Conditional Variance Estimation in Heteroscedastic Regression Models

Lu-Hung Chen[1], **Ming-Yen Cheng**[2] **and Liang Peng**[3]

Abstract

First, we propose a new method for estimating the conditional variance in heteroscedasticity regression models. For heavy tailed innovations, this method is in general more efficient than either of the local linear and local likelihood estimators. Secondly, we apply a variance reduction technique to improve the inference for the conditional variance. The proposed methods are investigated through their asymptotic distributions and numerical performances.

Keywords. Conditional variance, local likelihood estimation, local linear estimation, log-transformation, variance reduction, volatility.

Short title. Conditional Variance.

[1] Department of Mathematics, National Taiwan University, Taipei 106, Taiwan. Email: r91090@csie.ntu.edu.tw

[2] Department of Mathematics, National Taiwan University, Taipei 106, Taiwan. Email: cheng@math.ntu.edu.tw

[3] School of Mathematics, Georgia Institute of Technology, Atlanta GA 30332-0160, USA. Email: peng@math.gatech.edu

1 Introduction

Let $\{(Y_i, X_i)\}$ be a two-dimensional strictly stationary process having the same marginal distribution as (Y, X). Let $m(x) = E(Y|X = x)$ and $\sigma^2(x) = Var(Y|X = x)$ be respectively the regression function and the conditional variance, and $\sigma^2(x) > 0$. Write

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

Thus $E(\varepsilon_i|X_i) = 0$ and $Var(\varepsilon_i|X_i) = 1$. When $X_i = Y_{i-1}$, model (1.1) includes ARCH(1) time series (Engle, 1982) and AR (1) processes with ARCH(1) errors as special cases. See Borkovec (2001), Borkovec and Klüppelberg (2001) for probabilistic properties and Ling (2004) and Chan and Peng (2005) for statistical properties of such AR(1) processes with ARCH(1) errors.

There exists an extensive study on estimating the nonparametric regression function $m(x)$; see for example Fan and Gijbels (1996). Here we are interested in estimating the volatility function $\sigma^2(x)$. Some methods have been proposed in the literature; see Fan and Yao (1998) and references cited therein. More specifically, Fan and Yao (1998) proposed to first estimate $m(x)$ by the local linear technique, i.e., $\widehat{m}(x) = \widehat{a}$ if

$$(\widehat{a}, \widehat{b}) = argmin_{(a,b)} \sum_{i=1}^{n} \big\{Y_i - a - b(X_i - x)\big\}^2 K\Big(\frac{X_i - x}{h_2}\Big), \tag{1.2}$$

where K is a density function and $h_2 > 0$ is a bandwidth, and then estimate $\sigma^2(x)$ by $\widehat{\sigma}_1^2(x) = \widehat{\alpha}_1$ if

$$(\widehat{\alpha}_1, \widehat{\beta}_1) = argmin_{(\alpha_1,\beta_1)} \sum_{i=1}^{n} \big\{\widehat{r}_i - \alpha_1 - \beta_1(X_i - x)\big\}^2 W\Big(\frac{X_i - x}{h_1}\Big), \tag{1.3}$$

where $\widehat{r}_i = \{Y_i - \widehat{m}(X_i)\}^2$, W is a density function and $h_1 > 0$ is a bandwidth. The main drawback of this conditional variance estimator is that it is not always positive. Recently, Yu and Jones (2004) studied the local Normal likelihood estimation, mentioned in Fan and Yao (1998), with expansion for $\log \sigma^2(x)$ instead of $\sigma^2(x)$. The main advantage of doing this is to ensure that the resulting conditional variance estimator is always positive. More specifically, the estimator proposed by Yu and Jones (2004) is $\widehat{\sigma}_2^2(x) = \exp\{\widehat{\alpha}_2\}$, where

$$(\widehat{\alpha}_2, \widehat{\beta}_2) = argmin_{(\alpha_2,\beta_2)} \sum_{i=1}^{n} \Big[\widehat{r}_i \exp\big\{ -\alpha_2 - \beta_2(X_i - x) \big\} + \alpha_2 + \beta_2(X_i - x) \Big] W\Big(\frac{X_i - x}{h_1} \Big). \tag{1.4}$$

Yu and Jones (2004) derived the asymptotic limit of $\widehat{\sigma}_2^2(x)$ under very strict conditions which require independence of $(X_i, Y_i)'s$ and $\varepsilon_i \sim N(0,1)$.

Motivated by empirical evidences that financial data may have heavy tails, see for example Mittnik and Rachev (2000), we propose the following new estimator for $\sigma^2(x)$. Rewrite (1.1) as

$$\log r_i = \nu(X_i) + \log(\varepsilon_i^2/d) \tag{1.5}$$

where $r_i = \{Y_i - m(X_i)\}^2$, $\nu(x) = \log(d\sigma^2(x))$ and d satisfies $E\{\log(\varepsilon_i^2/d)\} = 0$. Based on the above equation, first we estimate $\nu(x)$ by $\widehat{\nu}(x) = \widehat{\alpha}_3$ where

$$(\widehat{\alpha}_3, \widehat{\beta}_3) = argmin_{(\alpha_3,\beta_3)} \sum_{i=1}^{n} \big\{ \log(\widehat{r}_i + n^{-1}) - \alpha_3 - \beta_3(X_i - x) \big\}^2 W\Big(\frac{X_i - x}{h_1} \Big). \tag{1.6}$$

Note that we employ $\log(\widehat{r}_i + n^{-1})$ instead of $\log \widehat{r}_i$ to avoid $\log(0)$. Next, by noting that $E(\varepsilon_i^2|X_i) = 1$ and $r_i = \exp\{\nu(X_i)\}\,\varepsilon_i^2/d$, we estimate d by

$$\widehat{d} = \Big[\frac{1}{n} \sum_{i=1}^{n} \widehat{r}_i \exp\{-\widehat{\nu}(X_i)\} \Big]^{-1}.$$

Therefore our new estimator for $\sigma^2(x)$ is defined as

$$\widehat{\sigma}_3^2(x) = \exp\{\widehat{\nu}(x)\}/\widehat{d}.$$

Intuitively, the log-transformation in (1.5) makes data less skewed, and thus the new estimator may be more efficient in dealing with heavy tailed errors than other estimators such as $\widehat{\sigma}_1^2(x)$ and $\widehat{\sigma}_2^2(x)$. Peng and Yao (2003) investigated this effect in least absolute estimation of the parameters in ARCH and GARCH models. Note that this new estimator $\widehat{\sigma}_3^2(x)$ is always positive.

We organize this paper as follows. In Section 2, the asymptotic distribution of this new estimator is given and some theoretical comparisons with $\widehat{\sigma}_1^2(x)$ and $\widehat{\sigma}_2^2(x)$ are addressed as well. In Section 3, we apply the variance reduction technique in Cheng, et al. (2007) to the conditional variance estimators $\widehat{\sigma}_1^2(x)$ and $\widehat{\sigma}_3^2(x)$ and provide the limiting distributions. Bandwidth selection for the proposed estimators is discussed in Section 4. A simulation study and a real application are presented in Section 5. All proofs are given in Section 6.

2 Asymptotic Normality

To derive the asymptotic limit of our new estimator, we impose the following regularity conditions. Denote by $p(\cdot)$ the marginal density function of X.

