This motivates a “bias-corrected” estimate:

$$\hat{\theta}_\nu = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j}(\nu)Y_iY_j - \hat{\sigma}^2 \text{tr}(A_\nu),$$  \hspace{1cm} (3.5)

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

5
4 Rates of convergence

Asymptotic expressions for the conditional bias and variance of $\hat{\theta}_\nu$ and $\hat{\nu}_\nu$ will be derived in this section. We shall use these to study how the mean square errors (MSE) 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 $f^{(\nu)}(\cdot)$ exists on $[-1,1]$.

(A3) The weight $w(\cdot)$ is a bounded and nonnegative 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. Condition (A3) is imposed to eliminate “boundary effects” so that deeper and cleaner results can be obtained. 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 $[-1,1]$,

$$|m^{(l)}(x) - m^{(l)}(y)| \leq M|x - y|^s, \quad \text{where } s = l + \alpha \text{ and } 0 < \alpha \leq 1. \quad (4.1)$$

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 $p \geq \nu$ and $s > \nu$. When estimating the $\nu$-th derivative of $m(\cdot)$, it is argued in Ruppert and Wand (1994) and Fan and Gijbels (1995) that $(p - \nu)$ should be odd. We assume that $(p - \nu)$ is 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 $(\phi_1 \ast \phi_2)(u) = \int \phi_1(t)\phi_2(u - t)dt$ denote the convolution of two real-valued functions $\phi_1$ and $\phi_2$.

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

$$E(\hat{\theta}_\nu) - \theta_\nu \equiv B_{n,\nu}$$
\[ \begin{cases} O_P(h^{s-[s]-2\nu}) + (C_2 + o_P(1))n^{-1}h^{-2\nu-1} & \text{if } s \leq (p + \nu + 1)/2, \\ (C_1 + o_P(1))h^{p+1-\nu} + (C_2 + o_P(1))n^{-1}h^{-2\nu-1} & \text{if } s > (p + \nu + 1)/2, \end{cases} \]

(4.2)

where

\[ C_1 = 2 \left( \int t^{p+1}K^*_\nu(t)dt \right) \left( \int \beta_{p+1}(x)\beta_{\nu}(x)w(x)dx \right), \]

\[ C_2 = \sigma^2 \left( \int [K^*_\nu(t)]^2 dt \right) \left( \int f(x)^{-1}w(x)dx \right); \]

(4.3)

(b) the asymptotic variance is

\[ \text{Var}(\hat{\theta}_\nu) \equiv V_{n,\nu} = \begin{cases} (D_1 + o_P(1))n^{-2}h^{-4\nu-1} + O_P(n^{-1}h^{-2\nu+s}) & \text{if } s \leq 2\nu, \\ (D_1 + o_P(1))n^{-2}h^{-4\nu-1} + (D_2 + o_P(1))n^{-1} & \text{if } s > 2\nu, \end{cases} \]

(4.4)

where

\[ D_1 = 2\sigma^2 \left( \int [(K^*_\nu \ast K^*_\nu)(z)]^2 dz \right) \left( \int f(x)^{-2}w^2(x)dx \right) + 2 \int \int m^2(x)f^{-1}(x)f^{-1}(y)w(x)w(y)dy, \]

\[ D_2 = \frac{4\sigma^2}{(\nu!)^2} \int [G^\nu(y)]^2 f(y)dy \]

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

**Remark 1:** Hall and Marron (1991) show that the bias for estimating integrated squared \( \nu \)-th density derivatives \( \int f^{(\nu)}(x)dx \) is of order \( O_P(h^{2(s-\nu)}) \) when \( s \leq (p+\nu+1)/2 \) with \( s \) the degree of smoothness of \( f(\cdot) \). Their results are based on a Lemma (Lemma 5.4 in Hall and Marron (1991)), which states as follows. Let \( H_q \) be the class of all Lipschitz functions with exponent \( q = s - [s] \).

Then for any function \( a(\cdot) \in H_q \),

\[ \int a(z)(a(z+\varepsilon) - a(\varepsilon))dz = -\int (a(z+\varepsilon) - a(z))^2dz/2 = O(\varepsilon^{2q}). \]

However, for estimating weighted integrals of squared density derivatives \( \int f^{(\nu)}(x)w(x)dx \), one has to evaluate a term of the form

\[ \int b(z)(a(z+\varepsilon) - a(z))dz, \quad a(\cdot) \in H_q, b(\cdot) \in H_q, \]
which is only of order $O(\varepsilon^9)$. Thus, when estimating weighted integrals of squared density derivatives, the bias is of order $O_P(h^{s+\lfloor s \rfloor - 2\nu})$ for $s \leq (p + \nu + 1)/2$. That also explains the bias in (4.2) for the current regression setting.

