Parameter Estimation of Heavy-Tailed AR Model With Missing Data Via Stochastic EM

Junyan Liu, Sandeep Kumar, and Daniel P. Palomar, *Fellow, IEEE*

***Abstract*—The autoregressive (AR) model is a widely used model to understand time series data. Traditionally, the innovation noise of the AR is modeled as Gaussian. However, many time series applications, for example, financial time series data, are non-Gaussian, therefore, the AR model with more general heavy-tailed innovations is preferred. Another issue that frequently occurs in time series is missing values due to system data record failure or unexpected data loss. Although there are numerous works about Gaussian AR time series with missing values, as far as we know, there does not exist any work addressing the issue of missing data for the heavy-tailed AR model. In this paper, we consider this issue for the first time, and propose an efficient framework for parameter estimation from incomplete heavy-tailed time series based on a stochastic approximation expectation maximization coupled with a Markov Chain Monte Carlo procedure. The proposed algorithm is computationally cheap and easy to implement. The convergence of the proposed algorithm to a stationary point of the observed data likelihood is rigorously proved. Extensive simulations and real datasets analyses demonstrate the efficacy of the proposed framework.**

***Index Terms*—AR model, heavy-tail, missing values, SAEM, Markov chain Monte Carlo, convergence analysis.**

I. Introduction

In THE recent era of data deluge, many applications collect and process time series data for inference, learning, parameter estimation, and decision making. The autoregressive (AR) model is a commonly used model to analyze time series data, where observations taken closely in time are statistically dependent on others. In an AR time series, each sample is a linear combination of some previous observations with a stochastic innovation. An AR model of order p, AR(p), is defined as

$$y_t = \varphi_0 + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t, \tag{1}$$

where y_t is the t-th observation, φ_0 is a constant, φ_i's are autoregressive coefficients, and ε_t is the innovation associated with the t-th observation. The AR model has been successfully used in many real-world applications such as DNA microarray data analysis [1], EEG signal modeling [2], financial time series analysis [3], and animal population study [4], to name but a few.

Traditionally, the innovation ε_t of the AR model is assumed to be Gaussian distributed, which, as a result of the linearity of the AR model, means that the observations are also Gaussian distributed. However, there are situations arising in applications of signal processing and financial markets where the time series are non-Gaussian and heavy-tailed, either due to intrinsic data generation mechanism or existence of outliers. Some examples are, stock returns [3], [5], brain fMRI [6], [7], and black-swan events in animal population [4]. For these cases, one may seek an AR model with innovations following a heavy-tailed distribution such as the Student's t-distribution. The Student's t-distribution is one of the most commonly used heavy-tailed distributions [8]. The authors of [9] and [10] have considered an AR model with innovations following a Student's t-distribution with a known number of degrees of freedom, whereas [11] and [12] investigated the case with an unknown number of degrees of freedom. The Student's t AR model performs well for heavy-tailed AR time series and can provide robust reliable estimates of the regressive coefficients when outliers occur.

Another issue that frequently occurs in practice is missing values during 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 [13]. Some real-world cases are: some stocks may suffer a lack of liquidity resulting in no transaction and hence no price recorded, observation devices like sensors may break down during measurement, and weather or other conditions disturb sample taking schemes. Therefore, investigation of AR time series with missing values is significant. Although there are numerous works considering Gaussian AR time series with missing values [14]–[17], less attention has been paid to heavy-tailed AR time series with missing values, since parameter estimation in such a case is complicated due to the intractable problem formulation. The frameworks for parameter estimation for heavy-tailed AR time series in [9]–[12] require complete data, and thereby, are not suited for scenarios with missing data. The objective of the current paper is to deal with this challenge and develop an efficient framework for parameter estimation from incomplete data under the heavy-tailed time series model via the expectation-maximization (EM) type algorithm.

The EM algorithm is a widely used iterative method to obtain the maximum likelihood (ML) estimates of parameters when there are missing values or unobserved latent variables. In each iteration, the EM algorithm maximizes the conditional expectation of the complete data likelihood to update the estimates.

Manuscript received September 12, 2018; revised January 7, 2019; accepted January 30, 2019. Date of publication February 21, 2019; date of current version March 15, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Vincent Y. F. Tan. This work was supported by the Hong Kong RGC 16208917 research grant. *(Corresponding author: Junyan Liu.)*

The authors are with the Hong Kong University of Science and Technology, Hong Kong (e-mail: jliubl@connect.ust.hk; eesandeep@ust.hk; palomar@ust.hk).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TSP.2019.2899816

Many variants of the EM algorithm have been proposed to deal with specific challenges in different missing value problems. For example, to tackle the problem posed by the intractability of the conditional expectation of the complete data log-likelihood, a stochastic variant of the EM algorithm, which approximates the expectation by drawing samples of the latent variables from the conditional distribution, has been proposed in [18], [19]. The stochastic EM has also been quite popular to curb the curse of dimensionality [14], [20], since its computation complexity is lower than the EM algorithm. The expectation conditional maximization (ECM) algorithm has been suggested to deal with the unavailability of the closed-form maximizer of the expected complete data log-likelihood [21]. The regularized EM algorithm has been used to enforce certain structures in parameter estimates like sparsity, low-rank, and network structure [22].

In this paper, we develop a provably convergent low cost algorithmic framework for parameter estimation of the AR time series model with heavy-tailed innovations from incomplete time series. As far as we know, there does not exist any convergent algorithmic framework for such problem. Following [9]–[11], here we consider the AR model with the Student's t distributed innovations. We formulate an ML estimation problem and develop an efficient algorithm to obtain the ML estimates of the parameters based on the stochastic EM framework. To tackle the complexity of the conditional distribution of latent variables, we propose a Gibbs sampling scheme to generate samples. Instead of directly sampling from the complicated conditional distribution, the proposed algorithm just need to sample from Gaussian distributions and gamma distributions alternatively. The convergence of the proposed algorithm to a stationary point is established. Simulations on real data and synthetic data show that the proposed framework can provide accurate estimation of parameters for incomplete time series, and is also robust against possible outliers. Although here we only focus on the Student's t distributed innovation, the idea of the proposed approach and the algorithm can also be extended to the AR model with other heavy-tailed distributions.

This paper is organized as follows. The problem formulation is provided in Section II. The review of the EM and its stochastic variants is presented in Section III. The proposed algorithm is derived in Section IV. The convergence analysis is carried out in Section V. Finally, Simulation results for the proposed algorithm applied to both real and synthetic data are provided in Section VI, and Section VII concludes the paper.

II. Problem Formulation

For simplicity of notations, we first introduce the AR(1) model. Suppose a univariate time series $y_1, y_2, \ldots, y_T$ follows an AR(1) model

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varepsilon_t, \tag{2}$$

where the innovations ε_t's follow a zero-mean heavy-tailed Student's t-distribution $\varepsilon_t \overset{i.i.d.}{\sim} t(0, \sigma^2, \nu)$. The Student's t-distribution is more heavy-tailed as the number of degrees of freedom ν decreases. Note that the Gaussian distribution is a special case of the Student's t-distribution with $\nu = +\infty$.

Given all the parameters φ_0, φ_1, σ^2 and ν, the distribution of y_t conditional on all the preceding data $\mathcal{F}_{t-1}$, which consists of $y_1, y_2, \ldots, y_{t-1}$, only depends on the previous sample y_{t-1}:

$$\begin{aligned} & p\left(y_t|\varphi_0, \varphi_1, \sigma^2, \nu, \mathcal{F}_{t-1}\right) \\ & = p\left(y_t|\varphi_0, \varphi_1, \sigma^2, \nu, y_{t-1}\right) \\ & = f_t\left(y_t; \varphi_0 + \varphi_1 y_{t-1}, \sigma^2, \nu\right) \\ & = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\sigma\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\left(y_t - \varphi_0 - \varphi_1 y_{t-1}\right)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}, \end{aligned} \tag{3}$$

where $f_t(\cdot)$ denotes the probability density function (pdf) of a Student's t-distribution.

In practice, a certain sample y_t may be missing due to various reasons, and it is denoted by $y_t = \mathsf{NA}$ (not available). Here we assume that the missing-data mechanism is ignorable, i.e., the missing does not depend on the value [13]. Suppose we have an observation of this time series with D missing blocks as follows:

$$y_1, \ldots, y_{t_1}, \mathsf{NA}, \ldots, \mathsf{NA}, y_{t_1+n_1+1}, \ldots y_{t_d}, \mathsf{NA}, \ldots, \mathsf{NA},$$
$$y_{t_d+n_d+1}, \ldots, y_{t_D}, \mathsf{NA}, \ldots, \mathsf{NA}, y_{t_D+n_D+1}, \ldots, y_T,$$

where, in the d-th missing block, there are n_d missing samples $y_{t_d+1}, \ldots, y_{t_d+n_d}$, which are surrounded from the left and the right by the two observed data y_{t_d} and $y_{t_d+n_d+1}$. We set for convenience $t_0 = 0$ and $n_0 = 0$. Let us denote the set of the indexes of the observed values by C_{o}, and the set of the indexes of the missing values by C_{m}. Also denote $\mathbf{y} = (y_t, 1 \leq t \leq T)$, $\mathbf{y}_{\mathsf{o}} = (y_t, t \in C_{\mathsf{o}})$, and $\mathbf{y}_{\mathsf{m}} = (y_t, t \in C_{\mathsf{m}})$.

Let us assume $\boldsymbol{\theta} = (\varphi_0, \varphi_1, \sigma^2, \nu) \in \Theta$ with $\Theta = \{\boldsymbol{\theta}|\sigma^2 > 0, \nu > 0\}$. Ignoring the marginal distribution of y_1, the log-likelihood of the observed data is

$$\begin{aligned} l\left(\boldsymbol{\theta}; \mathbf{y}_{\mathsf{o}}\right) &= \log\left(\int p\left(\mathbf{y}; \boldsymbol{\theta}\right) \mathsf{d}\mathbf{y}_{\mathsf{m}}\right) \\ &= \log\left(\int \prod_{t=2}^{T} p\left(y_t|\boldsymbol{\theta}, \mathcal{F}_{t-1}\right) \mathsf{d}\mathbf{y}_{\mathsf{m}}\right) \\ &= \log\left(\int \prod_{t=2}^{T} f_t\left(y_t; \varphi_0 + \varphi_1 y_{t-1}, \sigma^2, \nu\right) \mathsf{d}\mathbf{y}_{\mathsf{m}}\right). \end{aligned} \tag{4}$$

Then the maximum likelihood (ML) estimation problem for $\boldsymbol{\theta}$ can be formulated as

$$\underset{\boldsymbol{\theta}\in\Theta}{\mathsf{maximize}} \; l\left(\boldsymbol{\theta}; \mathbf{y}_{\mathsf{o}}\right). \tag{5}$$

The integral in (4) has no closed-form expression, thus, the objective function is very complicated, and we cannot solve the optimization problem directly. In order to deal with this, we resort to the EM framework, which circumvents such difficulty by optimizing a sequence of simpler approximations of the original objective function instead.

III. EM and Its Stochastic Variants

The EM algorithm is a general iterative algorithm to solve ML estimation problems with missing data or latent data. More specifically, given the observed data $\mathbf{X}$ generated from a statistical model with unknown parameter $\boldsymbol{\theta}$, the ML estimator of the parameter $\boldsymbol{\theta}$ is defined as the maximizer of the likelihood of the observed data

$$l\left(\mathbf{X};\boldsymbol{\theta}\right)=\log p(\mathbf{X}|\boldsymbol{\theta}). \tag{6}$$

In practice, it often occurs that $l(\mathbf{X};\boldsymbol{\theta})$ does not have manageable expression due to the missing data or latent data $\mathbf{Z}$, while the likelihood of complete data $p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$ has a manageable expression. This is when the EM algorithm can help. The EM algorithm seeks to find the ML estimates by iteratively applying these two steps [23]:

(E) *Expectation:* calculate the expected log-likelihood of the complete data set $(\mathbf{X},\mathbf{Z})$ with respect to the current conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$ and the current estimate of the parameter $\boldsymbol{\theta}^{(k)}$:

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)=\int\log p\left(\mathbf{X},\,\mathbf{Z}|\boldsymbol{\theta}\right)p\left(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(k)}\right)\mathsf{d}\mathbf{Z}, \tag{7}$$

where k is the iteration number.

(M) *Maximization:* find the new estimate

$$\boldsymbol{\theta}^{(k+1)}=\arg\max_{\boldsymbol{\theta}}Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right). \tag{8}$$

The sequence $\{l(\mathbf{X};\boldsymbol{\theta}^{(k)})\}$ generated by the EM algorithm is non-decreasing, and the limit points of the sequence $\{\boldsymbol{\theta}^{(k)}\}$ are proven to be the stationary points of the observed data log-likelihood under mild regularity conditions [24]. In fact, the EM algorithm is a particular choice of the more general majorization-minimization algorithm [25].

However, in some applications of the EM algorithm, the expectation in the E step cannot be obtained in closed-form. To deal with this, Wei and Tanner proposed the Monte Carlo EM (MCEM) algorithm, in which the expectation is computed by a Monte Carlo approximation based on a large number of independent simulations of the missing data [26]. The MCEM algorithm is computationally very intensive.

