PREDICTION AND NONPARAMETRIC ESTIMATION FOR TIME SERIES WITH HEAVY TAILS

Peter Hall[1] Liang Peng[1,2] Qiwei Yao[1,3]

Abstract. Motivated by prediction problems for time series with heavy-tailed marginal distributions, we consider methods based on 'local least absolute deviations' for estimating a regression median from dependent data. Unlike more conventional 'local median' methods, which are in effect based on locally fitting a polynomial of degree 0, techniques founded on local least absolute deviations have quadratic bias right up to the boundary of the design interval. And in contrast to local least-squares methods based on linear fits, the order of magnitude of variance does not depend on tail-weight of the error distribution. To make these points clear we develop theory describing local applications to time series of both least-squares and least-absolute-deviations methods, showing for example that in the case of heavy-tailed data the conventional local-linear least-squares estimator suffers from an additional bias term as well as increased variance.

Keywords. ARMA model, conditional median, heavy tail, least absolute deviation estimation, local-linear regression, prediction, regular variation, ρ-mixing, stable distribution, strong mixing, time series analysis.

Short title. Time series with heavy tails

[1]Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

[2]School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA

[3]Department of Statistics, London School of Economics, London, WC2A 2AE, UK

1 Introduction

Motivated by the problem of predicting a future value from observations of a particularly heavy-tailed but stationary time series, we consider robust nonparametric methods for estimating location from time-series data. When the marginal distribution has relatively light tails, and in particular finite variance, prediction is typically accomplished using the conditional mean. However, such techniques generally give poor performance when variance is infinite. We discuss the conditional median as an alternative in such cases, and suggest a local least-absolute-deviations estimator of the median, given the recent past.

Our method borrows ideas from robust approaches to nonparametric regression for independent data, where L_1 methods based on local polynomials have been considered by several authors; see Fan and Gijbels (1996, p. 199ff) for discussion, and the literature survey at the end of this section for further references. In the time-series case, related techniques based on local medians, equivalent to locally fitting a polynomial of order 0, have been considered by, for example, Truong (1989, 1991, 1992a,b) and Truong and Stone (1992). Our analysis differs substantially from that of other authors, in that we employ local-linear fits and focus sharply on the case where variance is infinite and data are correlated. The local-linear approach gives substantially reduced bias at the boundary.

We take the view that the virtues of robust methods are clearest when the conditions under which their more conventional, least-squares competitors are defined are violated. Thus, when we treat conventional least-squares local-linear methods (see section 3) we assume that the sampled data are generated by a mixing process for which the error distribution is in the domain of attraction of a stable law. In that context, consistency is only assured when $1 < \alpha < 2$, where α is the index of the stable law, and even then the limiting distribution is non-normal.

We show that when $1 < \alpha < 2$, and relative to the case where error variance is finite, the order of the error about the mean of the least-squares estimator is inflated by a factor $(nh)^{(2-\alpha)/(2\alpha)}\, \ell(n)$, where h represents bandwidth (depending on sample size, n) and ℓ is a generic function that is slowly varying at ∞. Since, in asymptotic terms, $nh \to \infty$ as n increases, then the inflation of stochastic error can be substantial. Moreover, an

additional term, which can be as large as $(nh)^{-(\alpha-1)/\alpha}\,\ell(n)$, is incorporated in the bias expansion. The net result is that both bias and error about the mean of least-squares estimators are liable to be substantially larger when variance is infinite than in the finite-variance case. Importantly, these two penalties can be virtually the size; even if the stochastic term could be suppressed, bias could render the estimator uncompetitive.

We prove that, by way of comparison, local-linear methods applied in the L_1 norm have the same orders of variance and bias, when applied to very heavy-tailed time series (in particular, to processes with stable marginals), as they do when variance is finite. See section 2. Of course, the 'target' function here is different from that in the least-squares case, although the two agree when the error distribution is symmetric.

There is a substantial literature on nonparametric regression with time-series data, including for example work of Robinson (1983), Truong (1991, 1992b) and Yakowitz (1985a, 1985b, 1987) on kernel and k-nearest neighbour methods for prediction, Truong (1994) on local polynomial techniques in the least-squares setting, Roussas, Tran and Ioannides (1992), Tran, Roussas, Yakowitz and Truong Van (1996) and Robinson (1997) on kernel nonparametric regression with fixed design, Roussas and Tran (1992) on recursive kernel methods for time series, Csörgő and Mielniczuk (1995a,b,c) on regression for both short- and long-range dependent data, Hall and Hart (1990) and Deo (1997) on the case of long-range dependence, and Chu and Marron (1991), Robinson (1994) and Ray and Tsay (1996) on bandwidth choice. Györfi, Härdle, Sarda and Vieu (1989) and Härdle (1990, Chapter 7) reviewed the literature up to the late 1980's, in addition to making their own contributions.

The practical importance of robust methods for parametric regression has motivated extensive study of related techniques in nonparametric contexts. Some have their roots in approaches suggested by Härdle (1984) and Härdle and Gasser (1984). Mallows (1980), Velleman (1980), Truong (1989), Härdle (1990, p. 69f) and Fan and Hall (1994) addressed local median smoothing for independent data, Tsybakov (1986) and Fan, Hu and Truong (1994) developed robust methods for fitting local polynomials, Truong and Stone (1992) and Truong (1991, 1992a,b) discussed robust nonparametric regression for time-series data, Wang and Scott (1994) developed L_1 methods for robust nonparametric regression, Leung, Marriott and Wu (1993) and Wang (1994) considered bandwidth choice in robust nonparametric settings, Smith and Kohn (1996) treated Bayesian methods for robust

nonparametric regression, and Welsh (1996) suggested robust quantile methods and M-estimation methods in the context of nonparametric regression.

2 Definition and parametric models

2.1 Conditional medians

Let $\{Y_t\}$ be a real-valued time series. Our goal is to predict future values Y_{n+m} for $m = 1, 2, \ldots$ from observed values $Y_n, Y_{n-1}, \ldots$. If Y_t has finite variance, the most frequently used point predictor is the conditional mean $E(Y_{n+m}|Y_n, Y_{n-1}, \ldots)$, which is optimal in the sense of minimising the mean squared error $E\{(Y_{n+m} - a)^2|Y_n, Y_{n-1}, \ldots\}$ over $a \in R$. When variance is infinite, one natural choice is the conditional median of Y_{n+m} given $Y_n, Y_{n-1}, \ldots$, which we denote by $M(Y_{n+m}|Y_n, Y_{n-1}, \ldots)$. When $E|Y_t| < \infty$, the predictor $Y_{n+m}(n) \equiv M(Y_{n+m}|Y_n, Y_{n-1}, \ldots)$ is optimal in the sense that it minimises the mean absolute deviation $E\{|Y_{n+m} - a||Y_n, Y_{n-1} \ldots\}$. Further, when $E|Y_t| = \infty$ but $E|Y_t|^\delta < \infty$ for some $\delta \in (0, 1)$, $Y_{n+m}(n)$ is still optimal in the sense that

$$\mathrm{sgn}\{Y_{m+n}(n)\}\,|Y_{m+n}(n)|^\delta = \underset{a \in R}{\mathrm{argmin}}\, E\left\{ |\mathrm{sgn}(Y_{n+m})\,|Y_{n+m}|^\delta - a| \,\middle|\, Y_n, Y_{n-1}, \ldots \right\}. \quad (2.1)$$

When $\{Y_t\}$ is Markovian of order p (*i.e.* Y_t is independent of $\{Y_{t-p+i}, i \geq 1\}$ given $Y_{t-1}, \ldots, Y_{t-p}$), $Y_{n+m} = M(Y_{n+m}|Y_n, \ldots, Y_{n-p+1})$.

