Statistica Sinica **10**(2000), 751-770

NONPARAMETRIC ESTIMATION OF RATIOS OF NOISE TO SIGNAL IN STOCHASTIC REGRESSION

Qiwei Yao and Howell Tong

London School of Economics

Abstract: In this paper, we study three different types of estimates for the noise-to-signal ratios in a general stochastic regression setup. The locally linear and locally quadratic regression estimators serve as the building blocks in our approach. Under the assumption that the observations are strictly stationary and absolutely regular, we establish the asymptotic normality of the estimates, which indicates that the residual-based estimates are to be preferred. Further, the locally quadratic regression reduces the bias when compared with the locally linear (or locally constant) regression *without* the concomitant increase in the asymptotic variance, if the same bandwidth is used. The asymptotic theory also paves the way for a fully data-driven undersmoothing scheme to reduce the biases in estimation. Numerical examples with both simulated and real data sets are used as illustration.

Key words and phrases: Absolutely regular, asymptotic normality, local polynomial regression, noise to signal ratio.

1. Introduction

It is often instructive to think of a stochastic dynamical system as consisting of two parts: the drift term and the diffusion term. The former is ordinarily endowed with substantial structure dominated by observable variables, and may be interpreted as the signal. By contrast, the latter is typically featureless and unobservable, and may be interpreted as the noise. An important problem of common interest to many different areas (examples will be given later) is the estimation of the noise-to-signal ratio. In this paper, we consider two (global) measures of the noise-to-signal ratio in a general setup.

We assume that $\{(Y_i, X_i)\}$ is a strictly stationary process having the same marginal distribution as (Y, X), where Y is a scalar and X is a d-dimensional vector. Let $m(x) = E\{Y|X = x\}$ and $\sigma^2(x) = \mathrm{Var}\,(Y|X = x) > 0$. We write a regression model of Y_i on X_i as

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i. \tag{1.1}$$

Then $E\{\epsilon_i|X_i\} = 0$, and $\mathrm{Var}\,(\epsilon_i|X_i) = 1$, although the conditional distribution of ϵ_i given $X_i = x$ may still depend on x. For $X_i = (Y_{i-1}, \ldots, Y_{i-d})$, (1.1) is an

autoregressive model with inhomogeneous noise. We define two measures of the noise-to-signal ratio as follows.

$$\zeta^2 = \frac{E\{\sigma^2(X)\}}{\mathrm{Var}\,\{m(X)\}}, \quad \text{and} \quad \xi^2 = \frac{E\{\sigma^2(X)\}}{E\{m^2(X)\}}. \tag{1.2}$$

Formally, they are motivated by the decompositions $\mathrm{Var}\,(Y) = \mathrm{Var}\,\{m(X)\} + E\{\sigma^2(X)\}$ and $E(Y^2) = E\{m^2(X)\} + E\{\sigma^2(X)\}$ respectively. A related measure is Pearson's correlation ratio

$$\eta^2 = \frac{\mathrm{Var}\,\{m(X)\}}{\mathrm{Var}\,(Y)}. \tag{1.3}$$

Obviously, $\eta^2 = 1/(1+\zeta^2)$. Doksum and Samarov (1995) reported some interesting results on estimating η^2 based on independent and identically distributed observations.

Our direct motivation to estimate noise-to-signal ratio comes from the need to detect *operational determinism* studied by Yao and Tong (1998a). In fact, a sufficiently small value of ξ^2 suggests that the system may be considered operationally deterministic, and therefore all the powerful techniques for deterministic systems, *e.g.* correlation dimension, Lyapunov exponents and so on, may then be brought to bear. (See *e.g.* Tong (1995)). On the other hand, the potential application of the two measures defined in (1.2) is diverse. For example, in the context of prediction, $|\zeta|$ can be considered an average relative error in prediction and used to guard against excessive claims in respect of any forecasting algorithm. In channel communications, we may use both ξ^2 and ζ^2 to assess the information loss in the transmission through a noisy channel if we take X as an input and Y as an output. (See *e.g.* Feinstein (1958)). In nonparametric regression, the measure ζ^2 is used as an indicator of the intrinsic difficulty of the problem of estimation (Fan and Gijbels (1995)). Noise-to-signal ratios have also played a role in quality control of experimental design (Box (1988)). In the vast engineering literature of signal processing, extraction of the signal by filtering out the noise is essentially equivalent to reducing the noise-to-signal ratio in the model. (See, *e.g.* Broomhead (1995), and the references therein.)

The purpose of this paper is to develop nonparametric estimates for noise-to-signal ratios based on locally linear and locally quadratic regression smoothers in such a way that the results are immediately applicable in practice.

In fact, we present and develop three alternative estimates for ζ^2 and ξ^2. Two simple data-driven bandwidths, derived from the extremal properties of the ratios, are applied. Further, we provide a simple but intuitively appealing undersmooth scheme to reduce the biases of the estimates, which is entirely data-determined. Asymptotic normality for the estimates is established at the

convergence rate $n^{\frac{1}{2}}$ which shows that the estimates based on locally quadratic smoothers are less biased than those based on locally linear smoothers *without* any concomitant increase in the variance, if the same bandwidth is used. In fact, estimates based on locally quadratic and locally linear (also locally constant) smoothers admit the same first-order asymptotic variances. (See Remark 2 in Section 3 below.) A close analogy to this phenomenon is the estimation of integrated squared density functions where the first order asymptotic variance does not depend on the kernel function (Hall and Marron (1987)). Both of them are consequences of the Hadamad differentiability which facilitates the convergence rate $n^{\frac{1}{2}}$; see Bickel and Ritov (1988) and Fan (1991).

Doksum and Samarov (1995) established the asymptotic normality for three estimates of η^2 defined in (1.3) based on the Nadaraya-Watson estimate with independent and identically distributed observations. In terms of estimation of η^2, our results could be viewed as further developments in several aspects (Remark 4 in Section 3 below). First, we have proved the asymptotic normality under a more general setup which includes both time series data and i.i.d. observations as special cases. Second, our estimators are based on locally linear and locally quadratic smoothers. More importantly, we have derived explicit formulas for asymptotic biases. These enable us to compare the different estimates qualitatively. In fact, from the qualitative comparison, we may single out the residual-based estimate as the best amongst the three types of estimates. (See Remark 3 in Section 3 below.) Simulation studies lend further support to this conclusion.

The paper is organized as follows. In Section 2, we propose estimates for ζ^2 and ξ^2, including procedures of bandwidth choice. In Section 3, we establish the asymptotic properties of the estimates. To save space, we state the results for ζ^2 only. In Section 4, we present a scheme for bias reduction by undersmoothing. Simulation studies are conducted as an illustration. Applications with three real data sets are also reported. Proofs are relegated to the Appendix.

2. Estimation of Noise-to-signal Ratios

2.1. Locally polynomial estimators of $m(.)$

Since estimators for $m(.)$ are the building blocks for our approach, we describe briefly the locally linear and locally quadratic regression estimators for $m(.)$. For more detailed discussion on local polynomial smoothing, see Fan and Gijbels (1996).

Given the observations $\{(X_i,\ Y_i);\ 1 \leq t \leq n\}$, one of the conventional nonparametric estimators of $m(x)$ is the Nadaraya-Watson kernel regression estimator, which can be viewed as the minimizer of the following least squares

problem

$$\sum_{i=1}^{n}\{Y_i - a\}^2 K\left(\frac{X_i - x}{h}\right),$$

where $K(.)$ is a kernel function on R^d and $h > 0$ is the bandwidth. However, if the derivative of m at the point x exists, by Taylor's expansion, we have $m(z) \approx m(x) + \dot{m}(x)(z - x)$. This suggests the locally linear regression estimator: $\hat{m}(x) = \hat{a}$, where $(\hat{a}, \hat{b})$ minimizes

$$\sum_{i=1}^{n}\{Y_i - a - b^\tau(X_i - x)\}^2 K\left(\frac{X_i - x}{h}\right). \tag{2.1}$$

It has been pointed out that the locally linear regression method has various advantages over the Nadaraya-Watson method (see, for example, Fan (1992), Hastie and Loader (1993)). Further, we consider the locally quadratic estimator: $\hat{m}(x) = \hat{a}$, where $(\hat{a}, \hat{b}, \hat{c})$ minimizes

$$\sum_{i=1}^{n}[Y_i - a - b^\tau(X_i - x) - c^\tau \mathrm{vec}\{(X_i - x)(X_i - x)^\tau\}]K\left(\frac{X_i - x}{h}\right). \tag{2.2}$$

In the above expression, $c \in R^{d(d+1)/2}$ and $\mathrm{vec}(A) = (a_{11}, a_{22}, \ldots, a_{d,d}, a_{12}, \ldots, a_{1,d}, a_{23}, \ldots, a_{d-1,d})^\tau$ for any $d \times d$ symmetric matrix $A = (a_{ij})$.

