Long-Term Prediction Intervals of Time Series

Zhou Zhou, Zhiwei Xu, and Wei Biao Wu

***Abstract*—We consider the problem of predicting aggregates or sums of future values of a process based on its past values. In contrast with the conventional prediction problem in which one predicts a future value given past values of the process, in our setting the number of aggregates can go to infinity with respect to the number of available observations. Consistency and Bahadur representations of the prediction estimators are established. A simulation study is carried out to assess the performance of different prediction estimators.**

***Index Terms*—Empirical quantiles, heavy tails, long-memory, long-run prediction, quenched central limit theory.**

I. Introduction

PREDICTION or forecasting of future values of a random process is one of the fundamental objectives in the study of time series. Let $(X_t)_{t\in\mathbb{Z}}$ be a stochastic process. Given the observations $X_1,\ldots,X_n$, one is interested in predicting future values $X_{n+j}, j\geq 1$. If $(X_t)_{t\in\mathbb{Z}}$ is stationary with $\mathbb{E}(X_t^2)<\infty$, then one can apply the celebrated Kolmogorov-Wiener theory to estimate the conditional mean $\mathbb{E}(X_{n+j}|X_1,\ldots,X_n)$. Since [11], there have been substantial progresses on the estimation theory of the conditional mean $\mathbb{E}(X_{n+j}|X_1,\ldots,X_n)$, the conditional distribution $[X_{n+j}|X_1,\ldots,X_n]$, or their variants; see, for example, [2], [3], [34], [36], [18], [19], [29]–[33], [37], and [48]. [30], [31] applied the tool of stopping times to estimate conditional expectations. Other contributions can be found in [6], [8], [17], [35] and [20] among others.

In this paper, we consider predicting $X_{n+1}+\cdots+X_{n+m}$, sum of future values, based on the past observations $X_1,\ldots,X_n$. In our setup we allow $m=m_n\to\infty$ as $n\to\infty$. Our formulation is attractive in situations in which one is interested in long-term prediction. For example, in telecommunication, engineers may have automatically collected by the minute or second time series data which represents a number of downloads by users every minute or second. However, at the management level, people are more interested in predicting numbers of downloads for a much longer time scale, say weekly, monthly, or even yearly. Let $X_1,X_2,\ldots,X_n$ be a minute observations and $m=7\times 24\times 60=10\,080$. Then $X_{n+1}+\cdots+X_{n+m}$ corresponds to the number of downloads in the upcoming week. Prediction of $X_{n+1}+\cdots+X_{n+m}$ may help implementing price policy of telecommunication services. Recently, the *Time Warner Cable, Inc.,* is planning to implement a price policy which is based on Internet usage and download volumes rather than a flat monthly fee [1]. To design a reasonable price policy, one needs to have a good prediction of $X_{n+1}+\cdots+X_{n+m}$.

Our framework is different from the one in [32] in that the latter paper deals with binary renewal processes and the number of terms can be random, while in our setting the number of terms is treated as nonrandom. On the other hand, we allow nonbinary processes which can be long memory and heavy-tailed.

As a mathematical framework, we consider the construction of prediction intervals for $X_{n+1}+\cdots+X_{n+m}$ given $X_1,\ldots,X_n$. Specifically, we need to find a random interval $[L,U]$, where L and U are functions of $X_1,\ldots,X_n$, such that

$$\mathbb{P}(L\leq X_{n+1}+\cdots+X_{n+m}\leq U|X_1,\ldots,X_n)=1-\alpha \tag{1}$$

where $1-\alpha$ is a preassigned coverage level. In practice, one typically uses $\alpha=0.01$ or 0.05. Clearly, the problem of finding such L and U involves the estimation of conditional distributions of $X_{n+1}+\cdots+X_{n+m}$ given $X_1,\ldots,X_n$. With the estimated interval $[L,U]$, one can assess uncertainty of future aggregates and then adopt appropriate price policies.

Note that the problem of constructing prediction interval $[L,U]$ is different from the one of finding confidence intervals for the conditional mean $\mathbb{E}(X_{n+1}+\cdots+X_{n+m}|X_1,\ldots,X_n)$. The former one reveals more detailed distributional information by estimating conditional distributions. See [21] and [44] for discussions of various statistical intervals. [12] argued that there is an urgent need to implement interval estimates instead of point estimates for sales forecasts of business firms. See also [9, Ch. 7].

In our setting, we let $m\to\infty$ as $n\to\infty$. Our framework is very different from the classical one in which one assumes $m=1$ as far as the methods of finding L and U are concerned. If $m=1$, then one needs to estimate the conditional distribution of X_{n+1} given $X_1,\ldots,X_n$. The latter problem is closely related to the Value-at-Risk (VaR) estimation problem; see [28] RiskMetrics Technical Report. In a typical conditional VaR estimation problem, analogously to (1), one seeks to find a number V which depends on $X_1,\ldots,X_n$, such that

$$\mathbb{P}(X_{n+1}>V|X_1,\ldots,X_n)=\alpha. \tag{2}$$

In other words, V is the conditional $(1-\alpha)$th quantile of X_{n+1} given $X_1,\ldots,X_n$. VaR is a very important measure for assets risks and it basically deals with the question that investors will be asking: "given historical information, how much can I lose with probability α over a preset horizon". It is a fundamental

Manuscript received February 23, 2008; revised March 27, 2009. Current version published March 10, 2010.

Z. Zhou is with the Department of Statistics, University of Toronto, Toronto, ON M5S3G3, Canada (e-mail: zhou@utstat.toronto.edu).

W. B. Wu is with the Department of Statistics, University of Chicago, Chicago, IL 60637 USA (e-mail: wbwu@galton.uchicago.edu).

Z. Xu is with the Department of Computer and Information Science, University of Michigan-Dearborn, Dearborn, MI 48128 USA (e-mail: zwxu@umich.edu).

Communicated by A. Krzyżak, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2009.2039158

quantity in financial risk managements. Estimation of VaR is a very challenging problem and one may generally need to apply nonlinear prediction theory; see [13, Ch. 10].

It turns out that, interestingly, estimation of L and U in the framework of (1) with $m \to \infty$ is relatively easier for certain class of processes. This is due to the so-called quenched or conditional central limit theory; see [43] for some recent developments. As argued in Section II-A , if the process (X_k) is weakly dependent, then the impact of $X_1, \ldots, X_n$ on the sum $X_{n+1} + \cdots + X_{n+m}$ is asymptotically negligible and one has the approximate relation

$$\mathbb{P}(X_{n+1} + \cdots + X_{n+m} \le x | X_1, \ldots, X_n) \approx \mathbb{P}(X_{n+1} + \cdots + X_{n+m} \le x) \quad (3)$$

when m is large. Let $l < u$ be two real numbers such that

$$\mathbb{P}(l \le X_{n+1} + \cdots + X_{n+m} \le u) = 1 - \alpha. \quad (4)$$

In many problems approximate solutions l and u can be obtained asymptotically or empirically; see Section II. Based on (3), we can choose L and U as l and u, respectively, so that they provide an approximate solution to (1).

We now impose structural assumptions on X_i so that we can interpret in what sense (3) holds and then utilize (3). In particular, we shall consider the long-term prediction for the linear model

$$X_i = w_i^T \beta + e_i \quad (5)$$

where ^{T} denotes the matrix transpose, (e_i) is a mean zero stationary process, β is a $p \times 1$ unknown regression coefficient vector and w_i are known $p \times 1$ covariates, explanatory variables or design vectors.

The rest of the paper is organized as follows. As a premier, Section II concerns the special case of model (5) in which $w_i^T \beta = 0$, namely there are no covariates involved. Prediction of the general linear model (5) is considered in Section III. A simulation study is carried out and we compare the performance of two different predicting estimators. In Section III-B we apply our estimation procedure to a telecommunication network traffic dataset. Proofs of results in Sections II and III are given in Section IV.

II. Quantiles of Sums of Stationary Processes

To illustrate the idea behind (3), we let $w_i^T \beta = 0$ and assume that (e_i) is a mean zero stationary process with finite second moment $\mathbb{E}(e_i^2) < \infty$. Let $S_m = e_1 + \cdots + e_m$ and $\mathcal{F}_i = (e_i, e_{i-1}, \ldots)$. Under this setting, by stationarity, it suffices to establish the following version of (3)

$$\mathbb{P}\left(\frac{S_m}{\sqrt{m}} \le x | \mathcal{F}_0\right) \approx \mathbb{P}\left(\frac{S_m}{\sqrt{m}} \le x\right) \quad (6)$$

when m is large. Let $\Phi(u) = \int_{-\infty}^{u} (2\pi)^{-1/2} e^{-x^2/2} dx$ be the standard normal distribution function. For a random variable X, we write $X \in \mathcal{L}^p$, $p > 0$, if $\|X\|_p := [\mathbb{E}(|X|^p)]^{1/p} < \infty$. For two distributions F and G on $\mathbb{R}$, define the Levy metric as shown in (7) at the bottom of the page.