2.2 ARMA(p, q) models

Suppose $\{Y_t\}$ is defined by an ARMA equation

$$Y_t - a_1 Y_{t-1} - \ldots - a_p Y_{t-p} = \varepsilon_t + b_1 \varepsilon_{t-1} + \ldots + b_q \varepsilon_{t-q}\,, \quad (2.2)$$

where $\{\varepsilon_t\}$ is a sequence of independent and identically distributed random variables. We assume that Y_t is causal, in the sense that it has an MA(∞) representation,

$$Y_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}, \quad \text{for all } t\,, \quad (2.3)$$

where $c_0 = 1$. We also suppose that the process is invertible, in the sense that

$$\varepsilon_t = \sum_{j=0}^{\infty} d_j Y_{t-j}, \quad \text{for all } t \quad (2.4)$$

where $d_0 = 1$. Explicit regularity conditions are given at (2.8) and (2.9) below. Then the following statements hold.

(a) Let $M(Z)$ denote the median of a random variable Z. Then,

$$Y_{n+1}(n) = M(\varepsilon_{n+1}) + \sum_{i=1}^{p} a_i Y_{n+1-i} + \sum_{j=1}^{q} b_j \varepsilon_{n+1-j}\,.$$

When $M(\varepsilon_t) \neq 0$, $Y_{n+1}(n)$ is not a homogeneous linear function of the sequence $Y_n, Y_{n-1}, \ldots$. Therefore, we deal with predictors different from those of Cambanis and Soltani (1984), and Cline and Brockwell (1985).

(b) It follows from (2.3) and (2.4) that for any $m \geq 1$,

$$\begin{aligned} Y_{n+m}(n) &= M\Big(\sum_{j=0}^{m-1} c_j \varepsilon_{n+m-j}\Big) + \sum_{i=0}^{\infty} c_{m+i}\varepsilon_{n-i} \qquad (2.5)\\ &= M\Big(\varepsilon_{n+m} - \sum_{j=1}^{m-1} d_j Y_{n+m-j}\Big| Y_n, Y_{n-1}, \ldots\Big) \;-\; \sum_{i=0}^{\infty} d_{m+i} Y_{n-i}\,. \end{aligned}$$

Further, Y_{n+m} can be expressed as $Y_{n+m} = \sum_{0\leq j\leq m-1} c_j \varepsilon_{n+m-j} + \sum_{i\geq 0} \varphi_i Y_{n-i}$. Correspondingly, $Y_{n+m}(n) = M(\sum_{1\leq j\leq m} c_j \varepsilon_{n+m-j}) + \sum_{i\geq 0} \varphi_i Y_{n-i}$, where $\{\varphi_j\}$ is a sequence of constants.

(c) If the distribution of ε_t is symmetric (about 0), so too is the distribution of Y_t. Further, the conditional distribution of Y_{n+m} given $Y_n, Y_{n-1}, \ldots$ is then symmetric about $Y_{n+m}(n)$, which is a linear combination of $Y_n, Y_{n-1}, \ldots$. In fact for any $m \geq 1$,

$$Y_{n+m}(n) = \sum_{j=1}^{p} a_j Y_{n+m-j}(n) + \sum_{i=1}^{q} b_i \varepsilon_{n+m-i}(n), \qquad (2.6)$$

where $Y_j(n) = Y_j$ and $\varepsilon_j(n) = \varepsilon_j$ for $j \leq n$, and $\varepsilon_j(n) = 0$ for $j > n$. (See model (2.2).) With the additional condition that $E|\varepsilon_t| < \infty$, it holds that $Y_{n+m}(n) = E(Y_{m+n}|Y_n, Y_{n-1}, \ldots)$.

2.3 Stable ARMA(p, q) models

A typical example of time series with infinite variance is an ARMA process defined in terms of a sequence of independent and identically distributed stable disturbances (see

Mikosch *et al.*, 1995). A random variable Z has a stable distribution, with shape (or index), scale, skewness and location parameters α, σ, β and μ respectively, denoted by $Z \sim S_\alpha(\sigma, \beta, \mu)$, if its log characteristic function has the form

$$\log E(e^{itZ}) = \begin{cases} i\mu t - \sigma^\alpha |t|^\alpha \{1 - i\beta \operatorname{sgn}(t) \tan(\pi\alpha/2)\} & \alpha \neq 1\,, \\ i\mu t - \sigma |t| \{1 + i\beta \operatorname{sgn}(t) (2/\pi) \log |t|\} & \alpha = 1\,, \end{cases} \tag{2.7}$$

where $\alpha \in (0, 2]$, $\sigma > 0$, $\mu \in (-\infty, \infty)$, and $\beta \in [-1, 1]$. It is easy to see that when $\alpha \neq 1$, $(Z - \mu)/\sigma \sim S_\alpha(1, \beta, 0)$. When $\alpha = 2$, Z is a normal random variable with mean μ and variance $2\sigma^2$. When $\alpha = 1$ and $\beta = 0$, Z has a Cauchy distribution. For further properties of stable distributions we refer to Zolotarev (1986) and Samorodnitsky and Taqqu (1994).

Let $\{\varepsilon_t\}$ be sequence of independent random variables with the common distribution $S_\alpha(\sigma, \beta, 0)$. For the sake of simplicity, we assume $\beta = 0$ if $\alpha = 1$. Given constants $c_0, \ldots c_q$, define $Y_t = \sum_{j \geq 0} c_j \varepsilon_{n-j}$. Then, $\{Y_t\}$ is a strictly stationary MA(q) process. Its marginal distribution is still stable with index α and location parameter 0. In fact, the doubly infinite series $\sum c_j \varepsilon_j$ converges absolutely, with probability 1, to a stable random variable with index α, provided $\sum |c_j|^\delta < \infty$ for some $0 < \delta < \min(\alpha, 1)$. Hence, the ARMA equation (2.2) defines a unique, strictly stationary time series $\{Y_t\}$ with an α-stable marginal distribution, under the condition that

$$1 - a_1 x - \ldots - a_p x^p \neq 0 \quad \text{for all } |x| \leq 1\,. \tag{2.8}$$

Further, Y_t is invertible if

$$1 + b_1 x + \ldots + b_q x^q \neq 0 \quad \text{for all } |x| \leq 1\,. \tag{2.9}$$

We assume that both (2.8) and (2.9) hold, and the two equations have no common roots. Below we discuss two special stable ARMA processes.

