Parameter Estimation of Heavy-Tailed AR(p) Model from Incomplete Data

Junyan Liu, Sandeep Kumar, and Daniel P. Palomar
Dept. of Electronic and Computer Engineering
Hong Kong University of Science and Technology
Hong Kong
Email: jliubl@connect.ust.hk; eesandeep@ust.hk; palomar@ust.hk

***Abstract*—The autoregressive (AR) model is a widely used model to represent the time series data from numerous applications, for example, financial time series, DNA microarray data, etc. In all such applications, issues with missing values frequently occur in the data observation or recording process. Traditionally, the parameter estimation for AR models of order p (AR(p)), from data with missing values has been considered under the Gaussian innovation assumption, and there does not exist any work addressing the issue of missing data for the heavy-tailed AR(p) model. This paper proposes an efficient framework for the parameter estimation from incomplete heavy-tailed AR(p) time series based on the stochastic approximation expectation maximization (SAEM) coupled with a Markov Chain Monte Carlo (MCMC) procedure. The proposed algorithm is computationally cheap and easy to implement. Simulation results demonstrate the efficacy of the proposed framework.**

***Index Terms*—AR model, heavy-tail, missing values, stochastic EM, MCMC**

I. Introduction

Time series data play an important role in data analysis, and its applications span across disciplines like science, engineering, and social science. The autoregressive (AR) model is a widely used model to understand time series data [1]–[3]. An AR model of order p, AR(p), is defined as

$$y_t = \varphi_0 + \boldsymbol{\varphi}^T \mathbf{x}_{t-1} + \varepsilon_t, \tag{1}$$

where y_t is the t-th observation, $\mathbf{x}_{t-1} = [y_{t-1}, \ldots, y_{t-p}]$ involves the previous p observations, φ_0 is a constant, $\boldsymbol{\varphi} = [\varphi_1, \ldots, \varphi_p]$ is the autoregressive coefficient, and ε_t is the innovation associated with the t-th observation. Basically, each sample in an AR time series is a linear combination of previous observations with a stochastic innovation.

Traditionally, the innovation ε_t is assumed to be Gaussian distributed [4]. However, in many real applications, the Gaussian assumption is no longer valid, as the data do not follow the Gaussian distribution, but instead show heavy tails due to the data heterogeneity, the existence of outliers, or simply the data generating process. Some examples are, the stock returns [1], [5], the brain fMRI [6], and the black-swan events in animal population [7]. For these cases, one may seek an AR model with the innovations following a heavy-tailed distribution such as the Student's t-distribution. The Student's t AR model performs well for the heavy-tailed AR time series [8]–[10].

This work was supported by the Hong Kong RGC 16208917 research grant.

In practice, missing values frequently appear in the data observation or recording process. There are various reasons that can lead to missing values: values may not be measured, values may be measured but get lost, or values may be measured but are considered unusable [11]. Therefore, the investigation of the AR time series with missing values is of significant interest. Although there are numerous works about Gaussian AR time series with missing values [12]–[14], less attention has been paid to heavy-tailed AR time series with missing values, since the parameter estimation in such a case is much more complicated due to the intractable problem formulation.

Towards this, in a recent work [15], we have developed an efficient algorithm for the parameter estimation of the Student's t AR(1) with missing values based on the stochastic approximation expectation maximization-Markov chain Monte Carlo (SAEM-MCMC) algorithm. In the present paper, we consider the more general case, the Student's t AR(p) time series with missing values. This extension from the AR(1) to the AR(p) is non-trivial. To apply the SAEM-MCMC algorithm, the sampling from the posterior distribution of missing data is required in each iteration. Since, in AR(p) time series, each sample depends on the previous p samples instead of just one, the posterior distribution of missing data is more complicated and difficult to sample from. To deal with this challenge, we propose an efficient Gibbs sampling scheme to draw realizations from the posterior distribution, and a low cost algorithmic framework for the parameter estimation of the Student's t AR(p) model from incomplete time series has been developed. Simulation results reveal the usefulness of the proposed framework.

II. Heavy-tailed AR(p) with missing values

A Student's t AR(p) model can be expressed as $y_t = \varphi_0 + \boldsymbol{\varphi}^T \mathbf{x}_{t-1} + \varepsilon_t$, with the innovations ε_t's following a zero-mean Student's t-distribution $\varepsilon_t \overset{\text{i.i.d.}}{\sim} t\left(0, \sigma^2, \nu\right)$. Suppose we have a Student's t AR(p) time series with length T, but some y_t's may be missing due to various reasons, and they are denoted by NA (not available). Here, we assume that the first p samples are observed, and the missing-data mechanism is ignorable, i.e., the missingness does not depend on the value [11]. Let us denote the set of the indexes of the observed y_t's except the first p observations by C_o, and the set of the indexes of the