(C1) For a given point x, the functions $E(Y^3|X = z)$, $E(Y^4|X = z)$ and $p(z)$ are continuous at the point x, and $\ddot{m}(z) = \frac{d^2}{dz^2}m(z)$ and $\ddot{\sigma}^2(z) = \frac{d^2}{dz^2}\sigma^2(z)$ are uniformly continuous on an open set containing the point x. Further, assume $p(x) > 0$;

(C2) $E(Y^{4(1+\delta)}) < \infty$ for some $\delta \in (0, 1)$;

(C3) The kernel functions W and K are symmetric density functions each with a bounded support in $(-\infty, \infty)$. Further, there exists $M > 0$ such that $|W(x_1) - W(x_2)| \leq M|x_1 - x_2|$ for all x_1 and x_2 in the support of W and $|K(x_1) - K(x_2)| \leq M|x_1 - x_2|$ for all x_1 and x_2 in the support of K;

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

$$\beta(j) := \sup_{j\geq 1} E\big\{ \sup_{A\in\mathcal{F}_{i+j}^{\infty}} |P(A|\mathcal{F}_1^i) - P(A)|\big\} \to 0 \quad \text{as } j \to \infty,$$

where $\mathcal{F}_i^j$ is the σ-field generated by $\{(Y_k, X_k) : k = i, \cdots, j\}, j \geq i$. Further, for the same δ as in **(C2)**, $\sum_{j=1}^{\infty} j^2 \beta^{\delta/(1+\delta)}(j) < \infty$;

(C5) As $n \to \infty$, $h_i \to 0$ and $\liminf_{n\to\infty} nh_i^4 > 0$ for $i = 1, 2$.

Our main result is as follows.

Theorem 1. Under the regularity conditions (C1)–(C5), we have

$$\sqrt{nh_1}\left\{\widehat{\sigma}_3^2(x)-\sigma^2(x)-\theta_{n3}\right\}\xrightarrow{d}N\Big(0,p(x)^{-1}\sigma^4(x)\lambda^2(x)R(W)\Big),$$

where $\lambda^2(x)=E\{(\log(\varepsilon^2/d))^2|X=x)$, $R(W)=\int W^2(t)\,dt$ and

$$\begin{aligned}\theta_{n3} &= \frac{1}{2}h_1^2\sigma^2(x)\ddot{\nu}(x)\int t^2W(t)\,dt\\ &\quad -\frac{1}{2}h_1^2\sigma^2(x)E\ddot{\nu}(X_1)\int t^2W(t)\,dt+o\big(h_1^2+h_2^2\big).\end{aligned}$$

Next we compare $\widehat{\sigma}_3^2(x)$ with $\widehat{\sigma}_1^2(x)$ and $\widehat{\sigma}_2^2(x)$ in terms of their asymptotic biases and variances. It follows from Fan and Yao (1998) and Yu and Jones (2004) that, for $i=1,2$,

$$\sqrt{nh_1}\big\{\widehat{\sigma}_i^2(x)-\sigma^2(x)-\theta_{ni}\big\}\xrightarrow{d}N\Big(0,p^{-1}(x)\sigma^4(x)\bar{\lambda}^2(x)R(W)\Big),$$

where $\bar{\lambda}^2(x)=E\{(\varepsilon^2-1)^2|X=x\}$,

$$\theta_{n1}=\frac{1}{2}h_1^2\ddot{\sigma}^2(x)\int t^2W(t)\,dt+o(h_1^2+h_2^2)$$

and $\theta_{n2}=\frac{1}{2}h_1^2\sigma^2(x)\frac{d^2}{dx^2}\{\log\sigma^2(x)\}\int t^2W(t)\,dt+o\big(h_1^2+h_2^2\big)$.

Remark 1. If $\hat{\nu}(X_i)$ in $\hat{d}$ is replaced by another local linear estimate with either a smaller order of bandwidth than h_1 or a higher order of kernel than W, then the asymptotic squared bias of $\widehat{\sigma}_3^2(x)$ is the same as that of $\widehat{\sigma}_2^2(x)$, which may be larger or smaller than the asymptotic squared bias of $\widehat{\sigma}_1^2(x)$; see Yu and Jones (2004) for detailed discussions.

Remark 2. Suppose that given $X=x$, ε has a t-distribution with degrees of freedom m. Then the ratios of $\lambda^2(x)$ to $\bar{\lambda}^2(x)$ are 0.269, 0.497, 0.674, 0.848, 1.001 for $m=5,6,7,8,9$, respectively. That is, $\widehat{\sigma}_3^2(x)$ may have a smaller variance than both $\widehat{\sigma}_1^2(x)$ and $\widehat{\sigma}_2^2(x)$ when ε has a heavy tailed distribution.

Remark 3. Note that the regularity conditions (C1)–(C5) were employed by Fan and Yao (1998) as well. However, it follows from the proof in Section 6 that we could replace condition (C2) by $E\{|\log(\varepsilon^2)|^{2+\delta}\} < \infty$ for some $\delta > 0$ in deriving the asymptotic normality of $\widehat{\nu}(x)$. Condition (C2) is only employed to ensure that $\widehat{d} - d = O_p(n^{-1/2})$. As a matter of fact we only need $\widehat{d} - d = o_p\big\{(nh_1)^{-1/2}\big\}$ to derive the asymptotic normality of $\widehat{\sigma}_3^2(x)$. In other words, the asymptotic normality of $\widehat{\sigma}_3^2(x)$ may still hold even when $E(Y^4) = \infty$. This is different from other conditional variance estimators such as $\widehat{\sigma}_1^2(x)$ and $\widehat{\sigma}_2^2(x)$, which require at least $E(Y^4) < \infty$ to ensure asymptotic normality.

3 Variance Reduced Estimation

Here we apply the variance reduction techniques proposed by Cheng, et al. (2007), which concerns nonparametric estimation of $m(x)$, to the conditional variance estimators $\widehat{\sigma}_1^2(x)$ and $\widehat{\sigma}_3^2(x)$. The reason why we do not consider $\widehat{\sigma}_2^2(x)$ here is that the asymptotic normality was derived by Yu and Jones (2004) under much more stringent conditions than those required by the other two estimators. The idea of our variance reduction strategy is to construct a linear combination of either $\widehat{\sigma}_1^2(z)$ or $\widehat{\sigma}_3^2(z)$ at three points around x such that the asymptotic bias is unchanged. The details are given below.

For any given point x, let $\big\{\beta_{x,0}, \beta_{x,1}, \beta_{x,2}\big\}$ be a grid of equally spaced points, with bin width $\gamma h_1 = \beta_{x,1} - \beta_{x,0}$, such that $x = \beta_{x,1} + l\gamma h_1$ for some $l \in [-1, 1]$. Then, like Cheng et al. (2007), our variance reduction estimators for $\sigma^2(x)$ are defined as

$$\widetilde{\sigma}_j^2(x) = \frac{l(l-1)}{2}\,\widehat{\sigma}_j^2(\beta_{x,0}) + (1-l^2)\,\widehat{\sigma}_j^2(\beta_{x,1}) + \frac{l(l+1)}{2}\,\widehat{\sigma}_j^2(\beta_{x,2}), \tag{3.1}$$

for $j = 1$ and 3. Suppose that $Supp(\sigma)$ is bounded, $Supp(\sigma) = [0, 1]$ say, since $\beta_{x,0} < x < \beta_{x,2}$, $\beta_{x,0}$ and $\beta_{x,2}$ would be outside $Supp(\sigma)$ if x is close to the endpoints. Therefore we take $\gamma(x) = \min\big\{\gamma, x/(1+l)h_1, (1-x)/(1-l)h_1\big\}$ so that $\big\{\beta_{x,0}, \beta_{x,1}, \beta_{x,2}\big\} \in Supp(\sigma) = [0, 1]$ all the time.