A. Quenched Central Limit Theory

Reference [43] proved the following conditional or quenched central limit theorem: Assume that $\mathbb{E}(|e_i|^p) < \infty$ for some $p > 2$ and, for some $q > 5/2$

$$\|\mathbb{E}(S_m | \mathcal{F}_0)\|_2 = O\left(\frac{\sqrt{m}}{\log^q m}\right). \quad (8)$$

Then we have the almost sure convergence

$$\Delta\left[N(0, \sigma^2), \mathbb{P}\left(\frac{S_m}{\sqrt{m}} \le \cdot | \mathcal{F}_0\right)\right] \to 0 \quad (9)$$

as $m \to \infty$, where $\sigma^2 = \lim_{m \to \infty} \|S_m\|_2^2 / m$ is the long-run variance. Namely, for almost all realizations of $\mathcal{F}_0$, if $\sigma > 0$, we have

$$\lim_{n \to \infty} \sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{S_m}{\sqrt{m}} \le x | \mathcal{F}_0\right) - \Phi\left(\frac{x}{\sigma}\right) \right| = 0. \quad (10)$$

Convergence in the stronger form of invariance principle is also valid; see [43, Corollary 3]. As argued in the latter paper, under (8), we also have the unconditional central limit theorem $S_m / \sqrt{m} \Rightarrow N(0, \sigma^2)$, or

$$\lim_{n \to \infty} \sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{S_m}{\sqrt{m}} \le x\right) - \Phi\left(\frac{x}{\sigma}\right) \right| = 0. \quad (11)$$

Clearly, (10) and (11) imply that not only (6) holds in the sense of (12) shown at the bottom of the page but also both

$$\Delta(F, G) = \inf\{\delta > 0 : F(x - \delta) - \delta \le G(x) \le F(x + \delta) + \delta \text{ holds for all } x \in \mathbb{R}\}. \quad (7)$$

$$\lim_{n \to \infty} \sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{S_m}{\sqrt{m}} \le x | \mathcal{F}_0\right) - \mathbb{P}\left(\frac{S_m}{\sqrt{m}} \le x\right) \right| = 0 \text{ almost surely} \quad (12)$$

$\mathbb{P}(S_m/\sqrt{m} \le x|\mathcal{F}_0)$ and $\mathbb{P}(S_m/\sqrt{m} \le x)$ can be approximated by $N(0,\sigma^2)$. The latter observation is in striking contrast with the construction of prediction intervals when $m = 1$, in which case the conditional and unconditional versions are quite different (see in [9, Sec. 7.4]). Additionally, (10) and (11) also suggest an approximate solution of L and U to (1) in the following form:

$$L, U = \pm\hat{\sigma} z_{\alpha/2}\sqrt{m} \tag{13}$$

where $z_{\alpha/2}$ is the $(\alpha/2)$-th quantile of the standard normal distribution and $\hat{\sigma}$ is an estimate of σ. A popular estimate of σ^2 is the lag window estimate

$$\hat{\sigma}^2 = \sum_{k=-k_n}^{k_n} \hat{\gamma}_k, \tag{14}$$

where k_n is the bandwidth sequence satisfying $k_n \to \infty$ and $k_n/n \to 0$, and $\hat{\gamma}_k$ is an estimate of γ_k:

$$\hat{\gamma}_k = \frac{1}{n}\sum_{i=1}^{n-|k|}(e_i - \bar{e})(e_{i+|k|} - \bar{e}), \quad \bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i. \tag{15}$$

For details see [4] or [8].

An interesting and useful feature of the prediction interval (13) is that one does not need to fit the underlying probability model for the process (e_i). On the other hand, however, the above normal approximation may fail if the process is strongly dependent or has heavy-tailed distributions. It is common that telecommunication time series may exhibit long-range dependence as well as heavy tails; see for example [27]. This is one of the major reasons that Internet Service Providers such as the *Time Warner Cable, Inc.,* are interested in imposing more charges on users who download files with very large sizes. To construct prediction intervals for processes with heavy tails, one way out is to resort to empirically based method which is discussed in detail in the section below. Our simulation study shows that the latter approach outperforms the one based on (13).

B. Quantile Estimates

Condition (8) ensures the normal approximation (9). For strongly dependent processes, however, (8) is violated [43] and (12) may be invalid. In this case, we propose to estimate quantiles of S_m by sample quantiles of $\sum_{j=i-m+1}^{i} e_j$, $i = m, m+1, \ldots$, via a moving window scheme. Specifically, let

$$\tilde{Y}_i = \frac{\sum_{j=i-m+1}^{i} e_j}{H_m}, \quad i = m, m+1, \ldots \tag{16}$$

where $H_m > 0$ is an appropriate normalizing constant such that $\tilde{Y}_i$ has a nondegenerate limiting distribution as $m \to \infty$. Note that $(\tilde{Y}_i)$, $i \in \mathbb{Z}$, is a (triangular array) stationary time series and we can calculate $(\tilde{Y}_i)_m^n$. In order to construct a $(1-\alpha)$ prediction interval for $\tilde{Y}_{n+m}$ based on $(\tilde{Y}_i)_m^n$, we shall estimate the $(1-\alpha/2)$th and $(\alpha/2)$th quantiles of this quantity. More specifically, let $\hat{Q}_n(\alpha/2)$ and $\hat{Q}_n(1-\alpha/2)$ be the $(1-\alpha/2)$th and $(\alpha/2)$th sample quantiles of $(\tilde{Y}_i)_m^n$, then $[\hat{Q}_n(\alpha/2), \hat{Q}_n(1-\alpha/2)]$ is a natural $(1-\alpha)$ prediction interval of $\tilde{Y}_{n+m}$. Therefore $[H_m\hat{Q}_n(\alpha/2), H_m\hat{Q}_n(1-\alpha/2)]$ is a $(1-\alpha)$ prediction interval for $\sum_{j=n+1}^{n+m} e_j$. Asymptotic properties of $\hat{Q}_n$ for short- and long-range dependent linear processes are dealt with in Sections II-B1 and II-B2, respectively.

Short-Range Dependent (SRD) Processes: To obtain asymptotic properties of $\hat{Q}_n(\alpha/2)$ and $\hat{Q}_n(1-\alpha/2)$, here we assume that (e_i) is a one-sided infinite order moving average $\mathrm{MA}(\infty)$ process:

$$e_i = \sum_{j=0}^{\infty} a_j \varepsilon_{i-j}, \tag{17}$$

where $(\varepsilon_j)_{-\infty}^{\infty}$ is an independent and identically distributed (i.i.d.) sequence having mean 0, and $(a_i)_0^{\infty}$ are real coefficients such that e_i exists almost surely. The existence of (17) can be checked by the Kolmogorov three-series theorem [10]).

The innovations $(\varepsilon_j)_{-\infty}^{\infty}$ can be either light or heavy-tailed. More precisely, we say ε_j is light-tailed if $\mathbb{E}(\varepsilon_j^2) < \infty$. For heavy-tailed processes, we consider ε_j which belongs to α-stable domain of attraction $\mathcal{D}(\alpha)$ for some $\alpha \in (1,2)$, namely the normalized partial sum process of ε_j converges to a stable distribution [10], [14]. For $\varepsilon_j \in \mathcal{D}(\alpha)$, it has the following characterization:

$$\begin{aligned} 1 - F_\varepsilon(t) &= (c_1 + o(1))t^{-\alpha}L(t) \text{ and } F_\varepsilon(-t) \\ &= (c_2 + o(1))t^{-\alpha}L(t) \end{aligned} \tag{18}$$

as $t \to \infty$, where $F_\varepsilon(\cdot)$ is the cumulative distribution function (cdf) of ε_j, $c_1,\ c_2 \ge 0$, $c_1 + c_2 > 0$ and L is a slowly varying function (s.v.f.), namely, $\lim_{x\to\infty} L(tx)/L(x) = 1$ for all $t > 0$; see [14]. Clearly, by (18)

$$g_n := \inf\left\{x : \mathbb{P}(|\varepsilon_i| > x) \le \frac{1}{n}\right\} = n^{1/\alpha}L_1(n)$$

where L_1 is also a s.v.f. Observe that $\mathbb{E}(|\varepsilon_i|^{\alpha'}) < \infty$ for all $\alpha' \in (0,\alpha)$, and α is called the heavy tail index, and $\mathbb{E}(\varepsilon_i^2) = \infty$.

We shall let the normalizing constant $H_m = \sqrt{m}$ if $\mathbb{E}(\varepsilon_j^2) < \infty$ and $H_m = g_m$ if $\varepsilon_j \in \mathcal{D}(\alpha)$, $1 < \alpha < 2$. By the central limit theorem of SRD linear processes (see, for example, [5]), we have

$$\tilde{Y}_i \Rightarrow Z \text{ as } m \to \infty \tag{19}$$

where Z is Gaussian if $\mathbb{E}(\varepsilon_j^2) < \infty$ and Z is α-stable if $\varepsilon_j \in \mathcal{D}(\alpha)$.