Remark 2: Note that in Theorem 4.1 (a), when $s \leq (p + 1)$, the convention of

$$\int \beta_{p+1}(x)\beta_{\nu}(x)w(x)dx,$$

which involves $m^{(p+1)}(x)$, is

$$\int \beta_{p+1}(x)\frac{dr}{dx} \beta_{\nu}(x)w(x)dx$$

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)) is that given the degree of smoothness $s$, what the best rate can be by choosing a sufficiently large $p$. 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 $h = O(n^{-1/(s+\lfloor s \rfloor + 1)})$ and $p \geq (2s - \nu - 1)$, then

$$E(\hat{\theta}_\nu - \theta_\nu)^2 = \begin{cases} 
O_P(n^{-2(s+\lfloor s \rfloor - 2\nu)/(s+\lfloor s \rfloor + 1)}) & \text{if } s \leq 2\nu, \\
(D_2 + o_P(1))n^{-1} + O_P(n^{-2(s+\lfloor s \rfloor - 2\nu)/(s+\lfloor s \rfloor + 1)}) & \text{if } s > 2\nu.
\end{cases}$$

In particular, when $s > 2\nu + 1$, $E(\hat{\theta}_\nu - \theta_\nu)^2 = D_2n^{-1} + o_P(n^{-1})$.

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

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

by taking

$$h_{OPT} = \begin{cases} 
\left(\frac{C_2}{C_1n}\right)^{1/(p+\nu+2)} & \text{if } C_1 < 0, \\
\left(\frac{C_2}{C_1n}\right)^{1/(p+\nu+2)} & \text{if } C_1 > 0.
\end{cases}$$

In particular, $E(\hat{\theta}_\nu - \theta_\nu)^2 = D_2n^{-1} + o_P(n^{-1})$, if $p > 3\nu - 1$ for the case of $C_1 < 0$, and $p > 3\nu$ for $C_1 > 0$. 

8
Remark 3: The bandwidth $h_{OPT}$ in (4.6) minimizes the mean square error of $\hat{\theta}_\nu$. Particularly, when $D_2 n^{-1}$ is the leading term in (4.5), $h_{OPT}$ minimizes the second-order terms of the MSE of $\hat{\theta}_\nu$. In this case, any $h$ satisfying

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

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

Bickel and Ritov (1988) give the efficient information bound for estimating integrals of squared density derivatives. For the current regression setting, we conjecture that $D_2 n^{-1}$ is the semiparametric information bound. In other words, $\hat{\theta}_\nu$ is efficient when the degree of smoothness is sufficiently large. But the construction of $\hat{\theta}_\nu$ 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 slower 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 favor 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 $\hat{\theta}_\nu$ defined in (3.5).

Suppose that $\sigma^2$ can be estimated at a $O(\epsilon_n)$ rate, i.e. $E[(\hat{\sigma}^2 - \sigma^2)^2] = O(\epsilon^2_n)$. 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 $\hat{\theta}_\nu$.

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

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

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

$$= \begin{cases} O_P \left(h^{s+|s|-2\nu} + \epsilon_n n^{-1} h^{-2\nu-1}\right) & \text{if } s \leq (p + \nu + 1)/2, \\ (C_1 + o_P(1)) h^{p+1-\nu} + O_P(\epsilon_n n^{-1} h^{-2\nu-1}) & \text{if } s > (p + \nu + 1)/2; \end{cases}$$

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

$$\text{Var}(\hat{\theta}_\nu) \equiv v_{n,\nu}$$

$$= \begin{cases} (D_1 + o_P(1)) n^{-2} h^{-4\nu-1} + O_P(n^{-1} h^{-2\nu+s} + \epsilon_n^2 n^{-2} h^{-4\nu-2}) & \text{if } s \leq 2\nu, \\ (D_1 + o_P(1)) n^{-2} h^{-4\nu-1} + (D_2 + o_P(1)) n^{-1} + O_P(\epsilon_n^2 n^{-2} h^{-4\nu-2}) & \text{if } s > 2\nu. \end{cases}$$
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 $\hat{\theta}_\nu$ are achieved as follows.

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

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

In particular, $E(\hat{\theta}_\nu - \theta_\nu)^2 = D_2n^{-1} + o_P(n^{-1})$ when $s > 2\nu + 0.5$.

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

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

by taking

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

In particular, $E(\hat{\theta}_\nu - \theta_\nu)^2 = D_2n^{-1} + o_P(n^{-1})$ when $p > 3\nu - 0.5$.

**Remark 4:** 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 MSE. In fact, when $D_2n^{-1}$ is the dominant term in (4.9), the $n^{-1/2}$ rate of convergence is attained for any bandwidth satisfying

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

In practical implementation one should attempt to estimate $\sigma^2$ and $h_{OPT}$ or $h_{opt}$. The parameter $\sigma^2$ can be estimated at a $n^{-1/2}$ rate using the estimators of Rice (1984) and Hall, Kay, and Titterington (1990), for example. The problem of estimating the ideal bandwidth for $\hat{\theta}_\nu$ or $\hat{\theta}_\nu$ is beyond the scope of this chapter, 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 2) by regarding the regression function as a polynomial function; estimate $f(x)$ and plug the resulting estimate in $C_2$ 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 $\hat{\theta}_\nu$ are robust against the choice of bandwidth (see Remarks 3 and 4).
Ruppert, Sheather, and Wand (1995) consider estimation of
\[ \theta_{rs} = \int m^{(r)}(x)m^{(s)}(x)f(x)dx, \quad r, s \geq 0 \text{ and } r + s \text{ even.} \] 
(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 \( \hat{m}_s(\cdot, g) \) are obtained via fitting a \( p \)-th 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 \( \hat{\theta}_\nu \) are different from \( \hat{\theta}_{rs} \).