The following theorem gives the asymptotic limits of the variance reduced estimators.

Theorem 2. Under the conditions of Theorem 1, for interior point x we have

$$\sqrt{nh_1}\big\{\widetilde{\sigma}_1^2(x)-\sigma^2(x)-\theta_{n1}\big\} \xrightarrow{d} N\Big(0,\big\{R(W)-l^2(1-l^2)\,C(\gamma)\big\}p(x)^{-1}\sigma^4(x)\bar{\lambda}^2(x)\Big)$$

and

$$\sqrt{nh_1}\big\{\widetilde{\sigma}_3^2(x)-\sigma^2(x)-\theta_{n3}\big\} \xrightarrow{d} N\Big(0,\big\{R(W)-l^2(1-l^2)\,C(\gamma)\big\}p(x)^{-1}\sigma^4(x)\lambda^2(x)\Big),$$

where $C(s,t)=\int W(u-st)W(u+st)\,du$ and $C(s)=\frac{3}{2}C(0,s)-2C(\frac{1}{2},s)+\frac{1}{2}C(1,s)$.

Hence $\widetilde{\sigma}_j^2(x)$ has the same asymptotic bias as $\widehat{\sigma}_j^2(x)$ for $j=1,3$. Note that $0\le l^2(1-l^2)\le 1/4$ for all $l\in[-1,1]$ and it attains the maximum at $l=\pm 2^{-1/2}$. Moreover, for symmetric kernel W the quantity $C(\gamma)$ is nonnegative for all $\gamma\ge 0$; $0\le C(\gamma)\le (3/2)R(W)$ and $C(\gamma)$ is increasing in γ if W is symmetric and concave; see Cheng et al. (2007). So, the variance reduction estimators have smaller asymptotic variances and asymptotic mean squared errors.

By choosing $l=\pm 2^{-1/2}$, we achieve the most variance reduction regardless what h_1, γ and W are and the resulting estimators are

$$\begin{aligned}\widetilde{\sigma}_{j,(1)}^2(x) \;=\;& \frac{1}{4}(1-2^{1/2})\,\widehat{\sigma}_j^2\big(x-(1+2^{-1/2})\gamma h_1\big)+\frac{1}{2}\,\widehat{\sigma}_j^2\big(x-2^{-1/2}\gamma h_1\big)\\ &+\frac{1}{4}(1+2^{1/2})\,\widehat{\sigma}_j^2\big(x-(2^{-1/2}-1)\gamma h_1\big) \qquad (3.2)\end{aligned}$$

and

$$\begin{aligned}\widetilde{\sigma}_{j,(2)}^2(x) \;=\;& \frac{1}{4}(1+2^{1/2})\,\widehat{\sigma}_j^2\big(x+(2^{-1/2}-1)\gamma h_1\big)+\frac{1}{2}\,\widehat{\sigma}_j^2\big(x+2^{-1/2}\gamma h_1\big)\\ &+\frac{1}{4}(1-2^{1/2})\,\widehat{\sigma}_j^2\big(x+(2^{-1/2}+1)\gamma h_1\big) \qquad (3.3)\end{aligned}$$

for $j=1$ and 3.

Either of the variance reduction estimators $\widetilde{\sigma}_{j,(1)}^2(x)$ and $\widetilde{\sigma}_{j,(2)}^2(x)$ uses more information from data points on one side of x than the other side; see (3.2) and (3.3). One way to balance this finite sample bias effect is to take the average

$$\widetilde{\sigma}_{j,(3)}^2(x)=\frac{1}{2}\big\{\widetilde{\sigma}_{j,(1)}^2(x)+\widetilde{\sigma}_{j,(2)}^2(x)\big\} \qquad (3.4)$$

for $j = 1$ and 3. When $Supp(\sigma) = [0, 1]$, to keep the points $\big\{\beta_{x,0}, \beta_{x,1}, \beta_{x,2}\big\}$ with $l = \pm 2^{-1/2}$ all within the data range $[0, 1]$ we let $\gamma(x) = \min\big\{\gamma, x/(1 + 2^{-1/2})h_1, (1 - x)/(1 + 2^{-1/2})h_1\big\}$ for a positive constant γ, $\gamma = 1$ say.

Theorem 3. Under the conditions of Theorem 1, for interior point x we have

$$\sqrt{nh_1}\{\widetilde{\sigma}^2_{1,(3)}(x) - \sigma^2(x) - \theta_{n1}\} \overset{d}{\to} N\Big(0, \big\{R(W) - C(\gamma)/4 - D(\gamma)/2\big\}p(x)^{-1}\sigma^4(x)\bar{\lambda}^2(x)\Big)$$

and

$$\sqrt{nh_1}\{\widetilde{\sigma}^2_{3,(3)}(x) - \sigma^2(x) - \theta_{n3}\} \overset{d}{\to} N\Big(0, \big\{R(W) - C(\gamma)/4 - D(\gamma)/2\big\}p(x)^{-1}\sigma^4(x)\lambda^2(x)\Big),$$

where

$$\begin{aligned} D(\gamma) \;\; = \;\; & R(W) - \frac{C(\gamma)}{4} - \frac{1}{16}\Big\{4\big(1 + \sqrt{2}\big)C\big(\sqrt{2} - 1, \gamma/2\big) + \big(3 + 2\sqrt{2}\big)C\big(2 - \sqrt{2}, \gamma/2\big) \\ & + 2C\big(\sqrt{2}, \gamma/2\big) + 4\big(1 - \sqrt{2}\big)C\big(\sqrt{2} + 1, \gamma/2\big) + \big(3 - 2\sqrt{2}\big)C\big(\sqrt{2} + 2, \gamma/2\big)\Big\}. \end{aligned}$$

Remark 4. Note that, for any kernel W, $0 \le D(\gamma) \le (5/8)R(W)$. Hence, $\widetilde{\sigma}^2_{j,(3)}(x)$ has a smaller asymptotic variance than both $\widetilde{\sigma}^2_{j,(1)}(x)$ and $\widetilde{\sigma}^2_{j,(2)}(x)$ for $j = 1$ and 3.

Remark 5. In Cheng et al. (2007) the variance reduction techniques are applied to nonparametric estimation of the regression $m(x)$. The results in Theorems 2 and 3 are nontrivial given the theory developed therein.

Remark 6. When estimating the conditional variance $\sigma^2(x)$, it does not provide any gain, in asymptotic terms, by replacing $\widehat{m}(X_i)$ with the variance reduced regression estimator of Cheng, et al. (2007) in the squared residuals $\widehat{r}_i = \big\{Y_i - \widehat{m}(X_i)\big\}^2$, $i = 1, \cdots, n$.

Remark 7. In (1.6), the term n^{-1} is added to avoid $\log 0$ and it can be replaced by $n^{-\eta}$, for any $\eta > 0$, without affecting the theoretical results in Theorems 1–3. However, in finite sample cases, a too small value of η would increase the bias and a too large value of η would increase the variability. In the simulation study summarized in Section 5.1, we also experimented with $\eta = 0.5$ and 2. We found that

$\eta = 0.5$ is undesirable as the MADE boxplot is always above the others even though it is narrower. When $\eta = 2$, besides the MADE boxplot is wider than the others, the performance is not stable with the MADE boxplot lower than the others only in some settings, not all.

