ASYMMETRIC LEAST SQUARES REGRESSION ESTIMATION: A NONPARAMETRIC APPROACH*

Qiwei Yao and Howell Tong

Institute of Mathematics and Statistics, University of Kent
Canterbury, Kent CT2 7NF, UK

Abstract

This paper considers the nonparametric estimation of regression expectiles and percentiles by using an asymmetric least squares (ALS) approach, in which the squared error loss function is given different weight depending on whether the residual is positive or negative. The kernel method based on locally linear fit is adopted, which also provides an estimator of the derivative of the regression function. Under the assumption that the observations are strictly stationary and ρ-mixing, the asymptotic normality for the estimators of conditional expectiles is established by using the convexity lemma. For a large class of regression models, the ALS approach can be adapted to estimate the conditional percentiles directly. Further, we show that these ALS estimators for conditional percentiles are consistent.

Abbreviated title: Asymmetric Least Squares Regression.

Some key words: Asymmetric least squares estimator; Expectile; Kernel estimation; Local linear regression; Percentile, ρ-mixing.

---

*Research partially supported by the Science and Engineering Research Council (U.K.).

1 Introduction

In the standard regression analysis, most of the methods developed so far are based on the mean regression function, which is an estimator of conditional expectation. Geometrically, the observations $\{(X_i,\ Y_i),\ i = 1, \ldots, n\}$ form a cloud of points in a Euclidean space. The mean regression function describes the middle of the point-cloud, in the Y direction, as a function of the covariate X (cf. Efron 1991). However, new insights about the underlying structure can be gained by investigating the higher or lower regions of the point-cloud. This leads us to study the estimation of the *conditional percentiles* of Y given X. Usually, the asymmetric least absolute deviations (ALAD) methods are used to estimate the regression curves (cf. Hogg 1975, Koenker and Bassett 1978, Bassett and Koenker 1982 etc.). Note that the mean regression function is virtually a (symmetric) least squares estimator. In this sense, a natural way to investigate the higher or lower regions of the point-cloud would be to consider the asymmetric least squares (ALS) regression. This leads to the concept of the so-called *conditional expectiles*, which was originally proposed by Newey and Powell (1987). See also Efron (1991) and Section 2 below. It turns out that similar to the conditional percentiles, the conditional expectiles also characterize the underlying conditional distribution. Therefore, it provides an effective diagnostic tool such as testing the heteroscedasticity of regression models and the conditional symmetry of the noise terms (cf. Efron 1991, Newey and Powell 1987). Although the ALS method is not as robust as the ALAD method against outliers, it has some desired features. For example, an ALS estimator is easier to compute and reasonably efficient under normality conditions (cf. Efron 1991). In terms of interval prediction, the conditional percentile is more appealing than the conditional expectile because of its conventional probability interpretation, whilst for general statistical diagnoses, the conditional expectile is a valuable alternative to the conditional percentile. Because neither is uniformly superior, the choice will usually depend on the particular application at hand. This situation is similar to the comparison between the conditional mean and the conditional median in conventional regression. On the other hand, we observe that for quite a large class of nonlinear regression models, the conditional expectiles as functions of X are in a one–one correspondence with the conditional percentiles. Therefore, the ALS approach can be adapted to estimate conditional percentiles directly.

All the above mentioned work concentrated on parametric models. Fan, Hu and Truong (1992) has studied a class of nonparametric function estimators based on the locally linear fit for i.i.d.

observations, which includes the ALAD estimators for conditional percentiles as a special case. Their method can also be applied to the ALS estimators for conditional expectiles although this was not stressed there explicitly. Fan (1992) has showed that the kernel estimator based on locally linear fit is superior to that based on locally constant fit.

In this paper, we estimate the conditional expectiles and the conditional percentiles in a special class of regression models by using the ALS method based on the locally linear fit for weakly dependent data. We highlight the fact that the locally linear fit offers a natural estimator of the derivative of the regression function. Yao and Tong (1994b) has indicated that the estimators of derivatives of regression curves play a very important role in monitoring the reliability of nonlinear prediction and detecting chaos. Under the assumption that the observations are strictly stationary and ρ-mixing, we establish the asymptotic normality for the ALS estimators of conditional expectiles as well as their derivatives by using the convexity lemma (cf. Pollard, 1991, for example). With a trivial extension of this interesting lemma, we also show that the estimator of conditional expectile converges weakly to the real curve uniformly on compact subsets. Further, the weak consistency of the ALS estimators of conditional percentiles is proved. Although the ALS estimator cannot be expressed in a closed form, a convenient iterative algorithm can be constructed easily. Simulation shows that this algorithm converges very fast.

The paper is organized as follows. Section 2 provides a brief description of conditional percentiles and expectiles, as well as their ALS estimators and iterative algorithms. Some numerical results are also reported. Section 3 states the main results, the technical proofs of which are deferred to Section 4.

2 Percentiles and Expectiles

Let $\{(X_i, Y_i), i = 1, \ldots, n\}$ be a sequence of two-dimensional real strictly stationary random vectors, each having the same distribution as (X, Y). Suppose that $E(Y^2|X = x)$ as a function of $x \in \mathbf{R}$ is continuous.

2.1 Expectiles

We begin our discussion with the conventional definition of percentiles. Let $\alpha \in [0, 1]$. The 100α-th conditional percentiles of Y given $X = x$ is defined as

$$\xi_\alpha(x) \;=\; \arg\min_{|a|<\infty} E\{\, R_\alpha(Y - a) \mid X = x \,\},$$

where the loss function

$$R_\alpha(y) = \begin{cases} (1-\alpha)|y| & y \le 0, \\ \alpha|y| & y > 0. \end{cases}$$

It is well known that the relation $\alpha = P\{Y \le \xi_\alpha(x)|X = x\}$ holds. To compare it with expectiles, we rewrite this relation as follows

$$\alpha = \frac{E\{\ I_{\{Y \le \xi_\alpha(x)\}} \mid X = x\ \}}{E\{\ 1\ \mid X = x\ \}} \tag{2.1}$$

It is easy to see that when $\alpha = \frac{1}{2}, \xi_\alpha(x)$ is the conditional median.

If we define the loss function as

$$Q_\omega(y) = \begin{cases} (1-\omega)y^2 & y \le 0, \\ \omega y^2 & y > 0, \end{cases} \tag{2.2}$$

for $\omega \in [0,1]$, the 100 ω-th conditional expectile of Y is defined as the minimizer

$$\tau_\omega(x) \ = \ \arg\min_{|a|<\infty}\ E\{\ Q_\omega(Y-a) \mid X = x\ \},$$

(cf. Neway and Powell 1987). Obviously, in the case $\omega = \frac{1}{2}$, this definition reduces to the conditional mean $E(Y|X = x)$. Since $Q_\omega(.)$ has a continuous first derivative, $\tau_\omega(x)$ satisfies the equation

$$E\{\ L_\omega(Y - \tau_\omega(x)) \mid X = x\ \} = 0,$$

where

$$L_\omega(y) = \begin{cases} (1-\omega)y & y \le 0, \\ \omega y & y > 0. \end{cases} \tag{2.3}$$

Consequently, we have

$$\omega = \frac{E\{\ |Y - \tau_\omega(x)|\ I_{\{Y \le \tau_\omega(x)\}} \mid X = x\}}{E\{\ |Y - \tau_\omega(x)|\ \mid X = x\ \}}. \tag{2.4}$$

Comparing this expression with (2.1), we can see that given $X = x$, the percentile $\xi_\alpha(x)$ specifies the position below which $100\alpha\%$ of the (probability) mass of Y lies; while the expectile $\tau_\omega(x)$ determines, again given $X = x$, the point such that $100\omega\%$ of the mean distance between it and Y comes from the mass below it. Further, it will be seen later that if the conditional distribution of Y given $X = x$ belongs to a location and scale family (with respect to $x \in \mathbf{R}$), then $\{\tau_\omega(.),\ \omega \in (0,1)\} \equiv \{\xi_\alpha(.),\ \alpha \in (0,1)\}$.