5 Asymptotic distributions of quadratic functionals

The asymptotic distribution of
\[ \hat{\theta} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} Y_i Y_j \] 
(5.1)

will be considered in this section. For simplicity of technical arguments, we restrict attention to the Gaussian model,
\[ Y_i = m(X_i) + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d. } \sim N(0, \sigma^2), \quad i = 1, \ldots, n. \] 
(5.2)

This Gaussian noise assumption can be removed with additional work on proofs, e.g. via applying the martingale central limit theorem. The \( a_{ij} \)'s in (5.1) possibly depend on \( n \) and the design points \( \{X_1, \ldots, X_n\} \). Let \( m \) be a \( n \times 1 \) column vector with entries \( m(X_i), i = 1, \ldots, n, \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \), and \( A = (a_{ij})_{n \times n} \). Then \( \hat{\theta} \) can be written as
\[ \hat{\theta} = Y^T A Y = m^T A m + 2m^T A \varepsilon + \varepsilon^T A \varepsilon. \] 
(5.3)

The mean and variance of \( \hat{\theta} \) can be calculated directly from (5.3):
\[ E\{\hat{\theta}|X_1, \ldots, X_n\} = m^T A m + \sigma^2 \text{tr}(A), \] 
(5.4)
\[ \text{Var}\{\hat{\theta}|X_1, \ldots, X_n\} = 2\sigma^4 \text{tr}(A^2) + 4\sigma^2 m^T A^2 m. \] 
(5.5)

Conditioned on \( \{X_1, \ldots, X_n\} \), \( 2m^T A \varepsilon \) and \( \varepsilon^T A \varepsilon \) are the stochastic terms in (5.3). It is easy to see that \( 2m^T A \varepsilon \) is linear in \( \varepsilon_i \)'s and contributes \( 4\sigma^2 m^T A^2 m \) to the variance in (5.5), while \( \varepsilon^T A \varepsilon \) is quadratic in \( \varepsilon_i \)'s with variance \( 2\sigma^4 \text{tr}(A^2) \).

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) \( \text{tr}(A^2)/m^T A^2 m \xrightarrow{P} 0 \) (\( m^T A \varepsilon \) dominates),

(ii) \( \text{tr}(A^2)/m^T A^2 m \xrightarrow{P} \infty \) (\( \varepsilon^T A \varepsilon \) dominates).

If the linear term dominates (case (i)), i.e. \( \hat{\theta} \) 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 “genuine-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 \( \text{tr}(A^2) \) and \( m^T A^2 m \) are of the same order. That 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:

\[
Y^T A Y \xrightarrow{C} N \left( m^T A m + \sigma^2 \text{tr}(A), \ 2\sigma^4 \text{tr}(A^2) + 4\sigma^2 m^T A^2 m \right),
\]

(5.6)

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

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

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

\[
P \left\{ \frac{Y^T A Y - m^T A m - \sigma^2 \text{tr}(A)}{(2\sigma^4 \text{tr}(A^2) + 4\sigma^2 m^T A^2 m)^{1/2}} \leq t | X_1, \ldots, X_n \right\} \xrightarrow{P} \Phi(t) \quad \forall t,
\]

(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:

\[
P \left\{ \frac{Y^T A Y - m^T A m - \sigma^2 \text{tr}(A)}{(2\sigma^4 \text{tr}(A^2) + 4\sigma^2 m^T A^2 m)^{1/2}} \leq t \right\} \xrightarrow{} \Phi(t) \quad \text{for all } t.
\]

6 Asymptotic normality

We now establish the asymptotic normality of \( \hat{\theta}_\nu \) and \( \tilde{\theta}_\nu \). In addition to Conditions (A), the bandwidth is assumed to satisfy

**Condition (B):**

\[
h = h(n) \to 0 \text{ and } nh + \log(h) \to \infty \text{ as } n \to \infty.
\]
The requirement (B) is minor in the sense that if a local smoothing neighborhood 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 \(\tilde{\theta}_\nu\).

**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 \(p_{n,j} \geq 0, j = 1, 2, \ldots\) and \(\max_j p_{n,j} < ch\), where \(c\) is a constant. Under Condition (B),

\[
P\{\max_j Z_{n,j} > nhb_n\} \rightarrow 0, \quad \text{as } n \rightarrow \infty,
\]

for any sequence \(b_n\) such that \(b_n \rightarrow \infty\).

**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),

\[
\frac{\lambda_{max}(\nu)}{\text{tr}(A_{\nu}^2)} \frac{P}{0.}
\]  

(6.1)

Since the design matrix \(X\) (see (2.3)) and the matrix \(A_\nu\) are the same for \((X_1, \psi(Y_1)), \ldots, (X_n, \psi(Y_n))\) with \(\psi(\cdot)\) a known function, Lemmas 6.1 and 6.2 are applicable to the case of estimating functionals of the form

\[
\int [m(\psi(x))^2 w(x)dx,
\]

where \(m_\psi(x) = E(\psi(y)|x)\).