The main difficulty when $\alpha = 1$ and $\beta \neq 0$ is that of certering the error distribution and hence the distribution of Y_t. For $\alpha > 1$ the mean is finite, and means may be taken as the centers of either distributions. When $\alpha < 1$, centering is necessary only for notational convenience, and does not play a substantive role. But for $\alpha = 1$ centering is important, the mean of ε_t is not well defined, and a suitable center for distributions of ε_t and Y_t will not necessarily be 0 unless $\beta = 0$.

(a) Symmetric ARMA processes

Let the distribution of ε_t be symmetric, namely $\varepsilon_t \sim S_\alpha(\sigma, 0, 0)$. Then $Y_t \sim S_\alpha(\sigma_x, 0, 0)$ where $\sigma_x = \sigma\,(\sum_{i\geq 0} |c_i|^\alpha)^{1/\alpha}$ (see (2.3)). For any linear form $\xi_n \equiv \sum_{j\geq 0} \psi_j Y_{n-j}$ such that $\sum_{j\geq 0} |\psi_j|^d < \infty$ for some $d \in (0, \alpha)$, $Y_{m+n} - \xi_n$ is an α-stable random variable with scale parameter

$$\Big(\sum_{j=0}^{m-1} |c_j|^\alpha + \sum_{i=0}^{\infty} |c_{m+i} - \psi_i|^\alpha \Big)^{1/\alpha}.$$

From (2.5) we may deduce that this parameter attains its minimum value when $\xi_n = Y_{n+m}(n)$. Therefore, for symmetric stable ARMA time series, the conditional median is also the optimal linear predictor in the sense of minimising the scaling parameter of the predictive error, which is also the minimum dispersion predictor studied by Cline and Brockwell (1985). Note too that for any $\delta \in (0, \alpha)$,

$$\Big\{ E|Y_{n+m} - Y_{n+m}(n)|^\delta \Big\}^{1/\delta} = \Big(\sum_{j=0}^{m-1} |c_j|^\alpha \Big)^{1/\alpha} \Big(E|\varepsilon_t|^\delta \Big)^{1/\delta},$$

which is the minimum value of $E(|Y_{n+m} - \xi_n|^\delta)$ over all linear ξ_n specified above. When $q = 0$ (*i.e.* when $\{Y_t\}$ is an AR(p) model), the above δ-norm predictive error can be expressed in terms of the autoregressive coefficients:

$$\Big\{ E|Y_{n+m} - Y_{n+m}(n)|^\delta \Big\}^{1/\delta} = \lambda_m \Big(E|\varepsilon_t|^\delta \Big)^{1/\delta}, \tag{2.10}$$

where $\lambda_m > 0$ is a constant depending on $a_1, \ldots, a_p$. In fact $\lambda_1 = 1$, $\lambda_m = (\lambda_{m-1}^\alpha + |\ell_{m-1}|^\alpha)^{1/\alpha}$ for $m > 1$, where $\ell_i = \sum_{1\leq j\leq \min(i,p)} a_j \ell_{i-j}$ for $i \geq 1$ and $\ell_0 = 1$.

(b) MA(∞)-representation with non-negative coefficients

Here we assume that $\varepsilon_t \sim S_\alpha(\sigma, \beta, 0)$ and all the coefficients c_j in (2.3) are non-negative. The latter property is implied by the condition that in model (2.2) all b_j's are non-negative and all roots of the equation $1 + \sum_{1\leq j\leq p} a_j x^j = 0$ are greater than 1. The assumed conditions imply that $Y_t \sim S_\alpha(\sigma_x, \beta, 0)$, with the same σ_x as before. It follows from (2.5) that

$$Y_{n+m} - Y_{n+m}(n) \sim S_\alpha\Big\{ \Big(\sum_{j=0}^{m-1} c_j^\alpha \Big)^{1/\alpha} \sigma,\ \beta,\ -M\Big(\sum_{j=0}^{m-1} c_j \varepsilon_{n+m-i} \Big) \Big\}.$$

Note that $M(\sum_{0\leq j\leq m-1} c_j \varepsilon_{n+m-i}) = (\sum_{0\leq j\leq m-1} c_j^\alpha)^{1/\alpha}\, M(\varepsilon_t)$. Hence, for any $\delta \in (0, \alpha)$,

$$\Big(E|Y_{n+m} - Y_{n+m}(n)|^\delta \Big)^{1/\delta} = \Big(\sum_{j=0}^{m-1} c_j^\alpha \Big)^{1/\alpha} \Big\{ E|\varepsilon_t - M(\varepsilon_t)|^\delta \Big\}^{1/\delta}.$$

For AR(p) models, the coefficient $(\sum_{0\le j\le m-1} c_j^\alpha)^{1/\alpha}$ can be replaced by λ_m, defined as in (2.10).

3 Nonparametric estimation

For parametric models, such as linear autoregressions, we predict Y_{n+m} by $Y_{n+m}(n) = \sum_{1\le j\le p} a_j X_{n+m-j}(n)$ (compare (2.6)) with all a_j's replaced by their estimates. For estimation of those autoregressive parameters we refer to Davis, Knight and Liu (1992), Resnick (1997) and references within. Although the linear autoregressive model offers attractive features for analysis, the class of linear autoregressive models is arguably too small for modelling real data with heavy tails. See Resnick (1997) and its discussion by R.J. Adler.

When we do not have sufficient knowledge about the underlying model, one possible approach is to use nonparametric methods. In the prediction context we would estimate $\mu(x_1,\ldots,x_p) \equiv M(Y_{n+m}|Y_n = x_1,\ldots,Y_{n-p+1} = x_p)$ from observed data $\{Y_1,\ldots,Y_n\}$ in order to predict Y_{n+m}. To avoid the difficulties associated with the 'curse of dimensionality', p would be chosen small. In the sequel we assume $p = 1$, for the sake of simple presentation. Similar results also hold for $p > 1$; see Remark 6.

Thus, we wish to estimate the conditional median $\mu(x) = M(Y_t|X_t = x)$ from observed values $\{(X_i, Y_i), 1 \le i \le n\}$ of a strictly stationary process $\{(X_t, Y_t)\}$. In the context of m-step ahead prediction for a time series $\{Y_t\}$, we would take $X_t = Y_{t-m}$.

3.1 Least absolute deviations estimation

Note that $\mu(x)$ is the minimiser of $E(|Y_t - a|\,|X_t = x)$ over a. Hence, we may define $\widehat{\mu}(x) = \widehat{\theta}_1$, where $(\widehat{\theta}_1, \widehat{\theta}_2)$ minimises

$$\sum_{i=1}^{n} |Y_i - \theta_1 - \theta_2(X_i - x)|\, K\left(\frac{X_i - x}{h}\right), \tag{3.1}$$

K is a kernel function and $h > 0$ is a bandwidth. This is the local linear least-absolute-deviations estimator.

Theorem 1 below states that the estimator $\widehat{\mu}(x)$ is asymptotically normal with mean

$\mu(x)+O(h^2)$ and variance of size $(nh)^{-1}$. To state the theorem we first introduce notation. Let $f(\cdot)$ denote the marginal density of X_t, let $g(\cdot|x)$ be the conditional density function of Y_t given $X_t = x$, and put $g_\mu(x) = g\{\mu(x)|x\}$. We write $\dot{m}(x) = (\partial/\partial x)\, m(x)$ and $\ddot{m}(x) = (\partial/\partial x)^2\, m(x)$. Let $\kappa_1 = \int K^2$, $\sigma_0 = \int u^2\, K(u)\, du$ and $\sigma(x)^2 = \kappa_1/\{4\, f(x)\, g_\mu(x)^2\}$. We shall use C to denote a generic positive constant. The theorem holds under the following regularity conditions.