2.2. Estimates for ζ^2 and ξ^2

Note that $\mathrm{Var}\{m(X)\} = \mathrm{Cov}\{m(X), Y\}$. We may characterize ζ^2 as follows:

$$\begin{aligned}
\zeta^2 &= \frac{E\{Y - m(X)\}^2}{\mathrm{Var}\{m(X)\}} = \frac{E\{Y - m(X)\}^2}{\mathrm{Var}(Y) - E\{Y - m(X)\}^2} \\
&= \inf_g \frac{E\{Y - g(X)\}^2}{\mathrm{Var}(Y) - E\{Y - g(X)\}^2} \qquad (2.3) \\
&= \frac{\mathrm{Var}(Y) - \mathrm{Cov}\{m(X), Y\}}{\mathrm{Cov}\{m(X), Y\}} = \frac{1 - \mathrm{Corr}^2\{g(X), Y\}}{\mathrm{Corr}^2\{g(X), Y\}} \\
&= \inf_g \frac{1 - \mathrm{Corr}^2\{g(X), Y\}}{\mathrm{Corr}^2\{g(X), Y\}}, \qquad (2.4)
\end{aligned}$$

where the infimum is taken over all real-valued functions $g(X)$ with finite second moments.

Let $\hat{m}(.)$ be an estimator of $m(.)$ constructed as in Section 2.1. From (2.3) and (2.4), we may define the estimates for ζ^2 as follows.

$$\hat{\zeta}_1^2 = \frac{\sum_{i=1}^n \{Y_i - \hat{m}(X_i)\}^2 w(X_i)}{\sum_{i=1}^n \hat{m}^2(X_i) w(X_i) - n\bar{m}^2},$$

$$\hat{\zeta}_2^2 = \frac{\sum_{i=1}^n \{Y_i - \hat{m}(X_i)\}^2 w(X_i)}{\sum_{i=1}^n Y_i^2 w(X_i) - n\bar{Y}_w^2 - \sum_{i=1}^n \{Y_i - \hat{m}(X_i)\}^2 w(X_i)},$$

$$\hat{\zeta}_3^2 = \frac{\sum_{i=1}^n Y_i^2 w(X_i) - n\bar{Y}_w^2 - \{\sum_{i=1}^n \hat{m}(X_i) Y_i w(X_i) - n\bar{m}\bar{Y}_w\}}{\sum_{i=1}^n \hat{m}(X_i) Y_i w(X_i) - n\bar{m}\bar{Y}_w},$$

where $w(.)$ is a nonnegative weight function, $\bar{Y}_w = n^{-1}\sum_{i=1}^n Y_i w(X_i)$ and $\bar{m} = n^{-1}\sum_{i=1}^n \hat{m}(X_i) w(X_i)$.

By reference to their relation with the estimator $\hat{m}(.)$, we call $\hat{\zeta}_1^2$, $\hat{\zeta}_2^2$ and $\hat{\zeta}_3^2$ the plug-in estimate, the residual-based estimate and the correlation estimate, respectively.

Due to the presence of the weight function $w(.)$ in their definitions, estimates $\hat{\zeta}_i^2$ are not necessarily consistent estimates for ζ^2. Instead they estimate

$$\zeta_w^2 = E\{\sigma^2(X) w(X)\} / [E\{m^2(X) w(X)\} - E^2\{m(X) w(x)\}] \tag{2.5}$$

consistently. Note that weight functions are invariably introduced when global measures of deviation are used in order to avoid large bias in the estimation near the boundary of the support of the density function $p(.)$ of X (cf. Marron and Härdle (1986), and Doksum and Samarov (1995)). Typically, the weight function $w(.)$ will be chosen to be 1 in the central part of the support of $p(.)$ and descend to 0 near the boundary of the support of $p(.)$.

Similar estimates for ξ^2 can be defined as follows.

$$\hat{\xi}_1^2 = \frac{\sum_{i=1}^n \{Y_i - \hat{m}(X_i)\}^2 w(X_i)}{\sum_{i=1}^n \hat{m}^2(X_i) w(X_i)},$$

$$\hat{\xi}_2^2 = \frac{\sum_{i=1}^n \{Y_i - \hat{m}(X_i)\}^2 w(X_i)}{\sum_{i=1}^n Y_i^2 w(X_i) - \sum_{i=1}^n \{Y_i - \hat{m}(X_i)\}^2 w(X_i)},$$

$$\hat{\xi}_3^2 = \frac{\sum_{i=1}^n Y_i^2 w(X_i) - n\bar{Y}_w^2 - \{\sum_{i=1}^n \hat{m}(X_i) Y_i w(X_i) - n\bar{m}\bar{Y}_w\}}{n\bar{Y}_w^2 + \sum_{i=1}^n \hat{m}(X_i) Y_i w(X_i) - n\bar{m}\bar{Y}_w}.$$

2.3. Bandwidth selection

As we shall see in Section 3 below, all the estimates $\hat{\zeta}_i^2$ (also $\hat{\xi}_i^2$) are $\sqrt{n}$-consistent and asymptotically normal. Therefore, the standard choice of the bandwidth h which minimizes the asymptotic mean squared error cannot be applied, unless higher-order asymptotics are entertained. Even so, plug-in estimates for some unknown quantities involved must be evaluated, which could be cumbersome. Instead, as a first attempt, we consider two simple and direct data-driven methods for the selection of h, namely the cross-validation estimate and the maximum correlation estimate. The latter has been used in the estimation of Pearson's correlation ratio by Doksum and Samarov (1995).

Based on the extremal property in (2.3) and (2.4), we may choose h as follows. Define

$$\hat{h}_2 = \arg\min_h \sum_{i=1}^{n} \{Y_i - \hat{m}_{-i}(X_i)\}^2 w(X_i), \tag{2.6}$$

$$\hat{h}_3 = \arg\max_h \frac{\left\{\sum_{i=1}^n \hat{m}_{-i}(X_i) Y_i w(X_i) - n\bar{m}_- \bar{Y}_w\right\}^2}{\sum_{i=1}^n \hat{m}_{-i}^2(X_i) w(X_i) - n\bar{m}_-^2}, \tag{2.7}$$

where $\hat{m}_{-i}(.)$ is the estimator of $m(.)$ without the observation (X_i, Y_i), and $\bar{m}_- = \frac{1}{n}\sum_{i=1}^n \hat{m}_{-i}(X_i) w(X_i)$. To prevent possible over-fitting, we use leave-one-out estimator for $m(.)$ in the above optimization. (It is hardly necessary to leave more than one out since X is random; see Yao and Tong (1998b).) We suggest using the cross validation estimate $\hat{h}_2$ in estimates $\hat{\zeta}_2^2$ and $\hat{\xi}_2^2$, and the maximum correlation bandwidth $\hat{h}_3$ in estimates $\hat{\zeta}_3^2$ and $\hat{\xi}_3^2$.

3. Asymptotic Properties

For simplicity of presentation, in this section we treat only univariate X, that is $d = 1$. We use $p(.)$ to denote the marginal density function of X, and $g(.|x)$ the conditional density function of Y given $X = x$. For any f_1 and f_2 which are functions of Y and X, we use the notation $E_w(f_1) = E\{f_1 w(X)\}$, $\mathrm{Cov}_w(f_1, f_2) = E_w(f_1 f_2) - E_w(f_1)E_w(f_2)$, and $\mathrm{Var}_w(f_1) = \mathrm{Cov}_w(f_1, f_1)$.