4 Bandwidth Selection

In the construction of our estimator $\widehat{\sigma}_3^2(x)$, two bandwidths are needed: h_2 is the bandwidth in (1.2) to get the squared residuals $\widehat{r}_1, \cdots, \widehat{r}_n$ and h_1 is the bandwidth in (1.6) to estimate the conditional variance. Since both (1.2) and (1.6) are local linear fittings based on data $\big\{(X_i, Y_i), i = 1, \cdots, n\big\}$ and $\big\{(X_i, \log(\widehat{r}_i + n^{-1}), i = 1, \cdots, n\big\}$ respectively, we suggest to employ the same bandwidth procedure in the two steps. Let $\hat{h}(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$ denote any data-driven bandwidth rule for the local linear fitting (1.2).

1. Take $h_2 = \hat{h}(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$ in the local linear regression (1.2) to obtain the regression estimates $\widehat{m}(X_i)$, $i = 1, \cdots, n$, and the squared residuals $\widehat{r}_i = \big\{Y_i - \widehat{m}(X_i)\big\}^2$, $i = 1, \cdots, n$.

2. Use bandwidth $h_1 = \hat{h}\big(X_1, \cdots, X_n; \log(\widehat{r}_1 + n^{-1}), \cdots, \log(\widehat{r}_n + n^{-1})\big)$ in the local linear fitting (1.6) to get $\widehat{\sigma}_3^2(x)$.

A simple modification of the above bandwidth procedure can be used to implement our variance reduction estimator $\widetilde{\sigma}_{3,(3)}^2(x)$. Comparing Theorem 1 and Theorem 3, the asymptotically optimal global (or local) bandwidths of $\widetilde{\sigma}_{3,(3)}^2(x)$ and $\widehat{\sigma}_3^2(x)$ differ by the constant multiplier $\big\{R(W) - C(\gamma)/4 - D(\gamma)/2\big\}^{1/5}$ which depend only on the known W and γ. Therefore, no matter whether $\hat{h}$ is a global bandwidth or a local bandwidth, the modification proceeds as, in step 2 above, using

$$h_1 = \big\{R(W) - C(\gamma)/4 - D(\gamma)/2\big\}^{1/5}\, \hat{h}\big(X_1, \cdots, X_n; \log(\widehat{r}_1 + n^{-1}), \cdots, \log(\widehat{r}_n + n^{-1})\big) \tag{4.1}$$

in (1.6) to obtain $\widehat{\sigma}_3^2(z)$, $z \in \{\beta_{x,0}, \beta_{x,1}, \beta_{x,2}\}$ with $l = \pm 2^{-1/2}$. Then one can form the linear combinations, specified in (3.2), (3.3) and (3.4), to get $\widetilde{\sigma}_{3,(3)}^2(x)$.

For the conditional variance estimator $\widehat{\sigma}_1^2(x)$, Fan and Yao (1998) recommended a bandwidth principle analogous to what we specify for $\widehat{\sigma}_3^2(x)$ in the above. To modify the bandwidth rule of $\widehat{\sigma}_1^2(x)$ for use in $\widetilde{\sigma}_{1,(3)}^2(x)$, apply the same constant factor adjustment as in (4.1) when computing $\widehat{\sigma}_1^2(z)$ for $z \in \{\beta_{x,0}, \beta_{x,1}, \beta_{x,2}\}$ with $l = \pm 2^{-1/2}$.

5 Numerical Study

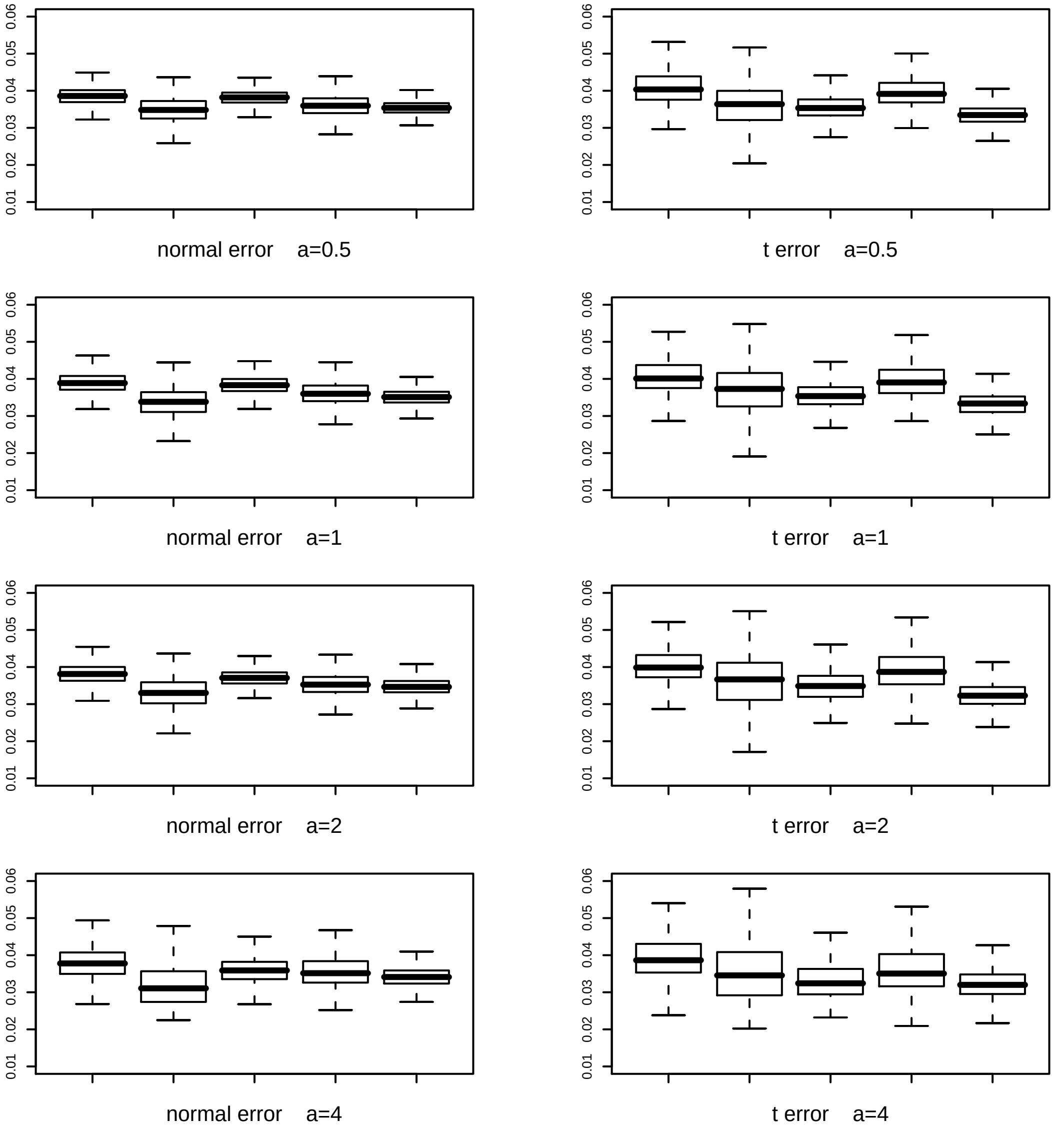
The five estimators $\widehat{\sigma}_1^2(x)$, $\widehat{\sigma}_2^2(x)$, $\widehat{\sigma}_3^2(x)$, $\widetilde{\sigma}_{1,(3)}^2(x)$ and $\widetilde{\sigma}_{3,(3)}^2(x)$ are compared based on their finite sample performances via a simulation study and an application to the motorcycle data set.