(C1) For fixed x, $f(x) > 0$ and $g(y|x) > 0$ is continuous at $y = \mu(x)$. In a neighbourhood of x, $\mu(\cdot)$ has continuous second derivative, both $f(\cdot)$ and $g(y|\cdot)$ have continuous first derivatives, and $E(|Y_t|^\delta|X_t = \cdot) < \infty$ for some $\delta > 0$.

(C2) The kernel K is a symmetric, bounded and non-negative function with compact support.

(C3) The process $\{(X_t, Y_t)\}$ is strong mixing , *i.e.*

$$\alpha(j) \equiv \sup_{A \in \mathcal{F}_{i+j}^{\infty},\, B \in \mathcal{F}_1^i} |P(A)\, P(B) - P(AB)| \to 0 \quad \text{as} \quad j \to \infty\,,$$

where $\mathcal{F}_i^j$ denotes the σ-field generated by $\{(X_k, Y_k) : i \le k \le j\}$. Further, $\sum_{j \ge 1} \alpha(j)^{\delta_0/(1+\delta_0)} < \infty$ for some $\delta_0 < 1$.

(C4) $h = h(n) \to 0$ as $n \to \infty$, and $\liminf_{n\to\infty} nh^{1+2\delta_0} > 0$.

Remark 1: *Mixing condition.* The strong mixing condition assumed in (C3) is mild in the sense that the mixing coefficients are permitted to decay only polynomially fast. Any process which admits representation (2.3) with exponentially decaying coefficients $\{c_i\}$ and continuous ε_t is absolutely regular, therefore also strong mixing, with exponentially decaying mixing coefficients; see Pham and Tran (1985).

Remark 2: *Motivation for bandwidth condition.* The method of proof of Theorem 1 involves approximating $\widehat{\mu}$ by a linear estimator, and the latter involves the derivative of $\mu(\cdot)$. Our demonstration that the linear approximation is sufficiently accurate requires the estimated derivative of μ to be $O_p(1)$, and for that, the condition $\liminf_{n\to\infty} nh^{1+2\delta_0} > 0$ is sufficient. The latter assumption is of little consequence in practice, however, since it will follow from Theorem 1 that the optimal bandwidth is asymptotic to a constant multiple of $n^{-1/5}$, for which $nh^{1+2\delta_0} \to \infty$.

Theorem 1. Suppose conditions (C1)–(C4) hold. Then as $n \to \infty$,

$$\sqrt{nh}\left\{\widehat{\mu}(x) - \mu(x) - \tfrac{1}{2}\,h^2\,\sigma_0\,\ddot{\mu}(x)\right\} \xrightarrow{d} N\left(0, \sigma(x)^2\right) .$$

Remark 3: *Local constant estimators.* Robinson (1984) discussed the local constant version of regression by least absolute deviations. There the function estimator is given by $\widetilde{\mu}(x) = \widetilde{\theta}_1$, and $\widetilde{\theta}_1$ is defined as the minimiser of (3.1) subject to $\theta_2 = 0$. Local constant estimators are more commonly encountered in a least-squares setting, where they are equivalent to the Nadaraya-Watson estimator. In both the local constant and local linear version of regressions by least squares, regularity conditions for a central limit theorem generally require that $E(Y_t^2|X_t = x) < \infty$, which is a substantial strengthening of the moment condition in (C1). This reflects the fact that the variance of the limiting distribution of a least-squares estimator if μ is proportional to the variance of the Y process.

Remark 4: *Bandwidth choice.* The theorem implies that, to first order, the asymptotically optimal bandwidth is $h_0(x) = \{\eta(x)\,n\}^{-1/5}$, where $\eta(x) = \sigma_0\,|\ddot{\mu}(x)|/\sigma(x)$. We may compute an empirical bandwidth $\hat{h}_0$ that is consistent for h_0, in the sense that $\hat{h}_0/h_0 \to 1$ in probability, by arguing as follows. Observe that

$$\eta(x) = \kappa_3\,|\ddot{\mu}(x)|\,f_{XY}\{\mu(x), x\}\,f(x)^{-1/2}\,,$$

where $\kappa_3 = \sigma_0/\kappa_1^{1/2}$ is known and f_{XY} denotes the joint density of X and Y. Statistically consistent, pilot estimators of f and f_{XY} may be computed using conventional kernel methods. Their properties for dependent data satisfying a mixing condition are well known; see *e.g.* Hart (1984), Hart and Vieu (1984), Györfi, Härdle, Sarda and Vieu (1989) and Wand and Jones (1995 section 6.2.1). Moreover, Theorem 1 shows that, in order to be consistent for μ, $\widehat{\mu}$ need only be computed using a bandwidth h that satisfies condition (C4), so optimal bandwidth choice is not essential at this step. Likewise it may be proved that under conditions (C1)–(C3), and (instead of (C4)) $h = h(n) \to \infty$ and $\liminf nh^7 > 0$, $\ddot{\mu}$ may be estimated consistently from a local quadratic or local cubic fit. Combining these properties we see that we may produce a consistent estimate $\hat{\eta}(x)$ of $\eta(x)$. Taking $\hat{h}_0 = \{\hat{\eta}(x)\,n\}^{-1/5}$ we obtain the desired plug-in bandwidth estimator.

Remark 5: *Comparison with local medians.* Local median estimators are defined by minimising $\sum_i |Y_i - \theta|\,K\{(X_i - x)/h\}$ with respect to θ, instead of minimising the series

at (3.1). In the time-series case, see for example the methods and results of Truong and Stone (1992), where K is in effect the uniform kernel. These estimators share with ours the orders h^2 and $(nh)^{-1}$ of bias and variance, respectively, at interior points of the design interval. However, the absolute value of their bias increases to $O(h)$ at the boundary, where it may be shown that our techniques continue to enjoy $O(h^2)$ bias. A proof of this property is similar to that of Theorem 1, and so is not given here.

Remark 6: *Generalisations to higher dimensions and higher degrees.* There are no difficulties extending Theorem 1 to the case of p-variate regressands X. There, asymptotic bias remains at order h^2, and asymptotic variance changes to order $(nh^p)^{-1}$. Conditions (C1)–(C3) should be modified in obvious ways, and (C4) should be altered by asking that $h \to 0$ and $\liminf nh^{p+2} > 0$. High-order generalisations of the method, to L_1 fitting of polynomials of arbitrary degree, produce estimators of $\mu(x)$ with biases of the same orders as in the L_2 case (see Ruppert and Wand (1994) for the latter) and variances of order $(nh)^{-1}$ (provided $p = 1$).