To estimate $\tau_\omega(x)$, a conventional method may be the kernel estimation based on the locally constant fit. More precisely, estimate $\tau_\omega(x)$ by the minimizer of the function

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

where $K(.)$ and h are respectively a density function and a bandwidth. In the special case $\omega = \frac{1}{2}$, the method leads to the popular Nadaraya-Watson estimator. It has been pointed out that the kernel estimation based on the locally constant fit can be rather deficient (cf. Chu and Marron 1991, Fan 1992). For example, the bias of the estimator can have an adverse effect when the derivative of the marginal density or that of the estimated function is large, which would be typically the case in chaotic systems (cf. Yao and Tong 1994a,b). The drawback can be repaired by using a locally linear fit instead (Fan 1992, Fan, Hu and Truong 1992). Further, in order to monitor the efficacy in the context of stochastic prediction, it is essential to estimate the derivative of the estimated functions besides the latter (cf. Yao and Tong 1994a,b).

Let $\lambda_\omega(x) = d\tau_\omega(x)/dx$. The idea of a locally linear fit is to approximate the unknown function $\tau_\omega(.)$ by a linear function $\tau_\omega(z) \approx \tau_\omega(x) + \lambda_\omega(x)(z-x) \equiv a + b(z-x)$ for all z near x. Locally, estimating $\tau_\omega(x)$ and $\lambda_\omega(x)$ is equivalent to estimating a and b. Thus we may estimate $\tau_\omega(x)$ by $\hat{a}$ and $\lambda_\omega(x)$ by $\hat{b}$, where $(\hat{a},\ \hat{b})$ minimizes the function

$$\sum_{i=1}^{n} Q_\omega\{Y_i - a - b(X_i - x)\}K\left(\frac{X_i - x}{h}\right). \tag{2.5}$$

It is easy to see that $\{\hat{\tau}_\omega(x), \hat{\lambda}_\omega(x)\}$ satisfies the following equation

$$\begin{cases} \sum_{i=1}^{n} L_\omega\{Y_i - \hat{\tau}_\omega(x) - \hat{\lambda}_\omega(x)(X_i - x)\}K\left(\frac{X_i - x}{h}\right) = 0, \\ \sum_{i=1}^{n} (X_i - x) L_\omega\{Y_i - \hat{\tau}_\omega(x) - \hat{\lambda}_\omega(x)(X_i - x)\}K\left(\frac{X_i - x}{h}\right) = 0. \end{cases} \tag{2.6}$$

Here, $L_\omega(.)$ is the piecewise linear function defined by (2.3). The following interative algorithm is available for computing $(\hat{\tau}_\omega(x), \hat{\lambda}_\omega(x))$.

For $a, b \in \mathbf{R}$, let

$$r_i(x,a,b) = \begin{cases} (1-\omega)K\left(\frac{X_i - x}{h}\right), & Y_i \le a + b(X_i - x), \\ \omega K\left(\frac{X_i - x}{h}\right), & Y_i > a + b(X_i - x), \end{cases} \tag{2.7}$$

and

$$S_k(x,a,b) = \sum_{i=1}^{n} (X_i - x)^k r_i(x,a,b), \quad k = 0,1,2;$$

$$T_k(x,a,b) = \sum_{i=1}^{n} Y_i (X_i - x)^k r_i(x,a,b), \quad k = 0,1.$$

Define

$$\begin{cases} \tau(x,a,b) = \frac{T_0(x,a,b)S_2(x,a,b) - T_1(x,a,b)S_1(x,a,b)}{S_0(x,a,b)S_2(x,a,b) - [S_1(x,a,b)]^2}, \\ \\ \lambda(x,a,b) = \frac{T_1(x,a,b)S_0(x,a,b) - T_0(x,a,b)S_1(x,a,b)}{S_0(x,a,b)S_2(x,a,b) - [S_1(x,a,b)]^2}. \end{cases} \tag{2.8}$$

It is easy to check from (2.6) that $(\hat{\tau}_\omega(x), \hat{\lambda}_\omega(x))$ are the stationary values of $(\tau(x,a,b), \lambda(x,a,b))$, i.e.

$$\begin{cases} \hat{\tau}_\omega(x) = \tau(x, \hat{\tau}_\omega(x), \hat{\lambda}_\omega(x)), \\ \hat{\lambda}_\omega(x) = \lambda(x, \hat{\tau}_\omega(x), \hat{\lambda}_\omega(x)). \end{cases} \tag{2.9}$$

The above algorithm can be explained as follows. Any trial value of (a,b) determines a straight line in the (X,Y) plane, say $\mathcal{L}(a,b) \equiv \{Y = a + bX;\ X \in \mathbf{R}\}$. This straight line determines weights on the data points as indicated in (2.7): assign weight ω if (X_i, Y_i) is above $\mathcal{L}(a,b)$, and weight $(1-\omega)$ otherwise. These weights produce a solution vector $(a', b') = (\tau(x,a,b), \lambda(x,a,b))$ by (2.8), and thus a new straight line $\mathcal{L}(a', b')$, which leads to new weights upon iteration. The final solution is $(\hat{\tau}_\omega(x), \hat{\lambda}_\omega(x)) = (a,b)$ at which $\mathcal{L}(a,b) = \mathcal{L}(a',b')$. We recommend the ordinary least squares estimate $(a,b) = (\hat{\tau}_{1/2}(x), \hat{\lambda}_{1/2}(x))$ as the initial value. Simulation shows that the convergence rate of this iterative algorithm for solving (2.9) is very fast. (See also Section 2.3 below.)

2.2 Percentiles

Using the same construction leading to $\hat{\tau}_\omega(x)$, we can formally construct an estimator of the conditional percentile $\xi_\alpha(x)$ by using the function $R_\alpha(.)$, which measures the error by absolute residuals instead of squared residuals. Fan, Hu and Truong (1992) has studied this method with i.i.d. observations. However, since $R_\alpha(x)$ is not differentiable at $x = 0$, we do not have any iterative algorithm like (2.6) — (2.9). Therefore, either further smooth approximation to $R_\alpha(.)$ (recall the scoring method) or more complicated software development seems necessary in order to compute estimators numerically (cf. Bloomfield and Steiger 1983). In what follows, we are going to explore the possibility of using the ALS approach to estimate $\xi_\alpha(x)$ for a special class of models.

Suppose that (X,Y) has the relation

$$Y = \mu(X) + \sigma(X)\epsilon, \tag{2.10}$$

where $\mu(.),\ \sigma(.) > 0$ are continuous functions on $\mathbf{R}$, ϵ is a random variable with zero mean and finite non-zero variance, and ϵ and X are independent. Let $\xi_\alpha^{(0)}$ denote the α-percentile of ϵ, i.e. $P\left\{\epsilon \le \xi_\alpha^{(0)}\right\} = \alpha$. It is easy to see that

$$\xi_\alpha(x) = \mu(x) + \sigma(x)\xi_\alpha^{(0)}. \tag{2.11}$$

Let $p(.)$ denote the marginal density function of X. For any $\alpha \in (0,1)$ and $x \in \{p(x) > 0\}$, define $\omega = \omega(\alpha, x) \in (0,1)$ for which

$$\tau_{\omega(\alpha,x)}(x) = \xi_\alpha(x). \tag{2.12}$$

From the definitions of $\tau_\omega(.)$ and $\xi_\alpha(.)$, such an $\omega(\alpha, x)$ exists.

Proposition 1. For model (2.10), $\omega(\alpha, x)$ is independent of x. Specifically, it can be expressed as

$$\omega(\alpha, x) \equiv \omega(\alpha) = \frac{\alpha\xi_\alpha^{(0)} - E\{\epsilon\, I_{\{\epsilon \le \xi_\alpha^{(0)}\}}\}}{2E\{\epsilon\, I_{\{\epsilon > \xi_\alpha^{(0)}\}}\} - (1-2\alpha)\xi_\alpha^{(0)}}. \tag{2.13}$$

Further, $\omega(\alpha)$ is monotonically increasing, and invariant with respect to scaling transformations on ϵ.

