Biometrika (1998), **85**, 3, *pp.* 645–660
Printed in Great Britain

Efficient estimation of conditional variance functions in stochastic regression

BY JIANQING FAN

Department of Statistics, University of California, Los Angeles, California 90095, U.S.A.
jfan@stat.cla.edu

AND QIWEI YAO

Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent CT2 7NF, U.K.
q.yao@ukc.ac.uk

SUMMARY

Conditional heteroscedasticity has often been used in modelling and understanding the variability of statistical data. Under a general set-up which includes nonlinear time series models as a special case, we propose an efficient and adaptive method for estimating the conditional variance. The basic idea is to apply a local linear regression to the squared residuals. We demonstrate that, without knowing the regression function, we can estimate the conditional variance asymptotically as well as if the regression were given. This asymptotic result, established under the assumption that the observations are made from a strictly stationary and absolutely regular process, is also verified via simulation. Further, the asymptotic result paves the way for adapting an automatic bandwidth selection scheme. An application with financial data illustrates the usefulness of the proposed techniques.

Some key words: Absolutely regular; ARCH; Conditional variance; Efficient estimator; Heteroscedasticity; Local linear regression; Nonlinear time series; Volatility.

1. INTRODUCTION

Many scientific studies depend on understanding the local variability of the data, which is often described by the conditional variance or the volatility function in a statistical model. It is of common interest to estimate conditional variance functions in a variety of statistical applications such as measuring volatility or risk in finance (Andersen & Lund, 1997; Gallant & Tauchen, 1997), monitoring the reliability of nonlinear prediction (Yao & Tong, 1994), identifying homoscedastic transforms in regression (Carroll & Ruppert, 1988, Ch. 4), choosing optimal design and understanding residual pattern (Müller & Stadtmüller, 1987; Gasser, Sroka & Jennen-Steinmetz, 1986), monitoring the signal-to-noise ratios in quality control of experimental design (Box, 1988) and so on. The problem can be mathematically formulated as follows.

Let $\{(Y_i, X_i)\}$ be a two-dimensional strictly stationary process having the same marginal distribution as (Y, X). Let $m(x) = E(Y \mid X = x)$ and $\sigma^2(x) = \text{var}(Y \mid X = x) \neq 0$. We write a regression model of Y_i on X_i as

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i. \tag{1·1}$$

Then $E(\varepsilon_i | X_i) = 0$ and $\text{var}(\varepsilon_i | X_i) = 1$, although the conditional distribution of ε_i given $X_i = x$ may still depend on x. For $X_i = Y_{i-1}$, (1·1) is an autoregressive conditional heteroscedastic (ARCH) time series model, and $\sigma(.)$ is called the volatility function (Engle, 1982). Connections of model (1·1) with the one-factor diffusion model in finance will be discussed in § 4. The aim of this paper is to derive an efficient fully-adaptive procedure for estimating $\sigma^2(.)$.

The simple decomposition $\sigma^2(x) = E(Y^2 | X = x) - m^2(x)$ motivates the following obvious and direct estimator:

$$\hat{\sigma}_d^2(x) = \hat{v}(x) - \{\hat{m}(x)\}^2, \quad (1·2)$$

where $\hat{m}(.)$ and $\hat{v}(x)$ are respectively estimators for $m(.)$ and $v(x) \equiv E\{Y^2 | X = x\}$ (Yao & Tong, 1994; Härdle & Tsybakov, 1997). However, $\hat{\sigma}_d^2(.)$ is not always nonnegative, especially if different smoothing parameters are used in estimating $m(.)$ and $v(.)$. Furthermore, such a direct method can create a very large bias; see § 3·1 below. Härdle & Tsybakov (1997) recognised these problems and used a common bandwidth and a common kernel to reduce the bias. While their idea is useful, the approach is still not fully adaptive to the unknown regression function $m(.)$. An alternative regression-adaptive approach is to apply the difference-based estimator (Rice, 1984; Gasser et al., 1986; Müller & Stadtmüller, 1987; see also § 3·2 below), which uses a high-pass filter to remove the regression function from the data sequence $\{Y_i\}$. Hall, Kay & Titterington (1990) demonstrated that the resulting estimator was inefficient even in homoscedastic models with optimal filters.

In this paper, we consider a residual-based estimator of the conditional variance. While the idea is not new (Hall & Carroll, 1989; Neumann, 1994), its implications and implementation are novel. In particular, we show that our estimator is fully regression-adaptive in the sense that, without knowing $m(.)$, we can estimate the conditional variance function $\sigma^2(.)$ asymptotically as well as if $m(.)$ were known. After we completed this paper, we found that this phenomenon is observed independently by Ruppert et al. (1997) in regression models with independent and identically distributed observations.

One interesting feature of our approach is that we do not need to undersmooth the regression function $m(.)$ in order to obtain a regression-adaptive estimator for the conditional variance $\sigma^2(.)$. In practice, this implies that we can use a data-driven bandwidth selector in estimating $m(.)$, and then apply the same bandwidth selector with the squared residuals to estimate $\sigma^2(.)$. This is in marked constrast with the previous methods, where new bandwidth, or filter length, selection problems are encountered.

The paper is organised as follows. In § 2, we propose and study the residual-based estimator of the conditional variance based on local linear regression. In § 3, we compare the performance of our estimator with various procedures in the literature and discuss their mutual relationship. In § 4, we present numerical applications with a financial data set and two simulated models. All the technical proofs are given in Appendix 2.

2. MAIN RESULTS

2·1. *The estimator*

If the regression function $m(.)$ is given, we can regard the problem of estimating $\sigma^2(.)$ as a nonparametric regression problem in view of the relation $E(r | X = x) = \sigma^2(x)$, where $r = \{Y - m(X)\}^2$. Given the observations $\{(Y_i, X_i), 1 \leqslant i \leqslant n\}$ from model (1·1), we write $r_i = \{Y_i - m(X_i)\}^2$. Then the local linear estimator of $\sigma^2(.)$ is $\hat{\sigma}_b^2(x) = \hat{\alpha}$, the subscript b

standing for 'benchmark', where

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \{r_i - \alpha - \beta(X_i - x)\}^2 W\left(\frac{X_i - x}{h_1}\right); \tag{2·1}$$

$W(.)$ is a density function on R and $h_1 > 0$ is a bandwidth (Fan & Gijbels, 1996, p. 58). The local linear estimators have several nice properties. They possess high statistical efficiency in an asymptotic minimax sense and are design-adaptive (Fan, 1993). Further, they automatically correct edge effects (Fan & Gijbels, 1992; Ruppert & Wand, 1994; Hastie & Loader, 1993). Therefore, $\hat{\sigma}_b^2(.)$ provides a benchmark for our problem.