3.2 Least-squares estimation

As in the parametric case we can also consider least-squares estimation of $\mu(\cdot)$. For example, suppose the process $\{(X_t, Y_t)\}$ is generated from model $Y_t = \varphi(X_t) + \varepsilon_t$, where $\{\varepsilon_t\}$ is a sequence of independent and identically distributed random variables, and ε_t is independent of $\{(X_i, Y_{i-1}) : i \le t\}$. Then $\mu(x) = \varphi(x) + \mu_0$, where μ_0 denotes the median of ε_t. The least-squares local-linear regression estimator is defined as $\widehat{\varphi}(x) = \widehat{\theta}_1$, where $(\widehat{\theta}_1, \widehat{\theta}_2)$ is now the minimiser of

$$\sum_{i=1}^{n} \{Y_i - \theta_1 - \theta_2(X_i - x)\}^2 K\left(\frac{X_t - x}{h}\right).$$

As a prelude to deriving the rate of convergence of $\widehat{\varphi}(x)$ we assume that the distribution of ε_t varies regularly at infinity with index $-\alpha$. That is,

$$P(|\varepsilon_t| > z) = z^{-\alpha} L(z)\,, \tag{3.2}$$

where the function L is slowly varying at infinity. It is easy to see that for such a distribution, the second moment is infinite when $\alpha < 2$, and the mean fails to exist when $\alpha < 1$. In the sequel we always assume $\alpha \in (0, 2)$. We also suppose that the tails of the

distribution of ε_t are balanced, in the sense that

$$\lim_{z\to\infty} P(\varepsilon_t > z)/P(|\varepsilon_t| > z) = p \in [0,1]\,. \tag{3.3}$$

Under (3.2) and (3.3), the normalised partial sum of $\{\varepsilon_t\}$ converges in distribution to a stable distribution with index α. See Section 17.5 of Feller (1971). Finally, we assume that

(C5) For fixed x, $f(x) > 0$ is continuous at x. In a neighbourhood of x, $\varphi(\cdot)$ has second continuous derivative.

(C6) The kernel $K(\cdot)$ is symmetric and non-negative with compact support. Further, $K(\cdot)$ is bounded away from both 0 and ∞ on its support.

(C7) The process $\{(X_t, Y_t)\}$ is ρ-mixing, *i.e.*

$$\rho(j) = \sup_{U\in L^2(\mathcal{F}_{i+j}^{\infty}),\, V\in L^2(\mathcal{F}_1^i)} |\mathrm{Corr}(U,V)| \to 0 \quad \text{as } j\to\infty,$$

where $\mathcal{F}_i^j$ denotes the σ-field generated by $\{(X_t,Y_t) \;:\; i \le t \le j\}$. Further, $\sum_{j\ge 1}\rho(j) < \infty$.

(C8) As $n\to\infty$, $h\to 0$ and $nh\to\infty$.

Put $K_t = K\{(X_t - x)/h\}$, let a_n be the infimum of values satisfying

$$P(|\varepsilon_t K_t| \ge a_n) \le 1/n\,, \tag{3.4}$$

and define $b_n = E\{\varepsilon_t\, K_t\, I(|\varepsilon_t K_t| \le a_n)\}$ when $\alpha \in [1,2)$, and $b_n = 0$ otherwise.

Theorem 2. Assume that $0 < \alpha < 2$, that conditions (C5) – (C8) hold, that the distribution of ε_t satisfies (3.2) and (3.3), that $p = 1/2$ when $\alpha = 1$, and that $E(\varepsilon_t) = 0$ when $\alpha > 1$. Then,

$$\frac{nh}{a_n} f(x)\left\{\widehat{\varphi}(x) - \varphi(x) - \tfrac{1}{2}\,h^2\,\sigma_0^2\,\ddot{\varphi}(x)\right\} - \frac{n}{a_n} b_n \xrightarrow{d} S^*$$

as $n\to\infty$, where the random variable S^* admits the representation

$$S^* = \sum_{j=1}^{\infty}\left[\delta_j \Gamma_j^{-1/\alpha} - (2p-1)\,E\left\{\Gamma_j^{-1/\alpha}\,\mathbf{1}(0 < \Gamma_j^{-1/\alpha} \le 1)\right\}\right]\,,$$

$\Gamma_j = \sum_{i=1}^{j} W_i$, $\{W_i\}$ is a sequence of exponential random variables with unit mean, $\{\delta_i\}$ is a sequence of 0–1 random variables with $P(\delta_i = 1) = p$, and the variables W_i, δ_i for $i \geq 1$ are totally independent.

Remark 7: *Consistency of $\hat{\varphi}$ for φ.* If $\alpha \in (1, 2)$ then $nh/a_n \to \infty$ and $b_n/h \to E\epsilon_t = 0$, implying that $\widehat{\varphi}(x)$ is a consistent estimator of $\varphi(x)$. However, $nh/a_n \to 0$ when $\alpha < 1$, and does not necessarily diverge to $+\infty$ when $\alpha = 1$. It follows from Theorem 2 that $\widehat{\varphi}(x)$ is not consistent in such cases.

Remark 8: *Rate of convergence of $\hat{\varphi}$ to φ.* To assess the size of a_n, let us suppose that h is asymptotic to a constant multiple of n^{-c} for some $0 < c < 1$, and $1 < \alpha < 2$. Then, $a_n = (nh)^{1/\alpha}\, \ell_1(n)$, where ℓ_j denotes a function that is slowly varying at ∞. Therefore, the error of the estimator $\hat{\varphi}$ about its mean is of order $(nh)^{-(\alpha-1)/\alpha}\ell_2(n)$, which increase from virtually order $(nh)^{-1/2}$ (the value it has when the error variance is finite) when α is close to 2, to $(nh)^{-\eta}$, for η arbitrarily small, when α is close to 1. Moreover, while b_n vanishes in the case of a symmetric error distribution, b_n/h can be as large as $(nh)^{-(\alpha-1)/\alpha}\ell_3(n)$. In view of Theorem 2, there is a contribution of this size to the asymptotic bias of $\hat{\varphi}$, and this is additional to the standard bias term, of order h^2. Moreover, modulo a slowly varying function, $(nh)^{-(\alpha-1)/\alpha}\ell_3(n)$ has the same behaviour as $(nh)^{-(\alpha-1)/\alpha}\ell_2(n)$, representing stochastic error. The net result is that, although $\hat{\varphi}$ is consistent for φ when $1 < \alpha < 2$, the rate of convergence can be particularly poor.

Remark 9: *Comparison with parametric cases.* Asymptotic properties of parametric estimators are radically different from those presented in Theorems 1 and 2 above. As an illustration, consider the simple AR(1) model $Y_t = aY_{t-1} + \varepsilon_t$, where $|a| < 1$ and ε_t satisfies both (3.2) and (3.3). A 'compensation' for the difficulties associated with heavy tails is the fact that a parametric estimator $\widehat{a}$ converges to a in probability at a rate faster than $n^{-1/\delta}$ for any $\delta > \alpha$, and in particular faster than the rate $n^{-1/2}$ in the case of finite variance. Here, $\widehat{a}$ could be either the least-squares estimator or the least absolute deviations estimator (Davis, Knight and Liu, 1992). Cox (1966) gave an intuitive explanation of this phenomenon. See also Hannan and Kanter (1977).