In order to reduce the amount of simulations required by the MCEM algorithm, the stochastic approximation EM (SAEM) algorithm replaces the E step of the EM algorithm by a stochastic approximation procedure, which approximates the expectation by combining new simulations with the previous ones [18]. At iteration k, the SAEM proceeds as follows:

(E-S1) *Simulation:* generate L realizations $\mathbf{Z}^{(k,l)}(l=1,2\ldots,L)$ from the conditional distribution $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(k)})$

(E-A) *Stochastic approximation:* update $\hat{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ according to

$$\begin{aligned}&\hat{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)\\&=\hat{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k-1)}\right)+\gamma^{(k)}\left(\frac{1}{L}\sum_{l=1}^{L}\log p\left(\mathbf{X},\,\mathbf{Z}^{(k,l)}|\boldsymbol{\theta}\right)\right.\\&\qquad\left.-\hat{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k-1)}\right)\right),\end{aligned} \tag{9}$$

where $\{\gamma^{(k)}\}$ is a decreasing sequence of positive step sizes.

(M) *Maximization:* find the new estimate

$$\boldsymbol{\theta}^{(k+1)}=\arg\max_{\boldsymbol{\theta}}\hat{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right). \tag{10}$$

The SAEM requires a smaller amount of samples per iteration due to the recycling of the previous simulations. A small value of L is enough to ensure satisfying results [27].

When the conditional distribution is very complicated, and the simulation step (E-S1) of the SAEM cannot be directly performed, Kuhn and Lavielle proposed to combine the SAEM algorithm with a Markov Chain Monte Carlo (MCMC) procedure, which yields the SAEM-MCMC algorithm [19]. Assume the conditional distribution $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$ is the unique stationary distribution of the transition probability density function $\Pi_{\boldsymbol{\theta}}$, the simulation step of the SAEM is replaced with

(E-S2) Simulation: draw realizations $\mathbf{Z}^{(k,l)}(l=1,2\ldots,L)$ based on the transition probability density function $\Pi_{\boldsymbol{\theta}^{(k)}}(\mathbf{Z}^{(k-1,l)},\cdot)$.

For each l, the sequence $\{\mathbf{Z}^{(k,l)}\}_{k\geq0}$ is a Markov chain with the transition probability density function $\{\Pi_{\boldsymbol{\theta}^{(k)}}\}$. The Markov Chain generation mechanism needs to be well designed so that the sampling is efficient and the computational cost is not too high.

IV. SAEM-MCMC FOR STUDENT'S t AR MODEL

For the ML problem (5), if we only regard $\mathbf{y}_{\mathsf{m}}$ as missing data and apply the EM type algorithm, the resulting conditional distribution of the missing data is still complicated, and it is difficult to maximize the expectation or the approximated expectation of the complete data log-likelihood. Interestingly, the Student's t-distribution can be regarded as a Gaussian mixture [28]. Since $\varepsilon_t\sim t(0,\sigma^2,\nu)$, we can present it as a Gaussian mixture

$$\varepsilon_t|\sigma^2,\tau_t\sim\mathcal{N}\left(0,\frac{\sigma^2}{\tau_t}\right), \tag{11}$$

$$\tau_t\sim Gamma\left(\nu/2,\,\nu/2\right), \tag{12}$$

where τ_t is the mixture weight. Denote $\boldsymbol{\tau}=\{\tau_t,\,1<t\leq T\}$. We can use the EM type algorithm to solve the above optimization problem by regarding both $\mathbf{y}_{\mathsf{m}}$ and $\boldsymbol{\tau}$ as latent data, and $\mathbf{y}_{\mathsf{o}}$ as observed data.

The resulting complete data likelihood is

$$\begin{aligned}&L\left(\boldsymbol{\theta};\mathbf{y},\boldsymbol{\tau}\right)\\&=p\left(\mathbf{y},\boldsymbol{\tau};\boldsymbol{\theta}\right)\\&=\prod_{t=2}^{T}\left\{f_N\left(y_t;\varphi_0+\varphi_1y_{t-1},\frac{\sigma^2}{\tau_t}\right)f_g\left(\tau_t;\frac{\nu}{2},\frac{\nu}{2}\right)\right\}\\&=\prod_{t=2}^{T}\left\{\frac{1}{\sqrt{2\pi\sigma^2/\tau_t}}\exp\left(-\frac{1}{2\sigma^2/\tau_t}\left(y_t-\varphi_0-\varphi_1y_{t-1}\right)^2\right)\right.\\&\qquad\left.\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\tau_t^{\frac{\nu}{2}-1}\exp\left(-\frac{\nu}{2}\tau_t\right)\right\}\\&=\prod_{t=2}^{T}\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}\tau_t^{\frac{\nu-1}{2}}}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\tau_t}{2\sigma^2}\left(y_t-\varphi_0-\varphi_1y_{t-1}\right)^2-\frac{\nu}{2}\tau_t\right),\end{aligned} \tag{13}$$

where $f_N(\cdot)$ and $f_g(\cdot)$ denote the pdf's of the Normal (Gaussian) and gamma distributions, respectively. Through some simple derivation, it is observed that the likelihood of complete data belongs to the curved exponential family [29], i.e., the pdf can be written as

$$L\left(\boldsymbol{\theta};\mathbf{y},\boldsymbol{\tau}\right)=h\left(\mathbf{y},\boldsymbol{\tau}\right)\exp\left(-\psi\left(\boldsymbol{\theta}\right)+\left\langle \mathsf{s}\left(\mathbf{y}_{\mathsf{o}},\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau}\right),\boldsymbol{\phi}\left(\boldsymbol{\theta}\right)\right\rangle\right), \tag{14}$$

where $\langle\cdot,\cdot\rangle$ is the inner product,

$$h\left(\mathbf{y},\boldsymbol{\tau}\right)=\prod_{t=2}^{T}\tau_t^{-\frac{1}{2}}, \tag{15}$$

$$\begin{aligned}\psi\left(\boldsymbol{\theta}\right)&=-\left(T-1\right)\left\{\frac{\nu}{2}\log\left(\frac{\nu}{2}\right)-\log\left(\Gamma\left(\frac{\nu}{2}\right)\right)\right.\\&\qquad\left.-\frac{1}{2}\log\left(\sigma^2\right)-\frac{1}{2}\log\left(2\pi\right)\right\},\end{aligned} \tag{16}$$

$$\boldsymbol{\phi}\left(\boldsymbol{\theta}\right)=\left[\frac{\nu}{2},-\frac{1}{2\sigma^2},-\frac{\varphi_0^2}{2\sigma^2},-\frac{\varphi_1^2}{2\sigma^2},\frac{\varphi_0}{\sigma^2},\frac{\varphi_1}{\sigma^2},-\frac{\varphi_0\varphi_1}{\sigma^2}\right], \tag{17}$$

and the minimal sufficient statistics

$$\begin{aligned}\mathsf{s}\left(\mathbf{y}_{\mathsf{o}},\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau}\right)=&\left[\sum_{t=2}^{T}(\log(\tau_t)-\tau_t),\sum_{t=2}^{T}\tau_t y_t^2,\sum_{t=2}^{T}\tau_t,\sum_{t=2}^{T}\tau_t y_{t-1}^2,\right.\\&\left.\sum_{t=2}^{T}\tau_t y_t,\sum_{t=2}^{T}\tau_t y_t y_{t-1},\sum_{t=2}^{T}\tau_t y_{t-1}\right].\end{aligned} \tag{18}$$

Then the expectation of the complete data log-likelihood can be expressed as

$$\begin{aligned}&Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)\\&=\iint\log\left(L\left(\boldsymbol{\theta};\mathbf{y},\boldsymbol{\tau}\right)\right)p\left(\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau}|\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}^{(k)}\right)\mathsf{d}\mathbf{y}_{\mathsf{m}}\mathsf{d}\boldsymbol{\tau}\\&=\iint\log\Big(h\left(\mathbf{y},\boldsymbol{\tau}\right)\exp\Big(-\psi\left(\boldsymbol{\theta}\right)+\left\langle\mathsf{s}\left(\mathbf{y}_{\mathsf{o}},\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau}\right),\boldsymbol{\phi}\left(\boldsymbol{\theta}\right)\right\rangle\Big)\Big)\\&\quad\times p\left(\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau}|\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}^{(k)}\right)\mathsf{d}\mathbf{y}_{\mathsf{m}}\mathsf{d}\boldsymbol{\tau}\\&=\iint\log\left(h\left(\mathbf{y},\boldsymbol{\tau}\right)\right)p\left(\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau}|\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}^{(k)}\right)\mathsf{d}\mathbf{y}_{\mathsf{m}}\mathsf{d}\boldsymbol{\tau}\\&\quad-\psi\left(\boldsymbol{\theta}\right)+\Big\langle\iint\mathsf{s}\left(\mathbf{y}_{\mathsf{o}},\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau}\right)p\left(\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau}|\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}^{(k)}\right)\mathsf{d}\mathbf{y}_{\mathsf{m}}\mathsf{d}\boldsymbol{\tau},\\&\quad\boldsymbol{\phi}\left(\boldsymbol{\theta}\right)\Big\rangle\\&=-\psi\left(\boldsymbol{\theta}\right)+\left\langle\bar{\mathsf{s}}\left(\boldsymbol{\theta}^{(k)}\right),\boldsymbol{\phi}\left(\boldsymbol{\theta}\right)\right\rangle+\text{const.},\end{aligned} \tag{19}$$

where

$$\bar{\mathsf{s}}\left(\boldsymbol{\theta}^{(k)}\right)=\iint\mathsf{s}\left(\mathbf{y}_{\mathsf{o}},\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau}\right)p\left(\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau}|\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}^{(k)}\right)\mathsf{d}\mathbf{y}_{\mathsf{m}}\mathsf{d}\boldsymbol{\tau}. \tag{20}$$

The EM algorithm is conveniently simplified by utilizing the properties of the exponential family. The E step of the EM algorithm is reduced to the calculation of the expected minimal sufficient statistics $\bar{\mathsf{s}}(\boldsymbol{\theta}^{(k)})$, and the M step is reduced to the maximization of the function (19).

A. E Step

The conditional distribution of $\mathbf{y}_{\mathsf{m}}$ and $\boldsymbol{\tau}$ given $\mathbf{y}_{\mathsf{o}}$ and $\boldsymbol{\theta}$ is:

$$\begin{aligned}&p\left(\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau}|\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}\right)\\&=\frac{p\left(\mathbf{y},\boldsymbol{\tau};\boldsymbol{\theta}\right)}{p\left(\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}\right)}\\&=\frac{p\left(\mathbf{y},\boldsymbol{\tau};\boldsymbol{\theta}\right)}{\iint p\left(\mathbf{y},\boldsymbol{\tau};\boldsymbol{\theta}\right)\mathsf{d}\mathbf{y}_{\mathsf{m}}\mathsf{d}\boldsymbol{\tau}}\\&\propto p\left(\mathbf{y},\boldsymbol{\tau};\boldsymbol{\theta}\right)\\&=\prod_{t=2}^{T}\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}\tau_t^{\frac{\nu-1}{2}}}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{\tau_t}{2\sigma^2}\left(y_t-\varphi_0-\varphi_1 y_{t-1}\right)^2-\frac{\nu}{2}\tau_t\Big)\\&\propto\prod_{t=2}^{T}\tau_t^{\frac{\nu-1}{2}}\exp\left(-\frac{\tau_t}{2\sigma^2}\left(y_t-\varphi_0-\varphi_1 y_{t-1}\right)^2-\frac{\nu}{2}\tau_t\right).\end{aligned} \tag{21}$$

Since the integral $\iint p(\mathbf{y},\boldsymbol{\tau};\boldsymbol{\theta})\mathsf{d}\mathbf{y}_{\mathsf{m}}\mathsf{d}\boldsymbol{\tau}$ does not have a closed-from expression, we only know $p(\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau}|\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta})$ up to a scalar. In addition, the proportional term is complicated, and we cannot get closed-form expression for the conditional expectations $\bar{\mathsf{s}}(\boldsymbol{\theta}^{(k)})$ or $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$. Therefore, we resort to the SAEM-MCMC algorithm, which generates samples from the conditional distribution using a Markov chain process, and approximates the expectation $\bar{\mathsf{s}}(\boldsymbol{\theta}^{(k)})$ and $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ by a stochastic approximation.