We shall impose the following conditions:

(SRD) $\sum_{i=0}^{\infty} |a_i| < \infty$;

(DEN) $\sup_{x\in\mathbb{R}}(f_\varepsilon(x) + |f_\varepsilon'(x)|) < \infty$, where $f_\varepsilon(\cdot)$ is the density of ε_i.

Condition (SRD) is a classic short-range dependence or stability condition for linear processes (see Box, [6]). Condition (DEN) appears quite mild. For example, by the inversion theorem, it allows the symmetric-α-stable distribution whose characteristic function is of the form $\mathbb{E}\exp(\sqrt{-1}\varepsilon_i u) = \exp(-|u/\sigma|^\alpha)$, $u \in \mathbb{R}$, where $\sqrt{-1}$ is the imaginary unit and $\sigma > 0$ is the scale parameter.

Let $\tilde{Q}_q, 0<q<1$, be the qth quantile of $\tilde{Y}_i$ and $f_m(\cdot)$ be the density of $\tilde{Y}_i$; let

$$\tilde{F}_n(x)=\frac{1}{n-m+1}\sum_{i=m}^{n} I\{\tilde{Y}_i\le x\} \tag{20}$$

be the empirical distribution function of $(\tilde{Y}_i)_m^n$. We have the following two theorems regarding the asymptotic behavior of $\hat{Q}_n$ in the light- and heavy-tailed cases, respectively.

Theorem 1: Suppose conditions (SRD) and (DEN) hold. Further assume $\mathbb{E}(\varepsilon_j^2)<\infty$ and $m^3/n\to 0$. Then we have for any fixed $q\in(0,1)$, (i) $|\hat{Q}_n(q)-\tilde{Q}_q|=O_p(m/\sqrt{n})$; (ii)

$$\left|\hat{Q}_n(q)-\tilde{Q}_q-\frac{q-\tilde{F}_n(\tilde{Q}_q)}{f_m(\tilde{Q}_q)}\right| = O_p\left(\frac{m^{5/2}}{n}+\left(\frac{m}{n}\right)^{3/4}\log^{1/2} n\right). \tag{21}$$

Theorem 2: Suppose conditions (SRD) and (DEN) hold. Further assume $\varepsilon_j\in\mathcal{D}(\alpha)$ for some $1<\alpha<2$ and $m=O(n^\gamma)$ for some $\gamma<(\alpha-1)/(\alpha+1)$. Then we have for any fixed $q\in(0,1)$, (i) $|\hat{Q}_n(q)-\tilde{Q}_q|=O_p(mn^v)$ for all $v>1/\alpha-1$; (ii) for all $v>1/\alpha-1$

$$\left|\hat{Q}_n(q)-\tilde{Q}_q-\frac{q-\tilde{F}_n(\tilde{Q}_q)}{f_m(\tilde{Q}_q)}\right| = O_p(H_m m^2 n^{2v}+\sqrt{H_m m n^{v-1}}\log^{1/2} n). \tag{22}$$

Theorems 1 and 2 establish convergence rates of the quantile estimate $\hat{Q}_n(q)$ under short memory. In particular, Theorems 1(i) and 2(i) assert consistency of $\hat{Q}_n(q)$ under $m=o(n^{1/3})$ and $m=O(n^\gamma)$ with $\gamma<(\alpha-1)/(\alpha+1)$, respectively. The Bahadur representations (21) and (22) reveal deep insights into the asymptotic properties of $\hat{Q}_n(q)$ by approximating it by additive forms.

Long-Range Dependent (LRD) Processes: It is common in the literature to call process (17) long memory or long range dependent if the coefficients $(a_i)_0^\infty$ are not absolutely summable or in other words, if $\sum_{i=0}^{\infty}|a_i|=\infty$. In this section, we shall consider the following decay of the series $(a_i)_0^\infty$:

(LRD) $a_i=i^{-\gamma}l(i)$ $i=1,2,\ldots$, $\lambda<\gamma<1$, where $l(\cdot)$ is a s.v.f. and $\lambda=1/2$ if $\mathbb{E}(\varepsilon_i^2)<\infty$ and $\lambda=1/\alpha$ if $\varepsilon_i\in\mathcal{D}(\alpha)$, $1<\alpha<2$.

That $\lambda<\gamma$ is necessary for the almost sure convergence of (17) and the other constraint $\gamma<1$ is to guarantee that $(a_i)_0^\infty$ are not absolutely summable and hence long memory of the series (e_i). An important class of models which satisfy condition (LRD) is the the fractionally integrated ARIMA (FARIMA) processes [23]).

In the long memory case, define the normalizing constants $H_m=m^{3/2-\gamma}l(m)$ if $\mathbb{E}(\varepsilon_i^2)<\infty$ and $H_m=g_m m^{1-\gamma}l(m)$ if $\varepsilon_i\in\mathcal{D}(\alpha)$ for some $\alpha\in(1,2)$. Then central limit results of (e_i) holds under the above normalization in the sense of (19). See [39] and [5]. Analogously to Theorems 1 and 2, we have the following theorems in the LRD case. The latter theorems are different from the ones in the SRD case in that γ, the parameter controlling the strength of dependence, is needed in the condition of m.

Theorem 3: Assume conditions (LRD) and (DEN), $\mathbb{E}(\varepsilon_j^2)<\infty$ and $m^{5/2-\gamma}n^{1/2-\gamma}l^2(n)\to 0$. Then for any fixed $q\in(0,1)$, we have 1) $|\hat{Q}_n(q)-\tilde{Q}_q|=O_p(mn^{1/2-\gamma}|l(n)|)$; and 2)

$$\left|\hat{Q}_n(q)-\tilde{Q}_q-\frac{q-\tilde{F}_n(\tilde{Q}_q)}{f_m(\tilde{Q}_q)}\right| = O_p\Big(\frac{m^{7/2-\gamma}|l(n)|^3}{n^{2\gamma-1}}+\frac{m^{5/4-\gamma/2}|l(n)|\log^{1/2} n}{n^{1/4+\gamma/2}}\Big). \tag{23}$$

Theorem 4: Assume that conditions (LRD) and (DEN) hold; that $\varepsilon_i\in\mathcal{D}(\alpha)$ for some $\alpha\in(1,2)$ and that $m=O(n^\kappa)$ for some $\kappa<(\alpha\gamma-1)/(2\alpha+1-\alpha\gamma)$. Then we have for any fixed $q\in(0,1)$, i) $|\hat{Q}_n(q)-\tilde{Q}_q|=O_p(mn^v)$ for all $v>1/\alpha-\gamma$; ii) for all $v>1/\alpha-\gamma$

$$\left|\hat{Q}_n(q)-\tilde{Q}_q-\frac{q-\tilde{F}_n(\tilde{Q}_q)}{f_m(\tilde{Q}_q)}\right| = O_p(H_m m^2 n^{2v}+\sqrt{H_m m n^{v-1}\log n}). \tag{24}$$

Remark 1: As discussed in the beginning of Section II-B, $H_m\hat{Q}_n(q)$ is an estimate of the qth quantile of $\sum_{i=n+1}^{n+m}e_i$, which equals to $H_m\tilde{Q}_q$. It is easy to check by i) of Theorems 1–4 that $H_m\hat{Q}_n(q)-H_m\tilde{Q}_q=o_p(1)$. In other words, $H_m\hat{Q}_n(q)$ is a weakly consistent estimate of the corresponding quantile of $\sum_{i=n+1}^{n+m}e_i$. On the other hand, if one constructs prediction interval for $\sum_{i=n+1}^{n+m}e_i$ based on (13), then one uses $\sqrt{m}\hat{\sigma}z_q$ as an estimate of the qth quantile of $\sum_{i=n+1}^{n+m}e_i$, where z_q is the qth quantile of the standard normal distribution. In this case it can be shown that $\sqrt{m}\hat{\sigma}z_q-H_m\tilde{Q}_q=O_p(1)$ based on Edgeworth's expansion. Therefore, when m is not growing too fast, it is more desirable to use the quantile estimation method. Simulation studies in Section III-A further confirm this claim.□

III. Prediction of Linear Models

Consider now model (5). We shall predict $X_{n+1}+\cdots+X_{n+m}$ based on $X_1,X_2,\ldots,X_n$. The latter observations can be used to estimate the unknown parameter vector β. Specifically, let $W=(w_1,\ldots,w_n)^T$ be the design matrix and $\mathbf{X}_n=(X_1,X_2,\ldots,X_n)^T$. Then the least squares estimate (LSE) of β has the form

$$\hat{\beta}_{ls}=(W^TW)^{-1}W^T\mathbf{X}_n. \tag{25}$$

When the errors (e_i) are heavy-tailed, it is more desirable to use robust estimation procedure [24, (Huber, 1981)]. Therefore in this situation we suggest using the least absolute deviation (LAD) estimate of β; namely let

$$\hat{\beta}_{\text{lad}}=\arg\min_{\beta}\sum_{i=1}^{n}|X_i-w_i^T\beta|. \tag{26}$$

The LAD estimation is equivalent to the median regression for which fast and stable algorithms are available; see [26].