The price paid for this fast convergence rate is the particularly slow rate at which the distribution of $\widehat{a}$ can converge to its stable limit; see Adler, Feldman and Gallagher (1997, section 3.4). The rate can be slower than $n^{-\eta}$ for any given $\eta > 0$; for example, this rate

obtains if the slowly-varying function $L(z)$ at (3.2) has the form $C_1 + C_2 z^{-\alpha\eta} + o(z^{-\alpha\eta})$, for constants $C_1 > 0$ and $C_2 \neq 0$. From a statistical viewpoint, this problem is brought about by the fact that the least-squares approach used to construct $\widehat{a}$ can be far from optimal when the marginal distribution of $\{Y_t\}$ is heavy-tailed. In marked contrast, it may be shown that under smoothness conditions (as distinct from tail conditions) on the distribution of ϵ_t, the rate of convergence of the distribution of $\widehat{\mu}$ to its limit is $O\{(nh)^{-1/2}\}$, in the sense of a Berry–Esseen bound. As a result, tests or confidence procedures based on local-linear methods in L_1 can have greater level or coverage accuracy than their parametric counterparts.

Appendix: *Outline Proofs of Theorems*

A.1 *Proof of Theorem 1*

We prove the theorem only in the case $\delta_0 > 0$, and divide the proof into two steps, addressing $\delta \geq 1$ and $\delta < 1$ respectively.

Step 1: When $\delta \geq 1$, $E|Y_t| < \infty$. The main idea of the proof is to approximate the series at (3.1) by a quadratic function whose minimiser is asymptotically normal, and then show that our estimator is close enough to the minimiser to share the latter's asymptotic behaviour, as implied by the convexity lemma of Pollard (1991). We sketch the proof below.

Let $K_i = K\{(X_i - x)/h\}$, $Z_i = (1, (X_i - x)/h)^T$, $Y_i^* = Y_i - \mu(x) - \dot{\mu}(x)\,(X_i - x)$ and $\widehat{\theta} = \sqrt{nh}\,(\widehat{\mu}(x) - \mu(x),\, h\,\{\widehat{\dot{\mu}}(x) - \dot{\mu}(x)\})^T$. For $\theta = (\theta_1, \theta_2)^T \in R^2$, define

$$G(\theta) = \sum_{i=1}^{n} \left\{ |Y_i^* - (\theta^T Z_i/\sqrt{nh})| - |Y_i^*| \right\} K_i\,, \tag{A.1}$$

$$R(\theta) = G(\theta) - f(x)\, g_\mu(x)\, (\theta_1^2 + \theta_2^2 \sigma_0) + \frac{1}{\sqrt{nh}}\, \theta^T \sum_{i=1}^{n} Z_i\, D(Y_i^*)\, K_i\,, \tag{A.2}$$

where $D(y) = I(y > 0) - I(y < 0)$. Then $\widehat{\theta}$ is the minimiser of $G(\theta)$.

We first prove that $R(\theta) \xrightarrow{P} 0$. In view of (A.1) and (A.2), $R(\theta) = \sum_{1 \leq i \leq n} T_i -$

$f(x)\, g_\mu(x)\, (\theta_1^2 + \theta_2^2 \sigma_0)$, where

$$T_i = \left[|Y_i^* - (\theta^T Z_i / \sqrt{nh})| - |Y_i^*| + \{\theta^T Z_i D(Y_i^*)/\sqrt{nh}\} \right] K_i \,.$$

After algebraic manipulation we obtain,

$$E(T_i) - n^{-1} f(x)\, g_\mu(x)\, (\theta_1^2 + \theta_2^2 \sigma_0) = o(n^{-1}) \,. \tag{A.3}$$

Hence, $E\{R(\theta)\} \to 0$. Therefore, in order to prove that $R(\theta) \xrightarrow{P} 0$ we need only show that $\sum_{1 \le i \le n} (T_i - ET_i) \xrightarrow{P} 0$. Note that

$$|\, |a+b| - |a| - D(a)\, b \,| \le 2|b|\, I(|a| < |b|) \,,$$

that $K(\cdot)$ has a compact support, and that $g_\mu(x) > 0$, whence it follows that for any $\gamma \ge 0$,

$$\begin{aligned} E\left(T_i^{2(1+\gamma)}\right) &\le \frac{C}{(nh)^{1+\gamma}} E\left\{ (\theta^T Z_i)^{2(1+\gamma)} K_i^{2(1+\gamma)} I(|Y_i^*| < C/\sqrt{nh}) \right\} \\ &= O\left(n^{-(3/2+\gamma)} h^{-(1/2+\gamma)} \right) . \end{aligned}$$

Applying Theorem 3 in Doukhan (1994, p. 9), with $r = 1 + \delta_0^{-1}$ and $p = q = 2(1+\delta_0)$, we may show that

$$\begin{aligned} P\left\{ \left| \sum_{i=1}^n (T_i - ET_i) \right| > \varepsilon \right\} &\le \sum_{i=1}^2 E(T_i^2) + 2 \sum_{i=1}^{n-1} (n-i)\, \mathrm{Cov}(T_1, T_{i+1}) \\ &\le C(nh)^{-1/2} + C \sum_{i=1}^{n-1} (n-i)\, \alpha(i)^{\delta_0/(1+\delta_0)} \left(n^{-(3/2+\delta_0)}\, h^{-(1/2+\delta_0)} \right)^{1/(1+\delta_0)} \\ &= O\left\{ (nh)^{-1/2} + n^{-1/\{2(1+\delta_0)\}}\, h^{-(1+2\delta_0)/\{2(1+\delta_0)\}} \right\} , \end{aligned}$$

which converges to 0 since the second term on the right-hand side is of smaller order than $(nh^3)^{-1/\{2(1+\delta_0)\}}$; see condition (C4).

Since $R(\theta) \xrightarrow{P} 0$ then the convex function $G(\theta) - (nh)^{-1/2}\, \theta^T \sum_{1 \le i \le n} Z_i\, D(Y_i^*)\, K_i$ converges to $f(x)\, g_\mu(x)\, (\theta_1^2 + \theta_2^2 \sigma_0^2)$. By the convexity lemma (Pollard, 1991), the convergence is uniform on compact sets in R^2. Using the arguments of Pollard (1991, p. 193) we may show that the difference between the minimiser of $\widehat{\theta}$ of $G(\theta)$ and the minimiser of

$$-\frac{1}{\sqrt{nh}}\, \theta^T \sum_{i=1}^n Z_i\, D(Y_i^*)\, K_i + f(x)\, g_\mu(x)\, (\theta_1^2 + \theta_2^2 \sigma_0^2)$$

converges to 0 in probability. This implies that

$$\sqrt{nh}\,\{\widehat{\mu}(x) - \mu(x)\} = \frac{1}{2\sqrt{nh}\,f(x)\,g_\mu(x)} \sum_{i=1}^{n} D(Y_i^*)\,K_i + o_p(1)\,. \tag{A.4}$$

Note too that

$$E\{D(Y_i^*)\,K_i\} = \sigma_0\,h^3\,f(x)\,g_\mu(x)\,\ddot{m}(x) + o(h^3)\,, \quad E\{D(Y_i^*)\,K_i\}^2 = f(x)\int K^2 + o(1)\,.$$

The required asymptotic normality now follows from Theorem 1 of Doukhan (1994, p. 46).