We propose to use the Gibbs sampling method to generate the Markov chains. The Gibbs sampler divides the latent variables $(\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau})$ into two blocks $\boldsymbol{\tau}$ and $\mathbf{y}_{\mathsf{m}}$, and then generates a Markov chain of samples from the distribution $p(\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau}|\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta})$ by drawing realizations from its conditional distributions $p(\boldsymbol{\tau}|\mathbf{y}_{\mathsf{m}},\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta})$ and $p(\mathbf{y}_{\mathsf{m}}|\boldsymbol{\tau},\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta})$ alternatively. More specifically, at iteration k, given the current estimate $\boldsymbol{\theta}^{(k)}$, the Gibbs sampler starts with $(\boldsymbol{\tau}^{(k-1,l)},\mathbf{y}_{\mathsf{m}}^{(k-1,l)})(l=1,2\ldots,L)$ and generate the next sample $(\boldsymbol{\tau}^{(k,l)},\mathbf{y}_{\mathsf{m}}^{(k,l)})$ via the following scheme:

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

Then the expected minimal sufficient statistics $\bar{\mathsf{s}}(\boldsymbol{\theta}^{(k)})$ and the expected complete data likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ are approximated by

$$\hat{\mathsf{s}}^{(k)}=\hat{\mathsf{s}}^{(k-1)}+\gamma^{(k)}\left(\frac{1}{L}\sum_{l=1}^{L}\mathsf{s}\left(\mathbf{y}_{\mathsf{o}},\mathbf{y}_{\mathsf{m}}^{(k,l)},\boldsymbol{\tau}^{(k,l)}\right)-\hat{\mathsf{s}}^{(k-1)}\right), \tag{22}$$

$$\hat{Q}\left(\boldsymbol{\theta},\hat{\mathsf{s}}^{(k)}\right)=-\psi\left(\boldsymbol{\theta}\right)+\left\langle\hat{\mathsf{s}}^{(k)},\boldsymbol{\phi}\left(\boldsymbol{\theta}\right)\right\rangle+\text{const.} \tag{23}$$

Lemmas 1 and 2 give the two conditional distributions $p(\boldsymbol{\tau}|\mathbf{y}_{\mathsf{m}},\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta})$ and $p(\mathbf{y}_{\mathsf{m}}|\boldsymbol{\tau},\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta})$. Basically, to sample from them, we just need to draw realizations from certain Gaussian distributions and gamma distributions, which is simple. Based on the above sampling scheme, we can get the transition probability density function of the Markov chain as follows:

$$\Pi_{\boldsymbol{\theta}}\left(\mathbf{y}_{\mathsf{m}},\boldsymbol{\tau},\,\mathbf{y}_{\mathsf{m}}',\boldsymbol{\tau}'\right)=p\left(\boldsymbol{\tau}'|\mathbf{y}_{\mathsf{m}},\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}\right)p\left(\mathbf{y}_{\mathsf{m}}'|\boldsymbol{\tau}',\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}\right). \tag{24}$$

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

$$p\left(\boldsymbol{\tau}|\mathbf{y}_{\mathsf{m}},\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}\right)=\prod_{t=2}^{T}p\left(\tau_t|\mathbf{y}_{\mathsf{m}},\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}\right). \tag{25}$$

In addition, τ_t follows a gamma distribution:

$$\begin{aligned}&\tau_t|\mathbf{y}_{\mathsf{m}},\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}\\&\sim\ Gamma\left(\frac{\nu+1}{2},\ \frac{\left(y_t-\varphi_0-\varphi_1y_{t-1}\right)^2/\sigma^2+\nu}{2}\right).\end{aligned} \tag{26}$$

Proof: See Appendix A-A. ■

Lemma 2: Given $\boldsymbol{\tau}$, $\mathbf{y}_{\mathsf{o}}$, and $\boldsymbol{\theta}$, the missing blocks $\mathbf{y}_d=[y_{t_d+1},y_{t_d+2},\ldots,y_{t_d+n_d}]^T$, where $d=1,2,\ldots,D$, are independent from each other, i.e.,

$$p\left(\mathbf{y}_{\mathsf{m}}|\boldsymbol{\tau},\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}\right)=\prod_{d=1}^{D}p\left(\mathbf{y}_d|\boldsymbol{\tau},\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}\right). \tag{27}$$

In addition, the conditional distribution of $\mathbf{y}_d$ only depends on the two nearest observed samples y_{t_d} and $y_{t_d+n_d+1}$ with

$$\mathbf{y}_d|\boldsymbol{\tau},\mathbf{y}_{\mathsf{o}};\boldsymbol{\theta}\sim\mathcal{N}\left(\boldsymbol{\mu}_d,\boldsymbol{\Sigma}_d\right), \tag{28}$$

where the i-th component of $\boldsymbol{\mu}_d$

$$\begin{aligned}\boldsymbol{\mu}_{d(i)}=&\sum_{q=0}^{i-1}\varphi_1^q\varphi_0+\varphi_1^iy_{t_d}+\frac{\sum_{q=1}^{i}\frac{\varphi_1^{i-2q}}{\tau_{t_d+q}}}{\sum_{q=1}^{n_d+1}\frac{\varphi_1^{n_d+1-2q}}{\tau_{t_d+q}}}\\&\times\left(y_{t_d+n_d+1}-\sum_{q=0}^{n_d}\varphi_1^q\varphi_0-\varphi_1^{n_d+1}y_{t_d}\right),\end{aligned} \tag{29}$$

and the component in the i-th column and the j-th row of $\boldsymbol{\Sigma}_d$

$$\begin{aligned}&\Sigma_{d(i,j)}\\&=\left(\sum_{q=1}^{\min(i,j)}\frac{\varphi_1^{i+j-2q}}{\tau_{t_d+q}}-\frac{\left(\sum_{q=1}^{i}\frac{\varphi_1^{i-2q}}{\tau_{t_d+q}}\right)\left(\sum_{q=1}^{j}\frac{\varphi_1^{j-2q}}{\tau_{t_d+q}}\right)}{\sum_{q=1}^{n_d+1}\frac{\varphi_1^{-2q}}{\tau_{t_d+q}}}\right)\sigma^2,\end{aligned} \tag{30}$$

where the sums of geometric progressions in $\boldsymbol{\mu}_{d(i)}$ can be simplified as

$$\sum_{q=0}^{i-1}\varphi_1^q\varphi_0=\begin{cases}i\varphi_0, & \varphi_1=1,\\ \frac{\varphi_0\left(\varphi_1^i-1\right)}{\varphi_1-1}, & \varphi_1\neq1,\end{cases} \tag{31}$$

and

$$\sum_{q=0}^{n_d}\varphi_1^q\varphi_0=\begin{cases}\left(n_d+1\right)\varphi_0, & \varphi_1=1,\\ \frac{\varphi_0\left(\varphi_1^{n_d+1}-1\right)}{\varphi_1-1}, & \varphi_1\neq1.\end{cases} \tag{32}$$

Proof: See Appendix A-B. ■

B. M Step

After obtaining the approximation $\hat{Q}(\boldsymbol{\theta},\hat{\mathrm{s}}^{(k)})$ in (23), we need to maximize it to update the estimates. The function $\hat{Q}(\boldsymbol{\theta},\hat{\mathrm{s}}^{(k)})$ can be rewritten as

$$\begin{aligned}&\hat{Q}\left(\boldsymbol{\theta},\hat{\mathrm{s}}^{(k)}\right)\\&=-\psi\left(\boldsymbol{\theta}\right)+\left\langle\hat{\mathrm{s}}^{(k)},\boldsymbol{\phi}\left(\boldsymbol{\theta}\right)\right\rangle+\mathrm{const.}\\&=(T-1)\left\{\frac{\nu}{2}\log\left(\frac{\nu}{2}\right)-\log\left(\Gamma\left(\frac{\nu}{2}\right)\right)-\frac{1}{2}\log\left(\sigma^2\right)\right\}\\&\quad+\frac{\nu}{2}\hat{s}_1^{(k)}-\frac{\hat{s}_2^{(k)}}{2\sigma^2}-\frac{\varphi_0^2\hat{s}_3^{(k)}}{2\sigma^2}-\frac{\varphi_1^2\hat{s}_4^{(k)}}{2\sigma^2}+\frac{\varphi_0\hat{s}_5^{(k)}}{\sigma^2}+\frac{\varphi_1\hat{s}_6^{(k)}}{\sigma^2}\\&\quad-\frac{\varphi_0\varphi_1\hat{s}_7^{(k)}}{\sigma^2}+\mathrm{const,}\end{aligned} \tag{33}$$

where $\hat{s}_i^{(k)}$ $(i=1,2,\ldots,7)$ is the i-th component of $\hat{\mathrm{s}}^{(k)}$.

The optimization of φ_0, φ_1, and σ^2 is decoupled from the optimization of ν. Setting the derivatives of $\hat{Q}(\boldsymbol{\theta},\hat{\mathrm{s}}^{(k)})$ with respect to to φ_0, φ_1, and σ^2 to 0 gives

$$\varphi_0^{(k+1)}=\frac{\hat{s}_5^{(k)}-\varphi_1^{(k+1)}\hat{s}_7^{(k)}}{\hat{s}_3^{(k)}}, \tag{34}$$

$$\varphi_1^{(k+1)}=\frac{\hat{s}_3^{(k)}\hat{s}_6^{(k)}-\hat{s}_5^{(k)}\hat{s}_7^{(k)}}{\hat{s}_3^{(k)}\hat{s}_4^{(k)}-\left(\hat{s}_7^{(k)}\right)^2}, \tag{35}$$

and

$$\begin{aligned}\left(\sigma^{(k+1)}\right)^2=&\frac{1}{T-1}\left(\hat{s}_2^{(k)}+\left(\varphi_0^{(k+1)}\right)^2\hat{s}_3^{(k)}+\left(\varphi_1^{(k+1)}\right)^2\hat{s}_4^{(k)}\right.\\&-2\varphi_0^{(k+1)}\hat{s}_5^{(k)}-2\varphi_1^{(k+1)}\hat{s}_6^{(k)}\\&\left.+2\varphi_0^{(k+1)}\varphi_1^{(k+1)}\hat{s}_7^{(k)}\right).\end{aligned} \tag{36}$$

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

$$\nu^{(k+1)}=\underset{\nu>0}{\arg\max}\ f\left(\nu,\hat{s}_1^{(k)}\right) \tag{37}$$

with $f(\nu,\hat{s}_1^{(k)})=\{\frac{\nu}{2}\log(\frac{\nu}{2})-\log(\Gamma(\frac{\nu}{2}))\}+\frac{\nu\hat{s}_1^{(k)}}{2(T-1)}$. According to Proposition 1 in [30], $\nu^{(k+1)}$ always exists and is unique. As suggested in [30], the maximizer $\nu^{(k+1)}$ can be obtained by one-dimensional search, such as half interval method [31].

The resulting SAEM-MCMC algorithm is summarized in Algorithm 1.

C. Particular Cases

In cases where some parameters in $\boldsymbol{\theta}$ are known, we just need to change the updates in M step accordingly, and the simulation and approximation steps remain the same. For example, if we know that the time series is zero mean [1], [12], i.e., $\varphi_0=0$, then the update for $\varphi_0^{(k+1)}$ and $\varphi_1^{(k+1)}$ should be replaced with

$$\varphi_0^{(k+1)}=0, \tag{38}$$

and

$$\varphi_1^{(k+1)}=\frac{\hat{s}_6^{(k)}}{\hat{s}_4^{(k)}}, \tag{39}$$

If the time series is known to follow the random walk model [14], which is a special case of AR(1) model with $\varphi_1=1$, then

Algorithm 1: SAEM-MCMC Algorithm for Student's t AR(1).

```
Initialize $\boldsymbol{\theta}^{(0)} \in \Theta$, $\hat{\mathbf{s}}^{(0)} = 0$, $k = 0$, and $\mathbf{y}_{\mathsf{m}}^{(0,l)}$ for $l = 1, 2 \ldots, L$..
for $k = 1, 2, \ldots$ do
    Simulation:
    for $l = 1, 2 \ldots, L$ do
        sample $\boldsymbol{\tau}^{(k,l)}$ from $p(\boldsymbol{\tau}|\mathbf{y}_{\mathsf{m}}^{(k-1,l)}, \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}^{(k)})$ using Lemma 1,
        sample $\mathbf{y}_{\mathsf{m}}^{(k,l)}$ for $p(\mathbf{y}_{\mathsf{m}}|\boldsymbol{\tau}^{(k,l)}, \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}^{(k)})$ using Lemma 2.
    end for
    Stochastic approximation: evaluate $\hat{\mathbf{s}}^{(k)}$ and $\hat{Q}(\boldsymbol{\theta}, \hat{\mathbf{s}}^{(k)})$ as in (22) and (23) respectively.
    Maximization: update $\boldsymbol{\theta}^{(k+1)}$ as in (34), (35), (36) and (37).
    if stopping criteria is met then
        terminate loop
    end if
end for
```

the update for $\varphi_0^{(k+1)}$ and $\varphi_1^{(k+1)}$ should be replaced with

$$\varphi_0^{(k+1)} = \frac{\hat{s}_5^{(k)} - \hat{s}_7^{(k)}}{\hat{s}_3^{(k)}}, \tag{40}$$

and

$$\varphi_1^{(k+1)} = 1. \tag{41}$$

D. Generalization to AR(p)

The above ML estimation method can be immediately generalized to the Student's t AR(p) model:

$$y_t = \varphi_0 + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t, \tag{42}$$

where $\varepsilon_t \overset{i.i.d.}{\sim} t(0, \sigma^2, \nu)$. Similarly, we can apply the SAEM-MCMC algorithm to obtain the estimates by considering $\boldsymbol{\tau}$ and $\mathbf{y}_{\mathsf{m}}$ as latent data, and $\mathbf{y}_{\mathsf{o}}$ as observed data. At each iteration, we draw some realizations of $\boldsymbol{\tau}$ and $\mathbf{y}_{\mathsf{m}}$ from the conditional distribution $p(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}|\mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}^{(k)})$ to approximate the expectation function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, and maximize the approximation $\hat{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ to update the estimates. The main difference is that the conditional distribution of the AR(p) will become more complicated than that of the AR(1), since each sample of the AR(p) has more dependence on the previous samples. To deal with this challenge, when applying the Gibbs sampling, we can divide the the latent data $(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau})$ into more blocks, $\boldsymbol{\tau}$ as a block and each $y_{i \in C_m}$ as a block, so that the distribution of each block of latent variables conditional on other latent variables will be easy to obtain and sample from. For limit of space, we do not go into details here, and we will consider this in our future work.