In practice, $m(.)$ is typically unknown. A natural approach is to substitute $m(.)$ by a nonparametric regression estimator. We choose the local linear estimator because of its aforementioned optimal properties. Let $\hat{m}(x) = \hat{a}$ be the local linear estimator that solves the following weighted least-squares problem:

$$(\hat{a}, \hat{b}) = \arg\min_{a,b} \sum_{i=1}^{n} \{Y_i - a - b(X_i - x)\}^2 K\left(\frac{X_i - x}{h_2}\right), \tag{2·2}$$

where $K(.)$ is a density function on R and $h_2 > 0$ is a bandwidth. Denote the squared residuals by $\hat{r}_i = \{Y_i - \hat{m}(X_i)\}^2$. This leads to the residual-based estimator $\hat{\sigma}^2(x) = \hat{\alpha}$ with kernel W and bandwidth h_1, where

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \{\hat{r}_i - \alpha - \beta(X_i - x)\}^2 W\left(\frac{X_i - x}{h_1}\right). \tag{2·3}$$

Although the above idea appears somewhat ad hoc, it has interesting implications. Specifically, while the bias for $\hat{m}$ itself is of order $O(h_2^2)$, its contribution to $\hat{\sigma}^2(.)$ is only of $o(h_2^2)$. This can be intuitively explained as follows. Observe that

$$\hat{r}_i - r_i = 2\{m(X_i) - \hat{m}(X_i)\}\sigma(X_i)\varepsilon_i + \{m(X_i) - \hat{m}(X_i)\}^2.$$

It is intuitively clear that the biases of the residuals are of order $O\{h_2^4 + (nh_2)^{-1}\}$, and this is the effect of the estimated regression function on the estimated variance; see Theorem 1 and Remark 1 below. This result also paves the way for adapting a fully data-driven bandwidth procedure in our estimation.

2·2. *Asymptotic normality*

THEOREM 1. *Suppose that Conditions* 1–5 *in Appendix* 1 *hold. Then*

$$(nh_1)^{\frac{1}{2}}\{\hat{\sigma}^2(x) - \sigma^2(x) - \theta_n\}$$

is asymptotically normal with mean 0 *and variance*

$$p^{-1}(x)\sigma^4(x)\lambda^2(x)\int W^2(t)\,dt,$$

where $p(.)$ denotes the marginal density function of X,

$$\lambda^2(x) = E\{(\varepsilon^2 - 1)^2 \mid X = x\}, \quad \varepsilon = \{Y - m(X)\}/\sigma(X), \quad \sigma_W^2 = \int t^2 W(t)\,dt,$$

$$\theta_n = \frac{h_1^2}{2}\,\sigma_W^2 \ddot{\sigma}^2(x) + o(h_1^2 + h_2^2). \tag{2·4}$$

The above adaptive result is obtained under the assumption that m is twice differentiable. This is not a minimal condition. The function $\sigma(.)$ can be estimated with optimal rates under weaker smoothness conditions on $m(.)$; see Hall & Carroll (1989) and Müller & Stadtmüller (1993). Note that the above asymptotic normality result complements the asymptotic approximations for conditional mean squared error obtained by Ruppert et al. (1997) for regression models with independent observations.

Remark 1. The bias and variance expressions given in Theorem 1 are exactly those which arise in the usual nonparametric regression analysis, considering the regression function to be $\sigma^2(x)$. In the bias of $\hat{\sigma}^2(x)$, the contribution from the error caused by estimating $m(x)$ is of smaller order than h_2^2, namely the order of the bias of $\hat{m}(x)$ itself. This permits us to use the optimal bandwidth to smooth $\hat{m}$; no undersmooth of $\hat{m}$ is needed. Our proof shows further that the second term on the right-hand side of (2·4) is $o(n^{-0\cdot 6})$ if we use bandwidths with optimal rates, that is $h_1 = O(n^{-1/5})$ and $h_2 = O(n^{-1/5})$.

2·3. *Efficiency*

It follows from local linear regression theory (Fan & Gijbels, 1996, § 6.2.2) that the benchmark estimator $\hat{\sigma}_b^2(.)$ derived from (2·1) is asymptotically normal. The leading terms in asymptotic bias and variance are exactly the same as those given in Theorem 1, provided h_2 used in estimator $\hat{\sigma}^2(.)$ converges to 0 no more slowly than h_1. This is a very minor requirement. It is well known that the best h_2 for estimating $m(.)$ should be of order $n^{-1/5}$. If we substitute such an h_2 in (2·4), the optimal h_1 which minimises the asymptotic mean squared error is also of the order $n^{-1/5}$. Therefore, the estimator $\hat{\sigma}^2(.)$ behaves asymptotically as well as $\hat{\sigma}_b^2(.)$ and hence is adaptive to the unknown regression function $m(.)$. Since the local linear estimator $\hat{\sigma}_b^2(.)$ is efficient in the sense of Fan (1993), then so is $\hat{\sigma}^2(.)$.

2·4. *Bandwidth selection*

The results given in § 2·2 permit us to take advantage of existing bandwidth selection methods for the local linear fit. Let $\hat{h}(X_1, \ldots, X_n; Y_1, \ldots, Y_n)$ be a data-driven bandwidth selection rule for local linear regression based on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$, such as the crossvalidation bandwidth rule, the pre-asymptotic substitution method of Fan & Gijbels (1995) or the plug-in approach of Ruppert, Sheather & Wand (1995). The last two methods have been demonstrated to be less variable and more effective. In all cases, $\hat{h}(X_1, \ldots, X_n; Y_1, \ldots, Y_n)$ is a consistent estimator of the asymptotic optimal bandwidth, which is $O(n^{-1/5})$. Our bandwidth selection rule is as follows.

ALGORITHM

1. *Use bandwidth* $h_2 = \hat{h}(X_1, \ldots, X_n; Y_1, \ldots, Y_n)$ *in local linear regression* (2·2) *to obtain the estimate* $\hat{m}(X_i)$ *for* $i = 1, \ldots, n$.
2. *Compute squared residuals* $\hat{r}_i = \{Y_i - \hat{m}(X_i)\}^2$ *for* $i = 1, \ldots, n$.
3. *Apply bandwidth* $h_1 = \hat{h}(X_1, \ldots, X_n; \hat{r}_1, \ldots, \hat{r}_n)$ *in local regression* (2·3) *to obtain* $\hat{\sigma}^2(.)$.

In the above algorithm, we keep the bandwidth selection method flexible. In our implementation, we use the pre-asymptotic substitution method by Fan & Gijbels (1995), since it has been demonstrated by L. S. Huang in a Ph.D. dissertation at University of North Carolina that the resulting estimator possesses fast relative rate of convergence.

3. Other estimators

3·1. *Direct estimators*

Härdle & Tsybakov (1997) proposed an improved version of the direct estimator $\hat{\sigma}_d^2(.)$, as given in (1·2), with local polynomial regression estimators $\hat{m}(.)$ and $\hat{v}(.)$ using the same kernel function and the same bandwidth, where $\hat{v}(x)$ is an estimator for $E(Y^2 | X = x)$. They also established the asymptotic normality of the estimator. If the local linear estimators are used with kernel $W(.)$ and bandwidth h_1, the leading terms in the asymptotic bias and the asymptotic variance of $\hat{\sigma}_d^2(x)$ are

$$\text{bias}\{\hat{\sigma}_d^2(x)\} : \frac{h_1^2}{2} \sigma_W^2 [\ddot{\sigma}^2(x) + 2\{\dot{m}(x)\}^2], \quad \text{var}\{\hat{\sigma}_d^2(x)\} : \frac{1}{nh_1} \sigma^4(x)\lambda^2(x)p^{-1}(x) \int W^2(t)\, dt.$$

If we compare this with Theorem 1, the direct estimator has the same asymptotic variance as the benchmark $\hat{\sigma}_b^2(.)$ and the residual-based estimator $\hat{\sigma}^2(.)$, but includes one more term in the bias. This extra term, $h_1^2 \sigma_W^2 \{\dot{m}(x)\}^2$, could have an adverse effect on the quality of estimation. For example, even when $m(.)$ is a linear function with a large slope, this direct method would have a large bias. Thus, the residual-based estimator $\hat{\sigma}^2(.)$ appears more appealing.