Proof. It follows from (2.4) and (2.12) that for any $\alpha \in (0,1)$ and $x \in \{p(x) > 0\}$,

$$\omega(\alpha, x) = \frac{E\{\, |Y - \xi_\alpha(x)|\, I_{\{Y \le \xi_\alpha(x)\}} \mid X = x\,\}}{E\{\, |Y - \xi_\alpha(x)| \mid X = x\,\}}.$$

Substituting (2.10) and (2.11) in the above expression, we have the relation (2.13). Consequently,

$$\frac{\omega(\alpha)}{1-\omega(\alpha)} = \frac{E\{\, \xi_\alpha(x) - Y\, I_{\{Y \le \xi_\alpha(x)\}} \mid X = x\,\}}{E\{\, Y - \xi_\alpha(x)\, I_{\{Y > \xi_\alpha(x)\}} \mid X = x\,\}}\,.$$

Since $\xi_\alpha(x)$ is monotonically increasing as α increases, $\omega(\alpha)/(1-\omega(\alpha))$, therefore also $\omega(\alpha)$, is a monotonically increasing function. The invariance is obvious. The proof is completed.

The above proposition shows that for model (2.10), any percentile $\xi_\alpha(x)$ is an expectile $\tau_\omega(x)$ with $\omega = \omega(\alpha)$ as given in (2.13). Further, if the distribution of ϵ is known, while the distribution of the error term (i.e. $\sigma(X)\epsilon$) of the model may still be unknown, the function $\omega = \omega(\alpha)$ is completely determined. For example, Fig.1 presents the curves of $\omega = \omega(\alpha)$ for normal, uniform, and symmetric exponential distributions. In fact, in the case of uniform distribution $U[-b, b]$, $\omega(\alpha) = \alpha^2/(2\alpha^2 - 2\alpha + 1)$. For exponential distribution with density $0.5a\exp\{-a|x|\}\,(a > 0)$, $\omega(\alpha)$ equals $\alpha/\{2\alpha - \log(2\alpha)\}$ for $\alpha \in (0, 1/2]$, and $\{1 - \alpha - \log(2-2\alpha)\}/\{2 - 2\alpha - \log(2-2\alpha)\}$ for $\alpha \in (1/2, 1]$, which is nearly a linear function (see Fig.1).

In most practical situations, the distribution of ϵ is unknown. The following estimator is constructed in which we adapt Newey and Powell's idea (1987) to estimate the function $\omega = \omega(\alpha)$ (also cf. Efron 1991).

Suppose the data $(X_i, Y_i), i = 1, \ldots, n$, satisfy the model

$$Y_i = \mu(X_i) + \sigma(X_i)\epsilon_i, \tag{2.14}$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. with zero mean, and for $i \geq 1, \epsilon_i$ is independent of $\{X_1, \ldots, X_i; Y_1, \ldots, Y_{i-1}\}$. Let $\eta_\alpha(x) = d\xi_\alpha(x)/dx$. We define the estimators $\hat{\xi}_\alpha(x) = \hat{a}$, and $\hat{\eta}_\alpha(x) = \hat{b}$, where $(\hat{a}, \hat{b})$ minimizes

$$\sum_{i=1}^{n} Q_{\hat{\omega}(\alpha)}\{Y_i - a - b(X_i - x)\} K\left(\frac{X_i - x}{h}\right), \tag{2.15}$$

where $Q_\omega(.)$ is defined as in (2.2), and $\hat{\omega}(\alpha) \in (0, 1)$ is determined in such a way that the proportion of the sample $\{(X_i, Y_i), 1 \leq i \leq n\}$ lying below the regression curve $\{y = \hat{a}(x) : x \in \mathbf{R}\}$ is $100\alpha\%$.

Notice that there are two distinct parts in the above definition: (i) for a given value of ω, the ALS method determines the regression curve; (ii) we choose the regression curve by varying the values of ω between 0 and 1 such that the proportion of data points lying below the curve is α. Note that when ω varies from 0 to 1, the proportion of the sample lying below the regression curve varies from 0% to 100%. Therefore, for given $\alpha \in (0, 1)$, such a $\hat{\omega}(\alpha)$ does exists.

Perhaps it might be worried that the above ALS approach is estimating something other than $\xi_\alpha(x)$, since we use the complete samples to determine $\hat{\omega}(\alpha)$ in order to estimate $\xi_\alpha(x)$ for a specified x. Under the model (2.14), Theorem 2 below shows that $\hat{\xi}_\alpha(x)$ is in fact consistent.

2.3 Examples

In order to get a rough idea how to use the ALS estimators to construct predictive intervals and so on, we are going to study the following nonlinear time series model

$$Z_t = 3.76\ Z_{t-1} - 0.235\ Z_{t-1}^2 + e_t, \quad t \geq 1, \tag{2.16}$$

where $e_1, e_2, \ldots,$ are i.i.d. with the same distribution as $(\eta_1 + \ldots + \eta_{48})$, and $\eta_1, \ldots, \eta_{48}$ are independent random variables with the same distribution $U[-0.075, 0.075]$. According to the central limit theorem, we can treat e_i as being nearly a normal random variable with mean 0 and variance 0.3^2. However, it has a bounded support $[-3.6, 3.6]$. Note that bounded support of e_t is necessary for the stationarity (see Chan and Tong 1994). We generate 1000 samples from model (2.16). We estimate conditional percentiles or expectiles in the following three cases: (i) $X_t = Z_t$, $Y_t = Z_{t+1}$; (ii) $X_t = Z_t$, $Y_t = Z_{t+2}$; and (iii) $X_t = Z_t$, $Y_t = Z_{t+3}$. We use Gaussian kernel in our estimation. All the bandwidths are adjusted subjectively based on the

cross-validation selection for the cases of 50% expectiles. The case (i) is a special case of model (2.14); we can use the ALS method to estimate $\xi_\alpha(x)$ directly. The estimated curves $\xi_\alpha(.)$ for five different values of α are plotted in Fig.2(i). To produce each of these curves, a C-program ran for 3.2 — 7.8 seconds on a SUN4 SPARC 2 workstation. The maximum number of iterations needed was 5. Fig.2(ii) shows the estimates of the same curves by the ALAD method. We can see that in this case, both methods give almost the same results. To effect the ALAD method, the Downhill Simplex method was used (cf. Press et.al. 1989, §10.4). The CPU time needed for each ALAD estimator was around 40 times of that for an ALS estimator. In cases (ii) and (iii), the condition (2.14) no longer holds. The estimated curves for $\tau_\omega(x)$ with five different values of ω are reported in Fig.3 for case (ii), and Fig.4 for case (iii). The maximum number of iteration used was 7 in these cases. Comparing all plots together, we can see that for case (i), the width of the predictive interval is almost uniform over the different values of x (see Fig.2). However, this is not the case in Fig.3 or Fig.4. For example in Fig.4, the three-step ahead prediction is at its worst when $x = 8$, and at its best when x is around 5 or 11. This observation also shows the potential of the conditional expectiles in testing conditional heteroscedasticity.

More examples on interval prediction can be found in Yao and Tong (1994a). Further research on testing the conditional heteroscedasticity and conditional symmetry will be reported elsewhere.

3 Asymptotic Properties

To discuss the asymptotic properties of the ALS estimators $\hat{\tau}_\omega(x)$ and $\hat{\xi}_\alpha(x)$, we need the following assumptions. We denote by $g(.|x)$ the conditional density function of Y given $X = x$, and by $p(.)$ the marginal density function of X. We use c to denote a generic constant which may be different at different places.

(A1) Let $\psi_i(x) = \int y^i g(y|x)dy$ for $i = 1, 2$. The marginal density function p, and ψ_1, ψ_2 have continuous derivatives.

(A2) The joint density of the distinct elements of (X_1, Y_1, X_k, Y_k) $(k > 0)$ is bounded by a constant independent of k.

(A3) The strictly stationary process $\{(X_i, Y_i), i \geq 1\}$ is ρ-mixing, i.e.

$$\rho_j \equiv \sup_{i \geq 1}\{\sup_{U \in \mathcal{F}_1^i, V \in \mathcal{F}_{i+j}^\infty} \text{Corr}(U, V)\} \to 0,$$

where $\mathcal{F}_i^j$ is the σ-field generated by $\{(X_k, Y_k) : k = i, \ldots, j\}(j \geq i)$. Further, we assume that $\sum_{k=1}^{\infty} \rho_k \leq \infty$

(A4) $K(.)$ is a bounded symmetric density function with a bounded support in $\mathbf{R}$. Further $\int xK(x)dx = 0$, $\int x^2K(x)dx = \sigma_0^2 > 0$, and $|K(x) - K(y)| \leq c|x - y|$ for any $x, y \in \mathbf{R}$.