From Lemmas 6.1 and 6.2, the asymptotic normality of \(\hat{\theta}_\nu\) and \(\tilde{\theta}_\nu\) follows.

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

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

\[
(\tilde{\theta}_\nu - \theta_\nu) \xrightarrow{C} N(b_{n,\nu}, v_{n,\nu}),
\]

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 \(p \geq (2s - \nu - 1)\),

\[
\sqrt{n}(\hat{\theta}_\nu - \theta_\nu) \xrightarrow{C} N(0, D_2).
\]
(b) Given \( p \) such that

\[
\begin{align*}
    p > 3\nu - 1 & \quad \text{if } C_1 < 0, \\
p > 3\nu & \quad \text{if } C_1 > 0,
\end{align*}
\]

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

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

(c) Given \( s > 2\nu + 0.5 \), taking \( h = O(n^{-2/(2s+2\lceil s \rceil+1)}) \) and \( p \geq (2s - \nu - 1) \), if \( \epsilon_n = o(h^{1/2}) \), then

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

(d) Given \( p > 3\nu - 0.5 \), taking \( h = h_{opt} \) in (4.10), if \( s > (p + \nu + 1)/2 \) and \( \epsilon_n = o(h^{1/2}) \), then

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

Note that the above \( n^{-1/2} \) rate of convergence is achieved when estimators are pseudo-quadratic, since \( D_2n^{-1} \) is one of the terms in the asymptotic variance of \( 2m^TA_\nu \varepsilon \).

Nonparametric estimation of \( \theta_\nu \) has important implications for the choice of the smoothing parameter of local regression. The optimal bandwidth for local linear regression obtained by Huang (1995) has an asymptotic representation similar to the optimal bandwidth in Hall, Sheather, Jones, and Marron (1991) for kernel density estimation. The key is to have higher order expansion for the mean integrated square error. Note that this expansion is necessary to achieve the \( n^{-1/2} \) relative rate of convergence. Also Chiu (1991a, b) constructs two \( n^{-1/2} \)-rate bandwidth selectors based on a similar idea. The optimal bandwidth in Huang (1995) involves functionals of the type \( \theta_\nu \). Based on the theory in this paper, Huang (1995) is able to construct a bandwidth selector for local linear regression and it has a relative rate of \( n^{-1/2} \). The procedure and the theoretical results appear to be new in the nonparametric regression literature.

7 Proofs

Proof of Lemma 2.1 : Proof: Stone (1980) shows that

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

Substituting \( f(x)\mu_j, \ j = 0, \ldots, 2p \), for the corresponding moments in the matrix \((X^TWX)\), we obtain

\[
||n^{-1}h^{-1}H^{-1}(X^TWX)H^{-1} - f(x)S||_\infty \xrightarrow{a.s.} 0,
\]

(7.1)
where \( H = \text{diag}(1, h, h^2, \ldots, h^p)_{(p+1) \times (p+1)} \). 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

\[
\text{tr}(A_\nu) = \sum_{i=1}^{n} \int \left( W^n_{\nu} \left( \frac{X_i - x}{h} \right) \right)^2 w(x) dx
\]

\[
= \frac{1}{(nh^{\nu+1})^2} \sum_{i=1}^{n} \int \left( \frac{1}{f(x)} K^*_\nu \left( \frac{X_i - x}{h} \right) + o_P(1) I_{[x_i - h, x_i + h]}(x) \right)^2 w(x) dx
\]

\[
= \frac{(1 + o_P(1))}{(nh^{\nu+1})^2} \sum_{i=1}^{n} \int \frac{1}{f(x)^2} K^{*2}_\nu \left( \frac{X_i - x}{h} \right) w(x) dx. \quad (7.2)
\]

Since

\[
\text{Var} \left( \sum_{i=1}^{n} K^{*2}_\nu \left( \frac{X_i - x}{h} \right) \right) = O(nh) = o \left( \left( n E \left\{ K^{*2}_\nu \left( \frac{X_i - x}{h} \right) \right\} \right)^2 \right),
\]

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

\[
\text{tr}(A_\nu) = \frac{1}{nh^{2\nu+1}} \left( \int K^{*2}_\nu(t) dt \right) \left( \int f(x)^{-1} w(x) dx \right) + o_P(n^{-1} h^{-2\nu-1}). \quad (7.3)
\]

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

\[
\sum_{i=1}^{n} W^n_{\nu} \left( \frac{X_i - x}{h} \right) m(X_i) = \sum_{i=1}^{n} W^n_{\nu} \left( \frac{X_i - x}{h} \right) [m(x) + m'(x)(X_i - x) + \ldots]
\]

\[
= \beta_\nu(x) + r(x), \quad (7.4)
\]

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

\[
r(x) = h^{p+1-\nu} \left( \int t^{p+1} K^*_\nu(t) dt \right) \beta_{p+1}(x) + o_P(h^{p+1-\nu}). \quad (7.5)
\]