5.1 Simulation

Consider the regression model

$$Y_i = a\left\{X_i + 2\exp\left(-16X_i^2\right)\right\} + \sigma\left(X_i\right)\varepsilon_i\,, \tag{5.1}$$

where $\sigma(x) = 0.4\exp\left(-2x^2\right) + 0.2$, $a = 0.5, 1, 2,$ or 4, $X_i \sim \text{Uniform}[-2, 2]$ and ε_i is independent of X_i and follows either the $N(0, 1)$ or the $(1/\sqrt{3})\,t_3$ distribution. For each of the settings, 1000 samples of size $n = 200$ were generated from model (5.1). The plug-in bandwidth of Ruppert et al. (1995) was employed as the bandwidth selector $\hat{h}$ in Section 4. To implement $\widehat{\sigma}_2^2(x)$, h_1 was taken as the data-driven bandwidth given in Yu and Jones (2004). Both K and W, kernels in the regression and conditional variance estimation stages, were taken as the Epanechnikov kernel $K(u) = (3/4)(1 - u^2)I(|u| < 1)$. The parameter γ in $\widehat{\sigma}_{1,(3)}^2(x)$ and $\widehat{\sigma}_{3,(3)}^2(x)$ was set to 1. Performance of an estimator $\widehat{\sigma}(\cdot)$ of $\sigma(\cdot)$ is measured by the mean absolute



normal error a=0.5

t error a=0.5

normal error a=1

t error a=1

normal error a=2

t error a=2

normal error a=4

t error a=4

Figure 1: *MADE boxplots under model (5.1).* In each panel, the MADE boxplots of $\widehat{\sigma}_1(\cdot)$, $\widehat{\sigma}_2(\cdot)$, $\widehat{\sigma}_3(\cdot)$, $\widetilde{\sigma}_{1,(3)}(\cdot)$ and $\widetilde{\sigma}_{3,(3)}(\cdot)$ are arranged from left to right. The left and right columns respectively give the results for the Normal and t errors. From the top, the rows correspond to $a = 0.5, 1, 2$ and 4.

deviation error or the mean squared deviation error, respectively defined by

$$MADE\left(\widehat{\sigma}\right) = \frac{1}{g}\sum_{i=1}^{g}\left|\widehat{\sigma}\left(x_i\right) - \sigma\left(x_i\right)\right|, \quad MSDE\left(\widehat{\sigma}\right) = \frac{1}{g}\sum_{i=1}^{g}\left\{\widehat{\sigma}\left(x_i\right) - \sigma\left(x_i\right)\right\}^2,$$

where $\{x_i, i = 1, \cdots, g\}$ is a grid on [-2,2] with $g = 101$. Here, we measure the performance by MADE or MSDE of estimating $\sigma(\cdot)$ instead of those of estimating $\sigma^2(\cdot)$ since the latter would seriously down-weight errors in estimating small values of $\sigma^2(\cdot)$.

Figure 1 presents the MADE boxplots of the five estimators $\widehat{\sigma}_j(\cdot)$, $j = 1, 2, 3$, and $\widetilde{\sigma}_{j,(3)}(\cdot)$, $j = 1, 3$. Under all of the configurations, $\widetilde{\sigma}_{j,(3)}(\cdot)$ outperforms $\widehat{\sigma}_j(\cdot)$ for $j = 1, 3$ and our log-tranform based methods improve on the Fan and Yao (1998) estimator. When the error distribution is Normal, the Yu and Jones (2004) estimator $\widehat{\sigma}_2(\cdot)$ is somehow the best since its MADE median is the lowest. The reason for this optimality is that $\widehat{\sigma}_2(\cdot)$ is derived from a local Normal likelihood model which now coincides with the true error model. However, its MADE boxplot is always much wider than those of the other four and this instability is intrinsic to local likelihood methods. Interestingly, our estimator $\widetilde{\sigma}_{3,(3)}(\cdot)$ is nearly optimal even under Normal errors: compared to $\widehat{\sigma}_2(\cdot)$, the MADE median is roughly the same and the MADE

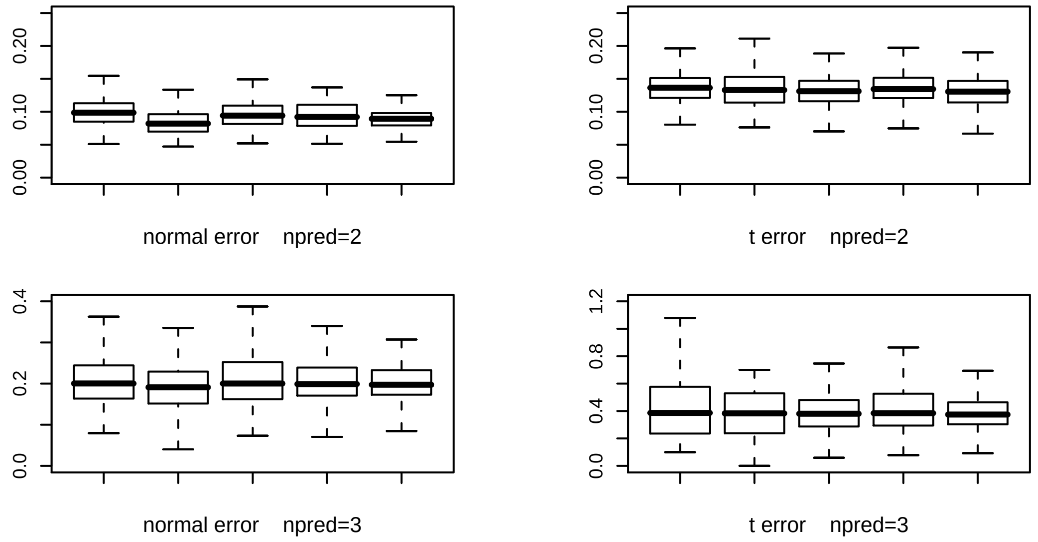


Figure 2: *MADE boxplots of predictions under model (5.2).* The layout is the same as in Figure 1, except that the top and bottom rows respectively represent the two- and three-step predictions here.

upper quartile is lower. Further, $\widehat{\sigma}_3(\cdot)$ and $\widetilde{\sigma}_{3,(3)}(\cdot)$ become decisively better than any of the others under t errors. Therefore, $\widehat{\sigma}_3(\cdot)$ and $\widetilde{\sigma}_{3,(3)}(\cdot)$ are very robust against heavy tailed errors, and $\widehat{\sigma}_2(\cdot)$ is not robust against departure from Normality. Although $\widetilde{\sigma}_{1,(3)}(\cdot)$ performs better than $\widehat{\sigma}_1(\cdot)$ all the time but its behaviors under different error distributions is predetermined by that of $\widehat{\sigma}_1(\cdot)$. The MSDE boxplots are not given here, but they provide similar conclusions.