We start with some regularity conditions.

(C1) The fourth order derivative of $m(.)$ is uniformly continuous on compact sets. $E(Y^{4(1+\delta)}) < \infty$, where $\delta \in [0, 1)$ is a constant. Further, $\frac{\partial^2 g(y|x)}{\partial x^2}$ is uniformly continuous in x on compact sets.

(C2) The kernel function K is a symmetric density function with a bounded support in R. Further, $|K(x_1) - K(x_2)| \le c|x_1 - x_2|$ and $|p(x_1) - p(x_2)| \le c|x_1 - x_2|$ for $x_1, x_2 \in R$.

(C3) The weight function $w(\cdot)$ is smooth, and has a compact support contained in $\{p(x) > 0\}$.

(C4) The process $\{(X_i, Y_i)\}$ is absolutely regular, i.e.

$$\beta(j) \equiv \sup_{i \ge 1} \mathrm{E}\left\{\sup_{A \in \mathcal{F}_{i+j}^{\infty}} |P(A|\mathcal{F}_1^i) - P(A)|\right\} \to 0, \quad \text{as } j \to \infty,$$

where $\mathcal{F}_i^j$ is the σ-field generated by $\{(X_k, Y_k) : k = i, \ldots, j\}$, $(j \ge i)$. Further, $\sum_{j=1}^{\infty} j^2 \beta^{\frac{\delta}{1+\delta}}(j) < \infty$ for the δ given in (C1). (We assume that $a^b = 0$ when $a = b = 0$.)

(C5) As $n \to \infty$, $h \to 0$ and $\liminf_{n\to\infty} nh^4 > 0$.

The condition on the boundedness of the support of $K(\cdot)$ in (C2) is imposed for brevity of proofs, and can be removed at the cost of lengthier ones. In particular, the Gaussian kernel is allowed. The assumption on the convergence rates of h in (C5) is also for technical convenience. It can be weakened by applying Collomb's inequality and involves more technical details. The assumption of the convergence rates of $\beta(j)$ is also not the weakest possible.

Remark 1. When $\{(X_t, Y_t)\}$ are independent, (C4) holds with $\delta = 0$ and the condition in (C1) reduces to $E(Y^4) < \infty$. On the other hand, if (C4) holds with $\delta = 0$, there are at most finitely many non-zero $\beta(j)'$s. This means that there exists an integer $j_0, 0 < j_0 < \infty$, for which (X_i, Y_i) is independent of $\{(X_j, Y_j),\ j \geq i + j_0\}$, for all $i \geq 1$.

Theorem 1. *Suppose conditions* (C1) — (C5) *hold. Let* $\hat{m}(.)$ *be the locally linear estimator of* $m(.)$ *derived from* (2.1). *Then for* $i = 1, 2$ *and* 3, *as* $n \to \infty$, $n^{\frac{1}{2}}(\hat{\zeta}_i^2 - \zeta_w^2 - \lambda_{n,i}) \xrightarrow{d} N\left(0, \mathrm{Var}\left(Z_\zeta\right)\right)$, *where*

$$\lambda_{n,1} = -h^2\zeta_w^2\sigma_0^2\frac{\mathrm{Cov}_w\{Y, \ddot{m}(X)\}}{\mathrm{Var}_w\{m(X)\}} + o(h^2),\ \lambda_{n,2} = \frac{h^4\sigma_0^4}{4}(1+\zeta_w^2)\frac{E_w\{\ddot{m}^2(X)\}}{\mathrm{Var}_w\{m(X)\}} + o(h^4),$$

$$\lambda_{n,3} = -\frac{h^2\sigma_0^2}{2}(1+\zeta_w^2)\frac{\mathrm{Cov}_w\{Y, \ddot{m}(X)\}}{\mathrm{Var}_w\{m(X)\}} + o(h^2),$$

$$Z_\zeta = \frac{w(X)}{\mathrm{Var}_w\{m(X)\}}[\sigma^2(X)\epsilon^2 - \zeta_w^2\{m^2(X) + 2m(X)\sigma(X)\epsilon - 2YE_w(Y)\}].$$

In the above expressions, $\epsilon = \{Y - m(X)\}/\sigma(X)$, $\ddot{m}(x) = \frac{d^2}{dx^2}m(x)$, $\sigma_0^2 = \int x^2K(x)dx$, and ζ_w^2 is defined as in (2.5).

Theorem 2. *Suppose conditions* (C1) — (C5) *hold. Let* $\hat{m}(.)$ *be the locally quadratic estimator of* $m(.)$ *derived from* (2.2). *Then for* $i = 1, 2$ *and* 3, *as* $n \to \infty$, $n^{\frac{1}{2}}(\hat{\zeta}_i^2 - \zeta_w^2 - \pi_{n,i}) \xrightarrow{d} N\left(0, \mathrm{Var}\left(Z_\zeta\right)\right)$, *where* Z_ζ *is the same as in Theorem* 1, *and*

$$\pi_{n,1} = -2h^4\zeta_w^2\mu_*\frac{\mathrm{Cov}_w\{Y, m^{(4)}(X)\}}{\mathrm{Var}_w\{m(X)\}} + o(h^4),$$

$$\pi_{n,2} = h^8\mu_*^2(1+\zeta_w^2)\frac{E_w\{m^{(4)}(X)\}^2}{\mathrm{Var}_w\{m(X)\}} + o(h^8),$$

$$\pi_{n,3} = -h^4\mu_*(1+\zeta_w^2)\frac{\mathrm{Cov}_w\{Y, m^{(4)}(X)\}}{\mathrm{Var}_w\{m(X)\}} + o(h^4).$$

In the above expressions, $\mu_* = \frac{\mu_4^2 - \mu_2\mu_6}{24(\mu_4 - \mu_2^2)}$, $\mu_i = \int x^iK(x)dx$, and ζ_w^2 is defined as in (2.5).

Remark 2. The estimate for the noise-to-signal ratios based on locally quadratic regression appears to be of higher order infinitesimal while the variance remain unchanged. However, a smaller bandwidth should be used in local linear estimation; see Section 4.1 and also Remark 5 below.

Remark 3. The $\hat{\zeta}_i^2$'s have the same asymptotic variance, but the biases are of different order. Note that the residual-based estimate $\hat{\zeta}_2^2$ depends on the estimator $\hat{m}(.)$ through $n^{-1}\sum_{i=1}^n\{Y_i - \hat{m}(X_i)\}^2$ only, which estimates $E_w\{Y - m(X)\}^2 = E_w\{\sigma^2(X)\}$ with a bias of the order r_n^2, where r_n denotes the order of the bias of $\hat{m}(.)$. However, the bias in estimating $E_w\{m^2(X)\}$ via $n^{-1}\sum_{i=1}^n\{\hat{m}(X_i)\}^2$ is of the order r_n. This explains why the bias of $\hat{\zeta}_2^2$ is of a higher order infinitesimal than those of $\hat{\zeta}_1^2$ and $\hat{\zeta}_3^2$. The same observation applies to the $\hat{\xi}_i^2$'s. Similar phenomenon has been observed in the estimation of conditional variance functions by Fan and Yao (1998).

Remark 4. For Pearson's correlation ratio (1.3), we define estimates $\hat{\eta}_i^2 = 1/(1+\hat{\zeta}_i^2)$ for $i = 1, 2, 3$. It follows from Theorems 1 and 2 that all the $\hat{\eta}_i^2$'s are asymptotically normal with common asymptotic variance $\mathrm{Var}\,(Z_\eta)$, where

$$Z_\eta = -\left\{\frac{\mathrm{Var}_w\{m(X)\}}{\mathrm{Var}_w(Y)}\right\}^2 Z_\zeta = \frac{E_w\{\sigma^2(X)\}}{\mathrm{Var}_w^2(Y)}\{Y - E_w(Y)\}^2 - \frac{1}{\mathrm{Var}_w(Y)}\sigma^2(X)\epsilon^2 + c.$$

The asymptotic bias of $\hat{\eta}_i^2$ is $-\{\frac{\mathrm{Var}_w\{m(X)\}}{\mathrm{Var}_w(Y)}\}^2\lambda_{n,i}$ when $\hat{m}(.)$ is a locally linear smoother, and is $-\{\frac{\mathrm{Var}_w\{m(X)\}}{\mathrm{Var}_w(Y)}\}^2\pi_{n,i}$ when $\hat{m}(.)$ is a locally quadratic smoother. The three estimates for η^2 proposed by Doksum and Samarov (1995) have the same asymptotic variance $\mathrm{Var}\,(Z_\eta)$ (their Proposition 2.2).