(A5) The bandwidth $h \to 0$, $nh^3 \to \infty$, and $(\log n)/(nh) \to 0$, as $n \to \infty$.

The condition of bounded support of kernel function is imposed for the brevity of proofs, which can be removed at the expense of a longer proof. In particular, Gaussian kernel is allowed. The assumption that the process is ρ-mixing is also for the technical convenience, which is not the weakest possible. An autoregressive process satisfying some mild conditions is ρ-mixing (cf. Theorem 3.4.10 of Györfi *et al* 1989). More detailed discussion on different mixing conditions can also be found in Bradley (1986).

Theorem 1 *Assume that conditions (A1) — (A5) hold. Then for* $x \in \{p(x) > 0\}$,

$$\sqrt{nh}\{\hat{\tau}_\omega(x) - \tau_\omega(x) - h^2\mu_\tau\} \xrightarrow{d} N(0, \sigma_\tau^2), \tag{3.1}$$

$$\sqrt{n}h^{3/2}\{\hat{\lambda}_\omega(x) - \lambda_\omega(x) - h\mu_\lambda\} \xrightarrow{d} N(0, \sigma_\lambda^2). \tag{3.2}$$

Further, $\hat{\tau}_\omega(x)$ *and* $\hat{\lambda}_\omega(x)$ *are asymptotically independent in the sense that the random variables on the RHS of (3.1) and (3.2) are jointly asymptotic normal with zero covariance. Here,*

$$\mu_\tau \equiv \mu_\tau(\omega, x) = \frac{1}{2}\dot{\lambda}_\omega(x)\sigma_0^2 \ + \ \circ(1), \qquad \mu_\lambda \equiv \mu_\lambda(\omega, x) = \frac{1}{2\sigma_0^2}\dot{\lambda}_\omega(x)\int u^3K(u)du \ + \ \circ(1),$$

$$\sigma_\tau^2 \equiv \sigma_\tau^2(\omega, x) = \frac{\int K^2(u)du \int \{\dot{Q}_\omega(y - \tau_\omega(x))\}^2 g(y|x)dy}{p(x)\{\gamma(\omega, x)\}^2},$$

$$\sigma_\lambda^2 \equiv \sigma_\lambda^2(\omega, x) = \frac{\int u^2K^2(u)du \int \{\dot{Q}_\omega(y - \tau_\omega(x))\}^2 g(y|x)dy}{p(x)\sigma_0^2\{\gamma(\omega, x)\}^2},$$

and

$$\gamma(\omega, x) = 2\omega P\{Y_1 \leq \tau_\omega(x)|X_1 = x\} + 2(1 - \omega)P\{Y_1 > \tau_\omega(x)|X_1 = x\}.$$

In the above expressions, σ_0^2 *is as given in (A4), and* $\dot{\varphi}(.)$ *denotes the derivative of the function* $\varphi(.)$.

Theorem 2. *Suppose that conditions (A1)-(A5) hold. We assume that* $\lambda_\omega(x) = \dot{\tau}_\omega(x)$ *is continuous in* $(\omega, x) \in (0, 1) \times \mathbf{R}$.

(i) For any compact subset $\mathbf{A} \times \mathbf{B} \subset (0,1) \times \{p(x) > 0\}$, $\hat{\tau}_\omega(x) \xrightarrow{P} \tau_\omega(x)$ *uniformly for* $\omega \in \mathbf{A}$ *and* $x \in \mathbf{B}$.

(ii) If equation (2.14) holds, $\hat{\xi}_\alpha(x) \xrightarrow{P} \xi_\alpha(x)$ *for* $\alpha \in (0,1)$ *and* $x \in \{p(x) > 0\}$.

Theorem 1 gives the asymptotic normality of the ALS estimator for the conditional expectile and its derivative. As shown in (3.1), the 'asymptotic bias' of $\hat{\tau}_\omega(x)$ is $\frac{1}{2}d^2\tau_\omega(x)/dx^2$, which is due to the local approximation of the underlying curve by a linear function (cf (2.5)). To use the locally quadratic fit will improve the estimate of $\lambda_\omega(x)$ considerably (cf. Fan et. al 1993). However, it creates further complications in practical implementation. Theorem 2 presents the uniform (weak) consistency of $\hat{\tau}_\omega(x)$ for both x and ω. On augmenting this result with the assumption of model (2.14), the weak consistency of the estimator $\hat{\xi}_\alpha(x)$ is proved. The reason that we need the uniform consistency of $\hat{\tau}_\omega(x)$ is that in order to estimate $\xi_\alpha(x)$ for a fixed x, we determine the function $\omega = \omega(\alpha)$ by using the complete sample scattering in the state space.

We are not going to discuss in any detail how to choose the bandwidth h, except to mention that, from (3.1), we can choose h such that $\int \{\mu_\tau^2(\omega, x)h^4 + \tau_\omega^2(x)/(nh)\}\ k(x)dx$ attains its minimum, where $k(x)$ is a weight function. However, practical implementation of this 'asymptotically optimal' bandwidth involves estimating some unknown functions. A lot of work has been done on choosing the bandwidth either subjectively or objectively. From the practical point of view, we believe that a good estimator should not be so 'approach dependent', i.e. the conclusions drawn from different approaches should not be substantially different if there is sufficient information in the data.

4 Proofs

The main idea of the proof of theorem 1 is to approximate the objective function in (2.5) by a quadratic function whose minimizer (vector) is asymptotically normal, and then to show that $(\hat{\tau}_\omega(x), \hat{\lambda}_\omega(x))$ lies close enough to the minimizer to share the latter's asymptotic behaviour. The convexity lemma plays a key role in the above approximation (cf. Pollard 1991). This idea has been used by Fan, Hu and Truong (1992) to study nonparametric regression with i.i.d. observations. By conditioning on the covariates, Fan, Hu and Truong (1992) has been able to exploit the convenience provided and obtain the conditional asymptotic normality of the regression estimators. Unfortunately, the idea of conditioning on the covariates does not apply to dependent

data. We have to prove the asymptotic normality directly. The proof of Theorem 2 is based on an extension of the convexity lemma.

In the sequel, we always assume that conditions (A1) — (A5) hold. Let $K_i = K\left(\frac{X_i - x}{h}\right)$, $Z_i = \left(1, \frac{X_i - x}{h}\right)'$, $Y_i^* = Y_i - \tau_\omega(x) - \lambda_\omega(x)(X_i - x)$, and $\hat{\theta} = \sqrt{nh}(\hat{\tau}_\omega(x) - \tau_\omega(x), h(\hat{\lambda}_\omega(x) - \lambda_\omega(x)))'$. For $\theta \in \mathbf{R}^2$, define

$$G_n(\theta) \equiv G_n(\theta; \omega, x) = \sum_{i=1}^{n} \{Q_\omega(Y_i^* - \theta' Z_i/\sqrt{nh}) - Q_\omega(Y_i^*)\} K_i, \tag{4.1}$$

and

$$R_n(\theta) \equiv R_n(\theta; \omega, x) = G_n(\theta) - \theta' D(\omega, x)\theta - \frac{1}{\sqrt{nh}} \theta' \sum_{i=1}^{n} Z_i \dot{Q}_\omega(Y_i^*) K_i, \tag{4.2}$$

where $D \equiv D(\omega, x) = \frac{1}{2} p(x) \gamma(\omega, x) \begin{pmatrix} 1 & 0 \\ 0 & \sigma_\circ^2 \end{pmatrix}$. We proceed to the proofs in a sequence of lemmas.

4.1 Proof of Theorem 1

Lemma 1. For $\theta \in \mathbf{R}^2$,

$$E[\{Q_\omega(Y_1^* - \theta' Z_1/\sqrt{nh}) - Q_\omega(Y_1^*)\} K_1] = \frac{1}{n} \theta' D(\omega, x)\theta - \frac{1}{\sqrt{nh}} \theta' E\{Z_1 \dot{Q}_\omega(Y_1^*) K_1\} + \circ\left(\frac{1}{n}\right). \tag{4.3}$$

Further, if $\lambda_\omega(x)$ is continuous in both ω and x, the convergence in (4.3) is uniform for (ω, x) in compact subsets of $(0,1) \times \mathbf{R}$.