In both the LSE and LAD cases, the estimated residuals can be written as

$$\hat{e}_i = X_i - w_i^T\hat{\beta}, \quad i = 1, 2, \ldots, n. \tag{27}$$

It is hoped that the procedures proposed in Sections II-A and II-B can be applied to the residuals and hence the prediction interval for $X_{n+1}+\cdots+X_{n+m}$ can be obtained. More precisely, we propose the following procedure:

i) Use the LAD procedure to obtain $\hat{\beta}$ and $\hat{e}_i$, $i = 1, 2, \ldots, n$.
ii) Let $\check{Y}_i = \sum_{j=i-m+1}^{i} \hat{e}_j$, $i = m, \ldots, n$. Obtain the $(\alpha/2)$th and $(1-\alpha/2)$th empirical quantiles of $(\check{Y}_i)_m^n$ and denote them by $\check{Q}_n(\alpha/2)$ and $\check{Q}_n(1-\alpha/2)$, respectively.
iii) A $(1-\alpha)$ prediction interval for $\sum_{i=n+1}^{n+m} X_i$ can be constructed as $\sum_{i=n+1}^{n+m} w_i^T\hat{\beta} + [\check{Q}_n(\alpha/2), \check{Q}_n(1-\alpha/2)]$.

Note that the LSE can be used in step i) when the errors are light-tailed in the hope of gaining some improvements. Residual quantile-quantile (QQ) plots can be used to check the tail behavior of the errors (e_i). Since in the regression case we have to estimate the errors (e_i) by the residuals $(\hat{e}_i)$, we need to investigate whether the consistency property of the empirical quantiles listed in Theorems 1–4 still hold in this case.

Let $\bar{Y}_i = \sum_{j=i-m+1}^{i} \hat{e}_j/H_m$, $i = m, \ldots, n$, where H_m is defined as in Section II-B; let $\bar{Q}_n(q)$ be the qth empirical quantile of $(\bar{Y}_i)_m^n$. We have the following theorem regarding the asymptotic behavior of $\bar{Q}_n(q)$.

Theorem 5: Let $\Sigma_n = W^TW$. Assume that (a) there exists constants $0 < C_s < C_l < \infty$, such that $C_s < \lambda_1 \le \lambda_n < C_l$ for all large n, where $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ are the eigenvalues of Σ_n/n; that (b) there exists a constant $C^* < \infty$, such that $\max_{1\le i\le n}|w_i| \le C^*$ for all large n; that (c) $f_e(0) > 0$, where $f_e(\cdot)$ is the density function of e_i. Then conclusion (i) in Theorems 1 to 4 holds with $\hat{Q}_n(q)$ therein replaced by $\bar{Q}_n(q)$.

For the linear process (17), we define

$$\alpha^* = \begin{cases} 2, & \text{if } \mathbb{E}(\varepsilon_i^2) < \infty \\ \alpha, & \text{if } \varepsilon_i \in \mathcal{D}(\alpha), 1 < \alpha < 2 \end{cases} \quad \text{and}$$
$$\gamma^* = \begin{cases} 1, & \text{if Condition (SRD) holds} \\ \gamma, & \text{if Condition (LRD) holds.} \end{cases} \tag{28}$$

From the proof of Theorem 5, we see that under the conditions of the latter theorem $\hat{\beta}_{lad} - \beta = O_p(n^v)$ for all $v > 1/\alpha^* - \gamma^*$. Hence $\sum_{i=n+1}^{n+m} w_i^T(\hat{\beta} - \beta) = O_p(mn^v) = o_p(1)$ under conditions of Theorems 1–4. On the other hand, by Theorem 5 and the discussions in Remark 1, $H_m\bar{Q}_n(q) - H_m\tilde{Q}_q = o_p(1)$. Therefore we conclude that the lower bound $\sum_{i=n+1}^{n+m} w_i^T\hat{\beta} + \check{Q}_n(\alpha/2)$ in iii) is a weakly consistent estimator of the $(\alpha/2)$th quantile of $\sum_{i=n+1}^{n+m} X_i$. Analogous conclusion holds for the upper bound.

Remark 2: Conditions (a) and (b) in Theorem 5 on covariates are general enough for many applications. For example, they are satisfied under the design $w_i = (1, (i/n)^a, \cos(i/b), \sin(i/b))^T$ for $a > -1$ and $b > 0$, which are the covariates chosen for the Motorola telecommunication network traffic dataset discussed in Section III-B. On the other hand, by changing condition (b) to $\max_{1\le i\le n}|w_i| \le C^* \log n$ for all large n, we can allow random design (w_i) with exponentially decaying tails. In this case conclusions of Theorem 5 continue to hold as long as we add an extra factor of $\log n$ into the probability bounds of $\bar{Q}_n(q) - \tilde{Q}_q$ there. □

A. A Simulation Study

In Sections II-A and II-B, methods based on quenched central limit theory and quantile estimation are proposed to construct prediction intervals for sums of future values. As discussed in Remark 1, asymptotically, the quantile estimation method is superior. In this section we shall conduct a simulation study and compare the finite sample performance of the above two methods.

Consider linear model (5) with predictors $w_i = (1, \sqrt{i/n}, \cos(2\pi i/24), \sin(2\pi i/24))^T$, $i = 1, 2, \ldots, n$, and regression coefficients $\beta = (3400, 2800, 3500, -771)^T$. The same design (w_i) will also be considered for the Motorola data example in the next section. Let $\sigma = 1000$. We shall investigate the following four scenarios of the error process (e_i):

1) $e_i = 0.6e_{i-1} + \sigma\varepsilon_i$, where ε_i, $i \in \mathbb{Z}$, are i.i.d. mixture normal $\frac{1}{2}N(0,1) + \frac{1}{2}N(0, 1.25)$.
2) $e_i = \sigma\sum_{j=0}^{\infty}(j+1)^{-0.8}\varepsilon_{i-j}$, where ε_i, $i \in \mathbb{Z}$, are i.i.d. mixture normal $\frac{1}{2}N(0,1) + \frac{1}{2}N(0, 1.25)$.
3) $e_i = 0.6e_{i-1} + \sigma\varepsilon_i$, where ε_i, $i \in \mathbb{Z}$, are i.i.d. symmetric α stable with heavy tail index $\alpha = 1.5$ and the scale parameter 1. So its characteristic function is $\exp(-|u|^{1.5})$.
4) $e_i = \sigma\sum_{j=0}^{\infty}(j+1)^{-0.8}\varepsilon_{i-j}$, where ε_i, $i \in \mathbb{Z}$, are i.i.d. symmetric α stable with heavy tail index $\alpha = 1.5$ and the scale parameter 1.

Scenarios (1)–(4) correspond to the cases of light-tailed and SRD, light-tailed and LRD, heavy-tailed and SRD and heavy-tailed and LRD error processes, respectively. In our simulations we choose $n = 6000$ and we are interested in predicting the sum of future $m = 168$ values. We choose the parameters β, σ, m and n in order to mimic the situation in the Motorola data example. For each of the four scenarios, we generate a data set of size $n + m$ from the model described above. It is worth mentioning that, for (2) and (4), using the convolution structure of the linear process (e_i), we can apply the circular embedding and the fast Fourier transform algorithm so that (e_i) can be generated very quickly; see [42]. We perform LAD regression on the first n data points, and then apply both the quenched central limit theory and the quantile estimation methods to the residuals to obtain prediction intervals for $\sum_{i=n+1}^{n+m} X_i$ at 90%, 95%, and 99% levels. Finally we check whether the true value of $\sum_{i=n+1}^{n+m} X_i$ is contained in those intervals. We repeat the above procedure for $N = 10\,000$ times and record the empirical coverage probabilities of the prediction intervals. The results are summarized in Table I.