4. Bias Correction and Examples

4.1. Bias correction

There are two obvious options to correct the bias in nonparametric estimation: (i) estimate the bias explicitly by using the asymptotic formulas derived in Theorems 1 and 2; (ii) undersmooth so that the bias is rendered negligible. The first approach involves estimating some unknown functions and is not pursued in this paper. The major difficulty in the second approach is to determine how much to undersmooth. However, for the problems concerned in this paper, we propose a simple and natural way of undersmoothing which is entirely determined by the data.

Theorems 1 and 2 show that we could use a locally quadratic estimator of $m(.)$ as the building block, since the derived estimates for noise-to-signal ratios have smaller biases if the same bandwidth is used. On the other hand, it is well

known that the bandwidth used in locally quadratic regression should be greater than that used in locally linear regression, simply because a quadratic function can accommodate more local variation in the data. In fact, the best bandwidth for estimating $m(.)$ is of the order $n^{-1/5}$ in the locally linear fitting, and is of the order $n^{-1/9}$ (with symmetric kernel) in the locally quadratic fitting. (See Section 3.2.3 of Fan and Gijbels (1996) and Section 4 of Hjellvik, Yao and Tjøstheim (1998).)

Based on the above observation, we propose the following scheme for undersmoothing: determine bandwidth $\hat{h}_2$ or $\hat{h}_3$ using either (2.6) or (2.7) with the locally linear estimator of $m(.)$, and estimate ξ^2 and ζ^2 using the locally quadratic estimator of $m(.)$ with either $\hat{h}_2$ or $\hat{h}_3$.

4.2. Simulation results

We illustrate our proposal through three numerical examples below. A Gaussian kernel has been used throughout. For illustration, we calculate the estimators $\hat{\zeta}_i^2$ and $\hat{\xi}_i^2$ with both locally linear and locally quadratic estimator of $m(.)$, and $i = 1, 2, 3$. The cross validation bandwidth $\hat{h}_2$ is used for the residual-based estimates as well as the plug-in estimates, and the maximum correlation bandwidth $\hat{h}_3$ is used for the correlation estimates.

The results from two simulated models (Examples 1 and 2) support the following general observations from Theorems 1 and 2.

- The bias of the residual-based estimate $\hat{\zeta}_2^2$ (resp. $\hat{\xi}_2^2$) is smaller than those of $\hat{\zeta}_1^2$ and $\hat{\zeta}_3^2$ (resp. $\hat{\xi}_1^2$ and $\hat{\xi}_3^2$). Therefore, the residual-based estimates are preferable.
- The variances of all the estimators $\hat{\zeta}_i^{2\prime}$s are about the same, and so are the variances of all the estimators $\hat{\xi}_i^{2\prime}$s.
- The estimators based on locally quadratic smoother of $m(.)$ are significantly less biased than those based on the locally linear smoother.

Remark 5. As a word of caution, our comparisons are based on the fact that we used the same bandwidth in both locally linear and locally quadratic estimates. For applications to high-dimensional X, locally quadratic fit involves estimating many local parameters and requires a reasonably large bandwidth in order to include enough data points. On the other hand, locally linear fit with a smaller bandwidth is easy to implement and will provide competitive performance in term of bias as well.

Example 1. Consider the model

$$Y_i = 2 - 5X_i + 5\exp\{-100(X_i - \frac{1}{2})^2\} + \tau\exp\{|X_i - \frac{1}{2}|\}\epsilon_i,$$

where $\{X_i\}$ and $\{\epsilon_i\}$ are two independent random series, X_i's are independent with the common distribution $U(0,1)$ and ϵ_i's are independent and standard normal. This is a generalized "bump" model considered by Härdle (1990) and Doksum and Samarov (1995). We simulated 400 random samples of size $n = 200$ for each of four different values of τ: 2, 1, $1/\sqrt{2}$, and $1/2$. We let $w(.) \equiv 1$. For each sample, we calculate the estimators $\hat{\zeta}_i^2$ and $\hat{\xi}_i^2$ for $i = 1, 2, 3$. The results are summarized in Table 1. We also calculated $\hat{\xi}_i^2$ $(i = 1, 2, 3)$ over a wide range of values of bandwidth h. Figure 1 plots the median over 400 Monte Carlo trials of estimators $\hat{\xi}_i^2$ against h. Note that $\hat{h}_2$ (with mean 0.034 and standard deviation 0.0076) and $\hat{h}_3$ (with mean 0.035 and standard deviation 0.0082) lie in a reasonably robust area. Further, the estimators based on locally quadratic smoother are less biased than those based on locally linear smoother over different values of h, and $\hat{\xi}_2^2$ has the smallest bias.

Table 1. The average biases and standard deviations (STDV) of the estimates in Example 1 based on locally linear (LL) and locally quadratic (LQ) estimators of $m(.)$ in Monte Carlo trials with 400 replications with sample size $n = 200$.

(ξ^2, ζ^2)		$\tau^2 = \frac{1}{4}$ $(0.094, 0.097)$		$\tau^2 = \frac{1}{2}$ $(0.188, 0.194)$		$\tau^2 = 1$ $(0.375, 0.388)$		$\tau^2 = 4$ $(1.502, 1.552)$	
		LL	LQ	LL	LQ	LL	LQ	LL	LQ
$\hat{\zeta}_1^2$ $(\hat{h}_2)$	Bias	0.020	0.008	0.045	0.019	0.103	0.044	0.543	0.175
	STDV	0.019	0.018	0.042	0.038	0.101	0.087	0.700	0.532
$\hat{\zeta}_2^2$ $(\hat{h}_2)$	Bias	0.015	0.007	0.031	0.017	0.063	0.034	0.221	0.063
	STDV	0.018	0.017	0.038	0.037	0.087	0.084	0.527	0.478
$\hat{\zeta}_3^2$ $(\hat{h}_3)$	Bias	0.042	0.012	0.070	0.026	0.125	0.050	0.401	0.115
	STDV	0.022	0.018	0.045	0.038	0.101	0.087	0.602	0.506
$\hat{\xi}_1^2$ $(\hat{h}_2)$	Bias	0.018	0.007	0.042	0.018	0.094	0.039	0.468	0.142
	STDV	0.019	0.018	0.041	0.037	0.094	0.083	0.591	0.475
$\hat{\xi}_2^2$ $(\hat{h}_2)$	Bias	0.014	0.007	0.029	0.015	0.058	0.030	0.186	0.041
	STDV	0.018	0.017	0.038	0.037	0.083	0.080	0.471	0.433
$\hat{\xi}_3^2$ $(\hat{h}_3)$	Bias	0.039	0.011	0.066	0.024	0.116	0.046	0.350	0.088
	STDV	0.021	0.018	0.043	0.038	0.095	0.083	0.529	0.457

(For $\tau^2 = \frac{1}{4}, \frac{1}{2}, 1$ and 4, the mean and STDV of $\hat{h}_2$ are $(0.025, 0.004)$, $(0.030, 0.006)$, $(0.035, 0.008)$ and $(0.050, 0.015)$ respectively; the mean and STDV of $\hat{h}_3$ are $(0.026, 0.005)$, $(0.030, 0.006)$, $(0.035, 0.008)$ and $(0.048, 0.016)$ respectively.)