Another setting we considered is the following nonlinear time series model

$$X_{t+1} = 0.235 X_t \left(16 - X_t\right) + \varepsilon_t \tag{5.2}$$

where $\varepsilon_t \sim N\left(0, 0.3^2\right)$ is independent of X_t. From model (5.2), 500 samples of size $n = 500$ were simulated. The conditional variance of X_{t+1} given the past data $\left\{X_t, X_{t-1}, \cdots\right\}$ is a constant function. Hence we investigate estimation of the conditional variances in two-step and three-step prediction problems with $Y_t = X_{t+2}$ and $Y_t = X_{t+3}$ respectively. Figure 2 presents the MADE boxplots, which convey similar conclusions as in the previous example. In particular, $\widetilde{\sigma}_{3,(3)}(\cdot)$ is quite reliable and robust.

In the construction of $\widehat{\sigma}_3(\cdot)$ and $\widetilde{\sigma}_{3,(3)}(\cdot)$, the term n^{-1} is added in (1.6) to avoid $\log 0$. We also experimented with the term n^{-1} replaced by $n^{-0.5}$ or n^{-2}. To save space, the results are not given here. We found that with $n^{-0.5}$ the MADE boxplot is always above the others even though it is narrower. When using n^{-2}, the MADE boxplot is wider than the others, but the performance is not stable with the MADE boxplot lower than the others only in some settings.

5.2 An Application

The estimators $\widehat{\sigma}_1^2(x)$, $\widehat{\sigma}_2^2(x)$, $\widehat{\sigma}_3^2(x)$, $\widetilde{\sigma}_{1,(3)}^2(x)$ and $\widetilde{\sigma}_{3,(3)}^2(x)$ were employed to estimate the conditional variance for the motorcycle data given by Schmidt et al. (1981). The covariate X is the time (in milliseconds) after a simulated impact on motorcycles and the response variable Y is the head acceleration (in gram) of a test object. The

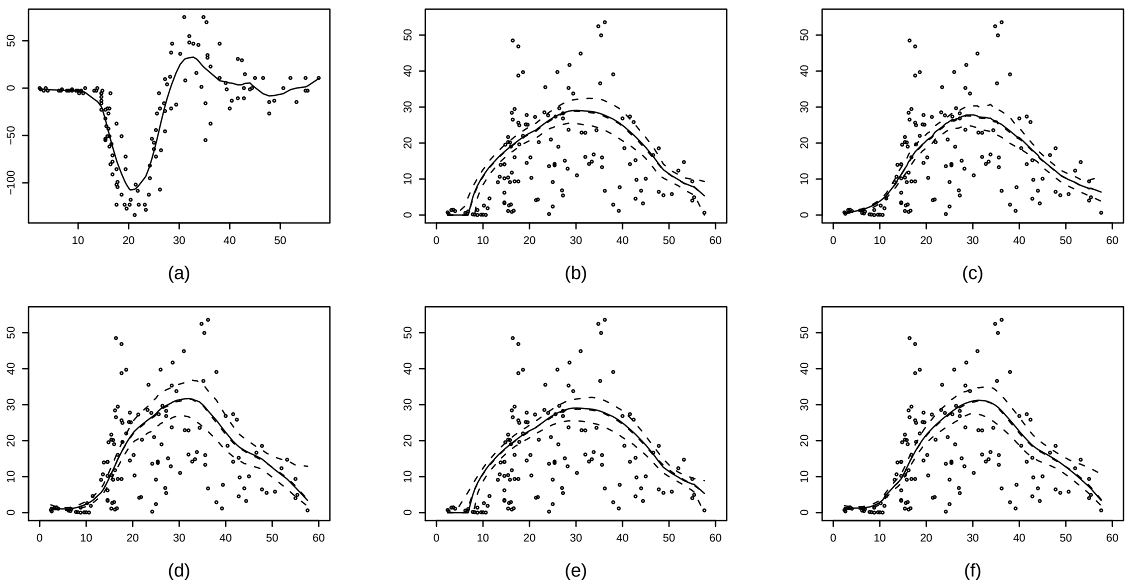


Figure 3: *Motorcycle data.* Panel (a) depicts the motorcycle data and the local linear regression estimate $\widehat{m}(\cdot)$. The absolute residuals are plotted against the design points in panels (b)–(f), which respectively show the estimates $\widehat{\sigma}_1(\cdot)$, $\widehat{\sigma}_2(\cdot)$, $\widehat{\sigma}_3(\cdot)$, $\widetilde{\sigma}_{1,(3)}(\cdot)$ and $\widetilde{\sigma}_{3,(3)}(\cdot)$ (solid lines) and 10%, 50% and 90% variability curves of their resampled versions (dashed lines).

sample size is 132. As before, the squared residuals $\widehat{r}_1, \cdots, \widehat{r}_n$ are used to estimate the conditional variance. We took $\hat{h}$ in Section 4 as the bandwidth selector of Rupprt, et al. (1995). Then the bandwidth h_2 in $\widehat{m}(\cdot)$ was 4.0145, and the bandwidth h_1 in $\widehat{\sigma}_1^2(\cdot)$, $\widehat{\sigma}_3^2(\cdot)$, $\widetilde{\sigma}_{1,(3)}^2(\cdot)$ and $\widetilde{\sigma}_{3,(3)}^2(\cdot)$ was respectively 6.1775, 4.5763, 5.1053 and 3.7821. The bandwidth h_1 in $\widehat{\sigma}_3^2(\cdot)$ was selected by the method of Yu and Jones (2004) and was 13.4188. In Figure 3, panel (a) depicts the original data and the local linear regression estimate $\widehat{m}(\cdot)$, and the solid lines in panels (b)–(f) are respectively the estimates $\widehat{\sigma}_1(\cdot)$, $\widehat{\sigma}_2(\cdot)$, $\widehat{\sigma}_3(\cdot)$, $\widetilde{\sigma}_{1,(3)}(\cdot)$ and $\widetilde{\sigma}_{3,(3)}(\cdot)$. In Figure 4, panel (a) plots the residuals, and panels (b)–(f) depicts the Normal Q-Q plots of the residuals divided by the estimates of the conditional standard deviations. Panels (b) and (e) suggest that the motorcycle data has a heavy left tail in the error distribution. Panels (d) and (f) show that our estimators $\widehat{\sigma}(\cdot)$ and $\widetilde{\sigma}_{3,(3)}(\cdot)$ effectively correct the heavy left tail. Panel (c) indicates a departure from normality when $\widehat{\sigma}_2(\cdot)$ is applied to this data