It is clear from the table that the quantile estimation method is very robust to model assumptions and the coverage probabilities are close to the nominal levels. When the error process is light-tailed and SRD (Scenario (1)), the assumptions of quenched CLT hold. Hence we can see that the latter method works reasonably well in this case. It is surprising to notice that the quenched CLT method gives reasonable results in Scenario

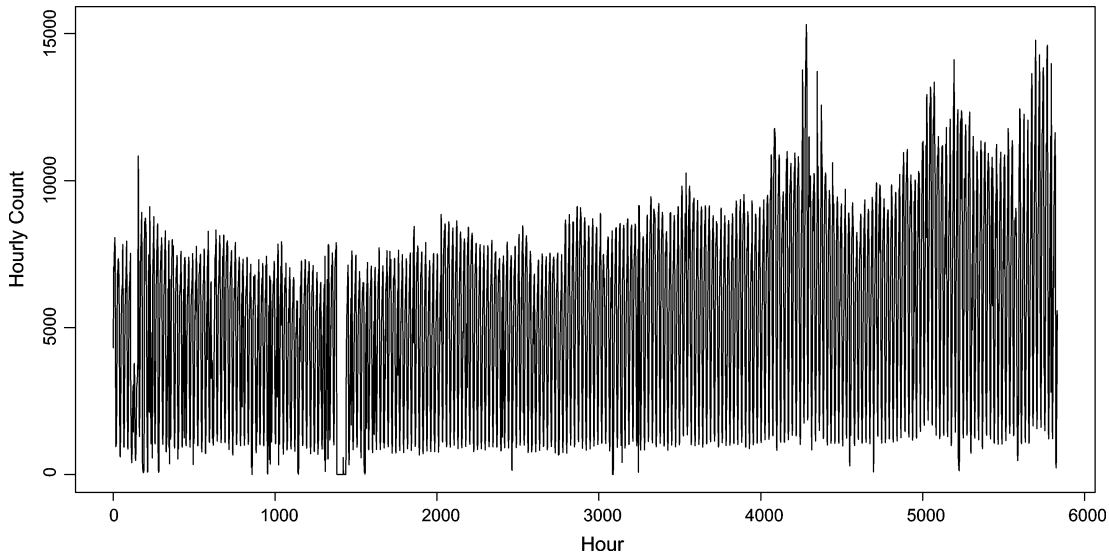

Fig. 1. Time series plot of the hourly counts of the Motorola telecommunication network traffic data.

TABLE I
SIMULATED COVERAGE PROBABILITIES OF THE PREDICTION INTERVALS AT 90%, 95% AND 99% NOMINAL LEVELS

	Quenched CLT			Quantile Estimation		
Scenario	90%	95%	99%	90%	95%	99%
(1)	.860	.922	.981	.867	.923	.971
(2)	.575	.657	.786	.852	.911	.961
(3)	.913	.936	.963	.878	.934	.961
(4)	.726	.786	.863	.860	.919	.958

(3) even though its assumptions are violated. In fact, due to the heavy tails of the error process, procedure (14) tend to overestimate the "true long-run variance". On the other hand, $\sqrt{m}$ in (13) is an underestimate of the true normalizing constant $m^{1/\alpha}L_1(m)$. A possible explanation is that in this very simulation design the over and under estimation effects cancel. The quenched CLT method does not capture the asymptotic distribution of $\sum_{i=n+1}^{n+m} e_i$ in the heavy-tailed case. In fact we conducted another simulation which constructed a 60% prediction interval under Scenario (3). It turned out the quenched CLT method gave a empirical coverage probability of 76% while the quantile estimation method covered 58% of the time. It should be emphasized that the quenched CLT method is not stable when the error process is heavy-tailed and therefore it is not recommended under such circumstances.

B. Prediction of Wireless Network Traffics

In this section we shall apply the quantile estimation procedure to the Motorola telecommunication network traffic dataset which was collected from a mobile infrastructure network deployed in Asia and the United States. The network is designed for mobile users to conduct voice communication and to download digital items (ring tones, wall paper, music, video, games, etc.) to their mobile devices. As multimedia cell phones and pocket PCs become popular, there is an increasing demand for digital items. However, mobile users cannot download digital items directly from the third party content provider (TPCP) network. They need to go through wireless access point provided by their wireless service provider to access TPCP websites. The wireless networks that handle both voice and digital items are more complex, expensive and they require more bandwidth than traditional networks that handle voice only. Knowing the traffic trend is critical to the management and long-term prediction will be useful for resource allocation, maintenance plan and price policy.

We collected the traffic data of the eight months period from 11:00 AM July 8, 2005 to 10:59 AM March 8, 2006, and we obtained 5832 hourly transaction counts (see Fig. 1 for a plot of the hourly counts). Our purpose is to construct a 95% prediction interval for the total usage of the week following this 8-month period. This amounts to predicting the sum of $m = 24 \times 7 = 168$ future values. Fig. 1 gives the time series plot of the hourly counts.

It is noticeable from Fig. 1 there is an abnormal period around hour 1400 (around September 5th 2005) with consecutive low counts. We checked the system log and found that it was due to system outage. Other outliers can be caused by system maintenance, upgrade, etc. Our LAD regression procedure and the quantile estimation method are resistent to the occurrence of outliers [26].

The hourly data exhibits strong periodicity of 24 as the wireless usage peaks at about 8 pm and minimizes at about 5 am every day. There is also a noticeable increasing pattern of the usage. We choose the following regression model:

$$X_i = \beta_1 + \beta_2\sqrt{\frac{i}{5832}} + \beta_3 \cos\left(\frac{2\pi i}{24}\right) + \beta_4 \sin\left(\frac{2\pi i}{24}\right) + e_i, \quad i = 1, 2, \ldots, 5832. \tag{29}$$

The LAD estimates are $\hat{\beta}_1 = 3429$, $\hat{\beta}_2 = 2811.9$, $\hat{\beta}_3 = 3498.7$ and $\hat{\beta}_4 = -771.1$. Fig. 2 below shows the residuals of model (29).

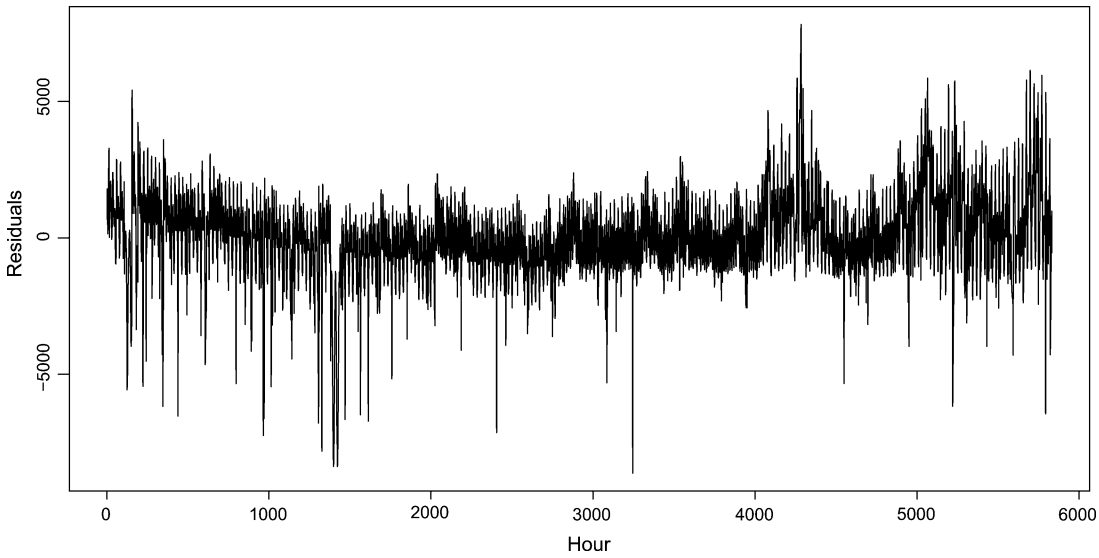

Fig. 2. Residual plot of model (29).

Following steps i)–iii) in Section III, we obtain the 95% prediction interval for the usage of the following week as $[771714.7, 1297852]$. Note again that the latter interval is not the 95% confidence interval for the mean of usage of the following week.

IV. Proofs

Without loss of generality, we fix $a_0 = 1$ in (17) for all the proofs in this section. We first introduce some notation. For a random variable X with finite pth moment, $p > 0$, write $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$ and $\|X\| = \|X\|_2$. For $k \in \mathbb{Z}$ define the projection operator as shown in the equation at the bottom of the page. Recall (16) for $\tilde{Y}_i$. Let $\tilde{Z}_{i-1} = \tilde{Y}_i - \varepsilon_i/H_m$, $i = m, m+1, \ldots, n$. Note that both $\tilde{Y}_i$ and $\tilde{Z}_i$ are $\mathcal{F}_i$ measurable and

$$\tilde{Z}_{i-1} = \frac{\sum_{j=1}^{\infty} \tilde{b}_j \varepsilon_{i-j}}{H_m} \tag{30}$$

where $\tilde{b}_j = a_0 + a_1 + \cdots + a_j$ if $1 \le j \le m-1$ and $\tilde{b}_j = a_{j-m+1} + a_{j-m+2} + \cdots + a_j$ if $j \ge m$.

Define

$$\tilde{F}_n^*(x) = \frac{1}{n-m+1} \sum_{i=m}^{n} F_\varepsilon(H_m(x - \tilde{Z}_{i-1}))$$

where $F_\varepsilon(\cdot)$ is the distribution function of ε. Let $\tilde{F}(x) = \mathbb{P}(\tilde{Y}_i \le x)$. Write

$$\tilde{F}_n(x) - \tilde{F}(x) := M_n(x) + N_n(x)$$

where $M_n(x) = \tilde{F}_n(x) - \tilde{F}_n^*(x)$ and $N_n(x) = \tilde{F}_n^*(x) - \tilde{F}(x)$.