Example 2. Consider the quadratic autoregressive model

$$X_{t+1} = 3.76X_t - 0.235X_t^2 + 0.3e_t,$$

where $\{e_t\}$ are independent with the common distribution $U[-\sqrt{3}, \sqrt{3}]$. We consider three cases: $Y_t = X_{t+m}$ for $m = 1, 2$ and 3. Note that for $m = 2$

or 3, the conditional variance functions are no longer constant. We evaluate the exact values of the conditional mean functions numerically for $m = 2$ and 3. Based on this, the true values of ζ_w^2 and ξ_w^2 can be easily obtained. We set $w(x) = I_{[2.5,14.5]}(x)$ which corresponds to about 80% inner sample range of the data. We simulated 400 random samples of size $n = 300$. The results from simulation are summarized in Table 2.

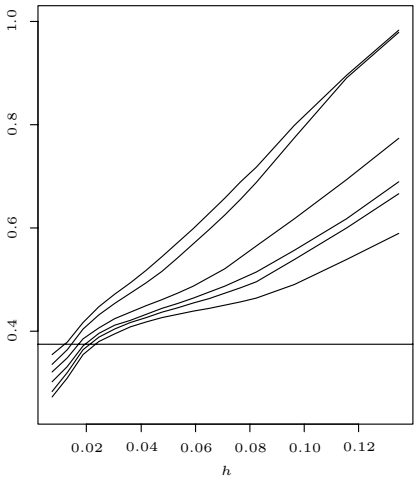

Figure 1. The medians over 400 Monte Carlo trials of six estimates for ξ^2 as functions of the bandwidth h. The horizontal line indicates the true value of $\xi^2 = 0.375$. From bottom to top: $\hat{\xi}_2^2$ with locally quadratic regression, $\hat{\xi}_1^2$ with locally quadratic regression, $\hat{\xi}_3^2$ with locally quadratic regression, $\hat{\xi}_2^2$ with locally linear regression, $\hat{\xi}_1^2$ with locally linear regression, and $\hat{\xi}_3^2$ with locally linear regression.

Example 3. For the deterministic model

$$X_{t+1} = 4X_t - 0.25X_t^2,$$

we consider the cases $Y_t = X_{t+m}$ for $m = 1, 3$ and 5. The data range from 0 to 1. It is easy to see that both ζ^2 and ξ^2 are 0. Let $w(x) \equiv 1$. We simulated 400 random samples of size $n = 200$. The results from simulation are summarized in Table 3. Although the asymptotic results stated in Section 3.1 do not strictly apply to purely deterministic models (i.e. with zero noise variance), the simulation results lend some support to the general conclusions. For example, the estimates based on locally quadratic regression are less biased than those based on the locally linear regression (with the same h), and the residual-based estimates are better than the other two types of estimates.

Table 2. The average biases and standard deviations (STDV), multiplying by 100, of the estimates in Example 2 based on locally linear (LL) and locally quadratic (LQ) estimators of $m(.)$ in Monte Carlo trials with 400 replications with sample size $n = 300$.

$(100 \times \xi^2, 100 \times \zeta^2)$		$m=1$ $(0.077, 0.853)$ LL	$m=1$ LQ	$m=2$ $(0.432, 2.538)$ LL	$m=2$ LQ	$m=3$ $(1.439, 9.249)$ LL	$m=3$ LQ
$\hat{\zeta}_1^2$ $(\hat{h}_2)$	Bias	-0.491	-0.402	-0.928	-0.960	-1.736	-1.001
	STDV	0.027	0.026	0.203	0.189	0.711	0.783
$\hat{\zeta}_2^2$ $(\hat{h}_2)$	Bias	-0.301	-0.291	-0.901	-0.901	-1.552	-0.865
	STDV	0.026	0.026	0.190	0.192	0.708	0.530
$\hat{\zeta}_3^2$ $(\hat{h}_3)$	Bias	0.419	-0.388	-1.006	0.676	-1.696	-1.299
	STDV	0.032	0.042	0.195	0.239	0.717	0.621
$\hat{\xi}_1^2$ $(\hat{h}_2)$	Bias	-0.016	-0.010	-0.082	-0.048	-0.206	-0.020
	STDV	0.005	0.004	0.059	0.062	0.234	0.233
$\hat{\xi}_2^2$ $(\hat{h}_2)$	Bias	-0.016	-0.004	-0.077	-0.045	-0.201	0.102
	STDV	0.004	0.003	0.059	0.060	0.233	0.242
$\hat{\xi}_3^2$ $(\hat{h}_3)$	Bias	0.013	-0.006	0.086	-0.067	0.251	-0.138
	STDV	0.006	0.009	0.077	0.057	0.358	0.237

(For $m = 1, 2$ and 3, the mean and STDV of $\hat{h}_2$ are $(0.353, 0.003)$, $(0.360, 0.004)$ and $(0.271, 0.003)$ respectively; the mean and STDV of $\hat{h}_3$ are $(0.491, 0.035)$, $(0.359, 0.006)$ and $(0.284, 0.003)$ respectively.)

Table 3. The average means and standard deviations (STDV) of the estimates in Example 3 based on locally linear (LL) and locally quadratic (LQ) estimators of $m(.)$ in Monte Carlo trials with 400 replications with sample size $n = 200$. The true values of both ξ^2 and ζ^2 are 0.

		$m=1$ LL	$m=1$ LQ	$m=3$ LL	$m=3$ LQ	$m=5$ LL	$m=5$ LQ
$\hat{\zeta}_1^2$ $(\hat{h}_2)$	Mean	0.0000	0.0000	0.0006	0.0000	0.1265	0.0458
	STDV	0.0000	0.0000	0.0006	0.0000	0.0608	0.0366
$\hat{\zeta}_2^2$ $(\hat{h}_2)$	Mean	0.0000	0.0000	0.0006	0.0000	0.1088	0.0429
	STDV	0.0000	0.0000	0.0006	0.0000	0.0471	0.0337
$\hat{\zeta}_3^2$ $(\hat{h}_3)$	Mean	0.0000	0.0000	0.0053	0.0011	0.1390	0.0635
	STDV	0.0000	0.0000	0.0048	0.0036	0.1100	0.0645
$\hat{\xi}_1^2$ $(\hat{h}_2)$	Mean	0.0000	0.0000	0.0002	0.0000	0.0344	0.0139
	STDV	0.0000	0.0000	0.0002	0.0000	0.0155	0.0111
$\hat{\xi}_2^2$ $(\hat{h}_2)$	Mean	0.0000	0.0000	0.0002	0.0000	0.0328	0.0135
	STDV	0.0000	0.0000	0.0002	0.0000	0.0145	0.0107
$\hat{\xi}_3^2$ $(\hat{h}_3)$	Mean	0.0000	0.0000	0.0017	0.0004	0.0402	0.0191
	STDV	0.0000	0.0000	0.0016	0.0012	0.0311	0.0195

(For $m = 1, 3$ and 5, the mean and STDV of $\hat{h}_2$ are $(0.097, 0.000)$, $(0.079, 0.006)$ and $(0.066, 0.000)$ respectively; the mean and STDV of $\hat{h}_3$ are $(0.106, 0.004)$, $(0.067, 0.001)$ and $(0.051, 0.001)$ respectively.)

4.3. Applications

Finally, we apply the procedure to the following three data sets.

(i) The Great Salt Lake (GSL) data from Utah. We have fitted the bi-weekly volume data (the first 3200 points) of the GSL with a nonlinear autoregressive model with sampling time 12 and order 4. The order was determined by the cross-validation method. For the information on this data set, we refer to Sangoyomi, Lall and Abarbanel (1996).

(ii) Wolf's annual sunspot numbers (1700-1994). We have fitted the data with the optimal subset regression model determined by the cross-validation method, and it consists of the lagged variables at lags 1, 2, and 4. (See Yao and Tong (1994)).

(iii) The monthly New York measles data. In order to avoid possible outliers, we use only the first 158 points. We have fitted the data, on the natural log base, with the optimal subset regression model determined by the the cross-validation method, and it consists of the lagged variables at lags 1, 4, and 7.

We divide the data by their standard deviation first for each data set. We apply both the residual-based estimate and the plug-in estimate with the cross validation bandwidth $\hat{h}_2$, and the correlation estimate with the correlation bandwidth $\hat{h}_3$. We use $w(.)$ as the indicator function of the set on which the (estimated) density function of X is not smaller than 0.01. The results are reported in Table 4. Note that the fact that the values of $\hat{\xi}_i^2$ for the GSL data are small lends further support to the suggestion that this data set might be treated as operationally deterministic. (See Sangoyomi, Lall and Abarbanel (1996), Yao and Tong (1998a)). It is also clear that the noise level in the sunspot data is higher, comparing to, say, log measles data.