Proof. Let $\theta = (\theta_1, \theta_2)'$, and $d \equiv d(x,u) = \tau_\omega(x) + \lambda_\omega(x) uh$. Then the LHS of (4.3) is equal to

$$h \int \left[Q_\omega \left\{ y - d(x,u) - \frac{\theta_1 + \theta_2 u}{\sqrt{nh}} \right\} - Q_\omega\{y - d(x,u)\} \right] K(u) g(y|x+hu) p(x+hu) dy du$$

$$= \frac{1}{n} \int (\theta_1 + \theta_2 u)^2 \left[\omega I_{\{y \le d + \frac{\theta_1 + \theta_2 u}{\sqrt{nh}}\}} + (1-\omega) I_{\{y > d + \frac{\theta_1 + \theta_2 u}{\sqrt{nh}}\}} \right] K(u) g(y|x+hu) p(x+hu) dy du$$

$$- \frac{2h}{\sqrt{nh}} \int (y - d)(\theta_1 + \theta_2 u) \left[\omega I_{\{y \le d + \frac{\theta_1 + \theta_2 u}{\sqrt{nh}}\}} + (1-\omega) I_{\{y > d + \frac{\theta_1 + \theta_2 u}{\sqrt{nh}}\}} \right]$$
$$\times K(u) g(y|x+hu) p(x+hu) dy du$$

$$+ h \int (y - d)^2 \left[\omega I_{\{y \le d + \frac{\theta_1 + \theta_2 u}{\sqrt{nh}}\}} + (1-\omega) I_{\{y > d + \frac{\theta_1 + \theta_2 u}{\sqrt{nh}}\}} \right.$$
$$\left. - \omega I_{\{y \le d\}} - (1-\omega) I_{\{y > d\}} \right] K(u) g(y|x+hu) p(x+hu) dy du.$$

Assumption (A1) implies that the first term on the RHS of the above expression is equal to $n^{-1}\theta' D(\omega, x)\theta + \circ\left(\frac{1}{n}\right)$, and the second term is equal to $-(nh)^{-\frac{1}{2}}\theta' E\{Z_1 \dot{Q}_\omega(Y_1^*)K_1\} + \circ\left(\frac{1}{n}\right)$. Similarly, it is easy to see that the third term is $\circ\left(\frac{1}{n}\right)$. The uniform convergence follows from the continuity assumption in (A1) and the fact that K has a bounded support.

Lemma 2. For any $x, y \in \mathbf{R}, \omega \in [0,1]$,

$$|Q_\omega(x+y) - Q_\omega(x) - \dot{Q}_\omega(x)y| \leq 4y^2,$$

$$|\dot{Q}_\omega(x+y) - \dot{Q}_\omega(x) - \ddot{Q}_\omega(x)y| \leq 4|y|,$$

where $\ddot{Q}_\omega(x) = d^2\{Q_\omega(x)\}/dx^2$ for $x \neq 0$, and $\ddot{Q}_\omega(0) = 0$.

The proof follows from some simple algebra operation, which is omitted here.

Lemma 3. For any $\theta \in \mathbf{R}^2, \omega \in (0,1)$, and $x \in \mathbf{R}$, $R_n(\theta) \equiv R_n(\theta; \omega, x) \xrightarrow{P} 0$. Further, if $\lambda_\omega(x)$ is continuous in ω and x, the convergence is uniform for (ω, x) in compact subsets of $(0,1) \times \mathbf{R}$.

Proof. By (4.1) and (4.2), we have the following expression

$$R_n(\theta) = \sum_{i=1}^n T_i - \theta' D(\omega, x)\theta, \tag{4.4}$$

where

$$T_i \equiv T_i(\theta; \omega, x) = \{Q_\omega(Y_i^* - \theta' Z_i/\sqrt{nh}) - Q_\omega(Y_i^*) - \dot{Q}_\omega(Y_i^*)\theta' Z_i/\sqrt{nh}\}K_i. \tag{4.5}$$

Hence

$$\mathrm{Var}\{R_n(\theta)\} \leq nE(T_i^2) + 2\sum_{i<j}\{E(T_iT_j) - (ET_1)^2\}. \tag{4.6}$$

It follows from Lemma 2 and (A4) that

$$ET_1^2 \leq 16E(\theta' Z_1 K_1/\sqrt{nh})^4 = O\left(\frac{1}{n^2 h}\right).$$

By the Cauchy-Schwarz inequality, we can ignore all the summands with $j - i \leq \log n$ in the second term of the RHS of (4.6), since $(\log n)/(nh) \to 0$. It follows from (A3) and (A5) that

$$\sum_{j-i>\log n}\{E(T_iT_j) - (ET_1)^2\} \leq \frac{c}{n^2 h}\sum_{k=\log n}^{n-1}(n-k)\rho_k \leq \frac{c}{nh}\sum_{k=\log n}^{n}\rho_k \to 0.$$

Hence, $\mathrm{Var}\{R_n(\theta)\} = O(\delta_n)$, where $\delta_n = \max\{(\log n)/(nh), \frac{1}{nh}\sum_{k=\log n}^n \rho_k\} \to 0$.

It follows from Lemma 1 that for any $\theta \in \mathbf{R}^2$, $\omega \in (0,1)$ and $x \in \mathbf{R}$, $E\{R_n(\theta;\omega,x)\} \to 0$. Therefore, for any constant $\epsilon > 0$, $|E\{R_n(\theta;\omega,x)\}| \le \epsilon/2$ for all sufficiently large n. Consequently,

$$P\{|R_n(\theta;\omega,x)| > \epsilon\} \le P\{|R_n(\theta;\omega,x) - E\{R_n(\theta;\omega,x)\}| > \epsilon/2\}$$

$$\le \frac{4}{\epsilon^2}\mathrm{Var}\{(R_n(\theta;\omega,x)\} = O(\delta_n). \tag{4.7}$$

To prove the uniform convergence, consider $[t_1,t_2] \subset (0,1)$, and $M > 0$. Let m be a large integer. Define

$$\omega_k = t_1 + \frac{k}{m}(t_2 - t_1), \quad k = 1,\ldots,m;$$

$$x_j = -M + j\frac{2M}{m}, \quad j = 1,\ldots,m.$$

For any $\omega \in [t_1,t_2]$ and $|x| \le M$, there exist $1 \le k \le m$ and $1 \le j \le m$, such that $|\omega - \omega_k| \le \frac{1}{m}$, and $|x - x_j| \le \frac{2M}{m}$. It follows from (4.4) and Lemma 1 that for any $\epsilon > 0$, we can choose $m > 0$ large enough such that for all sufficiently large n,

$$\sup_{t_1\le\omega\le t_2,|x|\le M} R_n(\theta;\omega,x) - \max_{1\le k,j\le m} R_n(\theta;\omega_k,x_j)$$

$$\le \sup_{t_1\le\omega\le t_2,|x|\le M} \sum_{i=1}^n T_i(\theta;\omega,x) - \max_{1\le k,j\le m} \sum_{i=1}^n T_i(\theta;\omega_k,x_j) + \epsilon/2, \tag{4.8}$$

when T_i is defined as in (4.5). Notice that $\lambda_\omega(x)$ is continuous in both x and ω, and $K(.)$ has a bounded support. Assumptions (A1), (A2) and (A4) imply that the difference of the first two terms on the RHS of (4.8) is less than a random variable $|\zeta_n|$ plus $\epsilon/2$, and $\zeta_n \xrightarrow{P} 0$ as $n \to \infty$. Thus

$$P\left\{\sup_{t_1\le\omega\le t_2,|x|\le M} |R_n(\theta;\omega,x)| > 3\epsilon\right\} \le P\left\{\max_{1\le k,j\le m} |R_n(\theta;\omega_k,x_j)| + |\zeta_n| > 2\epsilon\right\}$$

$$\le P\left\{\max_{1\le k,j\le m} |R_n(\theta;\omega_k,x_j)| > \epsilon,\ |\zeta_n| \le \epsilon\right\} + \circ(1) \le P\left\{\max_{1\le k,j\le m} |R_n(\theta;\omega_k,x_j)| > \epsilon\right\} + \circ(1) \to 0,$$

the limit follows from (4.7).