Hence

\[
\mathbf{m}^T A_\nu \mathbf{m} = \int \beta^2_\nu(x) w(x) dx + 2h^{p+1-\nu} \left( \int t^{p+1} K^*_\nu(t) dt \right)
\]

\[
\times \left( \int \beta_{p+1}(x) \beta_\nu(x) w(x) dx \right) + o_P(h^{p+1-\nu}). \quad (7.6)
\]

As noted in Remark 2, \( \int \beta_{p+1}(x) \beta_\nu(x) w(x) dx \) 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), page 170.
In the case that \( s \leq (p + \nu + 1)/2 \), 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,

\[
m(X_i) = m(x) + \sum_{k=1}^{l-1} (X_i - x)^k m^{(k)}(x)/k! + \frac{(X_i - x)^l}{(l-1)!} \int_0^1 (1 - t)^{l-1} m^{(l)}(x + t(X_i - x)) dt,
\]

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

\[
\mathbf{mA}_\nu \mathbf{m} = \int \beta_\nu^2(x) w(x) dx + \int \beta_\nu(x) r(x) w(x) dx + O_F(h^{2(s-\nu)}).
\]

The second term in the RHS of (7.8) can be written as

\[
\sum_{i=1}^{n} \int_0^1 \frac{1}{(l-1)!} \int \frac{d^{[s-\nu]}}{dx^{[s-\nu]}} \left( \beta_\nu(x) W_\nu \left( \frac{X_i - x}{h} \right) (X_i - x)^l w(x) \right)
\times \left( m^{([s])}(x + t(X_i - x)) - m^{([s])}(x) \right) dx dt
\]

\[
= h^{-\nu-1} \int_0^1 \frac{1}{(l-1)!} E \left\{ \int \frac{d^{[s-\nu]}}{dx^{[s-\nu]}} \left( \beta_\nu(x) K^*_\nu \left( \frac{X_i - x}{h} \right) (X_i - x)^l w(x) f^{-1}(x) \right)
\times \left( m^{([s])}(x + t(X_i - x)) - m^{([s])}(x) \right) dx \right\} dt + O_F(h^{l-\nu} a_n)
\]

\[
= h^{l-\nu} \int_0^1 \frac{1}{(l-1)!} \int \frac{d^{[s-\nu]}}{dx^{[s-\nu]}} \left( \beta_\nu(x) K^*_\nu(u) u^l w(x) f^{-1}(x) \right)
\times \left( m^{([s])}(x + hu) - m^{([s])}(x) \right) f(x + hu) du dx dt + O_F(h^{l-\nu} a_n),
\]

where \( a_n = h + \frac{1}{\sqrt{nh}} \) 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 \( s \leq (p + \nu + 1)/2 \),

\[
\mathbf{m}^T A_\nu \mathbf{m} = \int \beta_\nu^2(x) w(x) dx + O_P(h^{s+[s]-2\nu}).
\]

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 the sequel, we treat explicitly the three terms given in the RHS of (3.4). The first term (omitting the factor \( 4\sigma^2 \)) is

\[
\mathbf{m}^T A_\nu^2 \mathbf{m} = \sum_{j=1}^{n} \left( \sum_{i=1}^{n} a_{i,j}(\nu) m(X_i) \right)^2
\]

\[
= \sum_{j=1}^{n} \sum_{i=1}^{n} \sum_{k=1}^{n} a_{i,j}(\nu) m(X_i) a_{k,j}(\nu) m(X_k)
\]

16
\[ \sum_{i,j,k \text{ all different}} + \sum_{i=k \neq j} + 2 \sum_{i=j \neq k} + \sum_{i=j=k} \]  

(7.11)

Now,

\[ \sum_{i \neq j}^{k \neq j} = \sum_{i \neq j} \sum_{k \neq j} m^2(X_i) \left( \int W^\nu \left( \frac{X_i - x}{h} \right) W^\nu \left( \frac{X_j - x}{h} \right) w(x) dx \right)^2 \]

\[ = \sum_{i \neq j}^{k \neq j} \frac{m^2(X_i)}{n^4 h^{4 \nu + 4}} \left\{ \int K^*_{\nu} \left( \frac{X_i - x}{h} \right) K^*_{\nu} \left( \frac{X_j - x}{h} \right) w(x) f^{-2}(x) \right. \]

\[ \times (1 + o_P(1) I_{[-h + X_i, h + X_j]}(x) I_{[-h + X_j, h + X_j]}(x)) dx \right\} \]

\[ = n^{-2} h^{-4 \nu - 2} \int \left\{ \int K^*_{\nu}(u_1) K^*_{\nu}(u_1 + \frac{x - y}{h}) m^2(x + hu_1) f(x + hu_1) \right. \]

\[ \times (1 + o_P(1) I_{[-1,1]}(u_1) I_{[-(x-y)/h, 1-(x-y)/h]}(u_1)) du_1 \}

\[ \times \left\{ \int K^*_{\nu}(u_2) K^*_{\nu}(u_2 + \frac{x - y}{h}) f(y + hu_2) \left( 1 + o_P(1) I_{[-1,1]}(u_2) \right) \right. \]

\[ \times I_{[-(x-y)/h, 1-(x-y)/h]}(u_2) du_2 \right\} w(x) w(y) f^{-2}(x) f^{-2}(y) dx dy \]

\[ + O_P(n^{-3} h^{-4 \nu - 3/2}) \]

\[ = n^{-2} h^{-4 \nu - 1} (1 + o_P(1)) \left( \int (K^*_{\nu} \ast K^*_{\nu}(z))^2 dz \right) \]

\[ \times \int \int m^2(x) f^{-1}(x) f^{-1}(y) w(x) w(y) dx dy. \]  