The existence of one more term in the bias of the direct estimator can be understood through the following heuristic argument. Note that

$$\hat{\sigma}_d^2(x) - \sigma^2(x) = \{\hat{v}(x) - v(x)\} - 2m(x)\{\hat{m}(x) - m(x)\} - \{\hat{m}(x) - m(x)\}^2. \tag{3·1}$$

The first term on the right-hand side has bias

$$\frac{h_1^2}{2} \sigma_W^2 \ddot{v}(x) = \frac{h_1^2}{2} \sigma_W^2 \ddot{\sigma}^2(x) + h_1^2 \sigma_W^2 \{\dot{m}(x)\}^2 + h_1^2 \sigma_W^2 m(x) \ddot{m}(x), \tag{3·2}$$

in which the last term on the right-hand side will cancel the bias of the second term on the right-hand side of (3·1). Note that the bias from the third term on the right-hand side of (3·1) is of order h_1^4. Therefore, the term involving $\{\dot{m}(x)\}^2$ remains. This argument also shows that using different kernels or bandwidths in the estimators $\hat{m}(.)$ and $\hat{v}(.)$ could further increase the bias of $\hat{\sigma}_d^2(.)$.

Why can the residual-based estimator $\hat{\sigma}^2(.)$ give smaller bias? To gain some insight, let us consider the local constant smoother, obtained by setting β equal to 0 in (2·3). Then the resulting estimator is

$$\sum_{i=1}^{n} \{Y_i - \hat{m}(X_i)\}^2 W\left(\frac{X_i - x}{h_1}\right) \Bigg/ \sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right).$$

See also Florens-Zmirou (1993). This estimator will reduce to the direct estimator $\hat{\sigma}_d^2(x)$ if all the $\hat{m}(X_i)$'s in the above expression are replaced by $\hat{m}(x)$. Clearly, $\{Y_i - \hat{m}(x)\}^2$ is more biased for $E\{Y - m(X)\}^2$ than is $\{Y_i - \hat{m}(X_i)\}^2$. This explains why the residual-based estimator inherits less bias from $\hat{m}(.)$ than does the direct estimator.

3·2. *Difference-based estimators*

For a fixed design model

$$Y_i = m(x_i) + \sigma(x_i)\varepsilon_i,$$

in which $x_1 \leqslant \ldots \leqslant x_n$ are fixed, $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^2) = 1$, Müller & Stadtmüller (1987) proposed to estimate $\sigma^2(.)$ through a difference sequence. Their approach can be briefly described as follows. Form an initial local variance estimate

$$\tilde{\sigma}^2(x_i) = \left(\sum_{j=-m}^{m} w_j Y_{i+j} \right)^2, \tag{3·3}$$

where $m > 0$ is a prescribed integer, and the difference sequence $\{w_j\}$ satisfies the conditions

$$\sum_{j=-m}^{m} w_j = 0, \quad \sum_{j=-m}^{m} w_j^2 = 1. \tag{3·4}$$

If we write $\sigma^2(x_i) = \tilde{\sigma}^2(x_i) + \tilde{\varepsilon}_i$, a kernel smoother is applied to obtain the final estimator for $\sigma^2(.)$ based on the above regression relationship.

Estimators of this type have a long history in the time series context; see, for example, Anderson (1971, p. 66). Application to nonparametric homoscedastic regression includes Rice (1984), Gasser et al. (1986) and Hall et al. (1990). It is shown by Hall et al. (1990) that, if the optimal difference sequence of $\{w_i\}$ is employed for a Gaussian model, the efficiency of the estimator is $4m/(4m+1)$.

In fact, the residual-based estimator $\hat{\sigma}^2(.)$ can be regarded as a generalised difference-based estimator, and $\hat{r}_i$ serves as a crude estimate of $\sigma^2(X_i)$. To make such a connection, we express the local linear estimator of $m(.)$ as

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i.$$

Then it can be shown that $w_1(x) + \ldots + w_n(x) = 1$. Write

$$\hat{r}_i = \{Y_i - \hat{m}(X_i)\}^2 = \left(\sum_{j=1}^{n} w_{i,j} Y_i \right)^2,$$

where $w_{i,i} = 1 - w_i(X_i)$ and $w_{i,j} = -w_j(X_i)$ for $i \neq j$. Obviously, $\{w_{i,j}\}$ is a difference sequence satisfying $w_{i,1} + \ldots + w_{i,n} = 0$. However, such a sequence of $\{w_{i,j}\}$ does not exactly fulfil the second condition in (3·4), but

$$\sum_j w_{i,j}^2 = 1 + O_p\{(nh_2)^{-1}\}.$$

The effective length of the sequence is $2m = 2nh_2$, which tends to infinity. This also explains why the estimator $\hat{\sigma}^2(.)$ is efficient in contrast to the aforementioned results of Hall et al. (1990).

Estimation of variance functions with more general weights was discussed by Müller & Stadtmüller (1993). Rates of convergence were thoroughly investigated. In particular, Müller & Stadtmüller (1993) find that it requires only very mild smoothness conditions on the regression function in order to obtain optimal rates for variance estimation.

3·3. *Maximum locally likelihood estimators*

If the distribution of ε is known, the locally maximum likelihood approach could be more efficient; see § 4.9 of Fan & Gijbels (1996) and the references therein. For example, if $\{\varepsilon_i\}$ are independent and normal, the loglikelihood function can be expressed as

$$-\frac{1}{2} \sum_{i=1}^{n} L\{\sigma^2(X_i), Y_i - m(X_i)\},$$

where $L(\alpha, y) = \alpha^{-1}y^2 + \log \alpha$. The local maximum likelihood approach with the local constant smoother leads to the direct estimator $\hat{\sigma}_d^2(.)$ with both $\hat{m}(.)$ and $\hat{v}(.)$ being the local constant estimators. The approach with the local linear smoother requires the estimation of four functions. To make it more tractable, we substitute $m(.)$ directly by its local linear estimator $\hat{m}(.)$, derived from (2·2). Let $\hat{\alpha}$ and $\hat{\beta}$ be the minimisers of the residual-based likelihood function:

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha,\beta} \sum_{i=1}^{n} L\{\alpha + \beta(X_i - x),\ Y_i - \hat{m}(X_i)\} W\left(\frac{X_i - x}{h_1}\right).$$

Then the maximum local likelihood estimator is defined by $\hat{\sigma}_{ml}^2(x) = \hat{\alpha}$. The estimator is also residual-based, and is adaptive to all unknown regression functions in a similar way to $\hat{\sigma}^2(.)$. The local maximum likelihood estimator shares the same leading terms as those for $\hat{\sigma}^2(x)$ given in § 2·1, but is computationally more involved. See also Genot-Catelot & Jacod (1993) for a related idea.

4. Applications and simulations

In this section, we first apply the adaptive estimator $\hat{\sigma}^2(.)$ derived from (2·3) to an interest-rate dataset. The finding from the application includes the validation of an existing structural model. Then, extensive simulations are carried out to confirm the theoretical claim that the adaptive estimator works almost as well as the ideal estimator $\hat{\sigma}_b^2(.)$ defined in (2·1). We use two simulated models, one with independent observations and one with nonlinear time series.

Throughout this section, the two dashed curves around a solid curve always indicate two standard deviations above and below the estimated curve. The conditional variance functions are always estimated by the adaptive estimator $\hat{\sigma}^2(.)$ derived from (2·3) unless specified otherwise. We always use the Epanechnikov kernel in our calculations, and all bandwidths are selected using the pre-asymptotic substitution method of Fan & Gijbels (1995).