Lemma 4. As $n \to \infty$,

$$E\{\theta' Z_1 \dot{Q}_\omega(Y_1^*)K_1\} = \frac{h^3}{2}\dot{\lambda}_\omega(x)p(x)\gamma(\omega,x)(\theta_1\sigma_0^2 + \theta_2\int u^3K(u)du)\{1+\circ(1)\}, \tag{4.9}$$

$$E\{\theta' Z_1 \dot{Q}_\omega(Y_1^*)K_1\}^2 = h(\theta_1\sigma_1^2 + \theta_2\sigma_2^2)\{1+\circ(1)\}, \tag{4.10}$$

where σ_0^2 is given as in (A4), and

$$\sigma_1^2 = p(x) \int K^2(u)du \int \{\dot{Q}_\omega(y - \tau_\omega(x))\}^2 g(y|x)dy,$$

$$\sigma_2^2 = p(x) \int u^2 K^2(u)du \int \{\dot{Q}_\omega(y - \tau_\omega(x))\}^2 g(y|x)dy.$$

Proof. We prove only (4.9). (4.10) can be proved in the similar, and simpler way.

Since $K(.)$ has a bounded support, we need only to consider X_1 such that $|X_1 - x| \leq hM$ for some constant $M > 0$. let $v = \tau_\omega(X_1) - \tau_\omega(x) - \lambda_\omega(x)(X_1 - x)$. It is easy to see that $v = \frac{1}{2}\dot{\lambda}_\omega(x)(X_1 - x)^2 + O(h^3) = O(h^2)$. Noticing that $|\dot{Q}_\omega(x)| \leq 2|x|$ and $Y_1^* = Y_1 - \tau_\omega(X_1) + v$, we have that

$$E\{\ \dot{Q}_\omega(Y_1^*)K_1\ I_{\{|Y_1 - \tau_\omega(X_1)| \leq |v|\}}\ \} \ \leq\ ch^2 E\{\ K_1\ I_{\{|Y_1 - \tau_\omega(X_1)| \leq |v|\}}\} \ =\ o(h^3).$$

Note that $E\{\dot{Q}_\omega(Y_1 - \tau_\omega(X_1))|X_1\} = 0$. Therefore, the LHS of (4.9) can be expressed as

$$\begin{aligned} & E\{\ \theta' Z_1 \dot{Q}_\omega(Y_1^*)K_1\ I_{\{|Y_i - \tau_\omega(X_i)| > |v|\}}\} + o(h^3) \\ = \ & E\left[\theta' Z_1 \ddot{Q}_\omega(Y_1 - \tau_\omega(X_1))K_1 \left\{\frac{1}{2}\dot{\lambda}_\omega(x)(X_1 - x)^2 + O(h^3)\right\}\right] + o(h^3). \end{aligned}$$

It can be proved that the first term on the RHS of the above equality is asymptotically equivalent to the RHS of (4.9).

Lemma 5. As $n \to \infty$,

$$\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \theta'[Z_i \dot{Q}_\omega(Y_i^*)K_i - E\{Z_i \dot{Q}_\omega(Y_i^*)K_i\}] \xrightarrow{d} N(0, \theta_1\sigma_1^2 + \theta_2\sigma_2^2), \tag{4.11}$$

where σ_1^2 and σ_2^2 are given as in Lemma 4.

Proof. Let $U_i = h^{-\frac{1}{2}}\theta' Z_i \dot{Q}_\omega(Y_i^*)K_i$, and $S_n = \sum_{i=1}^n U_i - EU_i$. It follows from (4.10) and (4.9) that

$$\sum_{i=1}^{n} \{EU_i^2 - (EU_i)^2\} = n(\theta_1\sigma_1^2 + \theta_2\sigma_2^2)(1 + o(1)).$$

Note that

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n} \{EU_i^2 - (EU_i)^2\} + 2 \sum_{0 < j-i \leq h^{-1/2}} \mathrm{Cov}(U_i, U_j) + 2 \sum_{j-i > h^{-1/2}} \mathrm{Cov}(U_i, U_j).$$

By assumption (A2), $\mathrm{Cov}(U_i, U_j) = O(h)$ for $i \neq j$. Therefore, the second term on the RHS of the above equality is of the order $nh^{-1/2}h = o(n)$. It follows from (A3) that the third term on the RHS of the above equality is less than

$$2\mathrm{Var}(U_1) \sum_{k=h^{-1/2}+1}^{n-1} (n-k)\rho_k \leq cn \sum_{k=h^{-1/2}+1}^{n-1} \rho_k = o(n).$$

Hence, $\mathrm{Var}(S_n) = n(\theta_1\sigma_1^2 + \theta_2\sigma_2^2)(1 + o(1))$. Consequently, (4.11) follows from Theorem 2.4 of Peligrad (1986).

The Proof of Theorem 1. Note that $G_n(\theta) - \frac{1}{\sqrt{nh}}\theta' \sum_{i=1}^{n} Z_i \dot{Q}_\omega(Y_i^*)K_i$ is convex as a function of θ (cf. (4.1)). Lemma 3 shows that it converges to the convex function $\theta' D(\omega, x)\theta$. By the convexity lemma (cf. Pollard 1991), $\sup_{\theta\in\mathbf{K}} |R_n^*(\theta)| = \circ_p(1)$ for any compact subset $\mathbf{K} \subset \mathbf{R}^2$. Using the same arguments as in p.193 of Pollard (1991), we can show that the difference between the minimizer $\hat{\theta}_n$ of $G_n(\theta)$ and the minimizer of the quadratic function $\theta' D(\omega, x)\theta + \frac{1}{\sqrt{nh}}\theta' \sum_{i=1}^{n} Z_i \dot{Q}_\omega(Y_i^*)K_i$ converges to 0 in probability. Note that the minimizer of the function $\theta' D(\omega, x)\theta + \frac{1}{\sqrt{nh}}\theta' \sum_{i=1}^{n} Z_i \dot{Q}_\omega(Y_i^*)K_i$ can be explicitly expressed as

$$\sqrt{nh}\,\{\hat{\tau}_\omega(x) - \tau_\omega(x)\} = \frac{1}{\sqrt{nh}p(x)\gamma(\omega, x)} \sum_{i=1}^{n} \dot{Q}_\omega(Y_i^*)K_i + \circ_p(1),$$

$$\sqrt{n}h^{\frac{3}{2}} \left\{\hat{\lambda}_\omega(x) - \lambda_\omega(x)\right\} = \frac{h^{-1}}{\sqrt{nh}p(x)\sigma_0^2\gamma(\omega, x)} \sum_{i=1}^{n} \dot{Q}_\omega(Y_i^*)(X_i - x)K_i + \circ_p(1).$$

The theorem follows from the above equations, Lemma 5, and (4.9) immediately.

4.2 Proof of Theorem 2

First, we prove Theorem 2 (i), which is based on the following trivial extension of the convexity lemma (cf. Pollard 1991).

Lemma 6. Let $\{F_n(\theta, \nu) : \theta \in \Theta, \nu \in \mathbf{B}\}$ be a sequence of random continuous functions, Θ be a convex open subset of $\mathbf{R}^d$, $\mathbf{B}$ be a compact subset of $\mathbf{R}^k$, and $F_n(., \nu)$ is a convex function on Θ for each fixed $\nu \in \mathbf{B}$. Suppose that $F(\theta, \nu)$ is a continuous real-valued function on $\Theta \times \mathbf{B}$, and $\sup_{\nu\in\mathbf{B}} |F_n(\theta, \nu) - F(\theta, \nu)| \xrightarrow{P} 0$ for each fixed $\theta \in \Theta$. Then for any compact subset $\mathbf{K} \subset \Theta$,

$$\sup_{\theta\in\mathbf{K}, \nu\in\mathbf{B}} |F_n(\theta, \nu) - F(\theta, \nu)| \xrightarrow{P} 0.$$

The proof of Lemma 6 can proceed in almost the same way as in Section 6 of Pollard (1991) (also see Proof of Theorem 2(i) below), which is omitted here.