(7.12)

By (7.4) and Lemma 2.1,

\[ \sum_{i,j,k \text{ all different}} = n^{-1} h^{-2 \nu} \int \int (\beta_{\nu}(y) + r(y))(\beta_{\nu}(y + h z) + r(y + h z)) \left( K^*_{\nu}(t) \right) \]

\[ \times K^*_{\nu}(t + z) f(y + h(z + t)) w(y) w(y + h z) f^{-1}(y + h z) \]

\[ \times f^{-1}(y) dt dz dy + O_P(\sqrt{nh} n^{-2} h^{-2 \nu - 2}) \]

\[ = n^{-1} h^{-2 \nu} \int \int g_{\nu}(y) g_{\nu}(y + h z) \left( K^*_{\nu}(t) K^*_{\nu}(t + z) \right) \]

\[ \times f(y + h(z + t)) dt dz dy + O_P(\sqrt{nh} n^{-2} h^{-2 \nu - 2}) \]  

(7.13)

with \( g_{\nu}(y) = (\beta_{\nu}(y) + r(y)) w(y) f^{-1}(y) \). For the case that \( s > 2 \nu \), it is assumed initially that \( s > 3 \nu \).
Then write
\[ g_\nu(y + hz) = \sum_{i=0}^{2\nu} g^{(l)}_\nu(y)h^l z^l/l! + o(h^{2\nu}) \] (7.14)
and
\[ f(y + h(z + t)) = \sum_{k=0}^{\nu} f^{(k)}(y)h^k(z + t)^k/k! + o(h^\nu). \] (7.15)

Note that from (2.10), for nonnegative integers \( l, k, \) and \( l + k \leq 2\nu, \)
\[ \int \int z^l(z + t)^kK^*_\nu(t)K^*_\nu(t + z)dtdz = \begin{cases} \frac{(1)^{\nu-l}}{\nu(l-n)!} & \text{if } l \geq \nu, k \leq \nu, \\ 0 & \text{and } l + k = 2\nu, \\ 0 & \text{elsewhere}. \end{cases} \] (7.16)

Combining (7.13), (7.14), (7.15), and (7.16),
\[ \sum_{i,j,k \text{ all different}} = \frac{(-1)^\nu}{n(\nu)!^2} \int g_\nu(y) \left( \sum_{i=0}^{\nu} g^{(\nu+i)}_\nu(y)f^{(\nu-i)}(y)\frac{v^l}{i!(\nu - i)!} \right) + o_P(n^{-1}) \]
\[ = \frac{(-1)^\nu}{n(\nu)!^2} \int g_\nu(y) \frac{d^\nu}{dy^\nu}[g^{(\nu)}_\nu(y)f(y)]dy + o_P(n^{-1}). \] (7.17)