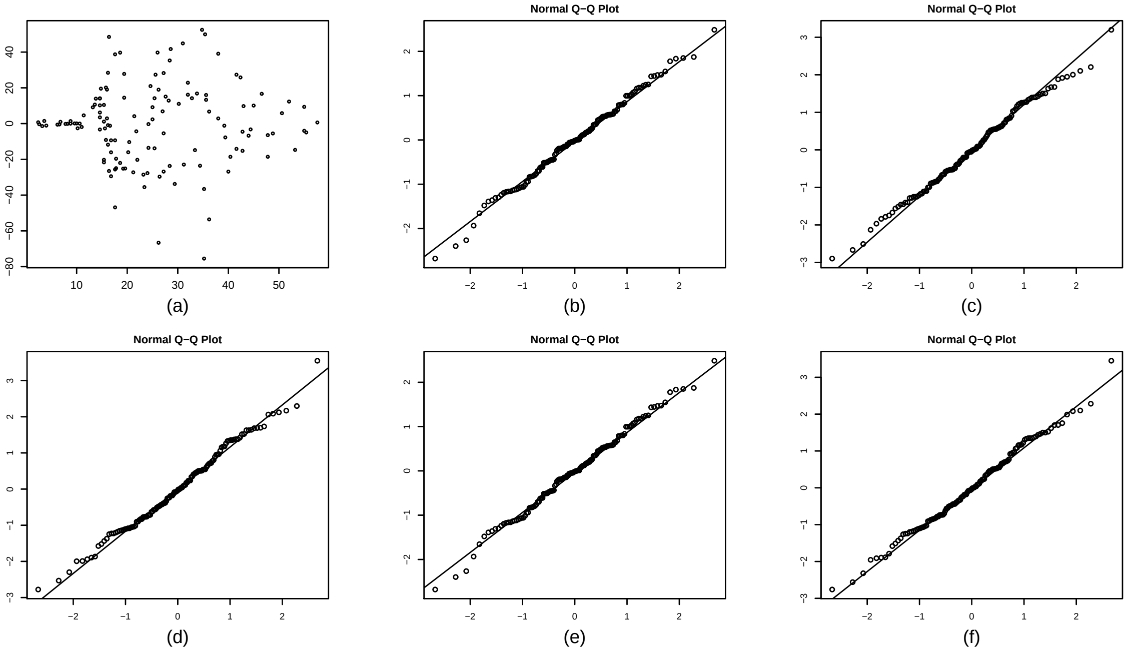


Figure 4: *Normal Q-Q plots of residuals.* Panel (a) depicts the residuals of the motorcycle data after the local linear regression. Panels (b)–(f) are respectively the Normal Q-Q plots of the residuals divided by $\widehat{\sigma}_1(\cdot)$, $\widehat{\sigma}_2(\cdot)$, $\widehat{\sigma}_3(\cdot)$, $\widetilde{\sigma}_{1,(3)}(\cdot)$ and $\widetilde{\sigma}_{3,(3)}(\cdot)$.

set.

To access the variability in each of the estimators, 500 simple random resamples of size 132 were drawn from $\big\{(X_i, \widehat{r}_i), i = 1, \cdots, 132\big\}$ and each of the estimators was applied to the resamples using the respective bandwidth h_1. The dashed lines in panels (b)–(f) of Figure 3 are the pointwise 10%, 50% and 90% curves for the respective estimators. Our estimator $\widetilde{\sigma}_{3,(3)}(\cdot)$ has the narrowest variability band while both the pointwise 50% curve and the estimate itself follow the trend in the residuals well. As expected, $\widehat{\sigma}^2_1(\cdot)$ (and hence $\widetilde{\sigma}^2_{1,(3)}(\cdot)$) suffers from the negativity problem.

6 Proofs

Proof of Theorem 1. We follow the lines of the proof in Fan and Yao (1998). Let $G \subset \{p(x) > 0\}$ be a compact set. Similar to the proof of Lemma 2 of Yao and Tong (2000), we can show that

$$\widehat{\nu}(x)-\nu(x) = \Big\{\frac{1}{np(x)}\sum_{i=1}^{n} W_{h_1}\big(X_i-x\big)\big[\log(\widehat{r}_i+n^{-1})-\nu(x)-\dot{\nu}(x)(X_i-x)\big]\Big\}\big\{1+o_p(1)\big\} \tag{6.1}$$

and

$$\widehat{m}(x)-m(x) = \frac{1}{np(x)}\sum_{i=1}^{n}\sigma(X_i)\,\varepsilon_i K_{h_2}\big(X_i-x\big)+\frac{h_2^2\ddot{m}(x)}{2}\int t^2K(t)\,dt+o_p\big(\frac{1}{\sqrt{nh_2}}+h_2^2\big) \tag{6.2}$$

uniformly in $x \in G$, and

$$\sup_{x\in G}\big|\widehat{m}(x)-m(x)\big| = O\big\{(nh_2)^{-1/2}(\log(h_2^{-1}))^{1/2}\big\} \quad \text{a.s.} \tag{6.3}$$

Here, $W_{h_1}(u) = h_1^{-1}W(u/h_1)$ and $K_{h_2}(u) = h_2^{-1}K(u/h_2)$. It follows from (6.1) that

$$\begin{aligned}
&\exp\{\widehat{\nu}(x)\}/d-\sigma^2(x)\\
= \; &\Big\{\frac{\sigma^2(x)}{np(x)}\sum_{i=1}^{n} W_{h_1}\big(X_i-x\big)\big[\log(\widehat{r}_i+n^{-1})-\nu(x)-\dot{\nu}(x)(X_i-x)\big]\Big\}\{1+o_p(1)\}\\
= \; &\Big\{\frac{\sigma^2(x)}{np(x)}\sum_{i=1}^{n} W_{h_1}\big(X_i-x\big)\big[\nu(X_i)-\nu(x)-\dot{\nu}(x)(X_i-x)\big]\\
&+\frac{\sigma^2(x)}{np(x)}\sum_{i=1}^{n} W_{h_1}\big(X_i-x\big)\log(\varepsilon_i^2/d)\\
&+\frac{\sigma^2(x)}{np(x)}\sum_{i=1}^{n} W_{h_1}\big(X_i-x\big)\big[\log(\widehat{r}_i+n^{-1})-\log(r_i)\big]\Big\}\{1+o_p(1)\}\\
= \; &\big\{I_1+I_2+I_3\big\}\{1+o_p(1)\}.
\end{aligned} \tag{6.4}$$

A direct application of ergodic theorem yields that

$$I_1 = \theta_{n3}+o_p(h_1^2). \tag{6.5}$$

Since $\{\log xI(x > e^3)\}^4$ is a concave function, condition (C2) implies that

$$E\big\{(\log\varepsilon)^4I(\varepsilon > e^3)\big|X = x\big\} \le \big\{\log\big(E[\varepsilon^2I(\varepsilon^2 > e^3)\big|X = x]\big)\big\}^4 < \infty. \tag{6.6}$$

By (6.6) and (C3), we have

$$\begin{aligned} E\Big\{W\big(\frac{X_i-x}{h_1}\big)\log(\varepsilon_i^2/d)\Big\}^4 &= E\Big\{W\big(\frac{X_i-x}{h_1}\big)\log(\varepsilon_i^2/d)I(\varepsilon_i^2\le e^3)\Big\}^4 \\ &\quad +E\Big\{W\big(\frac{X_i-x}{h_1}\big)\log(\varepsilon_i^2/d)I(\varepsilon_i^2> e^3)\Big\}^4 \\ &= E\Big\{W\big(\frac{X_i-x}{h_1}\big)^4 E\big((\log(\varepsilon_i^2/d))^4 I(\varepsilon_i^2\le e^3)\big|X_i\big)\Big\} \\ &\quad +E\Big\{W\big(\frac{X_i-x}{h_1}\big)^4 E\big((\log(\varepsilon_i^2/d))^4 I(\varepsilon_i^2> e^3)\big|X_i\big)\Big\} \\ &< \infty. \end{aligned} \tag{6.7}$$

Like Fan and Yao (1998), it follows from (6.7), (C4) and Theorem 1.7 of Peligrad (1986) that

$$\sqrt{nh_1}I_2 \xrightarrow{d} N\Big(0, p^{-1}(x)\,\sigma^4(x)\lambda^2(x)\int W^2(t)\,dt\Big). \tag{6.8}$$