V. Convergence

In this section, we provide theoretical guarantee for the convergence of the proposed algorithm. The convergence of the simple deterministic EM algorithm has been addressed by many different authors, starting from the seminal work in [23], to a more general consideration in [24]. However, the convergence analysis of stochastic variants of the EM algorithm, like the MCEM, SAEM and SAEM-MCMC algorithms, is challenging due to the randomness of sampling. See [18], [19], [32]–[35] for a more general overview of these stochastic EM algorithms and their convergence analysis. Of specific interest, the authors in [18] introduced the SAEM algorithm, and established the almost sure convergence to the stationary points of the observed data likelihood under mild additional conditions. The authors in [19] coupled the SAEM framework with an MCMC procedure, and they have given the convergence conditions for the SAEM-MCMC algorithm when the complete data likelihood belongs to the curved exponential family. The given set of conditions in our case is as follows.

(M1) For any $\boldsymbol{\theta} \in \Theta$,

$$\iint \|\mathsf{s}(\mathbf{y}_{\mathsf{o}}, \mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau})\| p(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}|\mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta})\, \mathsf{d}\mathbf{y}_{\mathsf{m}} \mathsf{d}\boldsymbol{\tau} < \infty. \tag{43}$$

(M2) $\psi(\boldsymbol{\theta})$ and $\boldsymbol{\phi}(\boldsymbol{\theta})$ are twice continuously differentiable on Θ.

(M3) The function

$$\bar{\mathsf{s}}(\boldsymbol{\theta}) = \iint \mathsf{s}(\mathbf{y}_{\mathsf{o}}, \mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau})\, p(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}|\mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta})\, \mathsf{d}\mathbf{y}_{\mathsf{m}} \mathsf{d}\boldsymbol{\tau} \tag{44}$$

is continuously differentiable on Θ.

(M4) The objective function

$$l(\boldsymbol{\theta}; \mathbf{y}_{\mathsf{o}}) = \log \left(\iint p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta})\, \mathsf{d}\mathbf{y}_{\mathsf{m}} \mathsf{d}\boldsymbol{\tau} \right) \tag{45}$$

is continuously differentiable on Θ, and

$$\partial_{\boldsymbol{\theta}} \iint p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta})\, \mathsf{d}\mathbf{y}_m \mathsf{d}\boldsymbol{\tau} = \iint \partial_{\boldsymbol{\theta}} p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta})\, \mathsf{d}\mathbf{y}_m \mathsf{d}\boldsymbol{\tau}. \tag{46}$$

(M5) For $Q(\boldsymbol{\theta}, \bar{\mathsf{s}}) = -\psi(\boldsymbol{\theta}) + \langle \bar{\mathsf{s}}, \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle$ + const., there exists a function $\tilde{\boldsymbol{\theta}}(\bar{\mathsf{s}})$ such that $\forall \bar{\mathsf{s}}$ and $\forall \boldsymbol{\theta} \in \Theta, Q(\tilde{\boldsymbol{\theta}}(\bar{\mathsf{s}}), \bar{\mathsf{s}}) \geq Q(\boldsymbol{\theta}, \bar{\mathsf{s}})$. In addition, the function $\tilde{\boldsymbol{\theta}}(\bar{\mathsf{s}})$ is continuously differentiable.

(SAEM1) For all k, $\gamma^{(k)} \in [0, 1]$, $\sum_{k=1}^{\infty} \gamma^{(k)} = \infty$ and there exists $\frac{1}{2} < \lambda \leq 1$ such that $\sum_{k=1}^{\infty} (\gamma^{(k)})^{1+\lambda} < \infty$.

(SAEM2) $l(\boldsymbol{\theta}; \mathbf{y}_{\mathsf{o}})$ is d times differentiable on Θ, where $d = 7$ is the dimension of $\mathsf{s}(\mathbf{y}_{\mathsf{o}}, \mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau})$, and $\tilde{\boldsymbol{\theta}}(\mathsf{s})$ is d times differentiable.

(SAEM3)

1) The chain takes its values in a compact set Ω.
2) The $\mathsf{s}(\mathbf{y}_{\mathsf{o}}, \mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau})$ is bounded on Ω, and the sequence $\{\hat{\mathsf{s}}^{(k)}\}$ takes its values in a compact subset.
3) For any compact subset V of Θ, there exists a real constant L such that for any $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ in V^2

$$\sup_{(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') \in \Omega^2} \Big| \Pi_{\boldsymbol{\theta}}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') - \Pi_{\boldsymbol{\theta}'}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') \Big| \leq L|\boldsymbol{\theta} - \boldsymbol{\theta}'|. \tag{47}$$

4) The transition probability $\Pi_{\boldsymbol{\theta}}$ generates a uniformly ergodic chain whose invariant probability is the conditional distribution $p(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau} | \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta})$.

In summary, the conditions (M1)–(M5) are all about the model, and are conditions for the convergence of the deterministic EM algorithm. The conditions (M1) and (M3) require the boundedness and continuous differentiability of the expectation of the sufficient statistics. The conditions (M2) and (M4) guarantee the continuous differentiability of the complete data log-likelihood $l(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{\tau})$, the expectation of the complete data likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, and the observed data log-likelihood $l(\boldsymbol{\theta}; \mathbf{y}_{\mathsf{o}})$. The condition (M5) indicates the existence of a global maximizer for $Q(\boldsymbol{\theta}, \bar{\mathbf{s}})$.

The conditions (SAEM1)-(SAEM3) are additional requirements for the SAEM-MCMC convergence. The condition (SAEM1) is about the step sizes $\{\gamma^{(k)}\}$. This condition can be easily satisfied by choosing the step sizes properly. It is recommended to set $\gamma^{(k)} = 1$ for $1 \leq k \leq K$ and $\gamma^{(k)} = \frac{1}{k-K}$ for $k \geq K + 1$, where K is a positive integer, since the initial guess $\boldsymbol{\theta}^{(0)}$ may be far from the ML estimates we are looking for, and choosing the first K step sizes equal to 1 allows the sequence $\{\boldsymbol{\theta}^{(k)}\}$ to have a large variation and then converge to a neighborhood of the maximum likelihood [27]. The condition (SAEM2) requires $d = 7$ times differentiability of $l(\boldsymbol{\theta}; \mathbf{y}_{\mathsf{o}})$ and $\hat{\boldsymbol{\theta}}(\bar{\mathbf{s}}^{(k)})$. The condition (SAEM3) imposes some constraints on the generated Markov chains.

In [19], the authors have established the convergence of the SAEM-MCMC algorithm to the stationary points. However, their analysis assumes that complete data likelihood belongs to the curved exponential family, and all these conditions (M1)-(M5) and (SAEM1)-(SAEM3) are satisfied. These assumptions are very problem specific, and do not hold trivially for our case, since our conditional distribution of the latent variable is extremely complicated. To comment on the convergence of our proposed algorithm, we need to establish the conditions (M1)-(M5) and (SAEM1)-(SAEM3) one by one. Finally, we have the convergence result about our proposed algorithm summarized in the following theorem.

Theorem 1: Suppose that the parameter space Θ is set to be a sufficiently large bounded set[1] with the parameter $\nu > 2$, and the Markov chain generated from (25) and (27) takes values in a compact set[2], the sequence $\{\boldsymbol{\theta}^{(k)}\}$ generated by Algorithm 1 has the following asymptotic property: with probability 1, $\lim_{k \to +\infty} d(\boldsymbol{\theta}^{(k)}, \mathcal{L}) = 0$, where $d(\boldsymbol{\theta}^{(k)}, \mathcal{L})$ denotes the distance from $\boldsymbol{\theta}^{(k)}$ to the set of stationary points of observed data log-likelihood $\mathcal{L} = \{\boldsymbol{\theta} \in \Theta, \frac{\partial l(\boldsymbol{\theta}; \mathbf{y}_{\mathsf{o}})}{\partial \boldsymbol{\theta}} = 0\}$.

Proof: Please refer to Appendix VII-B for the proof of the conditions (M1)-(M5) and (SAEM2)-(SAEM3). The condition (SAEM1) can be be easily satisfied by choosing the step sizes properly as mentioned before. Upon establishing these conditions, the proof of this theorem follows straightforward from the analysis of the work in [19]. ■

VI. Simulations

In this section, we conduct a simulation study of the performance of the proposed ML estimator and the convergence of the proposed algorithm. First, we show that the proposed estimator is able to make good estimates of parameters from the incomplete time series which have been synthesized to fit the model. Second, we show its robustness to innovation outliers. Finally, we test it on a real financial time series, the Hang Seng index.

A. Parameter Estimation

In this subsection, we show the convergence of the proposed SAEM-MCMC algorithm and the performance of the proposed estimator on incomplete Student's t AR(1) time series with different numbers of samples and missing percentages. The estimation error is measured by the mean square error (MSE):

$$\mathsf{MSE}\,(\theta) := \mathsf{E}\left[\left(\hat{\theta} - \theta^{\mathsf{true}}\right)^2\right],$$

where $\hat{\theta}$ is the estimate for the parameter θ, and θ^{true} is its true value. The parameter θ can be φ_0, φ_1, σ^2, and ν. The expectation is approximated via Monte Carlo simulations using 100 independent incomplete time series.

We set $\varphi_0^{\mathsf{true}} = 1$, $\varphi_1^{\mathsf{true}} = 0.5$, $(\sigma^{\mathsf{true}})^2 = 0.01$, and $\nu^{\mathsf{true}} = 2.5$. For each incomplete data set $\mathbf{y}_{\mathsf{o}}$, we first randomly generate a complete time series $\{y_t\}$ with T samples based on the Student's t AR(1) model. Then n_{mis} number of samples are randomly deleted to obtain an incomplete time series. The missing percentage of the incomplete time series is $\rho := \frac{n_{\mathsf{mis}}}{T} \times 100\%$.

In Section V, we have established the convergence of the proposed SAEM-MCMC algorithm to the stationary points of the observed data likelihood. However, it is observed that the estimation result obtained by the algorithm can be sensitive to initializations due to the existence of multiple stationary points. This is an inevitable problem since it is a non-convex optimization problem. Interestingly, it is also observed that when we initialize our algorithm using the ML estimates assuming the Gaussian AR(1) model, the final estimates are significantly improved, in comparison to random initializations. The ML estimation of the Gaussian AR model from incomplete data has been introduced in [13], and the estimates can be easily obtained via the deterministic EM algorithm. We initialize $\varphi_0^{(0)}$, $\varphi_1^{(0)}$, and $(\sigma^{(0)})^2$ use the estimates from the Gaussian AR(1) model $(\varphi_0)_{\mathsf{g}}$, $(\varphi_1)_{\mathsf{g}}$, and $(\sigma^2)_{\mathsf{g}}$, and initialize $\mathbf{y}_{\mathsf{m}}^{(0,l)}$ using the mean of the conditional distribution $p(\mathbf{y}_{\mathsf{m}}; \mathbf{y}_{\mathsf{o}}, (\varphi_0)_{\mathsf{g}}, (\varphi_1)_{\mathsf{g}}, (\sigma^2)_{\mathsf{g}})$, which is a Gaussian distribution. The parameter $\nu^{(0)}$ is initialized as a random positive number. In each iteration, we draw $L = 10$ samples. For the step sizes, we set $\gamma^{(k)} = 1$ for $1 \leq k \leq 30$ and $\gamma^{(k)} = \frac{1}{k-30}$ for $k \geq 31$. Figure 1 gives an example of applying the proposed SAEM-MCMC algorithm to estimate the parameters on a synthetic AR(1) data set with $T = 300$ and a missing percentage $\rho = 10\%$. We can see that the algorithm converges in less than 100 iterations, where each iteration just needs $L = 10$ runs of Gibbs sampling, and also the final estimation error is small. Table I compares the estimation results of the Student's t AR model and the Gaussian AR model. This testifies our

[1] This means that the unconstrained maximizer of (33) (given by (34), (35), (36), and (37)) lies in this bounded set.

[2] Theoretically, the Markov chain generated from (25) and (27) takes its values in an unbounded set. However, in practice, the chain will not take very large values, and we can consider the chain takes values in a very large compact set [19], [27].

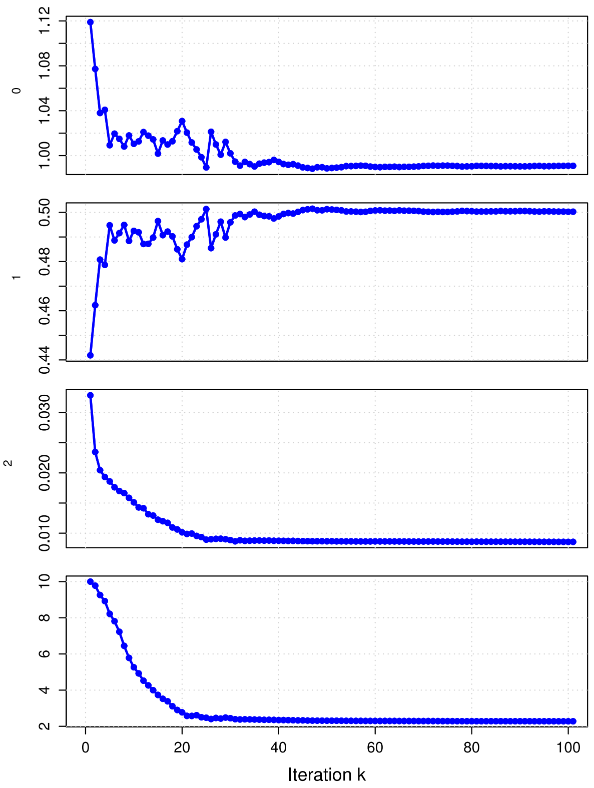

Fig. 1. Estimates versus iterations.