Applying integration by parts and Condition (A3) to (7.17), it follows that
\[ \sum_{i,j,k \text{ all different}} = \frac{1}{n(\nu)!^2} \int [g^{(\nu)}_\nu(y)]^2 f(y)dy + o_P(n^{-1}) \]
\[ = \frac{1}{n(\nu)!^2} \int [G^{(\nu)}_\nu(y)]^2 f(y)dy + o_P(n^{-1}) \] (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 \( 2\nu \)-th, and hence must hold for \( s > 2\nu \). To modify arguments from (7.13) to (7.18) for the case of \( s \leq 2\nu \), the change required is in (7.14) and (7.15). The initial assumption is that \( m(\cdot) \) has \( 2[\nu - \nu] \) bounded derivatives, and Taylor expansions of orders \( 2([\nu - \nu]) \) and \( (2\nu - [\nu]) \) for \( (g_\nu(y + hz) - g_\nu(y)) \) and \( f(y + h(z + t)) \) respectively. Then, applying integration by parts \( ([\nu - \nu]) \) 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 RHS of (3.4) (omitting the factor $2\sigma^4$) is
\[
\text{tr}(A_\nu^2) = \sum_{i=1}^{n} a_{i,i}^2(\nu) + \sum_{i \neq j} a_{i,j}^2(\nu)
\]
\[
= I_1 + I_2,
\]
where $I_1$ and $I_2$ denote the diagonal and non-diagonal terms respectively. Using similar arguments as in establishing the term $m^T A_\nu^2 m$, we obtain that $I_1 = O_P(n^{-3} h^{-4\nu-2})$ and
\[
I_2 = \frac{n(n-1)(1 + o_P(1))}{n^4 h^{4\nu+4}} \sum_{i \neq j} E \left\{ \int K_\nu^*(\frac{X_i - x}{h}) K_\nu^*(\frac{X_j - x}{h}) f(x)^{-2} w(x) dx \right\}^2
\]
\[
= n^{-2} h^{-4\nu-1} \int \int [(K_\nu^* * K_\nu^*)(z)]^2 dz f(x)^{-2} w^2(x) dx + o_P(n^{-2} h^{-4\nu-1}).
\]
Consequently,
\[
\text{tr}(A_\nu^2) = n^{-2} h^{-4\nu-1} (1 + o_P(1)) \left( \int [(K_\nu^* * K_\nu^*)(z)]^2 dz \right) \left( \int f(x)^{-2} w^2(x) dx \right).
\]  
(7.19)
Observe that the last term (omitting the factor $(E(\epsilon_1^4) - 3\sigma^4)$) in the RHS of (3.4) is
\[
\sum_{i=1}^{n} a_{i,i}^2(\nu) = I_1 = o_P(\text{tr}(A_\nu^2)).
\]  
(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.
\[
E(\hat{\theta}_\nu) = E(\hat{\theta}_\nu) - E(\hat{\sigma}^2 \text{tr}(A_\nu))
\]
\[
= E(\hat{\theta}_\nu) - \sigma^2 \text{tr}(A_\nu) + E \left( (\hat{\sigma}^2 - \sigma^2) \text{tr}(A_\nu) \right)
\]
\[
= E(\hat{\theta}_\nu) - \sigma^2 \text{tr}(A_\nu) + O_P(\epsilon_n n^{-1} h^{-2\nu-1}).
\]
\[
\text{Var}(\hat{\theta}_\nu) = \text{Var}(\hat{\theta}_\nu) - (\hat{\sigma}^2 - \sigma^2) \text{tr}(A_\nu))
\]
\[
= \text{Var}(\hat{\theta}_\nu) + O_P(\epsilon_n^2 n^{-2} h^{-4\nu-2}).
\]  
□

**Proof of Theorem 5.1:** If the linear term dominates, i.e. $\text{tr}(A^2)/m^T A^2 m \overset{P}{\to} 0$, then $\epsilon^T A \epsilon - E(\epsilon^T A \epsilon) = o_P \left( (m^T A^2 m)^{1/2} \right)$. Conditioning on $\{X_1, \ldots, X_n\}$, it is trivial that $m^T A \epsilon$ is normally distributed with mean 0 and variance $\sigma^2 m^T A^2 m$. This implies
\[
(4\sigma^2 m^T A^2 m)^{1/2}(\hat{\theta} - E(\theta))
\]
\[
= (4\sigma^2 m^T A^2 m)^{1/2} \{2m^T A \epsilon + (\epsilon^T A \epsilon - E(\epsilon^T A \epsilon))\} \overset{C}{\to} \mathcal{N}(0,1).
\]  
19
For case (ii), \( \text{tr}(A^2)/m^T A^2 m \overset{P}{\to} \infty \), we need only to consider \( \varepsilon^T A \varepsilon \). Factor \( A \) into a product \( \Gamma^T \Lambda \Gamma \), where \( \Gamma \) is an orthonormal matrix and \( \Lambda \) is a diagonal matrix with eigenvalues of \( A \), \( \lambda_i \)'s as its entries. Now

\[
\varepsilon^T A \varepsilon = \delta^T \Lambda \delta = \sum_{i=1}^{n} \lambda_i \delta_i^2
\]

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

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

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:

\[
\frac{\sum_{i=1}^{n} E^C|\lambda_i \delta_i^2 - \lambda_i|^4}{[\text{Var}(\sum_{i=1}^{n} \lambda_i \delta_i^2 |X_1, \ldots, X_n)]^2} = \frac{15\sigma^4 \sum_{i=1}^{n} \lambda_i^4}{[\text{tr}(A^2)^2]} \overset{P}{\to} 0,
\]

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

\[\square\]

**Proof of Lemma 6.1:** Note that

\[
P(\max_j Z_{n,j} > nhb_n) \leq \sum_j P(Z_{n,j} > nhb_n). \tag{7.21}
\]

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

Given an arbitrary positive constant \( d \),

\[
P(Z_{n,j} > nhb_n) = P(\exp(Z_{n,j}d) > \exp(nhb_n d))
\]

\[\leq \exp(-nhb_n d) E\{\exp(Z_{n,j}d)\}
\]

\[= \exp(-nhb_n d)(1 - p_{n,j} + e^d p_{n,j})^n
\]

\[\leq \exp(-nhb_n d + np_{n,j} e^d). \tag{7.22}
\]

Take \( d = \log(\frac{h b_n}{p_{n,j}}) \) in (7.22). If \( n \) is sufficiently large, then

\[
P(Z_{n,j} > nhb_n) \leq \left( \frac{p_{n,j} e}{h b_n} \right)^{nhb_n}. \tag{7.23}
\]