Proof of Theorem 2(i). By Lemmas 3 and 6, the remainder $R_n(\theta; \omega, x)$ converges to 0 in probability uniformly for $\theta \in \mathbf{K}$, $\omega \in \mathbf{B}_1$, and $x \in \mathbf{B}_2$, where $\mathbf{K}, \mathbf{B}_1$ and $\mathbf{B}_2$ are the compact subsets of $\mathbf{R}^2, (0, 1)$, and $\{p(x) > 0\}$ respectively. We rewrite (4.2) as follows

$$G_n(\theta; \omega, x) = G_n^*(\theta; \omega, x) + R_n(\theta; \omega, x), \tag{4.12}$$

where

$$G_n^*(\theta;\omega,x) = \theta' D(\omega,x)\theta + \frac{1}{\sqrt{nh}}\theta' \sum_{i=1}^n Z_i \dot{Q}_\omega(Y_i^*)K_i. \tag{4.13}$$

Let

$$\theta_n^* \equiv \theta_n^*(\omega,x) = -\frac{2}{\sqrt{nh}} D^{-1}(\omega,x) \sum_{i=1}^n Z_i \dot{Q}_\omega(Y_i^*)K_i,$$

which is the minimizer of $G_n^*(\theta;\omega,x)$ over $\theta \in \mathbf{R}^2$. We have

$$G_n^*(\theta;\omega,x) = (\theta-\theta_n^*)' D(\omega,x)(\theta-\theta_n^*) + G_n^*(\theta_n^*;\omega,x).$$

Let $B_n(\delta;\omega,x)$ be the closed ball with the center $\theta_n^*(\omega,x)$ and the radius $\delta > 0$. By Lemmas 1 and 3, we can choose the compact set $\mathbf{K} \subset \mathbf{R}^2$ such that it contains all $B_n(\delta;\omega,x)$ for all $\omega \in B_2$ with probability arbitrarily close to 1, i.e.

$$P\left\{\sup_{\omega\in\mathbf{B}_1, x\in\mathbf{B}_2} ||\hat{\theta}_n - \theta_n^*|| > \delta\right\}$$

$$\leq P\left\{\sup_{\omega\in\mathbf{B}_1, x\in\mathbf{B}_2} ||\hat{\theta}_n - \theta_n^*|| > \delta, B_n(\delta;\omega,x) \subset \mathbf{K} \text{ for all } \omega\in\mathbf{B}_1, x\in\mathbf{B}_2\right\} + \epsilon \tag{4.14}$$

and thereby also implying that

$$\Delta_n \equiv \sup_{\omega\in\mathbf{B}_1, x\in\mathbf{B}_2, \theta\in\mathbf{K}} |R_n^*(\theta;\omega,x)| = \circ_p(1).$$

For any $\theta \notin B_n(\delta;\omega,x), \theta = \theta_n^*(\omega,x) + \beta\upsilon$ with $\beta > \delta$ and υ a unit vector. Define $\widetilde{\theta}=\widetilde{\theta}(\theta;\omega,x)$ as the boundary point of $B_n(\delta;\omega,x)$ that lies on the linear segment from $\theta_n^*(\omega,x)$ to θ, i.e. $\widetilde{\theta}= \theta_n^*(\omega,x) + \delta\upsilon$. Convexity of G_n and (4.12), (4.13) imply that

$$\frac{\delta}{\beta}\ G_n(\theta;\omega,x) + \left(1-\frac{\delta}{\beta}\right) G_n(\theta_n^*;\omega,x) \geq G_n(\widetilde{\theta};\omega,x)$$

$$\geq G_n^*(\widetilde{\theta};\omega,x) - \Delta_n \geq (\delta^*)' D(\omega,x)\delta^* + G_n(\theta_n^*;\omega,x) - 2\Delta_n,$$

where $\delta^* = (\delta,\delta)'$. Let $c_0 = \inf_{\omega\in\mathbf{B}_1, x\in\mathbf{B}_2} (\delta^*)' D(\omega,x)\delta^* > 0$, the above inequality entails that

$$\inf_{\omega\in\mathbf{B}_1, x\in\mathbf{B}_2} \left\{\inf_{||\theta-\theta_n^*||>\delta} G_n(\theta) - G_n(\theta_n^*)\right\} \geq \frac{\beta}{\delta}c_0 + \circ_p(1).$$

Hence the first term on the LHS of (4.14) tends to zero. Consequently, $\sup_{\omega\in\mathbf{B}_1, x\in\mathbf{B}_2} |\hat{\tau}_\omega(x) - \tau_\omega^*(\omega)| \xrightarrow{P} 0$, where

$$\tau_\omega^*(x) = \tau_\omega(x) + \frac{1}{nhp(x)\gamma(\omega,x)} \sum_{i=1}^n \dot{Q}_\omega(Y_i^*)K_i.$$

Similar to Lemma 3, we can prove that the second term on the RHS of the above expression tends to zero in probability uniformly for $\omega \in \mathbf{B}_1$ and $x \in \mathbf{B}_2$. The proof is completed.

Proof of Theorem 2(ii). From (2.15), we can see that $\hat{\xi}_\alpha(x) = \hat{\tau}_{\hat{\omega}(\alpha)}(x)$. Proposition 1 ensures that $\xi_\alpha(x) = \tau_{\omega(\alpha)}(x)$, where $\omega(\alpha)$ is a continuous function given as in (2.13). Since $\hat{\tau}_\omega(x)$ is continuous in $\omega \in (0,1)$, and converges to $\tau_\omega(x)$, we need only to prove that for $\alpha^* \equiv \alpha^*(\alpha, X_1, Y_1, \ldots, X_n, Y_n)$ such that $\omega(\alpha^*) = \hat{\omega}(\alpha), \alpha^* \xrightarrow{P} \alpha$.

From the definition of $\hat{\omega}(\alpha)$ (cf. (2.15) and the statements therewith), the proportion of the sample $\{(X_i, Y_i), 1 \leq iln\}$ lying below the regression curve $\{y = \hat{\xi}_\alpha(x) : \ x \in \mathbf{R}\}$ is within $\alpha \pm \frac{1}{n}$. Therefore, we have that

$$|\alpha^* - \alpha| \leq |\alpha^* - \frac{1}{n}\sum_{i=1}^n I_{\{Y_i \leq \hat{\xi}_\alpha(X_i)\}}| + \frac{1}{n}$$

$$\leq \left|\alpha^* - \frac{1}{n}\sum_{i=1}^n I_{\{Y_i \leq \xi_{\alpha^*}(X_i)\}}\right| + \frac{1}{n}\left|\sum_{i=1}^n (I_{\{Y_i \leq \tau_{\omega(\alpha^*)}(X_i)\}} - I_{\{Y_i \leq \hat{\tau}_{\omega(\alpha^*)}(X_i)\}})\right| + \frac{1}{n}. \tag{4.15}$$

By (2.11), the first term on the RHS of the above expression is less than $\sup_{\alpha \in (0,1)} \left|\alpha - \frac{1}{n}\sum_{i=1}^n I_{\{\epsilon_i \leq \xi_\alpha^{(0)}\}}\right|$, which converges to zero in probability according to Glivenko-Cantelli's Theorem (cf. Theorem 8.2.2 of Chow and Teicher 1978, for example).

It is easy to see that for given $\alpha \in (0,1)$, there exist t_1, t_2 satisfying $0 < t_1 < \omega(\alpha) < t_2 < 1$, such that $P\{t_1 \leq \omega(\alpha^*) \leq t_2\} \to 1$ (cf. Propositions 1 and (2.15)). Therefore the second term on the RHS of (4.15) is less than

$$\sup_{\omega \in [t_1, t_2]} \frac{1}{n}\left|\sum_{i=1}^n (I_{\{Y_i \leq \tau_\omega(X_i)\}} - I_{\{Y_i \leq \hat{\tau}_\omega(X_i)\}})\right| + o_p(1) \equiv V_{n,1} + o_p(1).$$

Let us choose a compact $\mathbf{D}_2 \subset \mathbf{D}_1 \equiv \{p(x) > 0\}$ such that the probability of the event $\{X \in \mathbf{D}_1 - \mathbf{D}_2\}$ is less than an arbitrarily given constant $\epsilon > 0$. Then

$$V_{n,1} \leq \sup_{\omega \in [t_1, t_2]} \frac{1}{n}\left|\sum_{i=1}^n (I_{\{Y_i \leq \tau_\omega(X_i), X_i \in \mathbf{D}_2\}} - I_{\{Y_i \leq \hat{\tau}_\omega(X_i), X_i \in \mathbf{D}_2\}})\right| + \frac{2}{n}\sum_{i=1}^n I_{\{X_i \in \mathbf{D}_1 - \mathbf{D}_2\}}.$$

Theorem 2 (i) entails that the first term on the RHS of the above expression converges to 0 in probability. By the standard ergodic theorem, the second term has the limit $2\mathrm{P}(X_1 \in \mathbf{D}_1 - \mathbf{D}_2)$ which is smaller than 2ϵ. Thus $V_{n,1} \xrightarrow{P} 0$. Consequently, by (4.15), we have that $|\alpha^* - \alpha| \xrightarrow{P} 0$. The proof is completed.