Table 4. Estimated noise-to-signal ratios for three data sets.

data set	GSL	Sunspot numbers	Measles data
regressors	Y_{t-12}, Y_{t-24}, Y_{t-36}, Y_{t-48}	Y_{t-1}, Y_{t-2}, Y_{t-4}	Y_{t-1}, Y_{t-4}, Y_{t-7}
n	3152	291	151
$\hat{\zeta}_1^2$	0.655	1.172	0.100
$\hat{\zeta}_2^2$	0.667	1.211	0.097
$\hat{\zeta}_3^2$	0.687	1.210	0.145
$\hat{\xi}_1^2$	0.028	0.608	0.073
$\hat{\xi}_2^2$	0.029	0.627	0.072
$\hat{\xi}_3^2$	0.032	0.627	0.068
$\hat{h}_2$	0.18	0.82	0.58
$\hat{h}_3$	0.22	1.34	1.02

Acknowledgements

We thank Professor Jianqing Fan for his critical reading and insightful comments of an earlier version of the paper, Professor Alex Samarov for helpful discussions, and anonymous reviewers for helpful comments. This research was partially supported by the EPSRC Grant L16358.

Appendix

We present the proof of Theorem 1 only, the proof for Theorem 2 is similar. We consider only the cases with $\delta > 0$. When $\delta = 0$, the proof is more direct and simpler (see Remark 1).

We use the same notation as in Section 3. We say $B_n(x) = B(x) + o_p(b_n)$ (resp. $O_p(b_n)$) uniformly for $x \in G$ if $\sup_{x\in G} |B_n(x) - B(x)| = o_p(b_n)$(resp.$O_p(b_n)$).

Let $\hat{\beta} \equiv (\hat{m}(x), \hat{\dot{m}}(x))^\tau$ be the locally linear estimators of $m(x)$ and its derivative $\dot{m}(x) = \frac{d}{dx}m(x)$ derived from (2.1). In the case $d = 1$, the least squares theory gives

$$\hat{\beta} = (\mathbf{X}^\tau \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\tau \mathbf{W}\mathbf{Y}, \tag{A.1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^\tau$, $\mathbf{W} = \mathrm{diag}(K(\frac{X_1-x}{h}), \ldots, K(\frac{X_n-x}{h}))$, and $\mathbf{X}$ is an $n \times 2$ matrix with $(1, X_i - x)$ as the ith row. More specifically,

$$\hat{m}(x) = \frac{1}{nh}\sum_{i=1}^{n} W_n\left(\frac{X_i - x}{h}, x\right) Y_i, \tag{A.2}$$

where

$$W_n(t, x) = (1, 0)S_n^{-1}(x)(1, t)^\tau K(t), \tag{A.3}$$

and $S_n(x)$ is a 2×2 matrix with the (i, j)-th element $s_{i+j-2}(x)$, and

$$s_j(x) = \frac{1}{nh}\sum_{i=1}^{n}\left(\frac{X_i - x}{h}\right)^j K\left(\frac{X_i - x}{h}\right). \tag{A.4}$$

Lemma 1. *Assume that conditions* (C2), (C4) *and* (C5) *hold. For any bounded subset* $G \subset R$, *as* $n \to \infty$, $\sup_{x\in G} |s_j(x) - E\{s_j(x)\}| = o_p(h)$, $0 \le j \le 4$.

Proof. We prove only the case with $j = 0$. First we apply an exponential type inequality (Theorem 1.3(2) in Bosq (1996)) to prove that for any $\varepsilon > 0$ and $x \in G$,

$$P\{|s_0(x) - \mathrm{E}\, s_0(x)| \ge h\varepsilon\} \le 4e^{-c_1 h^3 n} + c_2 h^{-1} n^{-(\frac{3}{2\delta}+\frac{5}{4})} \equiv \pi_n, \tag{A.5}$$

where $c_1, c_2 > 0$ are constants independent of x.

Let $X_i^* = K_h(X_i - x) - E\{K_h(X_i - x)\}$, where $K_h(x) = h^{-1}K(x/h)$. By (C2), $|X_i^*| \le ch^{-1}$. Using Bosq's notation with $q = [n^{1/4}]$ and $p = n/(2q)$, we have

$$\begin{aligned}\sigma^2(q) =& \max_{0\le j\le 2q-1} E\{([jp]+1-jp)X^*_{[jp]+1} + X^*_{[jp]+2} + \cdots + X^*_{[(j+1)p]} \\ &+((j+1)p - [(j+1)p])X^*_{[(j+1)p+1]}\}^2 \\ \le& \max_{0\le j\le 2q-1}\left\{\sum_{[jp]<i\le[(j+1)p+1]} E(X_i^*)^2 + 2p^* \sum_{[jp]+1<i\le[(j+1)p+1]} |E(X^*_{[jp]+1}X_i^*)|\right\} \\ \le& \max_{0\le j\le 2q-1}\left\{cp^*h^{-1} + 2p^*h^{-(2+\frac{2\delta}{1+\delta})}\sum_{j=1}^{p^*-1}\beta^{\frac{\delta}{1+\delta}}(j)\right\} = O(ph^{-1}),\end{aligned}$$

where $p* = [(j+1)p+1] - [jp] + 1$. The last inequality follows from Lemma 1 of Yoshihara (1976). Therefore, $\nu^2(q) = \frac{2}{p^2}\sigma^2(q) + c = O(p^{-1}h^{-1})$. By Theorem 1.3(2) of Bosq (1996), the LHS of (A.5) is not greater than

$$\begin{aligned}&4\exp\left\{-\frac{\varepsilon^2h^2}{8\nu^2(q)}q\right\} + 22\left(1+\frac{8c}{\varepsilon^2h^2}\right)^{1/2} q\beta\left(\left[\frac{n}{2q}\right]\right) \\ &= 4e^{-c_1h^3n} + 22\left(1+\frac{8c}{\varepsilon^2h^2}\right)^{1/2} n^{\frac14}\beta\left(\left[\frac{n^{3/4}}{2}\right]\right).\end{aligned}$$

It follows from (C4) that $\beta(n) = o(n^{-2(1+\delta)/\delta})$. Therefore, the second term on the RHS of the above expression is bounded above by $c_2h^{-1}n^{1/4}n^{-\frac{3(1+\delta)}{2\delta}} = c_2h^{-1}n^{-\frac{3}{2\delta}-\frac54}$. Thus, (A.5) holds.

Now we cover G by a finite number of open intervals B_k centered at x_k in such a way that

$$G \subset \bigcup_{k=1}^{l_n} B_k, \qquad \sup_{x\in B_k}|x - x_k| \le h^{3+\varepsilon_0}, \qquad l_n \le ch^{-(3+\varepsilon_0)}, \tag{A.6}$$

where $\varepsilon_0 \in (0,1)$ is a constant. Consequently, for $x \in B_k$, $|K_h(X_i - x) - K_h(X_i - x_k)| \le ch^{1+\varepsilon_0}$ for all X_i, where c is independent of k. Therefore, $\sup_{x\in B_k}|s_0(x) - s_0(x_k) - Es_0(x_k) + Es_0(x_k)| \le ch^{1+\varepsilon_0}$. It follows from the above arguments that

$$\begin{aligned}&P\{\sup_{x\in G}|s_0(x) - \mathrm{E}\, s_0(x)| \ge h\varepsilon\} \\ \le& P\{\max_{1\le k\le l_n}|s_0(x_k) - \mathrm{E}\, s_0(x_k)| + \max_{1\le k\le l_n}\sup_{x\in B_k}|s_0(x) - \mathrm{E}\, s_0(x) - s_0(x_k) + \mathrm{E}\, s_0(x_k)| \ge h\varepsilon\} \\ \le& P\{\max_{1\le k\le l_n}|s_0(x_k) - \mathrm{E}\, s_0(x_k)| + ch^{1+\varepsilon_0} \ge h\varepsilon\} \le l_n\pi_n,\end{aligned}$$

which converges to 0 (see condition (C5)). The proof is complete.