Example 1. This example concerns the yields of the three-month Treasury Bill from the secondary market rates, on Fridays. The secondary market rates are annualised using a 360-day year of bank interest and are quoted on a discount basis. The data consist of 1735 weekly observations, from 5 January 1962 to 31 March 1995, and are presented in Fig. 1(a). The data were previously analysed by various authors, including Andersen & Lund (1997) and Gallant & Tauchen (1997).

Let z_t denote the time series presented in Fig. 1(a). We first fitted an autoregressive model with order selected by the Akaike information criterion. This yields the AR(5) model

$$z_t = 1{\cdot}0733 z_{t-1} - 0{\cdot}0423 z_{t-2} + 0{\cdot}0165 z_{t-3} + 0{\cdot}0228 z_{t-4} - 0{\cdot}0773 z_{t-5} + Y_t.$$

The selection of an AR(5) model coincides with that used by T. G. Andersen and J. Lund in a technical report. The 'residuals' Y_t are plotted against $X_t \equiv z_{t-1}$ in Fig. 1(b). Figure 1(c) depicts the estimator of the mean regression function $m(x) \equiv E\{Y_t | z_{t-1} = x\}$. The nonlinearity with a slightly increasing trend, up to $z_{t-1} = 14$, can be noted. The bandwidth selected by our software is 1·9535. The residual-based estimate for the conditional variance of Y_t given $z_{t-1} = x$ is denoted by $\hat{\sigma}^2(x)$, with automatically selected bandwidth of 3·1461. The estimated volatility function $\hat{\sigma}(x)$ is presented in Fig. 1(d). The overall fitted model is

$$z_t = \hat{m}(z_{t-1}) + 1{\cdot}0733 z_{t-1} - 0{\cdot}0423 z_{t-2} + 0{\cdot}0165 z_{t-3} + 0{\cdot}0228 z_{t-4} - 0{\cdot}0773 z_{t-5} + \hat{\sigma}(z_{t-1})\varepsilon_t,$$

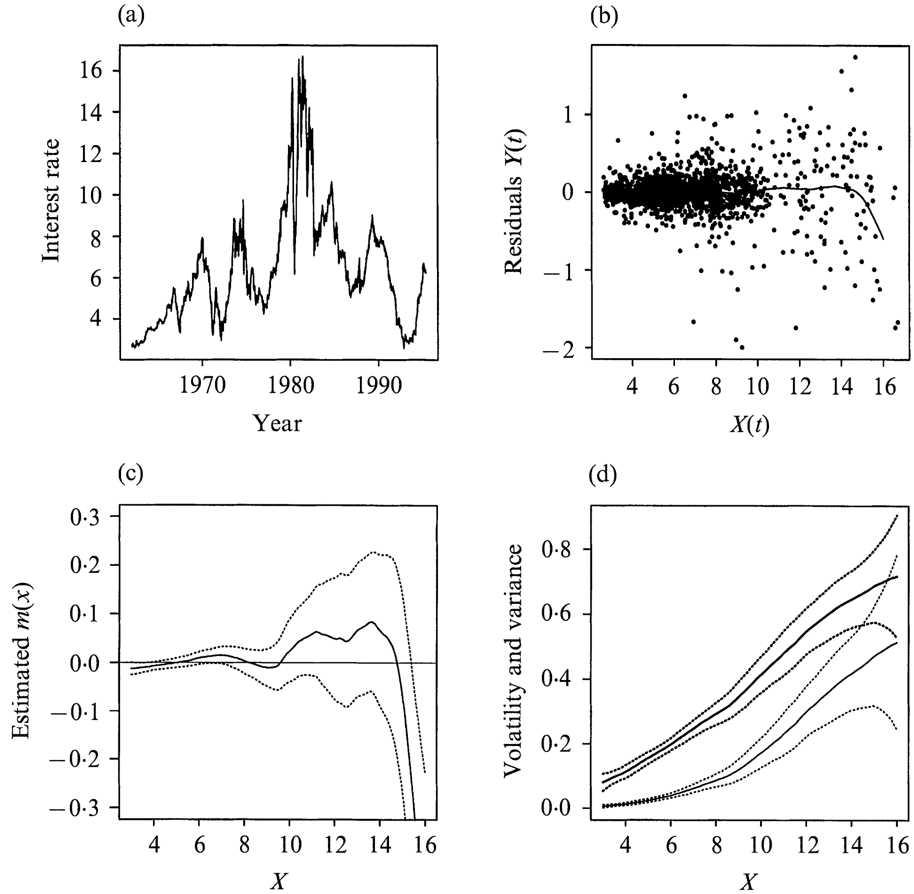

Fig. 1. Three-month Treasure Bill data, Example 1: (a) raw data z_t; (b) residuals after an AR(5) fit is plotted against $X_t \equiv z_{t-1}$, solid curve being the regression curve; (c) the regression curve for the data in (b); (d) the estimated volatility curve, thick curve, and the conditional variance function, thin curve.

in which $E(\varepsilon_t | z_{t-1}) = 0$ and $\text{var}(\varepsilon_t | z_{t-1}) = 1$. Note that the correlation coefficient between the logarithm of z_{t-1} and the logarithm of $\hat{\sigma}(z_{t-1})$ is 0·999. This lends strong support to the structural volatility model

$$\sigma(z_{t-1}) = \alpha z_{t-1}^{\beta},$$

which was considered by T. G. Andersen and J. Lund in their technical report. Applying least-squares fit to the log-transformed data, we found that $\alpha = 0{\cdot}0169$ and $\beta = 1{\cdot}380$.

A commonly-used model for asset pricing admits the following form: the value S_t of an underlying asset at time t satisfies

$$dS_t = \mu(S_t)\, dt + \sigma(S_t)\, dW_t, \tag{4·1}$$

where μ is the instantaneous expected rate of return, σ is the price volatility, and W_t is the standard Wiener process. This nonparametric model was recently used to model term structure dynamics by, for example, Aït-Sahalia (1996) and Stanton (1998). It includes the famous interest rate models of Cox, Ingersoll & Ross (1985) and Chan et al. (1992), among others. We now briefly relate this continuous model to our nonparametric regression model. Suppose that the data are sampled at time $i\Delta$ for $i=1, \ldots, T-1$. Set $Y_i = (S_{(i+1)\Delta} - S_{i\Delta})/\Delta$ and $X_i = S_{i\Delta}$. Model (4·1) can be written as

$$Y_i \simeq \mu(X_i) + \sigma(X_i)\varepsilon_i/\surd\Delta, \quad (4{\cdot}2)$$

where $\{\varepsilon_i\}$ are Gaussian white noise. Therefore, our method can be directly used to estimate functions $\mu(.)$ and $\sigma(.)$. For the short interest rate dataset, our Y_t is basically the same as the difference $z_t - z_{t-1}$. Therefore, functions in Figs 1(c) and (d) are respectively a scaled version of the estimated expected rate of return and price volatility in the continuous model (4·1). In fact, similar estimates to Figs 1(c) and (d) were independently obtained by Stanton (1988). Our method differs from that of Stanton (1988) in the following three important respects: squared residuals are used instead of the squared responses Y_i^2 to estimate the volatility, the local linear approach is used instead of the kernel method for nonparametric regression, more sophisticated bandwidth selection techniques are implemented. The first two modifications can reduce considerably the biases in the estimates and the last feature enables us to impose the correct amount of smoothing.