TABLE I
ESTIMATION RESULTS FOR INCOMPLETE STUDENT'S t AR(1)

	$\hat{\varphi}_0$	$\hat{\varphi}_1$	$(\hat{\sigma})^2$	$\hat{\nu}$
True value	1.000	0.500	0.010	2.5
Gaussian AR(1)	1.119	0.442	0.033	$+\infty$
Student's t AR(1)	0.989	0.501	0.009	2.234

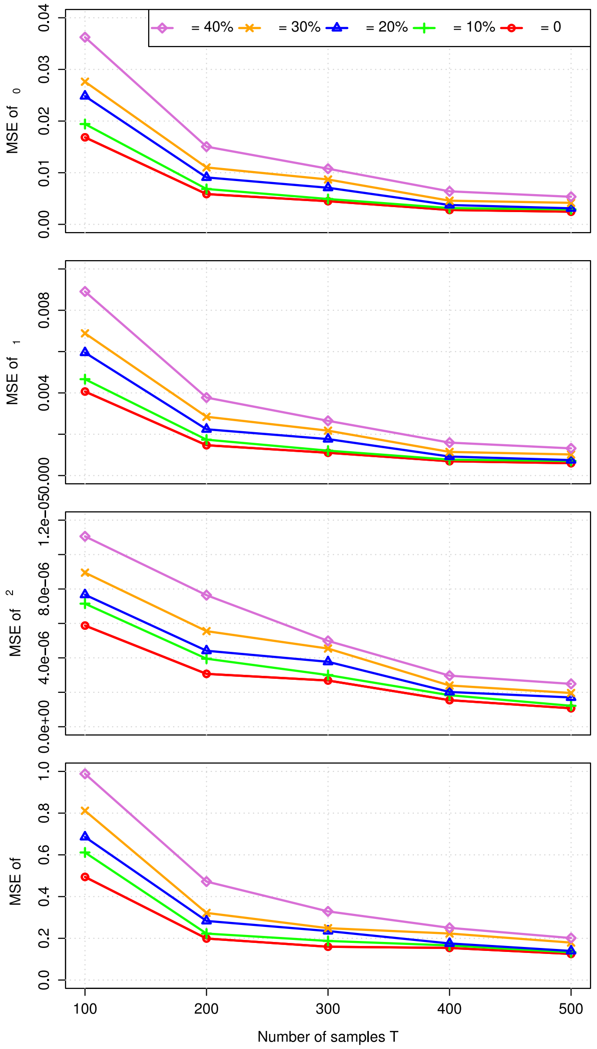

Fig. 2. MSEs for the incomplete time series with different number of samples and missing percentages.

argument that, for incomplete heavy-tailed data, the traditional method for incomplete Gaussian AR time series is too inefficient, and significant performance gain can be achieved by designing algorithms under heavy-tailed model.

Figure 2 shows the estimation results with the numbers of samples $T = 100, 200, 300, 400, 500$ and the missing percentages $\rho = 10\%, 20\%, 30\%, 40\%$. For reference, we have also given the ML estimation result from the complete data sets ($\rho = 0$), which is obtained using the algorithm in [11]. We can observe that our method performs satisfactorily well even for high percentage of missing data, and, with increasing sample sizes, the estimates with missing values match with the estimates of the complete data.

B. Robustness to Outliers

A useful characteristic of the Student's t is its resilience to outliers, which is not shared by the Gaussian distribution. Here we illustrate that the Student's t AR model can provide robust estimation of autoregressive coefficients under innovation outliers.

An innovation outlier is an outlier in the ε_t process, and it is a typical kind of outlier in AR time series [36], [37]. Due to the temporal dependence of AR time series data, an innovation outlier will affect not only the current observation y_t, but also subsequent observations. Figure 3 gives an example of a Gaussian AR(1) time series contaminated by four innovation outliers.

When an AR time series is contaminated by outliers, the traditional ML estimation of autoregressive coefficients based on the Gaussian AR model, which is equivalent to least squares fitting, will provide unreliable estimates. Although, for complete time series, there are numerous works about the robust estimation of autoregressive coefficients under outliers, unfortunately, less attention was paid to robust estimation from incomplete time series. As far as we know, only Kharin and Voloshko have considered robust estimation with missing values [16]. In their paper, they assume that ϕ_0 is known and equal to 0. To be consistent with Kharin's method, in this simulation, we also assume φ_0^{true} is known and $\varphi_0^{\text{true}} = 0$, although our method can also be applied to the case where φ_0^{true} is unknown.

We let $\varphi_1^{\text{true}} = 0.5$ and $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.01)$. Note here the innovations follow a Gaussian distribution. We randomly generate an incomplete Gaussian AR(1) time series with $T = 100$

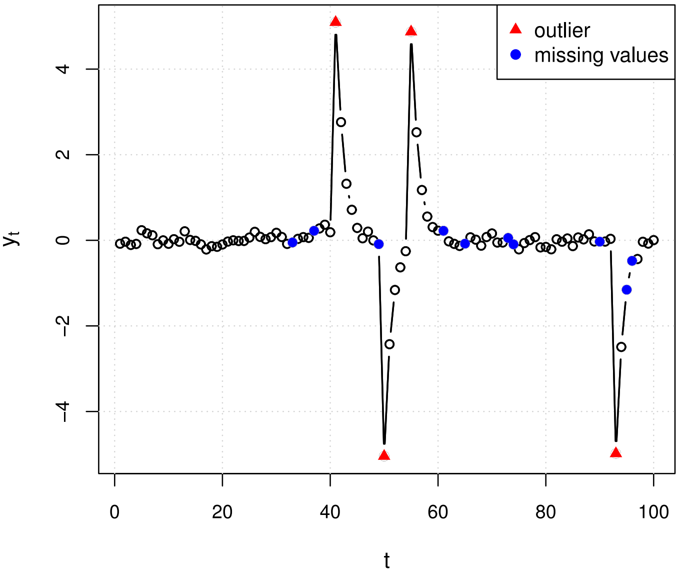

Fig. 3. Incomplete AR(1) time series with four innovation outliers.

TABLE II
ESTIMATION AND PREDICTION RESULTS FOR INCOMPLETE GAUSSIAN AR(1) TIME SERIES WITH OUTLIERS

	$\hat{\varphi}_1$ ($\varphi_1^{\text{true}} = 0.5$)	Averaged prediction error
Gaussian AR(1)	0.5337	0.0121
Student's t AR(1)	0.4947	0.0110
Kharin's method	0.4210	0.0212

samples and a missing percentage $\rho = 0.1$, and it is contaminated by four innovation outliers. The values of the innovation outliers are set to be 5, -5, 5, -5, and the positions are selected randomly. See Figure 3 for this incomplete contaminated time series. The Gaussian AR(1) model, the Student's t AR(1) model, and Kharin's method are applied to estimate the autoregressive coefficient φ_1. After obtaining the estimate $\hat{\varphi}_1$, we compute the one-step-ahead predictions $\hat{y}_t = \hat{\varphi}_1 y_{t-1}$ and the prediction error $(\hat{y}_t - y_t)^2$ for $t \in C_o$ and $t - 1 \in C_o$. It is not surprising that the outliers are poorly predicted, so we omit it when computing the averaged prediction error. Table II shows the estimation results and the one-step-ahead prediction errors. It is clear that the ML estimator based on the Gaussian AR(1) has been significantly affected by the presence of the outliers, while the Student's t AR(1) model is robust to them, since the outliers cause the innovations to have a heavy-tailed distribution, which can be modeled by the Student's t distribution. Kharin's method does not perform well, either, as this method is designed for addictive outliers and replacement outliers, rather than innovation outliers.

C. Real Data

Here we consider the returns of the Hang Seng index over 260 working days from Jan. 2017 to Nov. 2017 (excluding weekends and public holidays). Figure 4 shows the quantile-quantile (QQ) plot of the innovations obtained by fitting this time series to the Student's t AR(1) model. The deviation from the straight red line indicates that the innovations are significantly non-Gaussian and indeed heavy-tailed.

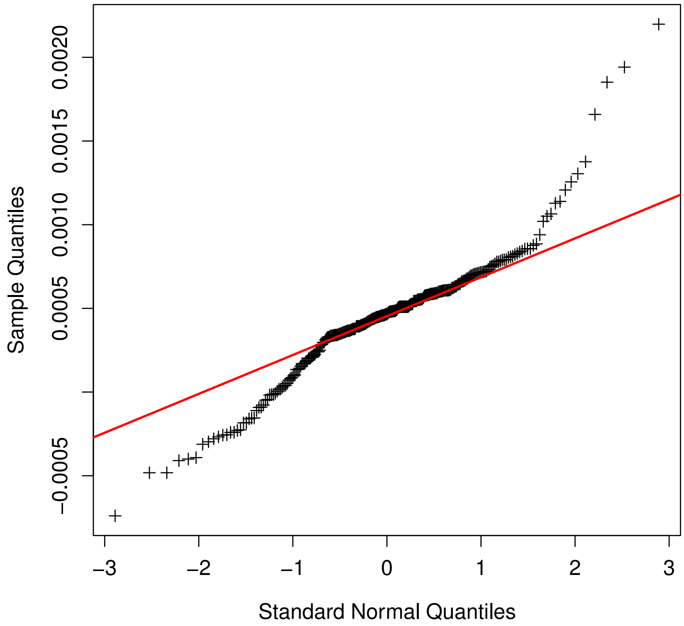

Fig. 4. Quantile-quantile plot of the innovations of the Hang Seng index returns.

We divide the 260 returns into two parts: the estimation data, which involves the first 250 samples, and the test data, which involves the remaining 10 samples. First, we fit the estimation data to the Gaussian AR(1) model and the Student's t AR(1) model, and estimate the parameters. Then we predict the test data using the one-step-ahead predication method based on the estimates, and compute the averaged prediction errors. Next, we randomly delete 10 of the estimation data, and estimate the parameters of the Gaussian AR(1) model and the Student's t AR(1) model from this incomplete data set. Finally, we also make predictions and compute the averaged prediction errors based on these estimates of the parameters. The result is summarized in Table III. We have the following conclusions: i) the Student's t AR(1) model performs better than the Gaussian AR(1) model for this heavy-tailed time series, ii) the proposed parameter estimation method for incomplete Student's t AR(1) time series can provide similar estimates to the result of complete data.

VII. CONCLUSIONS

In this paper, we have considered parameter estimation of the heavy-tailed AR model with missing values. We have formulated an ML estimation problem and developed an efficient approach to obtain the estimates based on the stochastic EM. Since the conditional distribution of the latent data in our case is complicated, we proposed a Gibbs sampling scheme to draw realizations from it. The convergence of the proposed algorithm to the stationary points has been established. Simulations show that the proposed approach can provide reliable estimates from incomplete time series with different percentages of missing values, and is robust to outliers. Although in this paper we only focus on the univariate AR model with the Student's t distributed innovations due to the limit of the space, our method can be

TABLE III
ESTIMATION AND PREDICTION RESULTS FOR THE HANG SENG INDEX RETURNS

	$\hat{\varphi}_0$	$\hat{\varphi}_1$	$(\hat{\sigma})^2$	$\hat{\nu}$	Averaged prediction error
Complete data assuming Gaussian innovations	7.548×10^{-4}	-1.058×10^{-1}	1.702×10^{-5}	$+\infty$	9.141×10^{-6}
Incomplete data assuming Gaussian innovations	8.618×10^{-4}	-1.253×10^{-1}	1.665×10^{-5}	$+\infty$	9.455×10^{-6}
Complete data assuming Student's t innovations	5.440×10^{-4}	-9.580×10^{-2}	6.524×10^{-6}	2.622	8.836×10^{-6}
Incomplete data assuming Student's t innovations	5.538×10^{-4}	-9.459×10^{-2}	6.331×10^{-6}	2.671	8.831×10^{-6}

extended to multivariate AR model and also other heavy-tailed distributed innovations.