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

missing y_t's by C_{m}. Also denote $\mathbf{y} = (y_t,\, p+1 \le t \le T)$, $\mathbf{y}_{\mathsf{o}} = (y_t,\, t \in C_{\mathsf{o}})$, and $\mathbf{y}_{\mathsf{m}} = (y_t,\, t \in C_{\mathsf{m}})$. Ignoring the marginal distribution of the first p observations (treat them as deterministic), the log-likelihood of the observed data is

$$\begin{aligned} l\left(\boldsymbol{\theta}; \mathbf{y}_{\mathsf{o}}\right) &= \log\left(\int p\left(\mathbf{y}; \boldsymbol{\theta}\right) \mathrm{d}\mathbf{y}_{\mathsf{m}}\right) \qquad (2)\\ &= \log\left(\int \prod_{t=p+1}^{T} f_t\left(y_t; \varphi_0 + \boldsymbol{\varphi}^T \mathbf{x}_{t-1}, \sigma^2, \nu\right) \mathrm{d}\mathbf{y}_{\mathsf{m}}\right), \end{aligned}$$

where $f_t\left(y; \mu, \sigma^2, \nu\right) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\sigma\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)$ is the probability density function (pdf) of the Student's t-distribution, and $\boldsymbol{\theta} = \left(\varphi_0, \boldsymbol{\varphi}, \sigma^2, \nu\right)$. Then the maximum likelihood (ML) estimation problem for $\boldsymbol{\theta}$ can be formulated as

$$\underset{\boldsymbol{\theta}\in\Theta}{\text{maximize}}\ l\left(\boldsymbol{\theta}; \mathbf{y}_{\mathsf{o}}\right). \qquad (3)$$

where $\Theta = \left\{\boldsymbol{\theta} | \sigma^2 > 0,\, \nu > 0\right\}$. The log-likelihood $l\left(\boldsymbol{\theta}; \mathbf{y}_{\mathsf{o}}\right)$ involves integrals and has no closed-form expression, therefore, we cannot solve this optimization problem directly. To handle this, we resort to the expectation maximization (EM) type algorithm, which solves this ML problem by optimizing a sequence of simpler sub-problems iteratively.

III. Parameter estimation via SAEM-MCMC

The EM algorithm is a general iterative algorithm to solve ML estimation problems with missing data or latent variables. In each iteration, the EM algorithm first performs an expectation (E) step, which computes the expectation function of the complete data log-likelihood with respect to the posterior distribution of latent variables given the current estimates, and then a maximization (M) step, which updates the estimates by maximizing the expected complete data log-likelihood [16].

For the ML problem (3), if we only regard $\mathbf{y}_{\mathsf{m}}$ as the latent variable and apply the EM type algorithm, it is difficult to obtain the expectation of the complete data log-likelihood and also maximize it. Interestingly, the Student's t-distribution can be regarded as a Gaussian mixture [17]. Since $\varepsilon_t \sim t\left(0, \sigma^2, \nu\right)$, we can present it as the Gaussian mixture

$$\varepsilon_t | \sigma^2, \nu, \tau_t \sim \mathcal{N}\left(0, \frac{\sigma^2}{\tau_t}\right), \tau_t \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \qquad (4)$$

where τ_t is the mixture weight. Then we can use the EM type algorithm to solve the problem by regarding both $\mathbf{y}_{\mathsf{m}}$ and $\boldsymbol{\tau} = \{\tau_t,\, p+1 \le t \le T\}$ as latent variables. The resulting expectation of the complete data log-likelihood at iteration k can be expressed as

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \bar{\mathbf{S}}^{(k)}\right) &= \mathsf{E}_{p\left(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau} | \mathbf{y}_{\mathsf{o}}, \boldsymbol{\theta}^{(k)}\right)}\left[\log\left(p\left(\mathbf{y}_{\mathsf{o}}, \mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau} \mid \boldsymbol{\theta}\right)\right)\right]\\ &= \frac{\nu}{2}\bar{s}_1^{(k)} - \frac{\bar{s}_2^{(k)}}{2\sigma^2} - \frac{\varphi_0^2 \bar{s}_3^{(k)}}{2\sigma^2} - \frac{\boldsymbol{\varphi}^T \bar{\mathbf{S}}_4^{(k)} \boldsymbol{\varphi}}{2\sigma^2} + \frac{\varphi_0 \bar{s}_5^{(k)}}{\sigma^2}\\ &\quad + \frac{\boldsymbol{\varphi}^T \bar{\mathbf{s}}_6^{(k)}}{\sigma^2} - \frac{\varphi_0 \boldsymbol{\varphi}^T \bar{\mathbf{s}}_7^{(k)}}{\sigma^2} + (T-p)\left(\frac{\nu\log\left(\frac{\nu}{2}\right)}{2}\right.\\ &\quad \left. - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) - \frac{\log\left(\sigma^2\right)}{2}\right) + const., \qquad (5) \end{aligned}$$

where the expectation

$$\begin{aligned} \bar{\mathbf{S}}^{(k)} &= \left(\bar{s}_1^{(k)}, \bar{s}_2^{(k)}, \bar{s}_3^{(k)}, \bar{\mathbf{S}}_4^{(k)}, \bar{s}_5^{(k)}, \bar{\mathbf{s}}_6^{(k)}, \bar{\mathbf{s}}_7^{(k)}\right)\\ &= \mathsf{E}_{p\left(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau} | \mathbf{y}_{\mathsf{o}}, \boldsymbol{\theta}^{(k)}\right)}\left[\mathbf{S}\left(\mathbf{y}_{\mathsf{o}}, \mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}\right)\right], \qquad (6) \end{aligned}$$

with

$$\begin{aligned} &\mathbf{S}\left(\mathbf{y}_{\mathsf{o}}, \mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}\right)\\ &= \left(\sum_{t=p+1}^{T}\left(\log\left(\tau_t\right) - \tau_t\right), \sum_{t=p+1}^{T}\tau_t y_t^2, \sum_{t=p+1}^{T}\tau_t, \right. \qquad (7)\\ &\quad \left. \sum_{t=p+1}^{T}\tau_t \mathbf{x}_{t-1}\mathbf{x}_{t-1}^T, \sum_{t=2}^{T}\tau_t y_t, \sum_{t=2}^{T}\tau_t y_t \mathbf{x}_{t-1}, \sum_{t=2}^{T}\tau_t \mathbf{x}_{t-1}\right). \end{aligned}$$