Note that $\tilde{F}(x) = \mathbb{E}\tilde{F}_n(x)$. Observe that

$$M_n(x) = \frac{1}{n-m+1} \sum_{i=m}^{n} \mathcal{P}_i I\{\tilde{Y}_i \le x\}$$

and hence it is a sum of triangular array martingale differences with respect to $\mathcal{F}_i$. On the other hand, note that $N_n(x)$ is differentiable with respect to x. The above properties of $M_n(x)$ and $N_n(x)$ are useful in investigating their oscillation rates. (See Lemmas 1 and 2). In the sequel the symbol C will denote a positive finite constant which may vary from place to place.

For presentational simplicity, here we will only prove Theorems 1 and 4 in Section II-B. Theorems 2 and 3 can be similarly proved without essential extra difficulties.

Lemma 1: Under conditions of Theorems 1, 2, 3, and 4, we have

$$\sup_{|u| \le b_n} |M_n(x+u) - M_n(x)| = O_p\left(\sqrt{\frac{H_m b_n}{n}} \log^{1/2} n + n^{-3}\right) \tag{31}$$

where (b_n) is a positive bounded sequence such that $\log n = o(H_m n b_n)$.

Proof: Let $c_0 = \sup_x |f_\varepsilon(x)| < \infty$. Since $\mathbb{P}(\tilde{Y}_i \le x|\mathcal{F}_{i-1}) = F_\varepsilon(H_m(x - \tilde{Z}_{i-1}))$, we have $\mathbb{P}(x \le \tilde{Y}_i \le x+u|\mathcal{F}_{i-1}) \le H_m c_0 u$ for all $u > 0$. Therefore for any $u \in [-b_n, b_n]$, see (32) at the bottom of the next page. Applying Freedman's martingale inequality [15] and implementing a chaining argument, we have (31). See, for instance,

$$\mathcal{P}_k \mathbf{V} = \mathbb{E}[\mathbf{V}|\mathcal{F}_k] - \mathbb{E}[\mathbf{V}|\mathcal{F}_{k-1}], \quad \mathbf{V} \in \mathcal{L}^1, \quad \text{where } \mathcal{F}_k = (\ldots, \varepsilon_{k-1}, \varepsilon_k).$$

Lemma 5 in [40], Lemma 4 in [41] and Lemma 6 in [47] for more details. Details are omitted.

Lemma 2: Under conditions of Theorem 1, we have

$$\| \sup_{|u|\le b_n} |N_n(x+u) - N_n(x)|\| = O\left(\frac{b_n m^{3/2}}{\sqrt{n}}\right). \quad (33)$$

Proof: Since $N_n(x) = \tilde{F}_n^*(x) - \tilde{F}(x)$, we have

$$N_n(x+u) - N_n(x) = \sqrt{m}\frac{\int_{t=0}^{u} R_n(x+t)dt}{n-m+1}$$

where we recall (30) for $\tilde{Z}_{i-1}$, and (34) at the bottom of the page. Hence $\|\sup_{|u|\le b_n}|N_n(x+u) - N_n(x)|\| \le \sqrt{m}b_n \sup_{|u|\le b_n}\|R_n(x+u)\|/(n-m+1)$. Therefore we only need to prove

$$\|R_n(x+u)\| \le Cm\sqrt{n} \quad \text{for all} \quad u \in [-b_n, b_n]. \quad (35)$$

Let $(\varepsilon_i')_{-\infty}^{\infty}$ be an i.i.d. copy of $(\varepsilon_i)_{-\infty}^{\infty}$ and $\tilde{Z}_{i-1,k}^* = \tilde{Z}_{i-1} - \tilde{b}_k\varepsilon_{i-k}/\sqrt{m} + \tilde{b}_k\varepsilon_{i-k}'/\sqrt{m}$. Note that for $k \ge 1$, see (36) at the bottom of the page, where $c_1 = \sup_{v\in\mathbb{R}}|f_\varepsilon'(v)|\|\varepsilon_0 - \varepsilon_0'\| < \infty$. Further note that

$$R_n(x+u) = \sum_{k=1}^{\infty}\sum_{i=m}^{n}\mathcal{P}_{i-k}f_\varepsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1}))$$

and, by the orthogonality of $\mathcal{P}_{i-k}$, $i = m, \ldots, n$,

$$\begin{aligned}\left\|\sum_{i=m}^{n}\mathcal{P}_{i-k}f_\varepsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1}))\right\|^2 &= \sum_{i=m}^{n}\|\mathcal{P}_{i-k}f_\varepsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1}))\|^2 \\ &\le c_1^2(n-m+1)\tilde{b}_k^2.\end{aligned}$$

Therefore, for all $u \in [-b_n, b_n]$, by the short-range dependence condition as

$$\begin{aligned}\|R_n(x+u)\| &\le \sum_{k=1}^{\infty}\left\|\sum_{i=m}^{n}\mathcal{P}_{i-k}f_\varepsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1}))\right\| \\ &\le c_1\sqrt{n}\sum_{k=1}^{\infty}|\tilde{b}_k| \le c_1 m\sqrt{n}\sum_{j=0}^{\infty}|a_j|.\end{aligned}$$

This lemma then follows by letting C in (35) be $c_1\sum_{j=0}^{\infty}|a_j|$. □

Proof of Theorem 1: [22] central limit theorem, $\tilde{Y}_i \Rightarrow N(0,\sigma^2)$, where $\sigma = \|\sum_{i=0}^{\infty}\mathcal{P}_0 e_i\| \le \sum_{i=0}^{\infty}\|\mathcal{P}_0 e_i\| < \infty$. Hence $\tilde{Q}_q$ is well-defined and it converges to the pth quantile of a $N(0,\sigma^2)$ distribution as $m \to \infty$. Furthermore, note that e_i is a weighted sum of i.i.d. random variables and the density $f_\varepsilon(\cdot)$ is bounded. Hence a classic characteristic function argument making use of the inversion formula leads to the local limit theorem for densities of the sequence $\tilde{Y}_i$; namely

$$\sup_x\left|f_m(x) - \frac{\varphi\left(\frac{x}{\sigma}\right)}{\sigma}\right| \to 0 \quad (37)$$

where $f_m(\cdot)$ is the density function of $\tilde{Y}_i$ and $\varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$ is the standard normal density. For more technical details, one can refer to [16] and [25, Th. 4.3.1] regarding the case of partial sums of i.i.d. random variables.

Let (c_n) be an arbitrary sequence of positive numbers that goes to infinity. Let $\bar{c}_n = \min(c_n, n^{1/4}/m^{3/4})$. Then $\bar{c}_n \to \infty$. Lemmas 1 and 2 above imply that

$$\begin{aligned}&|\tilde{F}_n(\tilde{Q}_q + B_n) - \tilde{F}(\tilde{Q}_q + B_n) - [\tilde{F}_n(\tilde{Q}_q) - \tilde{F}(\tilde{Q}_q)]| \\ &\quad = O_p\left(\frac{B_n m^{3/2}}{\sqrt{n}} + \sqrt[4]{m}\sqrt{\frac{B_n}{n}}\log^{1/2} n\right) \\ &\quad = o_p(B_n). \end{aligned} \quad (38)$$

$$\sum_{i=m}^{n}[\mathbb{E}(I\{x \le \tilde{Y}_i \le x+u\}|\mathcal{F}_{i-1}) - \mathbb{E}^2(I\{x \le \tilde{Y}_i \le x+u\}|\mathcal{F}_{i-1})] \le c_0(n-m+1)H_m b_n. \quad (32)$$

$$R_n(x) = \sum_{i=m}^{n}[f_\varepsilon(H_m(x-\tilde{Z}_{i-1})) - \mathbb{E}f_\varepsilon(H_m(x-\tilde{Z}_{i-1}))] \quad x \in \mathbb{R}. \quad (34)$$

$$\begin{aligned}\|\mathcal{P}_{i-k}f_\varepsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1}))\| &\le \|f_\varepsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1})) - f_\varepsilon(\sqrt{m}(x+u-\tilde{Z}_{i-1,k}^*))\| \\ &\le \sup_{v\in\mathbb{R}}|f_\varepsilon'(v)|\|\sqrt{m}(\tilde{Z}_{i-1,k} - \tilde{Z}_{i-1,k}^*)\| \\ &\le c_1|\tilde{b}_k| \end{aligned} \quad (36)$$

where $B_n = \bar{c}_n m/\sqrt{n}$. Furthermore, similar arguments as those in Lemmas 1 and 2 imply

$$|\tilde{F}_n(\tilde{Q}_q) - \tilde{F}(\tilde{Q}_q)| = O_p\left(\frac{m}{\sqrt{n}}\right) = o_p(B_n). \quad (39)$$

Using Taylor's expansion of $\tilde{F}(\cdot)$, we have

$$\tilde{F}(\tilde{Q}_q + B_n) - \tilde{F}(\tilde{Q}_q) = B_n f_m(\tilde{Q}_q) + O(B_n^2). \quad (40)$$

By (37), $f_m(\tilde{Q}_q) > 0$ for sufficiently large n. Plugging (39) and (40) into (38), we have $\mathbb{P}(\tilde{F}_n(\tilde{Q}_q + B_n) > q) \to 1$. Hence $\mathbb{P}(\hat{Q}_n(q) > \tilde{Q}_q + B_n) \to 0$ by the monotonicity of $\tilde{F}_n(\cdot)$. Similarly, we have $\mathbb{P}(\hat{Q}_n(q) < \tilde{Q}_q - B_n) \to 0$. Using the fact that c_n can approach ∞ arbitrarily slowly, Theorem 1 i) follows.