APPENDIX A
PROOF FOR LEMMAS 1 AND 2

A. *Proof for Lemma 1*

The conditional distribution of $\boldsymbol{\tau}|\mathbf{y}_{\mathsf{m}}, \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}$ is

$$
\begin{aligned}
& p\left(\boldsymbol{\tau}|\mathbf{y}_{\mathsf{m}}, \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}\right) \\
& = \frac{p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right)}{p\left(\mathbf{y}; \boldsymbol{\theta}\right)} \\
& \propto p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \\
& = \prod_{t=2}^{T} \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} \tau_t^{\frac{\nu-1}{2}}}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{\tau_t}{2\sigma^2}\left(y_t - \varphi_0 - \varphi_1 y_{t-1}\right)^2 - \frac{\nu}{2}\tau_t\Big) \\
& \propto \prod_{t=2}^{T} \tau_t^{\frac{\nu-1}{2}} \exp\left(-\left(\frac{\left(y_t - \varphi_0 - \varphi_1 y_{t-1}\right)^2}{2\sigma^2} + \frac{\nu}{2}\right)\tau_t\right),
\end{aligned}
\tag{48}
$$

which implies that $\{\tau_t\}$ are independent from each other with

$$
\begin{aligned}
& p\left(\tau_t|\mathbf{y}_{\mathsf{m}}, \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}\right) \\
& \propto \tau_t^{\frac{\nu-1}{2}} \exp\left(-\left(\frac{\left(y_t - \varphi_0 - \varphi_1 y_{t-1}\right)^2}{2\sigma^2} + \frac{\nu}{2}\right)\tau_t\right).
\end{aligned}
\tag{49}
$$

Comparing this expression with the pdf of the gamma distribution, we get that $\tau_t|\mathbf{y}_{\mathsf{m}}, \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}$ follows a gamma distribution:

$$
\begin{aligned}
& \tau_t|\mathbf{y}_{\mathsf{m}}, \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta} \\
& \sim Gamma\left(\frac{\nu+1}{2}, \frac{\left(y_t - \varphi_0 - \varphi_1 y_{t-1}\right)^2/\sigma^2 + \nu}{2}\right).
\end{aligned}
\tag{50}
$$

B. *Proof for Lemma 2*

According to the Gaussian mixture representation (11) and (12), given $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$, ε_t follows a Gaussian distribution: $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{\tau_t})$. From equation (2), we can see that, given $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$, the distribution of y_t conditional on all the preceding data $\mathcal{F}_{t-1}$, only depends on the previous sample y_{t-1}:

$$
p\left(y_t|\boldsymbol{\tau}, \mathcal{F}_{t-1}; \boldsymbol{\theta}\right) = p\left(y_t|\boldsymbol{\tau}, y_{t-1}; \boldsymbol{\theta}\right). \tag{51}
$$

In addition, the distribution of y_t conditional on all the preceding observed data $\mathcal{F}_{t-1}^{o}$, $\boldsymbol{\tau}$, and $\boldsymbol{\theta}$, only depends on the nearest observed sample:

$$
\begin{aligned}
& p\left(y_t|\boldsymbol{\tau}, \mathcal{F}_{t-1}^{o}; \boldsymbol{\theta}\right) \\
& = \begin{cases} p\left(y_t|\boldsymbol{\tau}, y_{t-1}; \boldsymbol{\theta}\right) & t = t_d + n_d + 2, \ldots, t_{d+1}, \\ & \text{for } d = 0, 1, \ldots, D, \\ p\left(y_t|\boldsymbol{\tau}, y_{t-n_d-1}; \boldsymbol{\theta}\right) & t = t_d + n_d + 1, \\ & \text{for } d = 1, 2, \ldots, D. \end{cases}
\end{aligned}
\tag{52}
$$

The first case refers to the situation where the previous sample y_{t-1} is observed, while the second case is when y_{t-1} is missing.

Based on the above properties, we have

$$
\begin{aligned}
p\left(\mathbf{y}_{\mathsf{m}}|\boldsymbol{\tau}, \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}\right) & = \frac{\prod_{t=2}^{T} p\left(y_t|\boldsymbol{\tau}, \mathcal{F}_{t-1}; \boldsymbol{\theta}\right)}{\prod_{t\in C_o} p\left(y_t|\boldsymbol{\tau}, \mathcal{F}_{t-1}^{o}; \boldsymbol{\theta}\right)} && \text{(53a)} \\
& = \frac{\prod_{t=2}^{T} p\left(y_t|\boldsymbol{\tau}, y_{t-1}; \boldsymbol{\theta}\right)}{\prod_{d=0}^{D} \prod_{t=t_d+n_d+2}^{t_{d+1}} p\left(y_t|\boldsymbol{\tau}, y_{t-1}; \boldsymbol{\theta}\right)} \\
& \quad \times \frac{1}{\prod_{d=1}^{D} p\left(y_{t_d+n_d+1}|\boldsymbol{\tau}, y_{t_d}; \boldsymbol{\theta}\right)} && \text{(53b)} \\
& = \frac{\prod_{d=1}^{D} \prod_{t=t_d+1}^{t_d+n_d+1} p\left(y_t|\boldsymbol{\tau}, y_{t-1}; \boldsymbol{\theta}\right)}{\prod_{d=1}^{D} p\left(y_{t_d+n_d+1}|\boldsymbol{\tau}, y_{t_d}; \boldsymbol{\theta}\right)} && \text{(53c)} \\
& = \prod_{d=1}^{D} \frac{p\left(\mathbf{y}_d, y_{t_d+n_d+1}|\boldsymbol{\tau}, y_{t_d}; \boldsymbol{\theta}\right)}{p\left(y_{t_d+n_d+1}|\boldsymbol{\tau}, y_{t_d}; \boldsymbol{\theta}\right)} && \text{(53d)} \\
& = \prod_{d=1}^{D} p\left(\mathbf{y}_d|\boldsymbol{\tau}, y_{t_d}, y_{t_d+n_d+1}; \boldsymbol{\theta}\right), && \text{(53e)}
\end{aligned}
$$

where the equations (53a) and (53e) are from the definition of conditional pdf, the equation (53b) is from (51) and (52). The equation (53e) implies that the different missing blocks $\{\mathbf{y}_d\}$ are independent from each other, and the conditional distribution of $\mathbf{y}_d$ only depends on the two nearest observed samples y_{t_d} and $y_{t_d+n_d+1}$.

To obtain the pdf of the missing block $p(\mathbf{y}_d|\boldsymbol{\tau}, y_{t_d}, y_{t_d+n_d+1}; \boldsymbol{\theta})$, we first analyze the joint pdf of the missing block and next observed sample $\mathbf{y}_{cd} = [\mathbf{y}_d^T, y_{t_d+n_d+1}]^T = [y_{t_d+1}, y_{t_d+2}, \ldots, y_{t_d+n_d+1}]$: $p(\mathbf{y}_{cd}|\boldsymbol{\tau}, y_{t_d}; \boldsymbol{\theta})$. Given $\boldsymbol{\tau}$, y_{t_d}, and $\boldsymbol{\theta}$, from (2), we have

$$
\begin{aligned}
y_{t_d+i} & = \varphi_0 + \varphi_1 y_{t_d+i-1} + \varepsilon_{t_d+i} \\
& = \varphi_0 + \varphi_1\left(\varphi_0 + \varphi_1 y_{t_d+i-2} + \varepsilon_{t_d+i-1}\right) + \varepsilon_{t_d+i} \\
& = \varphi_0 + \varphi_1\varphi_0 + \varphi_1^2 y_{t_d+i-2} + \varphi_1\varepsilon_{t_d+i-1} + \varepsilon_{t_d+i} \\
& = \sum_{q=0}^{i-1} \varphi_1^q \varphi_0 + \varphi_1^i y_{t_d} + \sum_{q=1}^{i} \varphi_1^{(i-q)} \varepsilon_{t_d+q},
\end{aligned}
\tag{54}
$$

for $i = 1, 2, \ldots, n_d + 1$, which means that y_{t_d+i} can be expressed as the sum of the constant $\sum_{q=0}^{i-1} \varphi_1^q \varphi_0 + \varphi_1^i y_{t_d}$ and a linear combination of the independent Gaussian random variables $\varepsilon_{t_d+1}, \varepsilon_{t_d+2}, \ldots, \varepsilon_{t_d+i}$. Therefore, we can obtain that $\mathbf{y}_{cd}$ follows a Gaussian distribution as follows:

$$\mathbf{y}_{cd}|\boldsymbol{\tau}, y_{t_d}; \boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\mu}_{cd}, \boldsymbol{\Sigma}_{cd}\right), \tag{55}$$

where the i-th component of $\boldsymbol{\mu}_{cd}$

$$\begin{aligned}\mu_{cd(i)} &= \mathsf{E}\left[y_{t_d+i}\right] \\ &= \mathsf{E}\left[\sum_{q=0}^{i-1} \varphi_1^q \varphi_0 + \varphi_1^i y_{t_d} + \sum_{q=1}^{i} \varphi_1^{(i-q)} \varepsilon_{t_d+q}\right] \\ &= \sum_{q=0}^{i-1} \varphi_1^q \varphi_0 + \varphi_1^i y_{t_d} + \sum_{q=1}^{i} \varphi_1^{(i-q)} \mathsf{E}\left[\varepsilon_{t_d+q}\right] \\ &= \sum_{q=0}^{i-1} \varphi_1^q \varphi_0 + \varphi_1^i y_{t_d},\end{aligned} \tag{56}$$

and the component in the i-th column and the j-th row of $\boldsymbol{\Sigma}_{cd}$

$$\begin{aligned}\Sigma_{cd(i,j)} &= \mathsf{E}\left[\left(y_{t_d+i} - \mu_{cd(i)}\right)\left(y_{t_d+j} - \mu_{cd(j)}\right)\right] \\ &= \mathsf{E}\left[\left(\sum_{q_1=1}^{i} \varphi_1^{(i-q_1)} \varepsilon_{t_d+q_1}\right)\left(\sum_{q_2=1}^{j} \varphi_1^{(j-q_2)} \varepsilon_{t_d+q_2}\right)\right] \\ &= \sum_{q_1=1}^{i} \sum_{q_2=1}^{j} \varphi_1^{(i+j-q_1-q_2)} \mathsf{E}\left[\varepsilon_{t_d+q_1} \varepsilon_{t_d+q_2}\right] \\ &= \sigma^2 \sum_{q=1}^{\min(i,j)} \frac{\varphi_1^{(i+j-2q)}}{\tau_{t_d+q}}.\end{aligned} \tag{57}$$

with the last equation following from

$$\mathsf{E}\left[\varepsilon_{t_d+q_1} \varepsilon_{t_d+q_2}\right] = \begin{cases} \frac{\sigma^2}{\tau_{t_d+q_1}}, & q_1 = q_2; \\ 0, & q_1 \neq q_2. \end{cases}$$

Recall that $p(\mathbf{y}_d|\boldsymbol{\tau}, y_{t_d}, y_{t_d+n_d+1}; \boldsymbol{\theta})$ is a conditional pdf of $p(\mathbf{y}_d, y_{t_d+n_d+1}|\boldsymbol{\tau}, y_{t_d}; \boldsymbol{\theta})$. Since conditional distributions of a Gaussian distribution is Gaussian, we can get that $\mathbf{y}_d|\boldsymbol{\tau}, y_{t_d}, y_{t_d+n_d+1}; \boldsymbol{\theta}$ follows a Gaussian distribution as (28). The parameters of this conditional distribution can be computed based on

$$\boldsymbol{\mu}_d = \boldsymbol{\mu}_{cd(1:n_d)} + \frac{\boldsymbol{\Sigma}_{cd(1:n_d, n_d+1)}}{\Sigma_{cd(n_d+1, n_d+1)}}\left(y_{t_d+n_d+1} - \mu_{cd(n_d+1)}\right), \tag{58}$$

and

$$\boldsymbol{\Sigma}_d = \boldsymbol{\Sigma}_{cd(1:n_d, 1:n_d)} - \frac{\boldsymbol{\Sigma}_{cd(1:n_d, n_d+1)} \boldsymbol{\Sigma}_{cd(n_d+1, 1:n_d)}}{\Sigma_{cd(n_d+1, n_d+1)}}, \tag{59}$$

where $\boldsymbol{\mu}_{cd(a_1:a_2)}$ denotes the subvector consisting of the a_1-th to a_2-th component of $\boldsymbol{\mu}_{cd}$, and the $\boldsymbol{\Sigma}_{cd(a_1:a_2, b_1:b_2)}$ means the submatrix consisting of the components in the a_1-th to a_2-th rows and the $b1$-th to b_2-th columns of $\boldsymbol{\Sigma}_{cd}$. Plugging the equations (56) and (57) into the equations (58) and (59) gives the equations (29) and (30), respectively.

Appendix B
Proof for Conditions (M1)-(M5) and (SAEM2)-(SAEM3)

In this section, we will establish the listed conditions one by one. The observed data $\mathbf{y}_\mathsf{o}$ is known. We assume that $\mathbf{y}_\mathsf{o}$ is finite. Since the parameter space Θ is a large bounded set with $\nu > 2$, we can assume that $|\varphi_0| < \varphi_0^+, |\varphi_1| < \varphi_1^+, \sigma > \sigma^-$, and $\nu^- < \nu < \nu^+$, where φ_0^+, φ_1^+, and ν^+ are very large positive numbers, σ^- is a very small positive number, and ν^- is a very small positive number satisfying $\nu^- \geq 2$. We first prove the conditions (M1)-(M5), then prove the conditions (SAEM2) and (SAEM3).

A. Proof of (M1)-(M5)

The proof begins by establishing the following two intermediary lemmas.

Lemma 3: For any $\mathbf{y}_\mathsf{o}$ and $\boldsymbol{\theta} \in \Theta$, $p(\mathbf{y}_\mathsf{o}; \boldsymbol{\theta}) = \iint p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}) \mathsf{d}\mathbf{y}_\mathsf{m} \mathsf{d}\boldsymbol{\tau} = \int p(\mathbf{y}; \boldsymbol{\theta}) \mathsf{d}\mathbf{y}_\mathsf{m} < \infty$.