However, due to the intractable form of the posterior distribution $p\left(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau} | \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}^{(k)}\right)$, we cannot get closed-form expression for the expectations $\bar{\mathbf{S}}^{(k)}$ and $Q\left(\boldsymbol{\theta}, \bar{\mathbf{S}}^{(k)}\right)$, and thus, the E step is intractable. To solve the unavailability of the expectation in the E step, the SAEM algorithm, which draws samples of latent variables from the posterior distribution and approximates the expectation in the E step by a stochastic approximation procedure, has been proposed [18]. Nevertheless, in our case, since we only know the posterior distribution $p\left(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau} | \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}^{(k)}\right)$ up to a scalar, and the proportional term is complicated, we cannot sample from $p\left(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau} | \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}^{(k)}\right)$ directly. Therefore, we resort to the SAEM-MCMC algorithm, which generates the samples via constructing a Markov chain [19]. The success of the SAEM-MCMC framework crucially depends on the appropriate design of sampling scheme so that the sampling is efficient, and the computational cost is not too high.

A. E step

We propose a Gibbs sampling scheme to generate samples from $p\left(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau} | \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}^{(k)}\right)$. At iteration k, given the current estimate $\boldsymbol{\theta}^{(k)}$, the Gibbs sampler starts with $\left(\boldsymbol{\tau}^{(k-1,l)}, \mathbf{y}_{\mathsf{m}}^{(k-1,l)}\right)$ $(l = 1, 2 \ldots, L)$ and generates the next sample $\left(\boldsymbol{\tau}^{(k,l)}, \mathbf{y}_{\mathsf{m}}^{(k,l)}\right)$ via the following scheme:

- sample $\boldsymbol{\tau}^{(k,l)}$ from $p\left(\boldsymbol{\tau} | \mathbf{y}_{\mathsf{m}}^{(k-1,l)}, \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}^{(k)}\right)$,
- sample $\mathbf{y}_{\mathsf{m}}^{(k,l)}$ from $p\left(\mathbf{y}_{\mathsf{m}} | \boldsymbol{\tau}^{(k,l)}, \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}^{(k)}\right)$.

Then the expectation $\bar{\mathbf{S}}^{(k)}$ is approximated by a combination of the previous value and the L new samples:

$$\hat{\mathbf{S}}^{(k)} = \hat{\mathbf{S}}^{(k-1)} + \gamma^{(k)}\left(\frac{1}{L}\sum_{l=1}^{L}\mathbf{S}\left(\mathbf{y}_{\mathsf{o}}, \mathbf{y}_{\mathsf{m}}^{(k,l)}, \boldsymbol{\tau}^{(k,l)}\right) - \hat{\mathbf{S}}^{(k-1)}\right), \qquad (8)$$

where $\left\{\gamma^{(k)}\right\}$ is a nonincreasing sequence of positive step sizes, and $Q\left(\boldsymbol{\theta}, \bar{\mathbf{S}}^{(k)}\right)$ is approximated by $Q\left(\boldsymbol{\theta}, \hat{\mathbf{S}}^{(k)}\right)$.

In the following part, we give the detailed illustration about the Gibbs sampling process. Lemma 1 indicates that, to sample $\boldsymbol{\tau}$ from $p\left(\boldsymbol{\tau} | \mathbf{y}_{\mathsf{m}}, \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}\right)$, we just need to draw a realization of τ_t from certain gamma distribution for each t.

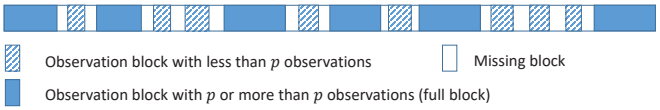

Fig. 1. AR(p) time series with missing values.

Lemma 1. *Given* $\mathbf{y}_\mathsf{m}$, $\mathbf{y}_\mathsf{o}$, *and* $\boldsymbol{\theta}$, *the mixture weights* $\{\tau_t\}$ *are independent from each other with*

$$\begin{aligned}&\tau_t|\mathbf{y}_\mathsf{m},\mathbf{y}_\mathsf{o};\boldsymbol{\theta} \\ &\sim \textit{Gamma}\left(\frac{\nu+1}{2},\ \frac{\left(y_t-\varphi_0-\boldsymbol{\varphi}^T\mathbf{x}_{t-1}\right)^2/\sigma^2+\nu}{2}\right).\end{aligned} \tag{9}$$

Before discussing the sampling of $\mathbf{y}_\mathsf{m}$ from the conditional pdf $p\left(\mathbf{y}_\mathsf{m}\mid\boldsymbol{\tau},\mathbf{y}_\mathsf{o};\boldsymbol{\theta}\right)$, we introduce some definitions. The consecutive missing values or a single missing value between two observations are defined as a missing block, and the consecutive observed values or a single observed value between two missing values are defined as an observation block (see Figure 1). For AR(1), given $\boldsymbol{\tau}$, $\mathbf{y}_\mathsf{o}$, and $\boldsymbol{\theta}$, each missing block is independent from others [15]. In addition, it is simple to get the distribution of each missing block. Thus, we can sample $\mathbf{y}_\mathsf{m}$ from $p\left(\mathbf{y}_\mathsf{m}\mid\boldsymbol{\tau},\mathbf{y}_\mathsf{o};\boldsymbol{\theta}\right)$ by sampling different missing blocks separately, and the computational cost is low. For AR(p), however, the missing blocks are not independent to each other anymore in general, and it is non-trivial to obtain the conditional pdf of $\mathbf{y}_\mathsf{m}$.