To prove ii), let $\tilde{\Delta}_n = \hat{Q}_n(q) - \tilde{Q}_q$. Then by (i), we have $\tilde{\Delta}_n = O_p(m/\sqrt{n})$. Let $\bar{B}_n = \min(c_n, \sqrt{m})m/\sqrt{n}$. We have by Lemmas 1 and 2 that

$$|\tilde{F}_n(\hat{Q}_n(q)) - \tilde{F}(\hat{Q}_n(q)) - [\tilde{F}_n(\tilde{Q}_q) - \tilde{F}(\tilde{Q}_q)]| = O_p\left(\frac{\bar{B}_n m^{3/2}}{\sqrt{n}+} \sqrt[4]{m}\sqrt{\frac{\bar{B}_n}{n}} \log^{1/2} n\right).$$

Note that $|\tilde{F}_n(\hat{Q}_n(q)) - q| = O(1/(n-m+1))$ and $\tilde{F}(\hat{Q}_n(q)) - \tilde{F}(\tilde{Q}_q) = \tilde{\Delta}_n f_m(\tilde{Q}_q) + O_p(B_n^2)$. Therefore, we have

$$|f_m(\tilde{Q}_q)\tilde{\Delta}_n - [q - \tilde{F}_n(\tilde{Q}_q)]| = O_p\left(\frac{\bar{B}_n m^{3/2}}{\sqrt{n}} + \sqrt[4]{m}\sqrt{\frac{\bar{B}_n}{n}} \log^{1/2} n\right). \quad (41)$$

Since $f_m(\tilde{Q}_p) > 0$ for sufficiently large m and $c_n \to \infty$ can be arbitrarily slowly, we have (ii). □

Lemma 3: Under conditions of Theorem 4, we have for any $\varrho \in (1/\gamma, \alpha)$ that

$$\| \sup_{|u| \le b_n} |N_n(x+u) - N_n(x)| \|_\varrho = O(H_m b_n m n^{1/\varrho - \gamma} |l(n)|). \quad (42)$$

Proof: Recall (34) for the definition of $R_n(\cdot)$. Similarly as in the proof of Lemma 2, to prove this lemma it suffices to show that, for some $0 < C < \infty$,

$$\|R_n(x+u)\|_\varrho \le C m n^{1/\rho + 1 - \gamma} |l(n)| \text{ for all } u \in [-b_n, b_n]. \quad (43)$$

Since $1 < \varrho < 2$, by Burkholder's inequality of martingales, we have (44) shown at the bottom of the page, where $C_\varrho = [18\varrho^{3/2}(\varrho-1)^{-1/2}]^\varrho$. Since $\mathbb{E}(|\varepsilon_i|^\varrho) < \infty$, similarly as (36) we have for $k \le i-1$ that

$$\|\mathcal{P}_k f_\varepsilon(H_m(x - Z_{i-1}))\|_\varrho \le c_1 |\tilde{b}_{i-k}| \quad (45)$$

where $c_1 = \sup_{v \in \mathbb{R}} |f'_\varepsilon(v)| \|\varepsilon_0 - \varepsilon'_0\|_\varrho < \infty$. Hence, by (44), (45) and Karamata's Theorem [14], we have (46) at the top of the next page. Similarly, since $\varrho > 1$ and $\varrho\gamma > 1$, we have by Hölder's inequality as shown in (47) at the top of the next page. Similarly, we have $II = O[m^\varrho n^{1+\varrho(1-\gamma)} |l(n)|^\varrho]$. Together with (46) and (47), we have (43). Hence the lemma follows. □

Proof of Theorem 4: It follows by Lemma 1 and Lemma 3, and similar arguments as those in the proof of Theorem 1. Since there are no essential extra technical difficulties, the details are omitted. A detailed proof is available upon request.

Proof of Theorem 5: Recall (28) for α^* and γ^*. Then for all four cases of tail and memory, it can be shown that

$$|\hat{\beta}_{lad} - \beta| = O_p(\pi(n)) \quad (48)$$

where $\pi(n) = n^{1/\alpha^* - \gamma^*} L_1^*(n) l^*(n)$, $L_1^*(n) = 1$ if $\mathbb{E}(\varepsilon_i^2) < \infty$, $L_1^*(n) = |L_1(n)|$ if $\varepsilon_i \in \mathcal{D}(\alpha)$, $l^*(n) = 1$ in the SRD case and $l^*(n) = |l(n)|$ in the LRD case. For a proof of (48) in the light-tailed and SRD case, see Theorem 1 in [41] and the proof of heavy-tailed and LRD case can be found in Theorem 1 of [46]. The other two cases can be proved with similar arguments

$$\begin{aligned}
\|R_n(x+u)\|_\varrho^\varrho &= \left\| \sum_{k=-\infty}^{n-1} \mathcal{P}_k \sum_{i=m}^{n} f_\varepsilon(H_m(x - \tilde{Z}_{i-1})) \right\|_\varrho^\varrho \\
&\le C_\varrho \sum_{k=-\infty}^{n-1} \left\| \mathcal{P}_k \sum_{i=m}^{n} f_\varepsilon(H_m(x - \tilde{Z}_{i-1})) \right\|_\varrho^\varrho \\
&\le C_\varrho \sum_{k=-\infty}^{n-1} \left(\sum_{i=m}^{n} \|\mathcal{P}_k f_\varepsilon(H_m(x - \tilde{Z}_{i-1}))\|_\varrho \right)^\varrho \\
&= C_\varrho \left(\sum_{k=-\infty}^{-n} + \sum_{k=-n+1}^{0} + \sum_{k=1}^{n-1} \right) \left(\sum_{i=m}^{n} \|\mathcal{P}_k f_\varepsilon(H_m(x - \tilde{Z}_{i-1}))\|_\varrho \right)^\varrho \\
&:= C_\varrho (I + II + III), \quad (44)
\end{aligned}$$

$$III \le c_1^{\varrho} \sum_{k=1}^{n-1} \left(\sum_{i=\max(m,k+1)}^{n} |\tilde{b}_{i-k}| \right)^{\varrho} \le c_1^{\varrho} \sum_{k=1}^{n-1} \left(m \sum_{i=0}^{n-k} |a_i| \right)^{\varrho}$$
$$= m^{\varrho} \sum_{k=1}^{n-1} O[(n-k)^{1-\gamma} |l(n-k)|]^{\varrho}$$
$$= O[m^{\varrho} n^{1+\varrho(1-\gamma)} |l(n)|^{\varrho}]. \tag{46}$$

$$I \le c_1^{\varrho} \sum_{k=-\infty}^{-n} \left(\sum_{i=m}^{n} |\tilde{b}_{i-k}| \right)^{\varrho} \le c_1^{\varrho} \sum_{k=n}^{\infty} \left(m \sum_{i=1}^{n} |a_{k+i}| \right)^{\varrho}$$
$$\le c_1^{\varrho} m^{\varrho} n^{\varrho-1} \sum_{k=n}^{\infty} \sum_{i=1}^{n} |a_{k+i}|^{\varrho}$$
$$= O[m^{\varrho} n^{1+\varrho(1-\gamma)} |l(n)|^{\varrho}]. \tag{47}$$

and the details are omitted. Note that $\hat{e}_i - e_i = w_i^T(\hat{\beta} - \beta)$. By (48) and condition (b) of Theorem 5 we have

$$\sup_{1 \le i \le n} |\hat{e}_i - e_i| = O_p(\pi(n)).$$

Therefore

$$\sup_{m \le i \le n} \left| \sum_{k=i-m+1}^{i} \hat{e}_k - \sum_{k=i-m+1}^{i} e_k \right| = O_p(m\pi(n)). \tag{49}$$

Hence

$$\bar{Q}_n(q) - \hat{Q}_n(q) = O_p\left(\frac{m\pi(n)}{H_m} \right). \tag{50}$$

Since the right-hand side of (50) is of smaller order than the bounds in (i) of Theorems 1–4, Theorem 5 follows. □

ACKNOWLEDGMENT

The authors are grateful to the referees and an Associate Editor for their many helpful comments.