Lemma 4: For any $\mathbf{y}_\mathsf{o}$, $\boldsymbol{\theta} \in \Theta$ and $1 < t \leq T$

$$\iint g\left(\mathbf{y}, \boldsymbol{\tau}\right) p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \mathsf{d}\mathbf{y}_\mathsf{m} \mathsf{d}\boldsymbol{\tau} < \infty, \tag{60}$$

where $g(\mathbf{y}, \boldsymbol{\tau})$ can be τ_t, τ_t^2, y_t^2, $\tau_t y_{t-1}^2$, $\tau_t y_t^2$, or $-\log(\tau_t)$

Lemma 3 indicates that the observed data likelihood $p(\mathbf{y}_\mathsf{o}; \boldsymbol{\theta})$ is bounded, and Lemma 4 shows that the expectation of $g(\mathbf{y}, \boldsymbol{\tau})$ is bounded. These lemmas provide the key ingredients required for establishing (M1)-(M5), and their usage for subsequent analysis is self-explanatory. Due to space limitations, we do not include their proofs here. Interested readers may refer to the supplementary material of the current paper.

(M1) For condition (M1), based on (18), we can get

$$\begin{aligned}&\iint \left\|\mathbf{s}\left(\mathbf{y}_\mathsf{o}, \mathbf{y}_\mathsf{m}, \boldsymbol{\tau}\right)\right\| p\left(\mathbf{y}_\mathsf{m}, \boldsymbol{\tau}|\mathbf{y}_\mathsf{o}; \boldsymbol{\theta}\right) \mathsf{d}\mathbf{y}_\mathsf{m} \mathsf{d}\boldsymbol{\tau} \\ &= \frac{\iint \left\|\mathbf{s}\left(\mathbf{y}_\mathsf{o}, \mathbf{y}_\mathsf{m}, \boldsymbol{\tau}\right)\right\| p\left(\mathbf{y}_\mathsf{o}, \mathbf{y}_\mathsf{m}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \mathsf{d}\mathbf{y}_\mathsf{m} \mathsf{d}\boldsymbol{\tau}}{p\left(\mathbf{y}_\mathsf{o}; \boldsymbol{\theta}\right)} \\ &\leq \frac{1}{p\left(\mathbf{y}_\mathsf{o}; \boldsymbol{\theta}\right)} \sum_{t=2}^{T} \iint \Big(\left|\log\left(\tau_t\right) - \tau_t\right| + \left|\tau_t y_t^2\right| + \left|\tau_t\right| \\ &\quad + \left|\tau_t y_{t-1}^2\right| + \left|\tau_t y_t\right| + \left|\tau_t y_t y_{t-1}\right| \\ &\quad + \left|\tau_t y_{t-1}\right|\Big) p\left(\mathbf{y}_\mathsf{o}, \mathbf{y}_\mathsf{m}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \mathsf{d}\mathbf{y}_\mathsf{m} \mathsf{d}\boldsymbol{\tau} \\ &\leq \frac{1}{p\left(\mathbf{y}_\mathsf{o}; \boldsymbol{\theta}\right)} \sum_{t=2}^{T} \iint \Big(\tau_t - \log\left(\tau_t\right) + \tau_t y_t^2 + \tau_t \\ &\quad + \tau_t y_{t-1}^2 + \frac{\tau_t^2 + y_t^2}{2} + \frac{\tau_t\left(y_t^2 + y_{t-1}^2\right)}{2} \\ &\quad + \frac{\tau_t^2 + y_{t-1}^2}{2}\Big) p\left(\mathbf{y}_\mathsf{o}, \mathbf{y}_\mathsf{m}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \mathsf{d}\mathbf{y}_\mathsf{m} \mathsf{d}\boldsymbol{\tau} \\ &< \infty,\end{aligned} \tag{61}$$

where the three inequalities follow from the triangular inequality, the property of squares $x_1 x_2 \leq \frac{x_1^2 + x_2^2}{2}$, and Lemma 4, respectively.

(M2) From the definition of $\psi(\boldsymbol{\theta})$ and $\boldsymbol{\phi}(\boldsymbol{\theta})$ in (16) and (17), their continuous differentiability can be easily verified.

(M3) For condition (M3),

$$\begin{aligned}\bar{\mathsf{s}}(\boldsymbol{\theta}) &= \iint \mathsf{s}\left(\mathbf{y}_{\mathsf{o}}, \mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}\right) p\left(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau} | \mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}\right) \mathsf{d}\mathbf{y}_{\mathsf{m}} \mathsf{d}\boldsymbol{\tau} \\ &= \iint \mathsf{s}\left(\mathbf{y}_{\mathsf{o}}, \mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}\right) \frac{p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right)}{p\left(\mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}\right)} \mathsf{d}\mathbf{y}_{\mathsf{m}} \mathsf{d}\boldsymbol{\tau} \\ &= \frac{\iint \mathsf{s}\left(\mathbf{y}_{\mathsf{o}}, \mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}\right) p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \mathsf{d}\mathbf{y}_{\mathsf{m}} \mathsf{d}\boldsymbol{\tau}}{\iint p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \mathsf{d}\mathbf{y}_{\mathsf{m}} \mathsf{d}\boldsymbol{\tau}}. \end{aligned} \tag{62}$$

Since $\iint p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}) \mathsf{d}\mathbf{y}_{\mathsf{m}} \mathsf{d}\boldsymbol{\tau} = p(\mathbf{y}_{\mathsf{o}}; \boldsymbol{\theta}) > 0$ and $p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta})$ is continuously differentiable, which can be easily checked from its definition (19), we can get that $\bar{\mathsf{s}}(\boldsymbol{\theta})$ is continuously differentiable.

(M4) Since $\iint p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}) \mathsf{d}\mathbf{y}_{\mathsf{m}} \mathsf{d}\boldsymbol{\tau} > 0$, and $p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta})$ is 7 times differentiable, $l(\boldsymbol{\theta}; \mathbf{y}_{\mathsf{o}}) = \log(\iint p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}) \mathsf{d}\mathbf{y}_{\mathsf{m}} \mathsf{d}\boldsymbol{\tau})$ is 7 times differentiable. For the verification of the equation (46), according to Leibniz integral rule, the equation (46) holds under the following three conditions:

1) $\iint p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}) \mathsf{d}\mathbf{y}_{\mathsf{m}} \mathsf{d}\boldsymbol{\tau} < \infty$,
2) $\frac{\partial p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ exists for all the $\boldsymbol{\theta} \in \Theta$,
3) there is an integrable function $g(\mathbf{y}, \boldsymbol{\tau})$ such that $|\frac{\partial p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}| \leq g(\mathbf{y}, \boldsymbol{\tau})$ for all $\boldsymbol{\theta} \in \Theta$ and almost every $\mathbf{y}$ and $\boldsymbol{\tau}$.

Since the first condition has been proved in Lemma 3, and the second condition can be easily verified from its definition, here we focus on the third condition.

From the equation (13), the derivative of $p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta})$ with respect to φ_0 is

$$\begin{aligned}&\left|\frac{\partial p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right)}{\partial \varphi_0}\right| \\ &= \left| p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \sum_{j=2}^{T} \frac{\tau_j \left(y_j - \varphi_0 - \varphi_1 y_{j-1}\right)}{\sigma^2} \right| \\ &\leq \frac{p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right)}{\sigma^2} \sum_{j=2}^{T} \left(\left|\tau_j y_j\right| + \left|\varphi_0 \tau_j\right| + \left|\varphi_1 \tau_j y_{j-1}\right|\right) \\ &\leq \frac{p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}^*\right)}{\left(\sigma^-\right)^2} \sum_{j=2}^{T} \left\{\left(\frac{\tau_j^2 + y_j^2}{2} + \varphi_0^+ \tau_j + \frac{\varphi_1^+ \left(y_{j-1}^2 + \tau_j^2\right)}{2}\right)\right\} \\ &= g_{\varphi_0}\left(\mathbf{y}, \boldsymbol{\tau}\right), \end{aligned} \tag{63}$$

where $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \Theta} p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta})$. The first inequality follows from the triangle inequality, and the second inequality follows from $p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}^*) \geq p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta})$, $|\varphi_0| < \varphi_0^+$, $|\varphi_1| < \varphi_1^+$, $\sigma > \sigma^-$, and the property of squares.

The derivative with respect to φ_1 is

$$\begin{aligned}&\left|\frac{\partial p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right)}{\partial \varphi_1}\right| \\ &= \left| p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \sum_{j=2}^{T} \frac{1}{\sigma^2} \tau_j y_{j-1} \left(y_j - \varphi_0 - \varphi_1 y_{j-1}\right) \right| \\ &\leq \frac{p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right)}{\sigma^2} \sum_{j=2}^{T} \left(\left|\tau_j y_j y_{j-1}\right| + \left|\varphi_0 \tau_j y_{j-1}\right| + \left|\varphi_1 \tau_j y_{j-1}^2\right|\right) \\ &\leq \frac{p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}^*\right)}{\left(\sigma^-\right)^2} \sum_{j=2}^{T} \left(\frac{\tau_j \left(y_j^2 + y_{j-1}^2\right)}{2} + \frac{\varphi_0^+ \left(\tau_j^2 + y_{j-1}^2\right)}{2} \right. \\ &\qquad \left. + \varphi_1^+ \tau_j y_{j-1}^2\right) \\ &= g_{\varphi_1}\left(\mathbf{y}, \boldsymbol{\tau}\right), \end{aligned} \tag{64}$$

where the first inequality follows from the triangle inequality, and the second inequality follows from $|\varphi_0| < \varphi_0^+$, $|\varphi_1| < \varphi_1^+$, $\sigma > \sigma^-$, and the property of squares.

The derivative with respect to σ^2 is

$$\begin{aligned}&\left|\frac{\partial p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right)}{\partial \sigma^2}\right| \\ &= p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \sum_{j=2}^{T} \left\{\frac{\tau_j}{2\sigma^4} \left(y_j - \varphi_0 - \varphi_1 y_{j-1}\right)^2 - \frac{1}{2\sigma^2}\right\} \Bigg| \\ &\leq p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \sum_{j=2}^{T} \left\{\frac{\tau_j}{2\sigma^4} \left(y_j - \varphi_0 - \varphi_1 y_{j-1}\right)^2 + \frac{1}{2\sigma^2}\right\} \\ &\leq p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \sum_{j=2}^{T} \left\{\frac{\tau_j}{2\sigma^4} \left(2\left(y_j - \varphi_0\right)^2 + 2\varphi_1^2 y_{j-1}^2\right) + \frac{1}{2\sigma^2}\right\} \\ &\leq p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \sum_{j=2}^{T} \left\{\frac{\tau_j}{2\sigma^4} \left(4y_j^2 + 4\varphi_0^2 + 2\varphi_1^2 y_{j-1}^2\right) + \frac{1}{2\sigma^2}\right\} \\ &\leq p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}^*\right) \sum_{j=2}^{T} \left\{\frac{\tau_j}{2\left(\sigma^-\right)^2} \left(4y_j^2 + 4\left(\varphi_0^+\right)^2 + 2\left(\varphi_1^+\right)^2 y_{j-1}^2\right) \right. \\ &\qquad \left. + \frac{1}{2\left(\sigma^-\right)^2}\right\} \\ &= g_{\sigma^2}\left(\mathbf{y}, \boldsymbol{\tau}\right), \end{aligned} \tag{65}$$

where the first inequality follows from the triangle inequality, the second and third inequalities follow from the property of squares $(x_1 - x_2)^2 \leq 2(x_1^2 + x_2^2)$, and the last inequality follows from $p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}^*) \geq p(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta})$, $|\varphi_0| < \varphi_0^+$, $|\varphi_1| < \varphi_1^+$, and $\sigma > \sigma^-$.

The derivative with respect to ν is

$$\begin{aligned}&\left|\frac{\partial p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right)}{\partial \nu}\right| \\ &= \left| p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \sum_{j=2}^{T} \frac{1}{2} \left(1 + \log\left(\frac{\nu}{2}\right) - \Psi\left(\frac{\nu}{2}\right) + \log\left(\tau_j\right) - \tau_j\right) \right| \\ &\leq \frac{1}{2} p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \sum_{j=2}^{T} \left\{\left|1 + \log\left(\frac{\nu}{2}\right) - \Psi\left(\frac{\nu}{2}\right)\right| + \left|\log\left(\tau_j\right) - \tau_j\right|\right\} \\ &\leq p\left(\mathbf{y}, \boldsymbol{\tau}; \boldsymbol{\theta}^*\right) \sum_{j=2}^{T} \left(\frac{1}{2} + \frac{1}{2} \log\left(\frac{\nu^-}{2}\right) - \frac{1}{2} \Psi\left(\frac{\nu^-}{2}\right) \right. \\ &\qquad \left. + \frac{1}{2} \tau_j - \frac{1}{2} \log\left(\tau_j\right)\right) \\ &= g_{\nu}\left(\mathbf{y}, \boldsymbol{\tau}\right), \end{aligned} \tag{66}$$

where $\Psi(\cdot)$ is the digamma function. The first inequality follows from the triangle inequality, and the second inequality is due to that $\log(\frac{\nu}{2}) - \Psi(\frac{\nu}{2})$ is positive and strictly decreasing for $\nu \geq \nu^-$ [30].