Interestingly, we find that, although missing blocks in AR(p) are not independent from each other, similar independence structure exists in another form. To illustrate this, we define an observation block that has p or more than p observations as a full block. Suppose there are D full blocks in an incomplete AR(p) time series:

$$\begin{aligned}&y_1,\ldots,y_{t_1},\mathsf{NA},\ldots,\mathsf{NA},y_{t_2-q_2+1},\ldots,y_{t_2},\mathsf{NA},\ldots,\mathsf{NA},\\&y_{t_d-q_d+1},\ldots,y_{t_d},\mathsf{NA},\ldots,\mathsf{NA},y_{y_{T-q_D+1}},\ldots,y_T.\end{aligned}$$

where the d-th full block $\mathbf{y}_{\mathsf{o}(d)}=(y_{t_d-q_d+1},\ldots,y_{t_d})$ has q_i consecutive observations with $q_1=t_1$ and $t_D=T$. We define all the missing blocks between two full blocks $\mathbf{y}_{\mathsf{o}(d)}$ and $\mathbf{y}_{\mathsf{o}(d+1)}$ as the d-th missing group denoted by $\mathbf{y}_{\mathsf{m}(d,d+1)}$. Note that, as in Figure 1, there may be some observation blocks that have less than p consecutive observations between two full blocks $\mathbf{y}_{\mathsf{o}(d)}$ and $\mathbf{y}_{\mathsf{o}(d+1)}$. They are denoted by $\mathbf{y}_{\mathsf{o}(d,d+1)}$. Lemma 2 indicates that each missing groups are independent to each other.

Lemma 2. *Given* $\boldsymbol{\tau}$, $\mathbf{y}_\mathsf{o}$, *and* $\boldsymbol{\theta}$, *the distribution of the missing group* $\mathbf{y}_{\mathsf{m}(d,d+1)}$ *is independent of the other missing values, and its distribution depends on the observed values* $\mathbf{y}_\mathsf{o}$ *only through* $\mathbf{y}_{\mathsf{o}(d,d+1)}$, $\mathbf{y}^{\mathsf{lp}}_{\mathsf{o}(d)}=\left[y_{t_d-p+1},\ldots,y_{t_d}\right]^T$*(the last* p *observations in the previous full block* $\mathbf{y}_{\mathsf{o}(d)}$*)*, $\mathbf{y}^{\mathsf{fp}}_{\mathsf{o}(d+1)}=\left[y_{t_{d+1}-q_{d+1}+1},\ldots,y_{t_{d+1}-q_{d+1}+p}\right]^T$*(the first* p *observations in the next full block* $\mathbf{y}_{\mathsf{o}(d+1)}$*)*:

$$\begin{aligned}&p\left(\mathbf{y}_\mathsf{m}|\boldsymbol{\tau},\mathbf{y}_\mathsf{o};\boldsymbol{\theta}\right)\\&=\prod_{i=1}^{c}p\left(\mathbf{y}_{\mathsf{m}(d,d+1)}|\boldsymbol{\tau},\mathbf{y}_{\mathsf{o}(d,d+1)},\mathbf{y}^{\mathsf{lp}}_{\mathsf{o}(d)},\mathbf{y}^{\mathsf{fp}}_{\mathsf{o}(d+1)};\boldsymbol{\theta}\right).\end{aligned} \tag{10}$$

To obtain $p\left(\mathbf{y}_{\mathsf{m}(d,d+1)}|\boldsymbol{\tau},\mathbf{y}_{\mathsf{o}(d,d+1)},\mathbf{y}^{\mathsf{lp}}_{\mathsf{o}(d)},\mathbf{y}^{\mathsf{fp}}_{\mathsf{o}(d+1)};\boldsymbol{\theta}\right)$, we first analyze the distribution of $\mathbf{y}_{(d+1)}|\boldsymbol{\tau},\mathbf{y}^{\mathsf{lp}}_{\mathsf{o}(d)};\boldsymbol{\theta}$ with $\mathbf{y}_{(d+1)}=\left[y_{t_d+1},\ldots,y_{t_{d+1}-q_{d+1}+p}\right]$ consisting of $\mathbf{y}_{\mathsf{m}(d,d+1)}$, $\mathbf{y}_{\mathsf{o}(d,d+1)}$, and $\mathbf{y}^{\mathsf{fp}}_{\mathsf{o}(d+1)}$. Through a recursive process based on (1), each component of $\mathbf{y}_{(d+1)}$ can be expressed as the sum of a linear combination of elements in $\mathbf{y}^{\mathsf{lp}}_{\mathsf{o}(d)}$, a linear combination of $\varepsilon_{t_d+1},\ldots,\varepsilon_{t_{d+1}-q_{d+1}+p}$, and a term only involving φ_0 and $\boldsymbol{\varphi}$:

$$y_{t_d+1}=\varphi_0+\sum_{i=1}^{p}\varphi_i y_{t_d+1-i}+\varepsilon_{t_d+1}, \tag{11}$$

$$\begin{aligned}y_{t_d+2}&=\varphi_0+\sum_{i=1}^{p}\varphi_i y_{t_d+2-i}+\varepsilon_{t_d+2}\\&=\varphi_0+\sum_{i=2}^{p}\varphi_i y_{t_d+2-i}+\varphi_1 y_{t_d+2-1}+\varepsilon_{t_d+2}\\&=\varphi_0+\sum_{i=1}^{p-1}\varphi_{i+1}y_{t_d+1-i}+\varphi_1\Big(\varphi_0+\sum_{i=1}^{p}\varphi_i y_{t_d+1-i}\\&\quad+\varepsilon_{t_d+1}\Big)+\varepsilon_{t_d+1}\\&=\varphi_0+\varphi_1\varphi_0+\sum_{i=1}^{p-1}\left(\varphi_{i+1}+\varphi_1\varphi_i\right)y_{t_d+1-i}\\&\quad+\varphi_1\varphi_p y_{t_d+1-p}+\varphi_1\varepsilon_{t_d+1}+\varepsilon_{t_d+2},\end{aligned} \tag{12}$$

$$\vdots$$

$$\begin{aligned}y_{t_d+r}&=a_r\left(\varphi_0,\boldsymbol{\varphi}\right)+\mathbf{b}_r^T\left(\varphi_0,\boldsymbol{\varphi}\right)\mathbf{y}^{\mathsf{lp}}_{\mathsf{o}(d)}\\&\quad+\mathbf{c}_r^T\left(\varphi_0,\boldsymbol{\varphi}\right)\left[\varepsilon_{t_d+1},\ldots,\varepsilon_{t_d+r}\right]^T,\end{aligned} \tag{13}$$

for $r=1,\ldots,t_{d+1}-q_{d+1}+p-t_d$, where $a_r\left(\varphi_0,\boldsymbol{\varphi}\right)$, $\mathbf{b}_r\left(\varphi_0,\boldsymbol{\varphi}\right)$, and $\mathbf{c}_r\left(\varphi_0,\boldsymbol{\varphi}\right)$ are a scalar function, a p-dimensional function, and a r-dimensional function of φ_0 and $\boldsymbol{\varphi}$, respectively.

From (4), given $\boldsymbol{\tau}$, $\mathbf{y}^{\mathsf{lp}}_{\mathsf{o}(d)}$, and $\boldsymbol{\theta}$, $\varepsilon_t\overset{\text{i.i.d.}}{\sim}\mathcal{N}\left(0,\frac{\sigma^2}{\tau_t}\right)$ for $t>t_d$. Since the linear combination of Gaussian variables is still Gaussian distributed, $\mathbf{y}_{(d+1)}|\boldsymbol{\tau},\mathbf{y}^{\mathsf{lp}}_{\mathsf{o}(d)};\boldsymbol{\theta}$ follows a Gaussian distribution $\mathcal{N}\left(\boldsymbol{\mu}_{(d)},\boldsymbol{\Sigma}_{(d)}\right)$ with r-the element of $\boldsymbol{\mu}_{(d)}$

$$\left(\boldsymbol{\mu}_{(d)}\right)_r=a_r\left(\varphi_0,\boldsymbol{\varphi}\right)+\mathbf{b}_r^T\left(\varphi_0,\boldsymbol{\varphi}\right)\mathbf{y}^{\mathsf{lp}}_{\mathsf{o}(d)} \tag{14}$$

and the element in the r-th row and j-th column of $\boldsymbol{\Sigma}_{(d)}$

$$\begin{aligned}\left(\boldsymbol{\Sigma}_{(d)}\right)_{r,j}&=\mathbf{c}_{r[1:u]}^T\left(\varphi_0,\boldsymbol{\varphi}\right)\mathsf{Cov}\left[\widetilde{\boldsymbol{\varepsilon}}_{(d,u)}\right]\mathbf{c}_{j[1:u]}\left(\varphi_0,\boldsymbol{\varphi}\right)\\&=\mathbf{c}_{r[1:u]}^T\left(\varphi_0,\boldsymbol{\varphi}\right)\mathsf{Diag}\left(\frac{\sigma^2}{\tau_{t_d+1}},\ldots,\frac{\sigma^2}{\tau_{t_d+u}}\right)\\&\quad\times\mathbf{c}_{j[1:u]}\left(\varphi_0,\boldsymbol{\varphi}\right),\end{aligned} \tag{15}$$

where $u=\min(r,j)$, $\widetilde{\boldsymbol{\varepsilon}}_{(d,u)}=\left[\varepsilon_{t_d+1},\ldots,\varepsilon_{t_d+u}\right]^T$, and $\mathbf{c}_{r[1:u]}\left(\varphi_0,\boldsymbol{\varphi}\right)$ $\left(\mathbf{c}_{j[1:u]}\left(\varphi_0,\boldsymbol{\varphi}\right)\right)$ is the vector of the first u