Step 2: In Step 1, the condition $E|Y_t| < \infty$ was used only in deriving (A.3). Hence, we need only show that $R(\theta) \xrightarrow{P} 0$ when $E|Y_t| = \infty$. To this end, define

$$Y_{t,n} = Y_t I(|Y_t| \le n^{2/\delta}), \quad Y_{t,n}^* = Y_{t,n} - \mu(x) - \dot{\mu}(x)\,(X_t - x)\,.$$

Further, let $G_n(\theta)$ and $R_n(\theta)$ be as at (A.1) and (A.2), with $\{Y_i^*\}$ there replaced by $\{Y_{i,n}^*\}$. It can be shown that $\mu(x) - M(Y_{i,n}|X_i = x) = O(n^{-2})$. Since $g(y|x)$ is positive and continuous at $y = \mu(x)$, and $E\{|Y_{i,n}|\,|X_i = x\} < \infty$, we may show that (A.3) still holds with $\{Y_i^*\}$ replaced by $\{Y_{i,n}^*\}$, and therefore $R_n(\theta) \xrightarrow{P} 0$.

On the set $\{|Y_i| < n^{2/\delta}\}$, $Y_i^* = Y_{i,n}^*$. Hence, for any $\eta > 0$ and sufficiently large n,

$$\begin{aligned} P\{|G(\theta) - G_n(\theta)| > \eta\} &\le n\,P\Big\{\,\big|\,|Y_i^* - (\theta^T Z_i/\sqrt{nh})| - |Y_i^*| \\ &\qquad - |Y_{i,n}^* - (\theta^T Z_i/\sqrt{nh})| + |Y_{i,n}^*|\,\big| > \eta/n\Big\} \\ &\le nP\left(|Y_i| \ge n^{2/\delta}\right) = O(n.n^{-2}) \to 0\,. \end{aligned}$$

In the same manner we may prove that $\sum_i Z_i\{D(Y_i^*) - D(Y_{i,n}^*)\}K_i = o_p\{(nh)^{1/2}\}$. Finally,

$$R(\theta) = R_n(\theta) + G(\theta) - G_n(\theta) - \frac{1}{\sqrt{nh}}\,\theta^T \sum_{i=1}^{n} Z_i\,\{D(Y_i^*) - D(Y_{i,n}^*)\}\,K_i \xrightarrow{P} 0\,,$$

as had to be shown.

A.2 *Proof of Theorem 2*

By standard arguments, based on the explicit formula for $\widehat{\varphi}$, it may be proved that

$$\widehat{\varphi}(x) - \varphi(x) - \tfrac{1}{2}\,h^2\,\sigma_0\,\ddot{\varphi}(x) = \frac{1 + o_p(1)}{nh\,f(x)} \sum_{i=1}^{n} \varepsilon_i\,K\left(\frac{X_i - x}{h}\right)\,. \tag{A.5}$$

Thus, we need only show that $a_n^{-1} \sum_i (\varepsilon_i K_i - b_n) \xrightarrow{d} S_\alpha$. This may be done using methods leading to Theorems 2 and 3 of Davis (1983), and to Theorem 2 of LePage, Woodroofe and Zinn (1981), on noting that

(a) for any $z > 0$, $\lim_{k\to\infty} \limsup_{n\to\infty} S_{k,n}(z) = 0$, where

$$S_{k,n}(z) = n \sum_{j=2}^{[n/k]} (t_{j1} + \ldots + t_{j4})$$

and t_{jk}, for $1 \le j \le 4$, runs over the values of $P(\pm\varepsilon_1 K_1 > a_n z,\, \pm\varepsilon_j K_j > a_n z)$ for the four different combinations of the $\pm$ signs; and

(b) it holds that

$$\lim_{\tau\to 0} \limsup_{n\to\infty} \frac{n}{a_n^2} \sum_{j=2}^{n} \left|\mathrm{Cov}\left\{\varepsilon_1 K_1\, I(|\varepsilon_1 K_1| \le a_n\tau),\ \varepsilon_j K_j\, I(|\varepsilon_j K_j| \le a_n\tau)\right\}\right| = 0\,.$$

To prove (a), note that by (3.2) and (3.4) we have for any $z > 0$,

$$\lim_{n\to\infty} n\, P(|\varepsilon_1 K_1| > a_n z) = z^{-\alpha}\,. \tag{A.6}$$

Hence, from the definition of $\rho(j)$ it follows that for $C > 0$ sufficiently large,

$$|S_{k,n}(z)| \le 4n \sum_{j=2}^{[n/k]} \{P(|\varepsilon_1 K_1| \ge a_n z)\}^2 \{1 + \rho(j-1)\} \le \frac{C}{n} \sum_{j=2}^{[n/k]} \{1 + \rho(j-1)\} \le 2C/k\,,$$

which implies (a) .

To prove (b), first note that in view of the regular variation of the distribution of ε_t,

$$\frac{a_n^2\, P(|\varepsilon_1 K_1| \ge a_n\tau)}{E\{\varepsilon_1^2 K_1^2\, I(|\varepsilon_1 K_1| \le a_n\tau)\}} \to \alpha/\{(2-\alpha)\,\tau^2\}$$

as $n \to \infty$. From this result, (A.6) and the fact that

$$\begin{aligned}\left|\mathrm{Cov}\left\{\varepsilon_1 K_1\, I(|\varepsilon_1 K_1| \le a_n\tau)\,,\ \varepsilon_j K_j\, I(|\varepsilon_j K_j| \le a_n\tau)\right\}\right| \\ \le\ \rho(j-1)\, E\left\{\varepsilon_1^2 K_1^2\, I(|\varepsilon_1 K_1| \le a_n\tau)\right\}\,,\end{aligned}$$

we may prove that as $n \to \infty$,

$$\frac{n}{a_n^2}\, E\left\{\varepsilon_1^2 K_1^2\, I(|\varepsilon_1 K_1| \le a_n\tau)\right\} \to \frac{2-\alpha}{\alpha}\,\tau^{2-\alpha}.$$

Claim (b) follows from this result and condition (C7).

References

Adler, R.J., Feldman, R.E. and Gallagher, C. (1997). Analysing stable time series. In: *A User's Guide to Heavy Tails: Statistical Techniques For Analysing Heavy Tailed Distributions and Processes*, Ed. R.J. Adler, R.E. Feldman and M. Taqqu. Birkhäuser, Boston.

Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1987). *Regular Variation.* Cambridge University Press.

Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods.* Springer, New York.

Cambanis, S. and Soltani, A.R. (1984). Prediction of stable processes: spectral and moving average representations. *Z. Wahrscheinlichkeitsteorie verw. Geb.* **66**, 593-612.

Chu, C.K. and Marron, J.S. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.* **19**, 1906–1918.

Cline, B.H. and Brockwell, P.J. (1985). Linear prediction of ARMA processes with infinite variance. *Stoch. Processes Appl.* **19**, 281-96.

Cox, D.R. (1966). The null distribution of the first serial correlation coefficient. *Biometrika* **53**, 623-26.

Csörgő, S. and Mielniczuk, J. (1995a). Close short-range dependent sums and regression estimation. *Acta Sci. Math. (Szeged)* **60**, 177–196.