Lemma 2. *Assume that conditions* (C2), (C4) *and* (C5) *hold, and* $G \subset \{p(x) > 0\}$ *is a compact set on which* $\ddot{m}(x)$ *is uniformly continuous. As* $n \to \infty$, *uniformly for* $x \in G$,

$$\begin{aligned}
&\hat{m}(x) - m(x) \\
&= \frac{1}{nhp(x)} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \{Y_i - m(x) - \dot{m}(x)(X_i - x)\} + O_p\{R_n(x)\} \quad \text{(A.7)} \\
&= \frac{1}{nhp(x)} \sum_{i=1}^{n} \sigma(X_i)\epsilon_i K\left(\frac{X_i - x}{h}\right) + \frac{h^2\sigma_0^2}{2}\ddot{m}(x) + O_p\{R_n(x)\}, \quad \text{(A.8)}
\end{aligned}$$

where

$$\begin{aligned}
R_n(x) = &\frac{1}{np(x)} \left\{ \left| \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \{Y_i - m(x) - \dot{m}(x)(X_i - x)\} \right| \right. \\
&\left. + \left| \sum_{i=1}^{n} \frac{X_i - x}{h} K\left(\frac{X_i - x}{h}\right) \{Y_i - m(x) - \dot{m}(x)(X_i - x)\} \right| \right\} \\
= &\frac{1}{np(x)} \left\{ \left| \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \sigma(X_i)\epsilon_i \right| + \left| \sum_{i=1}^{n} \frac{X_i - x}{h} K\left(\frac{X_i - x}{h}\right) \sigma(X_i)\epsilon_i \right| \right\} \\
&+ O(h^3).
\end{aligned}$$

Proof. Let $\beta = (m(x), \dot{m}(x))^\tau$. It follows from (A.1) that $(\hat{\beta} - \beta) = (\mathbf{X}^\tau \mathbf{W} \mathbf{X})^{-1} \mathbf{X} \mathbf{W}^\tau (\mathbf{Y} - \mathbf{X}\beta)$. Similar to (A.2), we have

$$\hat{m}(x) - m(x) = \frac{1}{nh} \sum_{i=1}^{n} W_n\left(\frac{X_i - x}{h}, x\right) \{Y_i - m(x) - \dot{m}(x)(X_i - x)\}, \quad \text{(A.9)}$$

where W_n is defined in (A.3). Let $S(x)$ be the 2×2 diagonal matrix with $p(x)$ and $p(x)\sigma_0^2$ as its two diagonal elements. It is easy to see that $ES_n(x) = S(x) + O(h)$ uniformly on compact sets. It follows from Lemma 1 that the following limits hold uniformly on compact sets:

$$S_n(x) = S(x) + O_p(h), \qquad \det\{S_n(x)\} = p^2(x)\sigma_0^2 + O_p(h). \quad \text{(A.10)}$$

Therefore,

$$S_n^{-1}(x) = S^{-1}(x) + O_p(h) \quad \text{(A.11)}$$

uniformly for $x \in G \subset \{p(x) > 0\}$. Let $Y_i^* = Y_i - m(x) - \dot{m}(x)(X_i - x)$. It is easy to see from (A.3) that

$$\left| \sum_{i=1}^{n} \left\{ W_n\left(\frac{X_i - x}{h}, x\right) - p^{-1}(x) K\left(\frac{X_i - x}{h}\right) \right\} Y_i^* \right|$$

$$
\begin{aligned}
&= \left| (1,0)\{S_n^{-1}(x) - S^{-1}(x)\} \sum_{i=1}^{n} \left(1, \frac{X_i - x}{h}\right)^{\tau} K\left(\frac{X_i - x}{h}\right) Y_i^* \right| \\
&\le \frac{1}{p(x)} \left[(1,0)\{S_n^{-1}(x) - S^{-1}(x)\}^2 (1,1)^{\tau} \right]^{1/2} \\
&\quad \times \left\{ \left| \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i^* \right|^2 + \left| \sum_{i=1}^{n} \frac{X_i - x}{h} K\left(\frac{X_i - x}{h}\right) Y_i^* \right|^2 \right\}^{1/2} \\
&\le \frac{2}{p(x)} \left[(1,0)\{S_n^{-1}(x) - S^{-1}(x)\}^2 (1,1)^{\tau} \right]^{1/2} \\
&\quad \times \left\{ \left| \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i^* \right| + \left| \sum_{i=1}^{n} \frac{X_i - x}{h} K\left(\frac{X_i - x}{h}\right) Y_i^* \right| \right\}. \qquad \text{(A.12)}
\end{aligned}
$$

It follows from (A.11) that $\sup_{x \in G} \left[(1,0)\{S_n^{-1}(x) - S^{-1}(x)\}^2 (1,1)^{\tau} \right]^{1/2} = O_p(h)$. Since $K(.)$ has a bounded support, Y_i^* on the RHS of (A.12) can be replaced by $\sigma(X_i)\epsilon_i + \frac{1}{2}\ddot{m}(x)(X_i - x)^2 + o(h^2)$. It follows from (A.10) that $\frac{1}{nh^3 p(x)} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(X_i - x)^2 = \sigma_0^2 + O_p(h)$ uniformly for $x \in G$. Now (A.7) follows from (A.9) and (A.12), and (A.8) follows from (A.7) consequently.

Lemma 3. *Let conditions* (C1) – (C5) *hold.*

(i) *Let $f(x,y)$ be a measurable function which is continuous in x. Further, $E\{|f(X,Y)|^{2(1+\delta)} + |f(X,Y)Y|^{2(1+\delta)}\} < \infty$, where $\delta \in [0,\infty)$ is a constant given in* (C4). *Then, as $n \to \infty$,*

$$
\begin{aligned}
&\sum_{i=1}^{n} f(X_i, Y_i)\{\hat{m}(X_i) - m(X_i)\} w(X_i) = \sum_{i=1}^{n} \sigma(X_i)\epsilon_i w(X_i) \int f(X_i, y) g(y|X_i) dy \\
&+ \frac{nh^2\sigma_0^2}{2} E_w\{f(X,Y)\ddot{m}(X)\} + o_p(nh^2 + h^{-1}), \qquad \text{(A.13)}
\end{aligned}
$$

where $g(.|x)$ denotes the conditional density function of Y given $X = x$.

(ii) *As $n \to \infty$,*

$$
\begin{aligned}
&\sum_{i=1}^{n} \{\hat{m}(X_i) - m(X_i)\}^2 w(X_i) \\
&= nh^4\sigma_0^4 E_w\{\ddot{m}^2(X)\}/4 + h^2\sigma_0^2 \sum_{i=1}^{n} \sigma(X_i)\epsilon_i \ddot{m}(X_i) w(X_i) + o_p(nh^4 + h^{-1}).
\end{aligned}
$$

Proof. We only prove (i) and note that (ii) can be proved in a similar manner.

First, assume $\int f(X,y)g(y|X)dy \neq 0$ almost surely. It follows from (A.8) that

$$
\sum_{i=1}^{n} f(X_i, Y_i)\{\hat{m}(X_i) - m(X_i)\} w(X_i) = (I_1 + I_2)\{1 + o_p(1)\}, \qquad \text{(A.14)}
$$

where

$$I_1 = \frac{h^2\sigma_0^2}{2}\sum_{i=1}^{n} f(X_i,Y_i)\ddot{m}(X_i)w(X_i) = \frac{nh^2\sigma_0^2}{2}E_w\{f(X,Y)\ddot{m}(X)\} + O_p(\sqrt{n}h^2), \tag{A.15}$$

$$\begin{aligned} I_2 &= \frac{1}{nh}\sum_{i,j=1}^{n} f(X_j,Y_j)w(X_j)K\left(\frac{X_i - X_j}{h}\right)\sigma(X_i)\epsilon_i/p(X_j) \\ &= \frac{K(0)}{nh}\sum_{i=1}^{n} f(X_i,Y_i)\sigma(X_i)\epsilon_i w(X_i)/p(X_i) + \frac{1}{nh}\sum_{1\le i<j\le n} K\left(\frac{X_i - X_j}{h}\right) \\ &\qquad \times\left\{f(X_j,Y_j)\sigma(X_i)\epsilon_i\frac{w(X_j)}{p(X_j)} + f(X_i,Y_i)\sigma(X_j)\epsilon_j\frac{w(X_i)}{p(X_i)}\right\} \\ &\equiv I_{21} + I_{22}, \qquad \text{say.} \end{aligned}$$

The limit in (A.15) follows from the ergodicity of the process $\{X_i, Y_i\}$ and Theorem 1.7 of Peligrad (1986). It is also easy to see that $I_{21} = O_p(\frac{1}{\sqrt{n}h})$.