element of $\mathbf{c}_r(\varphi_0, \boldsymbol{\varphi})$ ($\mathbf{c}_j(\varphi_0, \boldsymbol{\varphi})$). Then its conditional distribution $\mathbf{y}_{\mathsf{m}(d,d+1)}|\boldsymbol{\tau}, \mathbf{y}_{\mathsf{o}(d,d+1)}, \mathbf{y}^{\mathsf{lp}}_{\mathsf{o}(d)}, \mathbf{y}^{\mathsf{fp}}_{\mathsf{o}(d+1)}; \boldsymbol{\theta}$ also follows a Gaussian distribution $\mathcal{N}\left(\tilde{\boldsymbol{\mu}}_{(d)}, \tilde{\boldsymbol{\Sigma}}_{(d)}\right)$, and the parameters $\tilde{\boldsymbol{\mu}}_{(d)}$ and $\tilde{\boldsymbol{\Sigma}}_{(d)}$ can be computed easily from $\boldsymbol{\mu}_{(d)}$ and $\boldsymbol{\Sigma}_{(d)}$. Note that the coefficients $a_r(\varphi_0, \boldsymbol{\varphi})$, $\mathbf{b}_r(\varphi_0, \boldsymbol{\varphi})$, and $\mathbf{c}_r(\varphi_0, \boldsymbol{\varphi})$ do not depend on d (the number of the missing group), but only r (the relative position in the block), which means that they can be recycled for different missing groups. Finally, to sample $\mathbf{y}_{\mathsf{m}}$ from $p(\mathbf{y}_{\mathsf{m}}|\boldsymbol{\tau}, \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta})$, we just need to draw a realization of $\mathbf{y}_{\mathsf{m}(d,d+1)}$ from a Gaussian distribution $\mathcal{N}\left(\tilde{\boldsymbol{\mu}}_{(d)}, \tilde{\boldsymbol{\Sigma}}_{(d)}\right)$ for each missing group d.

Remark 3. In practice, there may be less than p observations in the last observation block. In this case, the missing blocks between this observation block and the last full block can still be regarded as a missing group, and its distribution can be computed similarly.

In summary, to draw realizations from the posterior distribution $p\left(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}|\mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}^{(k)}\right)$, the proposed Gibbs sampling scheme just need to sample from Gaussian distributions and gamma distributions alternatively, and the parameters of these distributions can be easily computed.

B. M step

Setting the derivatives of $Q\left(\boldsymbol{\theta}, \hat{\mathbf{S}}^{(k)}\right)$ with respect to to φ_0, φ_1 and σ^2 to 0 gives

$$\varphi_0^{(k+1)} = \left(\hat{s}_5^{(k)} - \left(\boldsymbol{\varphi}^{(k+1)}\right)^T \hat{s}_7^{(k)}\right) / \hat{s}_3^{(k)}, \tag{16}$$

$$\boldsymbol{\varphi}^{(k+1)} = \left(\hat{s}_3^{(k)} \hat{\mathbf{S}}_4^{(k)} - \hat{\mathbf{s}}_7^{(k)} \left(\hat{\mathbf{s}}_7^{(k)}\right)^T\right)^{-1} \left(\hat{s}_3^{(k)} \hat{\mathbf{s}}_6^{(k)} - \hat{s}_5^{(k)} \hat{\mathbf{s}}_7^{(k)}\right), \tag{17}$$

and

$$\begin{aligned}&\left(\sigma^{(k+1)}\right)^2 \\ =&\frac{1}{T-p}\left(\hat{s}_2^{(k)} + \left(\varphi_0^{(k+1)}\right)^2 \hat{s}_3^{(k)} + \left(\boldsymbol{\varphi}^{(k+1)}\right)^T \hat{\mathbf{S}}_4^{(k)} \boldsymbol{\varphi}^{(k+1)}\right. \\ &\left. - 2\varphi_0 \hat{s}_5^{(k)} - 2\left(\boldsymbol{\varphi}^{(k+1)}\right)^T \hat{\mathbf{s}}_6^{(k)} + 2\varphi_0^{(k+1)} \left(\boldsymbol{\varphi}^{(k+1)}\right)^T \hat{\mathbf{s}}_7^{(k)}\right).\end{aligned} \tag{18}$$

The $\nu^{(k+1)}$ can be found by one-dimensional search:

$$\nu^{(k+1)} = \underset{\nu>0}{\operatorname{argmax}} \frac{\nu \log\left(\frac{\nu}{2}\right)}{2} - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) + \frac{\nu \hat{s}_1^{(k)}}{2(T-p)}. \tag{19}$$

According to Proposition 1 in [20], $\nu^{(k+1)}$ always exists and is unique.