Put $\zeta_n=(nh_1)^{-1/2}\big\{\log(n)\big\}^{-2}$. Write

$$\begin{aligned} I_3 &= \frac{\sigma^2(x)}{np(x)}\sum_{i=1}^n W_{h_1}\big(X_i-x\big)\big[\log(\widehat{r}_i+n^{-1})-\log r_i\big]I(\varepsilon_i^2\le\zeta_n) \\ &\quad +\frac{\sigma^2(x)}{np(x)}\sum_{i=1}^n W_{h_1}\big(X_i-x\big)\big[\log(\widehat{r}_i+n^{-1})-\log r_i\big]I(\varepsilon_i^2>\zeta_n) \\ &= I_4+I_5. \end{aligned}$$

Note that

$$\widehat{r}_i=\sigma^2(X_i)\,\varepsilon_i^2+2\sigma(X_i)\,\varepsilon_i\big\{m(X_i)-\widehat{m}(X_i)\big\}+\big\{m(X_i)-\widehat{m}(X_i)\big\}^2.$$

When n is large enough, we have

$$\begin{aligned} |I_4| &= -\frac{\sigma^2(x)}{np(x)}\sum_{i=1}^n W_{h_1}\big(X_i-x\big)\log\big(\widehat{r}_i+n^{-1}\big)I(\varepsilon_i^2\le\zeta_n) \\ &\quad -\frac{\sigma^2(x)}{np(x)}\sum_{i=1}^n W_{h_1}\big(X_i-x\big)\log r_i\, I(\varepsilon_i^2\le\zeta_n) \\ &\le \frac{\sigma^2(x)}{np(x)}\sum_{i=1}^n W_{h_1}\big(X_i-x\big)\log n\, I(\varepsilon_i^2\le\zeta_n) \\ &\quad -\frac{\sigma^2(x)}{np(x)}\sum_{i=1}^n W_{h_1}\big(X_i-x\big)\log(\sigma^2(X_i)\,\varepsilon_i^2)I(\varepsilon_i^2\le\zeta_n) \\ &= o_p\big\{(nh_1)^{-1/2}\big\}. \end{aligned} \tag{6.9}$$

Since, for θ_i between $\widehat{r}_i + n^{-1}$ and r_i,

$$\begin{aligned}
\log\big(\widehat{r}_i + n^{-1}\big) - \log(r_i) &= \frac{1}{r_i}\big(\widehat{r}_i + n^{-1} - r_i\big) - \frac{1}{2\theta_i^2}\big(\widehat{r}_i + n^{-1} - r_i\big)^2 \\
&= \frac{1}{\sigma^2(X_i)\,\varepsilon_i^2}\big[2\sigma(X_i)\,\varepsilon_i\{m(X_i) - \widehat{m}(X_i)\} + \{m(X_i) - \widehat{m}(X_i)\}^2\big] \\
&\quad -\frac{1}{2\theta_i^2}\big[2\sigma(X_i)\,\varepsilon_i\{m(X_i) - \widehat{m}(X_i)\} + \{m(X_i) - \widehat{m}(X_i)\}^2\big]^2
\end{aligned}$$

and

$$\theta_i^{-2} \le \big\{\sigma^2(X_i)\,\varepsilon_i^2(X_i)\big\}^{-2} + \big\{\sigma^2(X_i)\,\varepsilon_i^2 + 2\sigma(X_i)\,\varepsilon_i(m(X_i) - \widehat{m}(X_i)) + (m(X_i) - \widehat{m}(X_i))^2\big\}^{-2},$$

we can show, in a way similar to Fan and Yao (1998), that

$$I_5 = o_p\big\{(nh_1)^{-1/2}\big\}. \tag{6.10}$$

By (6.9) and (6.10),

$$I_3 = o_p\big\{(nh_1)^{-1/2}\big\}. \tag{6.11}$$

Further we can show that

$$\begin{aligned}
&\widehat{d}^{-1} - d^{-1} \\
= \ & \tfrac{1}{n}\textstyle\sum_{i=1}^n(\hat{r}_i - r_i)\exp\{-\hat{\nu}(X_i)\} \\
& + \tfrac{1}{n}\textstyle\sum_{i=1}^n r_i \exp\{-\nu(X_i)\}(\exp\{-\hat{\nu}(X_i) + \nu(X_i)\} - 1) \\
= \ & \tfrac{1}{n}\textstyle\sum_{i=1}^n(\hat{r}_i - r_i)\exp\{-\hat{\nu}(X_i)\} \\
& + \tfrac{1}{n}\textstyle\sum_{i=1}^n d^{-1}\epsilon_i^2(\exp\{-\hat{\nu}(X_i) + \nu(X_i)\} - 1) \\
= \ & -\tfrac{1}{n}\textstyle\sum_{i=1}^n d^{-1}\epsilon_i^2\tfrac{1}{2}h_1^2\ddot{\nu}(X_i)\int t^2W(t)\,dt + o_p(\tfrac{1}{\sqrt{nh_1}} + h_1^2) \\
= \ & -\tfrac{1}{2d}h_1^2E\ddot{\nu}(X_1)\int t^2W(t)\,dt + o_p(\tfrac{1}{\sqrt{nh_1}} + h_1^2).
\end{aligned} \tag{6.12}$$

Hence, the theorem follows from (6.5), (6.8), (6.11) and (6.12).

Proofs of Theorems 2 and 3. Use (6.4) and the corresponding arguments in Cheng et al. (2007).

Acknowledgments. We thank a reviewer for his/her helpful comments. Cheng's research was supported in part by NSC grant NSC-94-2118-M-002-002 and Mathematics Division, National Center for Theoretical Sciences (Taipei Office). Peng's research was supported by NSF grants DMS-04-03443, SES-0631608, the Humboldt research fellowship and HKSAR-RGC grant No. 400306.

References

[1] M. Borkovec (2001). Asymptotic behavior of the sample autocovariance and auto correlation function of the AR(1) process with ARCH(1). *Bernoulli 6, 847 - 872.*

[2] M. Borkovec and C. Klüppelberg (2001). The tail of the stationary distribution of an autoregressive process with ARCH(1) errors. *Ann. Appl. Probab. 11, 1220 - 1241.*

[3] N.H. Chan and L. Peng (2005). Weighted least absolute deviations estimation for an AR(1) process with ARCH(1) errors. *Biometrika 92, 477 - 484.*

[4] M.-Y. Cheng, L. Peng and J.S. Wu (2007). Reducing variance in univariate smoothing. *Ann. Statist. 35, 522–542.*

[5] R.F. Engle (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica 50, 987 - 1008.*

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

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

[8] S. Ling (2004). Estimation and testing stationarity for double autoregressive models. *J. Roy. Statist. Soc. Ser. B 66, 63 - 78.*

[9] S. Mittnik and S.T. Rachev (2000). Stable Paretian Models in Finance. *Wiley, New York.*

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

[11] L. Peng and Q. Yao (2003). Least absolute deviations estimation for ARCH and GARCH models. *Biometrika 90, 967 - 975.*

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

[13] G. Schmidt, R. Mattern and F. Schüler (1981). Biomechanical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column with and without protective helmet under effects of impact. EED Research Program on Biomechanics of Impacts. Final Report Phase III, Project 65, Institut für Rechtsmedizin, Universität Heidelberg, Germany.

[14] Q. Yao and H. Tong (2000). Nonparametric estimation of ratios of noise to signal in stochastic regressions. *Statistica Sinica 10, 751 - 770.*

[15] K. Yu and M.C. Jones (2004). Likelihood-based local linear estimation of the conditional variance function. *J. Amer. Statist. Assoc. 99, 139 - 144.*