Csörgő, S. and Mielniczuk, J. (1995b). Nonparametric regression under long-range dependent normal errors. *Ann. Statist.* **23**, 1000–1014.

Csörgő, S. and Mielniczuk, J. (1995c). Distant long-range dependent sums and regression estimation. *Stoch. Processes Appl.* **59**, 143–155.

Davis, R.A. (1983). Stable limits for partial sums of dependent random variables. *Ann. Probab.* **11**, 262-69.

Davis, R.A., Knight, K. and Liu, J. (1992). M-estimation for autoregressions with infinite variances. *Stoch. Processes Appl.* **40**, 145-80.

Deo (1997). Nonparametric regression with long-memory errors. *Statist. Probab. Lett.* **18**, 385–393.

Doukhan, P. (1994). *Mixing.* Springer, New York.

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

Fan, J. and Hall, P. (1994). On curve estimation by minimizing mean absolute deviation and its implications. *Ann. Statist.* **22**, 867–885.

Fan, J., Hu, T.-C. and Truong, Y.K. (1994). Robust nonparametric function estimation. *Scand. J. Statist.* **21**, 433–446.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications.* (2nd Edition.) John Wiley and Sons, New York.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989). *Nonparametric Curve Estimation From Time Series.* Springer, Berlin.

Härdle, W. (1984). Robust regression function estimation. *J. Multivariate Anal.* **14**, 169–180.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Cambridge, UK.

Härdle, W. and Gasser, T. (1984). Robust nonparametric function fitting. *J. Roy. Statist. Soc. Ser. B* **46**, 42–51.

Hall, P. and Hart, J.D. (1990). Nonparametric regression with long-range dependence. *Stoch. Processes Appl.* **36**, 339–351.

Hannan, E.J. and Kanter, M. (1977). Autoregressive processes with infinite variance. *J. Appl. Probab.* **14**, 411-15.

Hart, J.D. (1984). Efficiency of a kernel density estimator under an autoregressive dependence model. *J. Amer. Statist. Assoc.* **79**, 110–117.

Hart, J.D. and Vieu, P. (1984). Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.* **18**, 873–890.

LePage, R., Woodroofe, M. and Zinn, J. (1981). Convergence to a stable distribution via order statistics. *Ann. Probab.* **9**, 624-32.

Leung, D.H.Y., Marriott, F.H.C. and Wu, E.K.H. (1993). Bandwidth selection in robust smoothing. *J. Nonparametr. Statist.* **2**, 333–339.

Mallows, C.L. (1980). Some theory of nonlinear smoothers. *Ann. Statist.* **8**, 695–715.

Mikosch, T., Gadrich, T., Kluppelberg, C. and Adler, R.J. (1995). Parametric estimation for ARMA models with infinite variance innovations. *Ann. Statist.* **23**, 305-326.

Pham, T.D. and Tran, L.T. (1985). Some mixing properties of time series models. *Stoch. Processes Appl.* **19**, 297-303.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators *Econometric Theory* **7**, 186-98.

Ray, B. and Tsay, R. (1996). Bandwidth selection for nonparametric regression with long range dependent errors. In: *Athens Conference in Applied Probability and Times Series Analysis*. Eds. P.M. Robinson and M. Rosenblatt **2**, 339–351. Springer, New York.

Resnick, S.I. (1997). Heavy tail modeling and teletraffic data (with discussion). *Ann. Statist.* **25,** 1805-1869.

Robinson, P.M. (1983). Nonparametric estimators for time series. *J. Time Series Anal.* **4**, 185–207.

Robinson, P.M. (1984). Robust nonparametric autoregression. In: *Robust and Nonlinear Time Series Analysis*. Eds. J. Franke, W. Härdle and D. Martin, 247-255. Springer, New York.

Robinson, P.M. (1994). Rates of convergence and optimal bandwidth choice for long-range dependence. *Probab. Theory Related Fields* **99**, 443–473.

Robinson, P.M. (1997). Large-sample inference for nonparametric regression with dependent errors. *Ann. Statist.* **25**, 2054–2083.

Roussas, G.G. and Tran, L.T. (1992). Asymptotic normality of the recursive estimates under dependence conditions, and time series. *Ann. Statist.* **20**, 98–120.

Roussas, G.G., Tran, L.T. and Ioannides, O.A. (1992). Fixed design regression for time series: asymptotic normality. *J. Multivariate Anal.* **40**, 262–291.

Ruppert, D. and Wand, M.P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–1370.

Samorodnitsky, G. and Taqqu, M.S. (1994). *Stable Non-Gaussian Random Processes.* Chapman & Hall, New York.

Smith, M. and Kohn, R. (1996). Bayesian robust nonparametric regression. *Proc. Bayesian Statist. Sci. Sect. Amer. Statist. Assoc.* 202–207.

Tran, L.T., Roussas, G.G., Yakowitz, S. and Truong Van, B. (1996). Fixed-design regression for linear time series. *Ann. Statist.* **24**, 975–991.

Truong, Y.K. (1989). Asymptotic properties of kernel estimators based on local medians. *Ann. Statist.* **17**, 606–617.

Truong, Y.K. (1991). Nonparametric curve estimation with time series errors. *J. Statist. Plann. Inf.* **28**, 167–183.

Truong, Y.K. (1992a). Robust nonparametric regression in time series. *J. Multivar. Anal.* **41**, 163–177.

Truong, Y.K. (1992b). A nonparametric approach for time series analysis. *IMA Volumes in Mathematics and its Applications* **45**, 371–386.

Truong, Y.K. (1994). Nonparametric time series regression. *Ann. Inst. Statist. Math.* **46**, 279–293.

Truong, Y.K. and Stone, C.J. (1992). Nonparametric function estimation involving time series. *Ann. Statist.* **20**, 77–97.

Tsybakov, A.B. (1986). Robust reconstruction of functions by local-approximation method. *Problems Inform. Transmission* **22**, 133–146.

Velleman, P.F. (1980). Definition and comparison of robust nonlinear data smoothing algorithms. *J. Amer. Statist. Assoc.* **75**, 609–615.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman & Hall, London.

Wang, F.T. (1994). Automatic smoothing parameter selection for robust nonparametric regression. *Proc. Statist. Comput. Sect. Amer. Statist. Assoc.* 106–111.

Wang, F.T. and Scott, D.W. (1994). The L_1 method for robust nonparametric regression. *J. Amer. Statist. Assoc.* **89**, 65–76.

Welsh, A.H. (1996). Robust estimation of smooth regression and spread functions and their derivatives. *Statistica Sinica* **6**, 347–366.

Yakowitz, S. (1985a). Nonparametric density estimation, prediction, and regression for Markov sequences. *J. Amer. Statist. Assoc.* **80**, 215–221.

Yakowitz, S. (1985b). Markov flow models and the flood warning problem. *Water Resources Res.* **21**, 81–88.

Yakowitz, S. (1987). Nearest neighbour methods for time series analysis. *J. Time Ser. Anal.* **18**, 1–13.

Zolotarev, V.M. (1986). *One-Dimensional Stable Distributions.* American Mathematical Society, Providence, RI.