Example 2. We simulated 400 random samples of size $n=200$ from the model

$$Y_i = a\{X_i + 2\exp(-16X_i^2)\} + \sigma(X_i)\varepsilon_i,$$

with $\sigma(x) = 0{\cdot}4\exp(-2x^2) + 0{\cdot}2$, where $\{X_i\}$ and $\{\varepsilon_i\}$ are two independent sequences of independent random variables, $X_i \sim \text{Un}[-2, 2]$ and $\varepsilon_i \sim N(0, 1)$. Four different values of a, namely $a = 0{\cdot}5, 1, 2, 4$, are used in the simulation. For each simulated sample, the performance of the estimator is evaluated by the mean absolute deviation error,

$$\mathscr{E}_{\text{MAD}} = n_{\text{grid}}^{-1} \sum_{i=1}^{n_{\text{grid}}} |\hat{\sigma}(x_j) - \sigma(x_j)|,$$

where $\{x_j, j=1, \ldots, n_{\text{grid}}\}$ are grid points on $[-1{\cdot}8, 1{\cdot}8]$ with $n_{\text{grid}} = 101$. The results are summarised in Fig. 2. Figure 2(a) compares the adaptive variance estimator with the ideal variance estimator $\hat{\sigma}_b^2(.)$ which does not vary with different values of a. Presented there are the boxplots of $\mathscr{E}_{\text{MAD}}$ based on 400 simulations. The first four boxplots are the $\mathscr{E}_{\text{MAD}}$ of the adaptive estimator $\hat{\sigma}^2(.)$ for $a = 0{\cdot}5, 1, 2, 4$, in that order, and the last one is that of the ideal estimator $\hat{\sigma}_b^2(.)$. As anticipated, the adaptive estimator performs almost as well as the ideal one.

To gain further insight, we consider the specific case $a=1$. The scenario is similar for other cases. Figure 2(b) plots the $\mathscr{E}_{\text{MAD}}$ based on the adaptive estimator versus the $\mathscr{E}_{\text{MAD}}$ based on the ideal estimator, using the same sample data. Clearly, there is about equal chance that each estimator beats the other. The marginal densities of $\mathscr{E}_{\text{MAD}}$ of the adaptive estimator, thick curve, and of the ideal estimator, thin curve, are also depicted in Fig. 2(b). This shows again that the performance of the two estimators is comparable. Figure 2(c) presents a typical simulated sample with its corresponding estimated regression function. The typical sample was selected in such a way that the corresponding $\mathscr{E}_{\text{MAD}}$ is equal to its median among the 400 simulations. The sample residuals and the estimated conditional standard deviations are plotted in Fig. 2(d). The bandwidths are automatically selected by the procedure outlined in § 2·4 and are 0·1867 for the mean regression and 0·4841 for the conditional variance function, respectively.

Example 3. Consider the nonlinear time series model

$$X_{t+1} = 0{\cdot}235X_t(16 - X_t) + e_t,$$

where $e_1, e_2, \ldots$ are independent with the common distribution $N(0, 0{\cdot}3^2)$. The skeleton

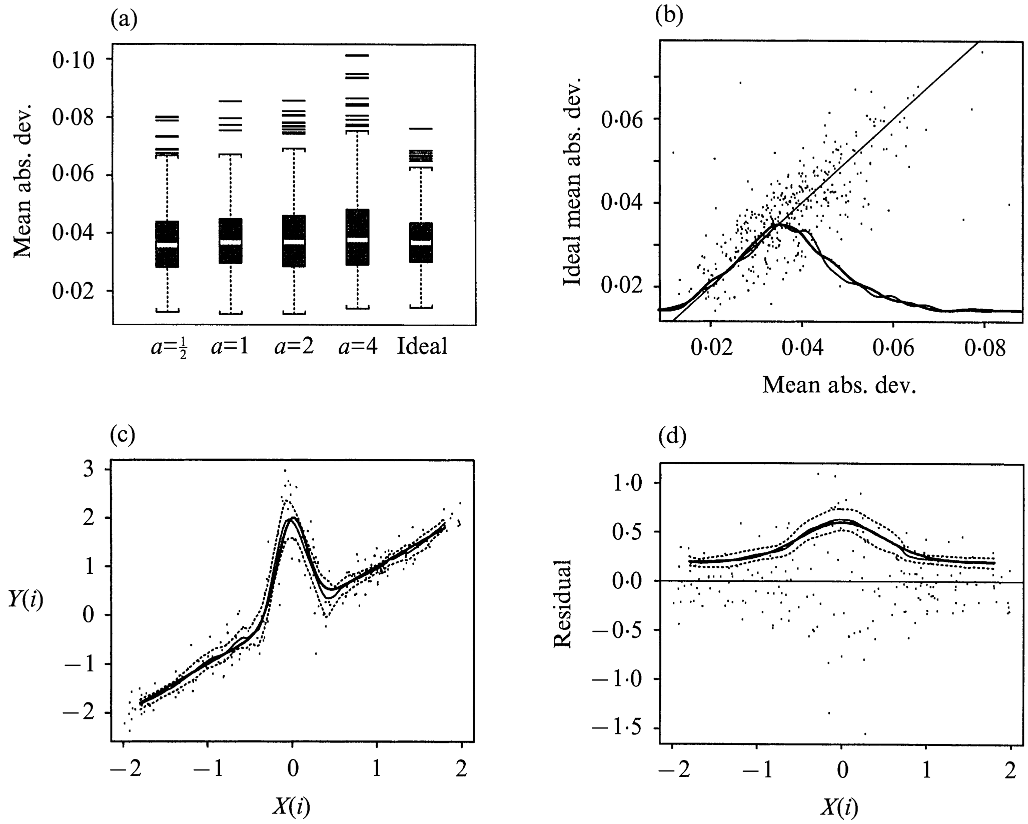

Fig. 2. Simulation results for Example 2. (a) Boxplots of the mean absolute deviation curve $\mathscr{E}_{\text{MAD}}$ based on 400 simulations for the adaptive estimator with $a = 0{\cdot}5, 1, 2, 4$ and for the ideal estimator, from left to right. (b) Scatter plot of the $\mathscr{E}_{\text{MAD}}$ of $\hat{\sigma}^2(.)$ versus the $\mathscr{E}_{\text{MAD}}$ of $\hat{\sigma}_b^2(.)$; the straight line marks the position where the two $\mathscr{E}_{\text{MAD}}$'s are equal. The thick curve is the estimated density function of the $\mathscr{E}_{\text{MAD}}$ of $\hat{\sigma}^2(.)$; the thin curve is the estimated density function of the $\mathscr{E}_{\text{MAD}}$ of $\hat{\sigma}_b^2(.)$. (c) A representative sample, with the corresponding estimated regression curve, thin curve, and the true regression curve, thick curve. (d) The sample residuals from (c), the estimated volatility, thin curve, and the true volatility, thick curve.

of this model exhibits chaos and has been used by Yao & Tong (1994) to illustrate the influence of the initial values on nonlinear prediction.

For this nonlinear time series, we consider two-step-ahead and three-step-ahead prediction by taking respectively $Y_t = X_{t+2}$ and $Y_t = X_{t+3}$. Note that the conditional variance functions concerned are not constant. On the other hand, the conditional variance of the one-step-ahead prediction is constant, and is therefore not presented here.