IV. Simulations

In this section, we conduct a simulation study of the proposed framework. First, we show the convergence and the good estimation performance of the proposed framework on synthetic data. We set $\varphi_0^{\text{true}} = 1$, $\boldsymbol{\varphi}^{\text{true}} = [0.90, 0.12, -0.16]$, $\left(\sigma^{\text{true}}\right)^2 = 0.01$, and $\nu^{\text{true}} = 1$. We randomly generate an incomplete Student's t AR(3) time series with length $T = 500$ and the missing rate 20% based on the model (1) with $\varepsilon_t \overset{\text{i.i.d.}}{\sim} t\left(0, \left(\sigma^{\text{true}}\right)^2, \nu^{\text{true}}\right)$. Then we apply the proposed algorithm to estimate the parameters. We use the estimates based the Gaussian AR(3) model as the initial estimates $\hat{\varphi}_0^{(0)}, \hat{\boldsymbol{\varphi}}^{(0)}$, and $\left(\hat{\sigma}^{(0)}\right)^2$, the resulting posterior mean of the missing values as the initial realization $\mathbf{y}_{\mathsf{m}}^{(0,l)}$, and a random $\nu^{(0)}$. As recommended in [21], we set $L = 10$, and $\gamma^{(k)} = 1$ for $1 \leq k \leq 30$ and $\gamma^{(k)} = \frac{1}{k-K}$ for $k \geq 31$. Figure 2 shows the estimation errors of the estimates versus iterations. The estimation error is measured by the normalized root square error (NRSE) $\text{NRSE}(\boldsymbol{\theta}_i) = \frac{\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^{\text{true}}\|}{\|\boldsymbol{\theta}_i^{\text{true}}\|}$, where $\boldsymbol{\theta}_i$ can be φ_0, $\boldsymbol{\varphi}$, σ^2, or ν, $\hat{\boldsymbol{\theta}}_i$ is its estimate, and $\boldsymbol{\theta}_i^{\text{true}}$ is its true value. We

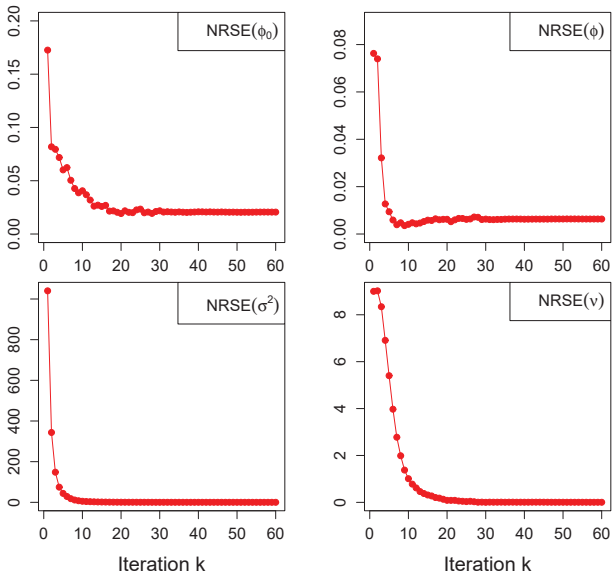

Fig. 2. The NRSE's of estimates versus iterations.

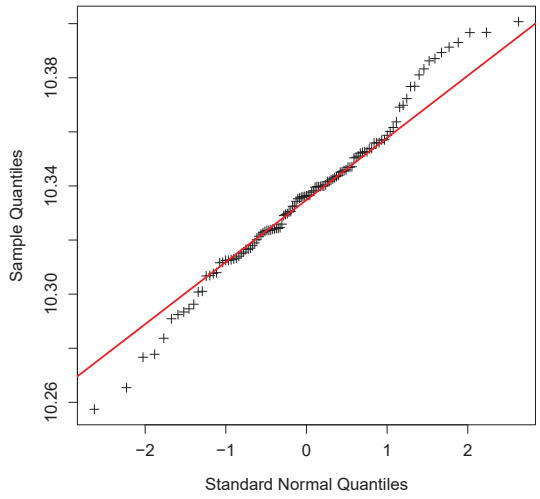

Fig. 3. Quantile-quantile plot of the innovations obtained by fitting logarithmic prices of the Hang Seng index to the Student's t AR model.

TABLE I
ESTIMATION RESULTS OF DIFFERENT METHODS ON SYNTHETIC DATA (THE NRSE'S ARE IN PARENTHESES).

	$\hat{\varphi}_0$	$\hat{\boldsymbol{\varphi}}$	$(\hat{\sigma})^2$	$\hat{\nu}$
True value	1.000	$[0.900, 0.120, -0.160]$	0.010	1.000
Student's t AR(3)	0.979 (0.021)	$[0.902, 0.124, -0.164]$ (0.006)	0.011 (0.100)	0.998 (0.002)
Gaussian AR(3)	0.827 (0.183)	$[0.838, 0.147, -0.141]$ (0.076)	10.411 (1040)	$+\infty$ ($+\infty$)
Kharin's method	0.080 (0.920)	$[1.398, 0.046, -0.460]$ (0.635)	0.181 (17.1)	$+\infty$ ($+\infty$)

TABLE II
ESTIMATION AND PREDICTION RESULTS OF DIFFERENT METHODS FOR REAL DATA.

	$\hat{\varphi}_0$	$\hat{\boldsymbol{\varphi}}$	$(\hat{\sigma})^2$	$\hat{\nu}$	averaged prediction errors
Student's t AR(4)	1.299	$[0.915, 0.006, 0.207, -0.253]$	1.139×10^{-4}	7.285	2.099×10^{-4}
Gaussian AR(4)	1.640	$[0.934, -0.017, 0.181, -0.257]$	1.433×10^{-4}	$+\infty$	2.291×10^{-4}
Kharin's method	-0.103	$[2.610, -0.178, -2.570, 1.148]$	1.310×10^{-5}	$+\infty$	3.677×10^{-4}

can see that the proposed algorithm converges in less than 60 iterations and the final estimation error is small. Table I compares the estimation results of the Student's t AR model, the Gaussian AR model [11], and Kharin's robust estimation method [13]. This testifies that for heavy-tailed AR time series, the traditional methods for Gaussian distributions are too inefficient, and significant performance gain can be achieved by designing algorithms for the heavy-tailed AR model.