REFERENCES

[1] Y. Adegoke, "Time Warner to test Internet billing based on usage," *Reuters*, 2008.
[2] P. Algoet, "Universal schemes for prediction, gambling and portfolio selection," *Ann. Probab.*, vol. 20, pp. 901–941, 1992.
[3] P. Algoet, "Universal schemes for learning the best nonlinear predictor given the infinite past and side information," *IEEE Trans. Inf. Theory*, vol. 45, pp. 1165–1185, 1999.
[4] T. W. Anderson, *The Statistical Analysis of Time Series*. New York: Wiley, 1971.
[5] F. Avram and M. S. Taqqu, "Weak convergence of moving averages with infinite variance," in *Dependence in Probability and Statistics: A Survey of Recent Results*, Eberlein and Taqqu, Eds. Boston, MA: Birkhauser, 1986, pp. 399–416.
[6] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis. Forecasting and Control*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 1994.
[7] F. J. Breidt, R. A. Davis, and W. Dunsmuir, "Improved bootstrap prediction intervals for autoregressions," *J. Time Ser. Anal.*, vol. 16, pp. 177–200, 1995.
[8] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed. New York: Springer-Verlag, 1991.
[9] C. Chatfield, *Time-Series Forecasting*. Boca Raton, FL: Chapman and Hall/CRC, 2000.
[10] Y. S. Chow and H. Teicher, *Probability Theory*. New York: Springer Verlag, 1988.
[11] T. M. Cover, "Open problems in information theory," in *Proc. 1975 IEEE Joint Workshop Inf. Theory*, New York, 1975, pp. 35–36.
[12] D. J. Dalrymple, "Sales forecasting practices: Results from a United States survey pages," *Int. J. Forecasting*, vol. 3, pp. 379–391, 1987.
[13] J. Fan and Q. Yao, *Nonlinear Time Series*. New York: Springer, 2003.
[14] W. Feller, *An Introduction to Probability Theory and its Applications*. New York: Wiley, 1971, vol. II.
[15] D. A. Freedman, "On tail probabilities for Martingales," *Ann. Prob.*, vol. 3, pp. 100–118, 1975.
[16] B. V. Gnedenko and B. V. Gnedenko, "A local limit theorem for densities," *Russian Doklady Akad. Nauk SSSR*, vol. 95, pp. 5–7, 1954.
[17] L. Györfi, W. Härdle, P. Sarda, and P. Vieu, *Nonparametric Curve Estimation From Time Series*. Berlin, Germany: Springer-Verlag, 1989, vol. 60, Lecture Notes in Statistics.
[18] L. Györfi, G. Lugosi, and G. Morvai, "A simple randomized algorithm for consistent sequential prediction of ergodic time series," *IEEE Trans. Inf. Theory*, vol. 45, pp. 2642–2650, 1999.
[19] L. Györfi, G. Morvai, and S. J. Yakowitz, "Limits to consistent on-line forecasting for ergodic time series," *IEEE Trans. Inf. Theory*, vol. 44, pp. 886–892, 1998.
[20] L. Györfi and G. Ottucsák, "Sequential prediction of unbounded stationary time series," *IEEE Trans. Inf. Theory*, vol. 53, pp. 1866–1872, 2007.
[21] G. J. Hahn and W. Meeker, *Statistical Intervals: A Guide for Practitioners*. New York: Wiley, 1991.
[22] E. J. Hannan, "Central limit theorems for time series regression," *Z. Wahrsch Verw. Gebiete*, vol. 26, pp. 157–170, 1973.
[23] J. R. M. Hosking, "Fractional differencing," *Biometrika*, vol. 68, pp. 165–176, 1981.
[24] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.
[25] I. A. Ibragimov and V. Y. Linnik, *Independent and Stationary Sequences of Random Variables*. Amsterdam, The Netherlands: Wolters-Noordhoff, 1971.
[26] R. Koenker, *Quantile Regression*. Cambridge, U.K.: Cambridge University Press, 2005.
[27] T. Mikosch, S. Resnick, H. Rootzén, and A. Stegeman, "Is network traffic approximated by stable Levy motion or fractional Brownian motion?," *Ann. Appl. Prob.*, vol. 12, pp. 23–68, 2002.

[28] J. P. Morgan, *Value at Risk, RiskMetrics Technical Document*. New York: Morgan Guaranty Trust, 1996.
[29] G. Morvai, "Guessing the output of a stationary binary time series," in *Foundations of Statistical Inference*, Y. Haitovsky, H. R. Lerche, and Y. Ritov, Eds. Berlin, Germany: Physika Verlag, 2003, pp. 205–213.
[30] G. Morvai and B. Weiss, "Intermittent estimation of stationary time series," *Test*, vol. 13, pp. 525–542, 2004.
[31] G. Morvai and B. Weiss, *Theor. Stochastic Process.*, vol. 11, pp. 112–120.
[32] G. Morvai and B. Weiss, "On universal estimates for binary renewal processes," *Ann. Appl. Probab.*, vol. 18, pp. 1970–1992, 2008.
[33] G. Morvai, S. Yakowitz, and P. Algoet, "Weakly convergent nonparametric forecasting of stationary time series," *IEEE Trans. Inf. Theory*, vol. 43, pp. 483–498, 1997.
[34] D. Ornstein, "Guessing the next output of a stationary process," *Israel J. Math.*, vol. 30, pp. 292–296, 1978.
[35] M. Pourahmadi, *Foundations of Time Series Analysis and Prediction Theory*. New York: Wiley, 2001.
[36] B. R. Ya, "Prediction of random sequences and universal coding," *Probl. Inf. Trans.*, vol. 24, pp. 3–14, 1988.
[37] D. Schäfer, "Strongly consistent online forecasting of centered Gaussian processes," *IEEE Trans. Inf. Theory*, vol. 48, pp. 791–799, 2002.
[38] R. L. Schmoyer, "Asymptotically valid prediction intervals for linear models," *Technometrics*, vol. 34, pp. 399–408, 1992.
[39] M. S. Taqqu, "Fractional brownian motion and long-range dependence," in *Theory and Applications of Long-range Dependence*, P. Doukhan, G. Oppenheim, and M. S. Taqqu, Eds. Boston, MA: Birkhauser, 2003.
[40] W. B. Wu, "On the Bahadur representation of sample quantiles for stationary sequences," *Ann. Stat.*, vol. 33, pp. 1934–1963, 2005.
[41] W. B. Wu, "M-estimation of linear models with dependent errors," *Ann. Stat.*, vol. 35, pp. 495–521, 2007.
[42] W. B. Wu, G. Michailidis, and D. Zhang, "Simulating sample paths of linear fractional stable motion," *IEEE Trans. Inf. Theory*, vol. 50, pp. 1086–1096, 2004.
[43] W. B. Wu and M. Woodroofe, "Martingale approximations for sums of stationary processes," *Ann. Prob.*, vol. 32, pp. 1674–1690, 2004.
[44] S. Vardeman, "What about the other intervals?," *Amer. Stat.*, vol. 46, pp. 193–197, 1992.
[45] A. Zagdański, "On the construction and properties of bootstrap-t prediction intervals for stationary time series," *Probab. Math. Statist.*, vol. 25, pp. 133–153, 2005.
[46] Z. Zhou and W. B. Wu, *On Linear Models With Long Memory and Heavy-Tailed Errors*, 2007.
[47] Z. Zhou and W. B. Wu, "Local linear quantile estimation of non-stationary time series," *Ann. Stat.*, 2009.
[48] G. Morvai, S. Yakowitz, and L. Györfi, "Nonparametric inference for ergodic, stationary time series," *Annals Statist.*, vol. 24, no. 1, pp. 370–379.

Zhou Zhou received the B.S. degree in mathematics in 2003 from Peking University, Beijing, China, in 2003 and the Ph.D. degree in statistics from the University of Chicago, Chicago, IL, in 2009.

Currently, he is an Assistant Professor at the Statistics Department, University of Toronto, Toronto, ON, Canada, which he joined in July 2009. His research interests lie in probability, time series analysis, non- and semiparametric estimation and inference, quantile regression, functional/longitudinal data analysis, and statistical methods in engineering and environmental sciences.

Zhiwei Xu received his Ph.D. in Computer Engineering in 2001 from Florida Atlantic University.

He is an Assistant Professor of the Department of Computer and Information Science at the University of Michigan–Dearborn. He was a Senior Staff Engineer and Research Scientist in Motorola Software and System Research Lab—Network and System Research Lab from 2001 to 2007. He joined the Department of Computer Information Science at the University of Michigan–Dearborn in September 2007. His research interests include software engineering, data mining and machine learning, optimization, and security.

Wei Biao Wu received the Ph.D. degree from University of Michigan, Ann Arbor, in 2001.

In 2001, he was an Assistant Professor and currently he is an Associate Professor at the Statistics Department, University of Chicago, IL. His research interests include probability theory, statistics, and econometrics.