Figure 3(a) compares the ideal estimator with the adaptive estimator based on 400 simulations with $n = 500$. As we can see, the adaptive estimator works almost as well as the ideal estimator. Figure 3(b) gives the scatter plot of $\mathscr{E}_{\text{MAD}}$ for the adaptive estimator and the ideal estimator for three-step-ahead prediction. A typical simulated dataset and the corresponding estimated curves are presented in Figs 3(b) and (c). The criterion used to choose a typical sample is again the one for which the $\mathscr{E}_{\text{MAD}}$ is equal to its median among the 400 simulations. Figure 3(c) presents the estimated regression function for three-step-ahead prediction, where bandwidth 0·5577 was selected by our procedure. The estimated volatility function is presented in Fig. 3(d) with data-driven bandwidth 0·8165.

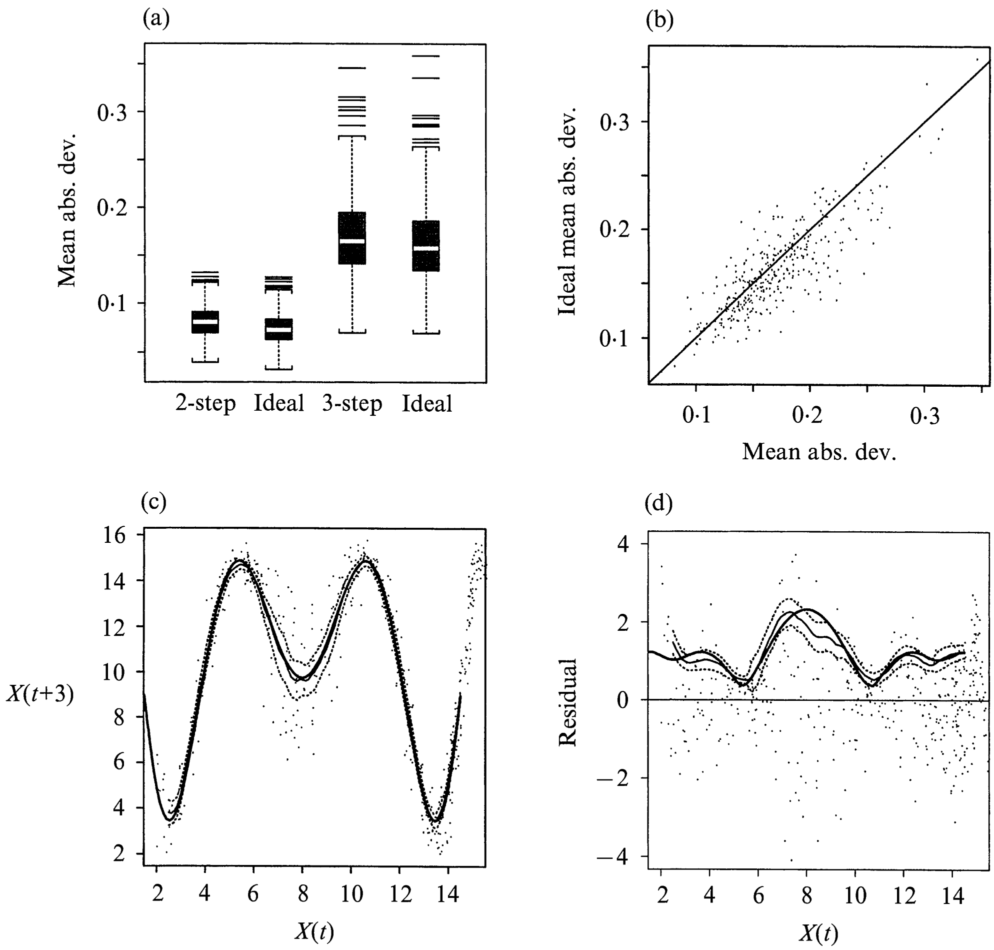

Fig. 3. Simulation results for Example 3. (a) Boxplots of the mean absolute deviation error $\mathscr{E}_{\text{MAD}}$ for the adaptive estimator and for the ideal estimator. (b) Scatter plot of the $\mathscr{E}_{\text{MAD}}$ versus the ideal $\mathscr{E}_{\text{MAD}}$ for three-step-ahead prediction. The straight line marks the position where the two $\mathscr{E}_{\text{MAD}}$'s are equal. (c) A representative sample and its estimated three-step-ahead regression curve. (d) The sample residuals of (c), the true volatility function, thick curve, and estimated volatility function, thin curve.

Similar results to Figs 3(b)–(d) were obtained for two-step-ahead prediction and are omitted for brevity.

Acknowledgement

We thank Professor Ron Gallant for kindly providing us the data analysed in Example 1. We are also grateful to the reviewers and the Editor for their insightful comments. J. Fan's research was partially supported by a National Science Foundation grant and a National Security Agency grant. Q. Yao's research was partially supported by a U.K. Engineering and Physical Sciences Research Council grant.

APPENDIX 1

Regularity conditions

We use the same notation as in § 2. We always use c to denote a generic constant which may be different at different places. We introduce the following regularity conditions.

Condition 1. For a given point x, $p(x) > 0$, $\sigma^2(x) > 0$ and the function $E\{Y^k | X = z\}$ is continuous at x for $k = 3, 4$. Further, $\ddot{m}(z) \equiv d^2m(z)/dz^2$ and $\ddot{\sigma}^2(z) \equiv d^2\{\sigma^2(z)\}/dz^2$ are uniformly continuous on an open set containing the point x.

Condition 2. We require $E\{Y^{4(1+\delta)}\} < \infty$, where $\delta \in [0, 1)$ is a constant.

Condition 3. The kernel functions W and K are symmetric density functions each with a bounded support in R. Further,

$$|W(x_1) - W(x_2)| \leqslant c|x_1 - x_2|, \quad |K(x_1) - K(x_2)| \leqslant c|x_1 - x_2|$$

and also $|p(x_1) - p(x_2)| \leqslant c|x_1 - x_2|$ for $x_1, x_2 \in R$.

Condition 4. The strictly stationary process $\{(X_i, Y_i)\}$ is absolutely regular, i.e.

$$\beta(j) \equiv \sup_{i \geqslant 1} E\left\{ \sup_{A \in \mathscr{F}_{i+j}^{\infty}} |\text{pr}(A | \mathscr{F}_1^i) - \text{pr}(A)| \right\} \to 0,$$

as $j \to \infty$, where $\mathscr{F}_i^j$ is the σ-field generated by $\{(X_k, Y_k) : k = i, \ldots, j\}$ $(j \geqslant i)$. Further, for the same δ as in Condition 2,

$$\sum_{j=1}^{\infty} j^2 \beta^{\delta/(1+\delta)}(j) < \infty.$$

We use the convention $0^0 = 0$.

Condition 5. As $n \to \infty$, $h_i \to 0$ and $\lim\inf_{n \to \infty} nh_i^4 > 0$ for $i = 1, 2$.

We impose boundedness on the supports of $K(.)$ and $W(.)$ for brevity of proofs; it may be removed at the cost of lengthier proofs. In particular, the Gaussian kernel is allowed. The assumption of the convergence rate of $\beta(j)$ is also for technical convenience. The assumption on the convergence rates of h_1 and h_2 is not the weakest possible.

Remark 2. When $\{(X_t, Y_t)\}$ are independent, Condition 4 holds with $\delta = 0$ and Condition 2 reduces to $E(Y^4) < \infty$. On the other hand, if Condition 4 holds with $\delta = 0$, there are at most finitely many nonzero $\beta(j)$'s. This means that there exists an integer $0 < j_0 < \infty$ for which (X_i, Y_i) is independent of $\{(X_j, Y_j), j \geqslant i + j_0\}$, for all $i \geqslant 1$.