Since \( \max_j p_{n,j} \leq c h \), the RHS of (7.23) is bounded by

\[
(ch)^{nhb_n-1} \left( \frac{e}{h b_n} \right)^{nhb_n} p_{n,j} = \frac{1}{c h} \left( \frac{e}{b_n} \right)^{nhb_n} p_{n,j}.
\]

20
Now \( b_n(\log(ce) - \log(b_n)) \leq -1 \) for sufficiently large \( n \), which together with Condition (B) give

\[
\sum_j P(Z_{n,j} > nhb_n) \leq \sum_j \frac{1}{ch} \left( \frac{1}{e} \right)^{nh} p_{n,j}
\]

\[
= \frac{1}{ch} \left( \frac{1}{e} \right)^{nh} \to 0.
\] (7.24)

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

**Proof of Lemma 6.2:** Observe that from (7.19)

\[
\text{tr}(A^2_x) \geq c_1 n^{-2\nu-1} + o_P(n^{-2\nu-1})
\] (7.25)

for some nonnegative constant \( c_1 \). It suffices to consider the rate of \( \lambda_{max}(\nu) \). By basic theory of linear algebra,

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

From Lemma 2.1, for \( n \) sufficiently large,

\[
\sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j}(\nu) b_i b_j \leq \frac{2}{(nh^{\nu+1})^2} \int f(x)^{-2} \left( \sum_{i=1}^{n} K^*(\frac{X_i - x}{h}) b_i \right)^2 w(x) dx.
\] (7.27)

Under Conditions (A), the RHS of (7.27) is further bounded by

\[
c_2 n^{-2\nu-2} \left( \sum_{i=1}^{n} |b_i| I_{[X_i-h, h+X_i]}(x) \right)^2 dx
\]

\[
= c_2 n^{-2\nu-2} \left( \sum_{i=1}^{n} b_i^2 \int I_{[X_i-h, X_i+h]}(x) dx \right)
\]

\[
+ \sum_{i \neq j} |b_i||b_j| \int I_{[X_i-h, X_i+h]}(x) I_{[X_j-h, X_j+h]}(x) dx
\]

\[
\leq c_2 n^{-2\nu-2} \left( 2h \sum_{i=1}^{n} b_i^2 + 2h \sum_{i \neq j} I_{[|X_i-X_j| \leq 2h]} |b_i||b_j| \right)
\] (7.28)

for some nonnegative 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 \( I_{[|X_i-X_j| \leq 2h]} = 0 \) for \( X_i \in I_{k_1}, X_j \in I_{k_2} \), and \( |k_1 - k_2| > 1 \),

\[
\sum_{i,j} I_{[|X_i-X_j| \leq 2h]} |b_i||b_j| \leq \sum_{k=-\infty}^{\infty} \left( \sum_{X_i, X_j \in I_k \cup I_{(k+1)}} |b_i||b_j| \right) = \sum_{k=-\infty}^{\infty} \left( \sum_{X_i \in I_k \cup I_{(k+1)}} |b_i|^2 \right).
\] (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 RHS of (7.29) over \( \|b\| = 1 \) is attained at \( |b_i| = 1/\sqrt{N} \) for those indices such that \( X_i \in J_{k_0} \cup I_{k_0 + 1} \). Therefore,

\[
\max_{\|b\| = 1} \sum_{i,j} \mathbf{1}_{\|X_i - X_j\| \leq 2h} |b_i| |b_j| \leq N.
\]

Note that \( \{n_k\} \) 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

\[
\lambda_{\max}(\nu) \leq c_2 n^{-2} h^{-2\nu - 2} O_P(h + nh^{7/4}).
\] (7.30)

Combine (7.25) and (7.30) to conclude that

\[
\frac{\lambda_{\max}^2(\nu)}{\text{tr}(A^2_\nu)} = O_P(n^{-2} h^{-1} + h^{1/2}) \to 0.
\]

\[\Box\]

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

\[
\frac{\text{tr}(A^4_\nu)}{\left(\text{tr}(A^2_\nu)\right)^2} \xrightarrow{p} 0, \quad \text{as } n \to \infty.
\]

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

\[
\frac{\text{tr}(A^4_\nu)}{\left(\text{tr}(A^2_\nu)\right)^2} = \frac{\sum_{i=1}^n \lambda^4_i(\nu)}{\left(\sum_{i=1}^n \lambda^2_i(\nu)\right)^2} \leq \frac{\lambda_{\max}^2(\nu)}{\text{tr}(A^2_\nu)} \xrightarrow{p} 0, \quad \text{as } n \to \infty.
\] (7.31)

Hence the asymptotic normality of \( \hat{\theta}_\nu \) and \( \tilde{\theta}_\nu \) follows. \[\Box\]

**Acknowledgments.** This work was part of Li-Shan Huang’s doctoral dissertation at University of North Carolina. She is very grateful to Jianqing Fan, her Ph.D. advisor, for the guidance and encouragements. Li-Shan Huang would also like to thank the Department of Statistics, University of North Carolina, National Institute of Statistical Sciences, Dr. M. R. Leadbetter and Dr. R. L. Smith for the financial support during her graduate studies.

**REFERENCES**