Acknowledgement. The authors wish to thank two referees for their helpful comments which let to improvements in this paper.

References

Bassett, G. and Koenker, R. (1982). An empirical quantile function for linear models with i.i.d. errors. *J. Ameri. Statist. Assoc.,* **77**, 407-415.

Bloomfield, P. and Steiger, W.L. (1983). *Least Absolute Deviations.* Birkhäuser, Boston.

Bradley, R.C. (1986). Basic properties of strong mixing conditions. *Dependence in Probability and Statistics*, Ed. E. Eberlein and M.S. Taqqu. Birkhäuser, Boston, 165-192.

Chan, K.S. and Tong, H. (1994). A note on noisy chaos. To appear in *J. R. Statist. Soc. B*, **56**.

Chow, Y.S. and Teicher, H. (1978). *Probability Theory.* Springer, New York.

Chu, C.K. and Marron, J.S. (1991). Choosing a kernel regression estimator. *Statist. Science,* **6**, 404-436.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica.* **1**, 93-125.

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

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1993). Local polynomial fitting: a standard for nonparametric regression. Technical Report, University of North-Carolina.

Fan, J., Hu, T.C. and Truong, Y.K. (1992). Robust nonparametric function estimation. Technical Report 035-92, Math.Science Research Inst., Berkeley.

Győrfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989). *Non-parametric Curve Estimation from Time Series.* Springer-Verlag, Berlin.

Hogg, R.V. (1975). Estimates of percentile regression lines using salary data. *J. Ameri. Statist. Assoc.,* **70**, 56-59.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica,* **46**, 33-50.

Neway, W.K. and Powell, J.K. (1987). Asymmetric least squares estimation and testing. *Econometrica*, **55**, 819-847.

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*, Ed. E. Eberlein and M.S. Taqqu. Birkhäuser, Boston, 193-223.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, **7**, 186-198.

Press, W.H., Flannery, B.P., Tenkolsky, S.A., and Vetterling, W.T. (1989). *Numerical Recipes*. Cambridge Univ.Press, Cambridge.

Yao, Q. and Tong, H. (1994a). On prediction and chaos in stochastic systems. *Phil. Trans. R. Soc. Land. A*, **348**, 357-369.

Yao, Q. and Tong, H. (1994b). Quantifying the influence of initial values on nonlinear prediction. *J. R. Statist. Soc. B*, **56**, 701-725.

Figure Captions

Fig.1 The curves of $\omega = \omega(\alpha)$ given as in (2.13) for three kinds of distributions: Solid curve — normal distribution $N(0, \sigma^2)$; dotted curve — exponential distribution with density function $0.5a \exp\{-a|x|\}$; dot-dashed curve — uniform distribution $U[-b, b]$.

Fig.2(i) The ALS estimated conditional percentiles of Z_{t+1} given Z_t (with $h = 0.31$). Solid curve — $\alpha = 0.05(\omega = 0.01)$; longer dashed curve — $\alpha = 0.25(\omega = 0.2)$; shorter dashed curve — $\alpha = 0.5(\omega = 0.54)$; dotted curve — $\alpha = 0.75(\omega = 0.88)$; dot-dashed curve — $\alpha = 0.95(\omega = 0.99)$.

Fig.2(ii) The ALAD estimated conditional percentiles of Z_{t+1} given Z_t (with $h = 0.35$). Solid curve — $\alpha = 0.05$; longer dashed curve — $\alpha = 0.25$; shorter dashed curve — $\alpha = 0.5$; dotted curve — $\alpha = 0.75$; dot-dashed curve — $\alpha = 0.95$.

Fig.3 The ALS estimated conditional expectiles of Z_{t+2} given Z_t (with $h = 0.28$). Solid curve — $\omega = 0.01$; longer dashed curve — $\omega = 0.2$; shorter dashed curve — $\omega = 0.5$; dotted curve — $\omega = 0.9$; dot-dashed curve — $\omega = 0.99$.

Fig.4 The ALS estimated conditional expectiles of Z_{t+3} given Z_t (with $h = 0.25$). Solid curve — $\omega = 0.01$; longer dashed curve — $\omega = 0.18$; shorter dashed curve — $\omega = 0.5$; dotted curve — $\omega = 0.9$; dot-dashed curve — $\omega = 0.99$.

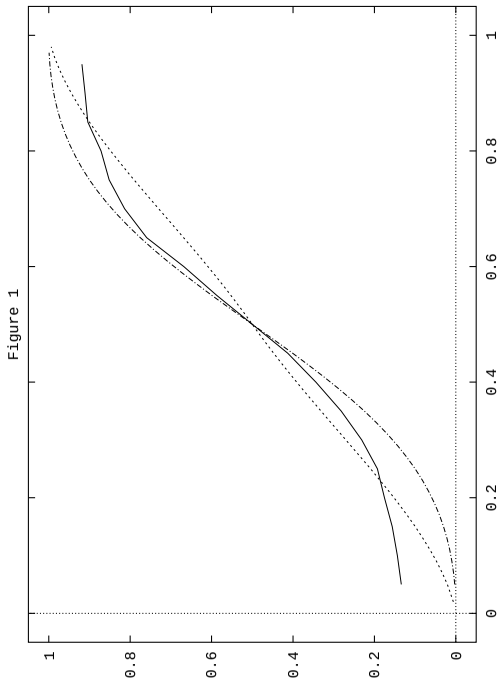
Figure 1
0
0.2
0.4
0.6
0.8
1
1
0.8
0.6
0.4
0.2
0

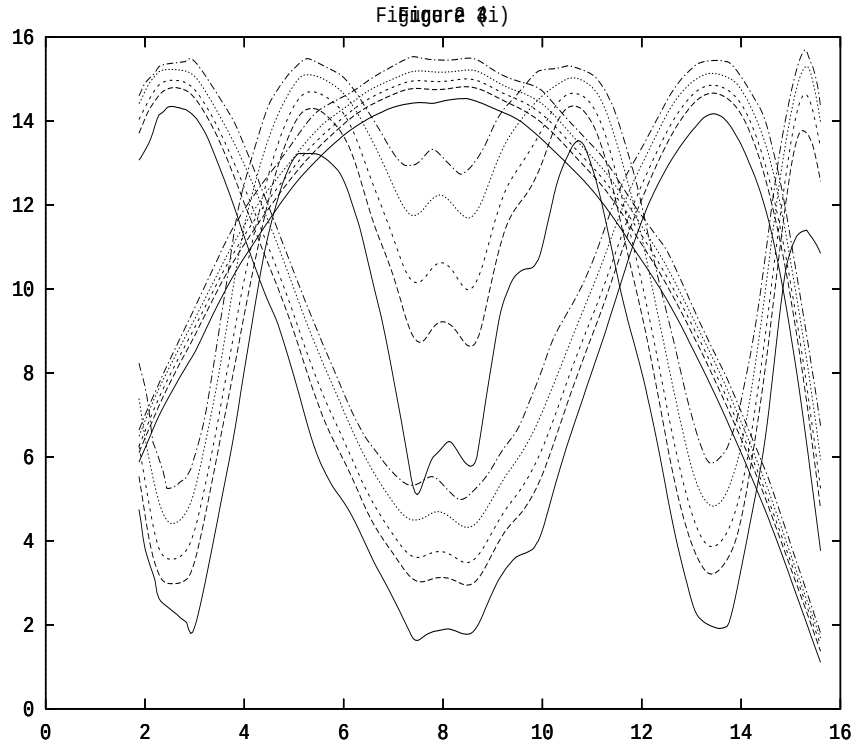
Figure 2 (i)
0
2
4
6
8
10
12
14
16

Figure 2 (ii)