APPENDIX 2

Proofs

In the sequel, $\hat{m}(.)$ denotes the local linear estimator derived from (2·2). We always assume that Conditions 1–5 hold. We say that $B_n(x) = B(x) + o_p(b_n)$, or $O_p(b_n)$, uniformly for $x \in G$ if $\sup_{x \in G} |B_n(x) - B(x)| = o_p(b_n)$, or $O_p(b_n)$. We only present the proof for the cases with $\delta > 0$. The case with $\delta = 0$ can be dealt with in a more direct and simpler way; see Remark 2 in Appendix 1.

The proof is based on the following lemma, which follows directly from Lemma 2 of a technical report by Q. Yao and H. Tong.

LEMMA 1. *Let $G \subset \{p(x) > 0\}$ be a compact subset. As $n \to \infty$, uniformly for $x \in G$,*

$$\hat{\sigma}^2(x) - \sigma^2(x) = \frac{1}{nh_1 p(x)} \sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right) \{\hat{r}_i - \sigma^2(x) - \dot{\sigma}^2(x)(X_i - x)\} + O_p\{R_{n,1}(x)\}, \quad (\text{A2·1})$$

$$\hat{m}(x) - m(x) = \frac{1}{nh_2 p(x)} \sum_{i=1}^{n} \sigma(X_i)\varepsilon_i K\left(\frac{X_i - x}{h_2}\right) + \frac{h_2^2 \sigma_K^2}{2} \ddot{m}(x) + O_p\{R_{n,2}(x)\}, \tag{A2·2}$$

where

$$R_{n,1}(x) = \frac{1}{np(x)} \left[\left| \sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right) \{\hat{r}_i - \sigma^2(x) - \dot{\sigma}^2(x)(X_i - x)\} \right| + \left| \sum_{i=1}^{n} \frac{X_i - x}{h_1} W\left(\frac{X_i - x}{h_1}\right) \{\hat{r}_i - \sigma^2(x) - \dot{\sigma}^2(x)(X_i - x)\} \right| \right],$$

$$R_{n,2}(x) = \frac{1}{np(x)} \left\{ \left| \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_2}\right) \sigma(X_i)\varepsilon_i \right| + \left| \sum_{i=1}^{n} \frac{X_i - x}{h_2} K\left(\frac{X_i - x}{h_2}\right) \sigma(X_i)\varepsilon_i \right| \right\} + O(h_2^3).$$

Proof of Theorem 1. Note that

$$\hat{r}_i = \{Y_i - \hat{m}(X_i)\}^2 = \{\sigma(X_i)\varepsilon_i + m(X_i) - \hat{m}(X_i)\}^2$$
$$= \sigma^2(X_i)\varepsilon_i^2 + 2\sigma(X_i)\varepsilon_i\{m(X_i) - \hat{m}(X_i)\} + \{m(X_i) - \hat{m}(X_i)\}^2.$$

It follows from (A2·1) that

$$\hat{\sigma}^2(x) - \sigma^2(x) = I_1 + I_2 - I_3 + I_4 + O_p(h_1)(|I_1 + I_2 - I_3 + I_4| + |I_1' + I_2' - I_3' + I_4'|),$$

where

$$\begin{aligned} I_1 &= \frac{1}{nh_1 p(x)} \sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right) \{\sigma^2(X_i) - \sigma^2(x) - \dot{\sigma}^2(x)(X_i - x)\}, \\ I_2 &= \frac{1}{nh_1 p(x)} \sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right) \sigma^2(X_i)(\varepsilon_i^2 - 1), \\ I_3 &= \frac{2}{nh_1 p(x)} \sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right) \sigma(X_i)\varepsilon_i\{\hat{m}(X_i) - m(X_i)\}, \\ I_4 &= \frac{1}{nh_1 p(x)} \sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right) \{\hat{m}(X_i) - m(X_i)\}^2, \end{aligned} \tag{A2·3}$$

and I_j' is defined in the same way as I_j with one more factor $h_1^{-1}(X_i - x)$ in the ith summand $(1 \leqslant j \leqslant 4)$. It is easy to see that the theorem follows directly from statements (a)–(d):

(a) $I_1 = \frac{1}{2} h_1^2 \ddot{\sigma}^2(x)\sigma_W^2 + o_p(h_1^2)$ and $I_1' = o_p(h_1^2)$;

(b) $(nh_1)^{\frac{1}{2}} I_2 \to N(0, \sigma^4(x)\lambda^2(x) \int W^2(t)\, dt/p(x))$, in distribution, and

$$(nh_1)^{\frac{1}{2}} I_2' \to N\left(0, \sigma^4(x)\lambda^2(x) \int t^2 W^2(t)\, dt/p(x)\right),$$

in distribution;

(c) $I_3 = o_p(h_1^2 + h_2^2)$ and $I_3' = o_p(h_1^2 + h_2^2)$;

(d) $I_4 = o_p(h_1^2 + h_2^2)$ and $I_4' = o_p(h_1^2 + h_2^2)$.

In the sequel, we establish the statements about I_j in (a)–(d) only. The cases with I_j' can be proved in the same manner.

It is easy to see that (a) follows from a Taylor expansion and a direct application of the ergodic theorem. Conditions 2 and 3 imply that

$$E\left\{W\left(\frac{X_i - x}{h_1}\right) \sigma^2(X_i)(\varepsilon_i^2 - 1)\right\}^{2+\delta/2} < \infty.$$

Note that the condition of absolutely regular implies α-mixing with $\alpha(j) \leqslant \beta(j)$. By Condition 4 and Theorem 1.7 of Peligrad (1986), I_2 is asymptotically normal with mean 0 and variance σ_*^2/nh_1, where

$$\sigma_*^2 = \frac{1}{h_1} E\left\{W\left(\frac{X-x}{h_1}\right)\frac{\sigma^2(X)}{p(X)}(\varepsilon^2-1)\right\}^2 + \frac{1}{h_1}\sum_{i=2}^{n} E\left\{W\left(\frac{X_1-x}{h_1}\right)\frac{\sigma^2(X_1)}{p(X_1)}(\varepsilon_1^2-1)W\left(\frac{X_i-x}{h_1}\right)\frac{\sigma^2(X_i)}{p(X_i)}(\varepsilon_i^2-1)\right\}. \quad (A2{\cdot}4)$$

It is easy to see that the first term in the above expression converges to

$$\sigma^4(x)\lambda^2(x)\int W^2(t)\,dt/p(x).$$

Note that, for any $i \geqslant 2$,

$$E\left\{W\left(\frac{X_1-x}{h_1}\right)\frac{\sigma^2(X_1)}{p(X_1)}(\varepsilon_1^2-1)W\left(\frac{X_i-x}{h_1}\right)\frac{\sigma^2(X_i)}{p(X_i)}(\varepsilon_i^2-1)\right\}^{1+\delta} = O(h_1^2),$$

$$E\left\{W\left(\frac{X-x}{h_1}\right)\frac{\sigma^2(X)}{p(X)}(\varepsilon^2-1)\right\} = 0, \quad E\left|W\left(\frac{X-x}{h_1}\right)\frac{\sigma^2(X)}{p(X)}(\varepsilon^2-1)\right|^{1+\delta} = O(h_1).$$

It follows from Condition 4 and Lemma 1 of Yoshihara (1976) that the absolute value of the second term on the right-hand side of (A2·4) is bounded above by

$$ch_1^{(1-\delta)/(1+\delta)}\{\beta^{\delta/(1+\delta)}(1) + \ldots + \beta^{\delta/(1+\delta)}(n-1)\} = o(1).$$

Hence (b) holds.