Then we test the performance of the proposed estimation method on a real data set. Here we consider the logarithm prices of the Hang Seng index over 121 working days from Jan. 1, 2018 to Jun. 30, 2018 (excluding weekends and public holidays). Through the computation of the PACF, an AR model of order 4 is proposed to fit this time series. Figure 3 shows the quantile-quantile (QQ) plot of the innovations obtained by fitting this time series to the Student's t AR(4) model. The deviation from the straight red line indicates that the innovations are significantly non-Gaussian and indeed heavy-tailed. We divide the 121 logarithm prices into two parts: the training data, which involves the first 115 samples, and the test data, which involves the remaining 6 samples. We randomly delete 10% of the training data, and apply the proposed framework to fit the Student's t AR(4) model to this incomplete training data. After obtaining the estimate $\hat{\varphi}_0$ and $\hat{\boldsymbol{\varphi}}$, we compute the one-step-ahead predictions for the test data $\hat{y}_t = \hat{\varphi}_0 + \hat{\boldsymbol{\varphi}}^T \mathbf{x}_{t-1}$,where $t = 116, \ldots, 121$, and the averaged prediction error $\frac{1}{6}\sum_{t=116}^{121}(\hat{y}_t - y_t)^2$. For comparison, we also apply the Gaussian AR(4) model and Kharin's method to estimate φ_0 and $\boldsymbol{\varphi}$, and do the prediction. Table II shows Student's t AR(4) model performs better than the Gaussian AR(4) model and Kharin's method for this heavy-tailed AR time series, and the proposed framework can provide reliable parameter estimation for real data with missing values.

REFERENCES

[1] R. S. Tsay, *Analysis of Financial Time Series*, 2nd ed. Hoboken, NJ: John Wiley & Sons, 2005.

[2] M. K. Choong, M. Charbit, and H. Yan, "Autoregressive-model-based missing value estimation for DNA microarray time series data," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 1, pp. 131–137, 2009.

[3] A. Schlögl and G. Supp, "Analyzing event-related EEG data with multivariate autoregressive parameters," *Prog. Brain Res.*, vol. 159, pp. 135–147, 2006.

[4] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag, 2005.

[5] S. T. Rachev, *Handbook of Heavy Tailed Distributions in Finance: Handbooks in Finance*. Amsterdam, Netherlands: Elsevier, 2003.

[6] F. Han and H. Liu, "ECA: High-dimensional elliptical component analysis in non-Gaussian distributions," *J. Am. Stat. Assoc.*, vol. 113, no. 521, pp. 252–268, 2018.

[7] S. C. Anderson, T. A. Branch, A. B. Cooper, and N. K. Dulvy, "Black-swan events in animal populations," *Proc. Natl. Acad. Sci.*, vol. 114, no. 12, pp. 3252–3257, 2017.

[8] M. L. Tiku, W.-K. Wong, D. C. Vaughan, and G. Bian, "Time series models in non-normal situations: Symmetric innovations," *J. Time Ser. Anal.*, vol. 21, no. 5, pp. 571–596, 2000.

[9] B. Tarami and M. Pourahmadi, "Multi-variate t autoregressions: Innovations, prediction variances and exact likelihood equations," *J. Time Ser. Anal.*, vol. 24, no. 6, pp. 739–754, 2003.

[10] J. Christmas and R. Everson, "Robust autoregression: Student-t innovations using variational Bayes," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 48–57, 2011.

[11] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. Hoboken, N.J.: John Wiley & Sons, 2002.

[12] J. Ding, L. Han, and X. Chen, "Time series AR modeling with missing observations based on the polynomial transformation," *Math. Comput. Modelling*, vol. 51, no. 5-6, pp. 527–536, 2010.

[13] V. A. V. Yuriy S. Kharin, "Robust estimation of AR coefficients under simultaneously influencing outliers and missing values," *J. Stat. Plan. Inference*, vol. 141, no. 9, pp. 3276–3288, 2011.

[14] J. Sargan and E. Drettakis, "Missing data in an autoregressive model," *Int. Econ. Rev.*, vol. 15, no. 1, pp. 39–58, 1974.

[15] J. Liu, S. Kumar, and D. P. Palomar, "Parameter estimation of heavy-tailed AR model with missing data via stochastic EM," *accepted in IEEE Trans. on Signal Process.*, Jan. 2019.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.

[17] C. Liu, "ML estimation of the multivariate t distribution and the EM algorithm," *J. Multivar. Anal.*, vol. 63, no. 2, pp. 296–312, 1997.

[18] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a stochastic approximation version of the EM algorithm," *Ann. Stat.*, vol. 27, no. 1, pp. 94–128, 1999.

[19] E. Kuhn and M. Lavielle, "Coupling a stochastic approximation version of EM with an MCMC procedure," *ESAIM Probab. and Statist.*, vol. 8, pp. 115–131, 2004.

[20] C. Liu and D. B. Rubin, "ML estimation of the t distribution using EM and its extensions, ECM and ECME," *Stat. Sin.*, vol. 5, no. 1, pp. 19–39, 1995.

[21] E. Kuhn and M. Lavielle, "Maximum likelihood estimation in nonlinear mixed effects models," *Comput. Stat. Data Anal.*, vol. 49, no. 4, pp. 1020–1038, 2005.