We denote the summand in I_{22} on the RHS of (A.16) by φ_{ij}, and write I_{22} as

$$I_{22} = \frac{1}{nh}\sum_{1\le i<j\le n}(\varphi_{ij} - \varphi_i - \varphi_j) + \frac{1}{h}\sum_{i=1}^{n}\varphi_i. \tag{A.16}$$

This is Hoeffding's projection decomposition of a U-statistic. Note that $K(.)$ has a compact support. Therefore in the above expression,

$$\begin{aligned} \varphi_i &= \sigma(X_i)\epsilon_i\int K\left(\frac{X_i - x}{h}\right)f(x,y)w(x)g(y|x)dxdy \\ &= h\sigma(X_i)\epsilon_i w(X_i)\int f(X_i,y)g(y|X_i)dy\{1 + o(h)\}. \end{aligned}$$

It follows from Lemma A(ii) of Hjellvik *et al.*(1998) that

$$\begin{aligned} &P\left\{\frac{1}{nh^{(\frac{1}{1+\delta}-\varepsilon_0)/2}}\left|\sum_{1\le i<j\le n}(\varphi_{ij} - \varphi_i - \varphi_j)\right| > \varepsilon\right\} \\ &\le \frac{ch^{\varepsilon_0}}{n^2}E\left\{\frac{1}{h^{\frac{1}{2(1+\delta)}}}\sum_{1\le i<j\le n}(\varphi_{ij} - \varphi_i - \varphi_j)\right\}^2 = o(h^{\varepsilon_0}). \end{aligned}$$

Therefore, the first term in (A.16) is of the order $o_p(h^{-\frac{1+2\delta+\varepsilon_0}{2(1+\delta)}}) = o_p\left(h^{-1}\right)$, provided $\varepsilon_0 < (1+\delta)^{-1}$. Therefore, overall we have that $I_2 = \sum_{i=1}^{n}\sigma(X_i)\epsilon_i w(X_i)\int f(X_i,y)g(y|X_i)dy + o_p(\sqrt{n}h + h^{-1})$. Together with (A.14) and (A.15), (i) holds.

In the case that $\int f(X,y)g(y|X)dy = 0$ almost surely, it is easy to see that $I_1 = o_p(nh^2)$, $I_{21} = O_p(n^{-1/2}h^{-1})$ and $I_{22} = o_p(h^{-1})$. Therefore, (A.13) still holds with the first two terms on its RHS being 0. The proof is complete.

Proof of Theorem 1. We proceed with the proof for $i = 3$ only, since the cases $i = 1$ and 2 are much simpler. Write $\hat{\zeta}_3^2 = D_2/D_1$, where

$$\begin{aligned} D_1 &= \frac{1}{n}\sum_{i=1}^n Y_i\hat{m}(X_i)w(X_i) - \bar{Y}_w\bar{m} \\ &= \frac{1}{n}\sum_{i=1}^n Y_i\{\hat{m}(X_i) - m(X_i)\}w(X_i) - \bar{Y}_w\frac{1}{n}\sum_{i=1}^n\{\hat{m}(X_i) - m(X_i)\}w(X_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^n m^2(X_i)w(X_i) + \frac{1}{n}\sum_{i=1}^n\{m(X_i) - \bar{m}\}\sigma(X_i)\epsilon_i w(X_i) \\ &\quad - \left\{\frac{1}{n}\sum_{i=1}^n m(X_i)w(X_i)\right\}^2, \end{aligned} \tag{A.17}$$

$$\begin{aligned} D_2 &= \frac{1}{n}\sum_{i=1}^n Y_i^2 w(X_i) - \bar{Y}_w^2 - \frac{1}{n}\sum_{i=1}^n Y_i\hat{m}(X_i)w(X_i) + \bar{Y}_w\bar{m} \\ &= \bar{Y}_w\frac{1}{n}\sum_{i=1}^n\{\hat{m}(X_i) - m(X_i)\}w(X_i) - \frac{1}{n}\sum_{i=1}^n Y_i\{\hat{m}(X_i) - m(X_i)\}w(X_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^n m(X_i)\sigma(X_i)\epsilon_i w(X_i) - \bar{Y}_w\frac{1}{n}\sum_{i=1}^n\sigma(X_i)\epsilon_i w(X_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^n\sigma^2(X_i)\epsilon_i^2 w(X_i). \end{aligned} \tag{A.18}$$

Note that

$$\begin{aligned} \sqrt{n}(\hat{\xi}_3^2 - \xi_w^2) &= \frac{\sqrt{n}D_2}{\mathrm{Var}_w\{m^2(X)\}D_1}[\mathrm{Var}_w\{m^2(X)\} - D_1] \\ &\quad + \frac{\sqrt{n}}{\mathrm{Var}_w\{m^2(X)\}}[D_2 - \mathrm{Var}_w\{\sigma^2(X)\}]. \end{aligned}$$

Substituting D_1 and D_2 from (A.17) and (A.18), the conclusion follows from Lemma 3, the Ergodic Theorem, and the Central Limit Theorem, which is implied by Theorem 1.7 of Peligrad (1986).

References

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

Bosq, D. (1996). *Nonparametric Statistics for Stochastic Processes*. Springer, New York.

Box, G. E. P. (1988). Signal-to-noise ratios, performance criteria, and transformations. *Technometrics* **30**, 1-17.

Broomhead, D. S. (1995). Nonlinear signal processing. In *Complex Stochastic Systems and Engineering* (Edited by D. M. Titterington), 13-28. Oxford University Press, Oxford.

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.

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

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87**, 998-1004.

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

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

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645-660.

Feinstein, A. (1958). *Foundations of Information Theory*. McGraw-Hill, New York.

Härdle, W. (1990). *Applied Nonparametric Regressions*. Combridge University Press, Combridge.

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

Hastie, T. J. and Loader, C. (1993). Local regression: automatic kernel carpentry (with discussion). *Statist. Sci.* **8**, 120-143.

Hjellvik, V., Yao, Q. and Tjøstheim, D. (1998). Linearity testing using local polynomial approximation. *J. Statist. Plann. Inference* **68**, 295-321.

Marron, J. S. and Härdle, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation. *J. Multivariate Anal.* **20**, 91-113.

Peligrad, M. (1986). Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables. *Dependence in Probability and Statistics* (Edited by E. Eberlein and M. S. Taqqu), 193-223. Birkhäuser, Boston.

Sangoyomi, T. B., Lall, U. and Abarbanel, H. D. I. (1996). Nonlinear dynamics of the great salt lake: dimension estimation. *Water Resources Research* **32**, 149-159.

Tong, H. (1995). A personal overview of nonlinear time series from a chaos perspective (with discussion). *Scand. J. Statist.* **22**, 399-445.

Yao, Q. and Tong, H. (1994). On subset selection in non-parametric stochastic regression. *Statist. Sinica* **4**, 51-70.

Yao, Q. and Tong, H. (1998a). A bootstrap detection for operational determinism. *Physica* D **115**, 49-55.

Yao, Q. and Tong, H. (1998b). Cross-validatory bandwidth selections for regression estimation based on dependent data. *J. Statist. Plann. Inference* **68**, 387-415.

Yoshihara, K. (1976). Limiting behaviour of U-statistics for stationary absolutely regular processes. *Z. Wahr. v. Gebiete* **35**, 237-252.

Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, U.K.

E-mail: Q.Yao@lse.ac.uk

(Received March 1998; accepted November 1999)