Note that $W(.)$ has a bounded support contained in the interval $[-s_w, s_w]$, say. Therefore, in the summation on the right-hand side of (A2·3), only those terms with $X_i \in [x - h_2 s_w, x + h_2 s_w]$ might not be 0. It follows from (A2·2) that we may write $I_3 = I_{31} + I_{32} + I_{33}$, where

$$I_{31} = \frac{1}{n^2 h_1 h_2 p(x)}\sum_{i,j=1}^{n} K\left(\frac{X_i - X_j}{h_2}\right)\sigma(X_i)\sigma(X_j)\varepsilon_i\varepsilon_j\left\{p^{-1}(X_i)W\left(\frac{X_i-x}{h_1}\right) + p^{-1}(X_j)W\left(\frac{X_j-x}{h_1}\right)\right\}$$

$$\equiv \frac{1}{n^2 h_1 h_2 p(x)}\sum_{i,j=1}^{n}\phi_{ij} = \frac{2}{n^2 h_1 h_2 p(x)}\sum_{1\leqslant i\leqslant j\leqslant n}\phi_{ij} + O_p\left(\frac{1}{nh_2}\right), \quad (A2{\cdot}5)$$

$$I_{32} = \frac{h_2^2\sigma_K^2}{nh_1 p(x)}\sum_{i=1}^{n} W\left(\frac{X_i-x}{h_1}\right)\sigma(X_i)\varepsilon_i\ddot{m}(X_i) = o_p(h_2^2), \quad (A2{\cdot}6)$$

$$|I_{33}| \leqslant \frac{O_p(1)}{n^2 h_1}\left|\sum_{i,j=1}^{n} W\left(\frac{X_i-x}{h_1}\right)K\left(\frac{X_i-X_j}{h_2}\right)\sigma(X_i)\sigma(X_j)|\varepsilon_i|\varepsilon_j/p(X_i)\right| + \frac{O_p(1)}{n^2 h_1}\left|\sum_{i,j=1}^{n}\frac{X_j - X_i}{h_2} W\left(\frac{X_i-x}{h_1}\right)K\left(\frac{X_i-X_j}{h_2}\right)\sigma(X_i)\sigma(X_j)|\varepsilon_i|\varepsilon_j/p(X_i)\right| + o_p(h_2^2).$$

It follows from Lemma A(ii) of a technical report of V. Hjellvik, Q. Yao and D. Tjøstheim that, for any $\varepsilon_0 > 0$ and $\varepsilon > 0$,

$$\text{pr}\left\{n^{-1}(h_1h_2)^{-\frac{1}{2}\{1/(1+\delta)-\varepsilon_0\}}\left|\sum_{i<j}\phi_{ij}\right| > \varepsilon\right\} \leqslant \frac{c(h_1h_2)^{\varepsilon_0}}{n^2} E\left\{(h_1h_2)^{-1/\{2(1+\delta)\}}\sum_{i<j}\phi_{ij}\right\}^2 = o\{(h_1h_2)^{\varepsilon_0}\}.$$

Therefore, the first term on the right-hand side of (A2·5) is $o_p\{n^{-1}(h_1h_2)^{-\frac{1}{2}\{(1+2\delta)/(1+\delta)+\varepsilon_0\}}\}$. Thus

$$I_{31} = o_p(n^{-1}(h_1h_2)^{-\frac{1}{2}\{(1+2\delta)/(1+\delta)+\varepsilon_0\}}) + O_p(n^{-1}h_2^{-1}).$$

Condition 5 implies that the terms on the right-hand side of the above expression are of order $o_p(h_1^2 + h_2^2)$ if we choose $\varepsilon_0 < (1+\delta)^{-1}$. Performing Hoeffding's projection decomposition of U-statistics, we can prove that $I_{33} = o_p(h_1^2 + h_2^2)$ in the same way.

The proof of (d) is similar to that of (c), and therefore is omitted here.

References

Aït-Sahalia, Y. (1996). Nonparametric pricing of interest rate derivative securities. *Econometrica* **64**, 527–60.

Andersen, T. G. & Lund, J. (1997). Estimating continuous time stochastic volatility models of the short term interest rate. *J. Economet.* **77**, 343–77.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. New York: Wiley.

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

Carroll, R. J. & Ruppert, D. (1988). *Transformations and Weighting in Regression*. London: Chapman & Hall.

Chan, K. C., Karolyi, A. G., Longstaff, F. A. & Sanders, A. B. (1992). An empirical comparison of alternative models of the short-term interest rate. *J. Finance* **47**, 1209–27.

Cox, J. C., Ingersoll, J. E. & Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica* **53**, 385–467.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of variance of U.K. inflation. *Econometrica* **50**, 987–1008.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Ann. Statist.* **21**, 196–216.

Fan, J. & Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20**, 2008–36.

Fan, J. & Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. R. Statist. Soc.* B **57**, 371–94.

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

Florens-Zmirou, D. (1993). On estimating the diffusion coefficient from discrete observations. *J. Appl. Prob.* **30**, 790–804.

Gallant, A. R. & Tauchen, G. (1997). Estimation of continuous time models for stock returns and interest rates. *Macroecon. Dyn.* **1**, 135–68.

Gasser, T., Sroka, L. & Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–33.

Genon-Catalot, V. & Jacod, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. Inst. H. Poincaré Prob. Statist.* **29**, 119–51.

Hall, P. & Carroll, R. J. (1989). Variance function estimation in regression: the effect of estimation of the mean. *J. R. Statist. Soc.* B **51**, 3–14.

Hall, P., Kay, J. W. & Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 521–8.

Härdle, W. & Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Economet.* **81**, 233–42.

Hastie, T. J. & Loader, C. (1993). Local regression: automatic kernel carpentry (with Discussion). *Statist. Sci.* **8**, 120–43.

Müller, H. C. & Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 610–25.

Müller, H. G. & Stadtmüller, U. (1993). On variance function estimation with quadratic forms. *J. Statist. Plan. Inf.* **35**, 213–31.

Neumann, M. H. (1994). Fully data-driven nonparametric variance estimators. *Statistics* **25**, 189–212.

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

Rice, J. (1984). Bandwidth choice for nonparametric kernel regression. *Ann. Statist.* **12**, 1215–30.

Ruppert, D., Sheather, S. J. & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Am. Statist. Assoc.* **90**, 1257–70.

Ruppert, D. & Wand, M. P. (1994). Multivariate weighted least squares regression. *Ann. Statist.* **22**, 1346–70.

Ruppert, D., Wand, M. P., Holst, U. & Hössjer, O. (1997). Local polynomial variance function estimation. *Technometrics* **39**, 262–73.

Stanton, R. (1998). A nonparametric model of term structure dynamics and the market price of interest rate risk. *J. Finance*. To appear.

YAO, Q. & TONG, H. (1994). Quantifying the influence of initial values on nonlinear prediction. *J. R. Statist. Soc.* B **56**, 701–25.
YOSHIHARA, K. (1976). Limiting behaviour of U-statistics for stationary absolutely regular processes. *Z. Wahr. verw. Geb.* **35**, 237–52.

[*Received October* 1996. *Revised November* 1997]