Based on Lemmas 3 and 4, we can obtain that $\iint g_{\varphi_0}(\mathbf{y}, \boldsymbol{\tau},)\, \mathsf{d}\mathbf{y}_{\mathsf{m}}\mathsf{d}\boldsymbol{\tau} < \infty, \iint g_{\varphi_1}(\mathbf{y}, \boldsymbol{\tau})\, \mathsf{d}\mathbf{y}_{\mathsf{m}}\mathsf{d}\boldsymbol{\tau} < \infty,\ \iint g_{\sigma^2}(\mathbf{y}, \boldsymbol{\tau})\mathsf{d}\mathbf{y}_{\mathsf{m}}\mathsf{d}\boldsymbol{\tau} < \infty$, and $\iint g_{\nu}(\mathbf{y}, \boldsymbol{\tau})\mathsf{d}\mathbf{y}_{\mathsf{m}}\mathsf{d}\boldsymbol{\tau} < \infty$. The condition (M4) is verified.

(M5) This condition requires the existence of the global maximizer $\tilde{\boldsymbol{\theta}}(\bar{\mathbf{s}})$ for $Q(\boldsymbol{\theta}, \bar{\mathbf{s}})$ and its continuous differentiability. Since $Q(\boldsymbol{\theta}, \bar{\mathbf{s}})$ takes the same form with $\hat{Q}(\boldsymbol{\theta}, \hat{\mathbf{s}}^{(k)})$, the maximizer will also take the same form. From (34)–(37), we have

$$\tilde{\varphi_0}(\bar{\mathbf{s}}) = \frac{\bar{s}_5 - \tilde{\varphi_1}(\bar{\mathbf{s}})\,\bar{s}_7}{\bar{s}_3}, \tag{67}$$

$$\tilde{\varphi_1}(\bar{\mathbf{s}}) = \frac{\bar{s}_3\bar{s}_6 - \bar{s}_5\bar{s}_7}{\bar{s}_3\bar{s}_4 - \bar{s}_7^2}, \tag{68}$$

$$\begin{aligned}(\tilde{\sigma}(\bar{\mathbf{s}}))^2 = \frac{1}{T-1}\Big(&\bar{s}_2 + (\tilde{\varphi_0}(\bar{\mathbf{s}}))^2\,\bar{s}_3 + (\tilde{\varphi_1}(\bar{\mathbf{s}}))^2\,\bar{s}_4 \\ &- 2\tilde{\varphi_0}(\bar{\mathbf{s}})\,\bar{s}_5 - 2\tilde{\varphi_0}(\bar{\mathbf{s}})\,\bar{s}_6 + 2\tilde{\varphi_0}(\bar{\mathbf{s}})\,\tilde{\varphi_1}(\bar{\mathbf{s}})\,\bar{s}_7\Big),\end{aligned} \tag{69}$$

and

$$\tilde{\nu}(\bar{\mathbf{s}}) = \underset{\nu^- < \nu < \nu^+}{\arg\max}\, f(\nu, \bar{s}_1), \tag{70}$$

where $\bar{s}_i\ (i = 1, \ldots 7)$ is the i-th component of $\bar{\mathbf{s}}$. It can be easily verified that $\tilde{\varphi_0}(\bar{\mathbf{s}})$, $\tilde{\varphi_1}(\bar{\mathbf{s}})$ and $(\tilde{\sigma}(\bar{\mathbf{s}}))^2$ are continuous functions of $\bar{\mathbf{s}}$, and are 7 times differentiable with respect to $\bar{\mathbf{s}}$. For $\tilde{\nu}(\bar{\mathbf{s}})$, the gradient of $f(\nu, \bar{s}_1)$ at $\tilde{\nu}$

$$\begin{aligned}g(\tilde{\nu}, \bar{s}_1) &= \left.\frac{\partial f(\nu, \bar{s}_1)}{\partial \nu}\right|_{\nu=\tilde{\nu}} \\ &= \frac{1}{2}\left(\log\left(\frac{\tilde{\nu}}{2}\right) - \Psi\left(\frac{\tilde{\nu}}{2}\right) + 1 + \frac{\bar{s}_1}{T-1}\right) \\ &= 0.\end{aligned} \tag{71}$$

According to the implicit function theorem [38], since $g(\tilde{\nu}, \bar{s}_1)$ is 7 times continuously differentiable and $\frac{\partial g(\tilde{\nu}, \bar{s}_1)}{\partial \tilde{\nu}} = \frac{1}{2}(\frac{1}{\tilde{\nu}} - \frac{1}{2}\Psi'(\frac{\tilde{\nu}}{2})) \neq 0$ for any $\tilde{\nu}$ and $\bar{s}_1$ [30], $\tilde{\nu}(\mathbf{s})$ is 7 times continuously differentiable with respect to $\bar{\mathbf{s}}$.

B. Proof of (SAEM2) and (SAEM3)

The condition (SAEM2) has been verified in the proof of the conditions (M4) and (M5). The condition (SAEM3.1) holds due to the compactness assumption of the chain in the theorem. The functions $\mathbf{s}(\mathbf{y}_{\mathsf{o}}, \mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau})$ and $\{\hat{\mathbf{s}}^{(k)}\}$ are continuous function of the chain, therefore, they also take values in a compact set according to the boundness theorem, which implies the condition (SAEM3.2) hold. Now we focus on the proof of the conditions (SAEM3.3) and (SAEM3.4).

From the definition of the transition probability $\Pi_{\boldsymbol{\theta}}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}')$ in (24), we can easily verify that the transition probability $\Pi_{\boldsymbol{\theta}}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}')$ is continuously differentiable with respect to $\boldsymbol{\theta}$. In addition, since the derivative is a continuous function of $\boldsymbol{\theta} \in V$ and $(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') \in \Omega^2$, where V and Ω^2 are compact set, according to the boundness theorem, the derivative is bounded. Therefore, $\Pi_{\boldsymbol{\theta}}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}')$ is Lipschitz continuous, i.e., for any $(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') \in \Omega^2$, there exists a real constant $K(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}')$ such that for any $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in V^2$,

$$\begin{aligned}&\left|\Pi_{\boldsymbol{\theta}}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') - \Pi_{\boldsymbol{\theta}'}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}')\right| \\ &\leq K(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}')\,|\boldsymbol{\theta} - \boldsymbol{\theta}'|.\end{aligned} \tag{72}$$

It follows that

$$\begin{aligned}&\sup_{(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') \in \Omega^2}\left|\Pi_{\boldsymbol{\theta}}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') - \Pi_{\boldsymbol{\theta}'}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}')\right| \\ &\leq L|\boldsymbol{\theta} - \boldsymbol{\theta}'|\end{aligned} \tag{73}$$

with $L = \max_{(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') \in \Omega^2} K(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}')$, which implies that the condition (SAEM3.3) is verified.

The condition (SAEM3.4) is about the uniform ergodicity of the Markov chain generated by the transition probability $\Pi_{\boldsymbol{\theta}}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}')$. According to Theorem 8 in [39], a Markov chain is uniformly ergodic, if the transition probability satisfies some minorization condition, i.e., there exists $\alpha \in N^+$ and some probability measure $\delta(\cdot)$ such that $\Pi^{\alpha}_{\boldsymbol{\theta}}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') \geq \epsilon\delta(\mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}')$ for any $(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') \in \Omega^2$. Recall our transition probability $\Pi_{\boldsymbol{\theta}}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}')$ is a continuous function for $(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}) \in \Omega$, according to the extreme value theorem, there must exist an infimum $g(\mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}', \boldsymbol{\theta}) = \inf_{(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}) \in \Omega} \Pi_{\boldsymbol{\theta}}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}')$. It follows that

$$\Pi_{\boldsymbol{\theta}}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') \geq \epsilon\delta(\mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') \tag{74}$$

with $\epsilon = \iint g(\mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}', \boldsymbol{\theta})\mathsf{d}\boldsymbol{\tau}'\mathsf{d}\mathbf{y}'_{\mathsf{m}}$, and $\delta(\mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}') = \epsilon^{-1}g(\mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}', \boldsymbol{\theta})$. Therefore, the minorization condition holds in our case, and thus, the Markov chain generated by $\Pi_{\boldsymbol{\theta}}(\mathbf{y}_{\mathsf{m}}, \boldsymbol{\tau}, \mathbf{y}'_{\mathsf{m}}, \boldsymbol{\tau}')$ is uniformly ergodic. The condition (SAEM3.4) is verified.

References

[1] 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, Jan. 2009.

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

[3] R. S. Tsay, *Analysis of Financial Time Series*, 2nd ed. Hoboken, NJ, USA: Wiley, 2005.

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

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

[6] D. Alexander, G. Barker, and S. Arridge, "Detection and modeling of non-gaussian apparent diffusion coefficient profiles in human brain data," *Magn. Reson. Med.*, vol. 48, no. 2, pp. 331–340, 2002.

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

[8] K. L. Lange, R. J. Little, and J. M. Taylor, "Robust statistical modeling using the t distribution," *J. Amer. Statistical Assoc.*, vol. 84, no. 408, pp. 881–896, 1989.

[9] 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.

[10] 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.

[11] U. C. Nduka, "EM-based algorithms for autoregressive models with t-distributed innovations," *Commun. Statist. Simul. Comput.*, vol. 47, no. 1, pp. 206–228, 2018.

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

[13] R. J. Little and D. B. Rubin, *Statistical Analysis With Missing Data*, 2nd ed. Hoboken, NJ, USA: Wiley, 2002.

[14] G. DiCesare, "Imputation, estimation and missing data in finance," Ph.D. dissertation, Dept. Statist. Actuarial Sci., Univ. Waterloo, Waterloo, ON, Canada, 2006.

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

[16] V. A. Voloshko and Y. S. Kharin, "Robust estimation of AR coefficients under simultaneously influencing outliers and missing values," *J. Statistical Planning Inference*, vol. 141, no. 9, pp. 3276–3288, 2011.

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

[18] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a stochastic approximation version of the EM algorithm," *Ann. Statist.*, 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 Probability Statist.*, vol. 8, pp. 115–131, 2004.

[20] S. F. Nielsen *et al.*, "The stochastic EM algorithm: Estimation and asymptotic results," *Bernoulli*, vol. 6, no. 3, pp. 457–489, 2000.

[21] X.-L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[22] X. Yi and C. Caramanis, "Regularized EM algorithms: A unified framework and statistical guarantees," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 1567–1575.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statistical Soc. Ser. B Statistical Methodology*, vol. 39, no. 1, pp. 1–38, 1977.

[24] C. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983.

[25] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794–816, Feb. 2017.

[26] G. C. Wei and M. A. Tanner, "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *J. Amer. Statistical Assoc.*, vol. 85, no. 411, pp. 699–704, 1990.

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

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

[29] A. DasGupta, *The Exponential Family and Statistical Applications*. New York, NY, USA: Springer, 2011, pp. 583–612.

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

[31] B. Carnahan and H. A. Luther, *Applied Numerical Methods*. New York, NY, USA: Wiley, 1969.

[32] K. Chan and J. Ledolter, "Monte Carlo EM estimation for time series models involving counts," *J. Amer. Statistical Assoc.*, vol. 90, no. 429, pp. 242–252, 1995.

[33] M. G. Gu and F. H. Kong, "A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems," *Proc. Nat. Acad. Sci.*, vol. 95, no. 13, pp. 7270–7274, 1998.

[34] G. Fort and E. Moulines, "Convergence of the Monte Carlo expectation maximization for curved exponential families," *Ann. Statist.*, vol. 31, no. 4, pp. 1220–1259, 2003.

[35] R. C. Neath *et al.*, "On convergence properties of the Monte Carlo EM algorithm," in *Advances in Modern Statistical Theory and Application*. Beachwood, OH, USA: Inst. Math. Statist., 2013, pp. 43–62.

[36] R. Maronna, R. D. Martin, and V. Yohai, "Time series" in *Robust Statistics: Theory and Methods*. New York, NY, USA: Wiley, 2006, pp. 247–323.

[37] C. Caroni and V. Karioti, "Detecting an innovative outlier in a set of time series," *Comput. Statist. Data Anal.*, vol. 46, no. 3, pp. 561–570, 2004.

[38] S. G. Krantz and H. R. Parks, *Introduction to the Implicit Function Theorem*. New York, NY, USA: Springer, 2013, pp. 1–12.

[39] G. O. Roberts and J. S. Rosenthal, "General state space Markov chains and MCMC algorithms," *Probability Surv.*, vol. 1, pp. 20–71, 2004.

Junyan Liu received the B.Eng. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2015. She is currently working toward the Ph.D. degree in electronic and computer engineering with the Hong Kong University of Science and Technology, Hong Kong. Her research interests include convex optimization and efficient algorithms, with applications in signal processing, financial engineering, and machine learning. She was the recipient of the Hong Kong Ph.D. Fellowship Scheme.

Sandeep Kumar received the B.tech. degree from the College of Engineering Roorkee, Roorkee, India, in 2007, and the M.tech. and Ph.D. degrees from the Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur, India, in 2013 and 2017, respectively. He is currently a Postdoctoral Researcher with the Department of Electronics and Communication Engineering, Hong Kong University of Science and Technology, Hong Kong. The overarching theme of his research is on algorithms, analysis, and applications of optimization, statistics, and signal processing for data science application.

Daniel P. Palomar (S'99–M'03–SM'08–F'12) received the electrical engineering and Ph.D. degrees from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1998 and 2003, respectively.

He was a Fulbright Scholar with Princeton University during 2004–2006. He is a Professor with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology (HKUST), Hong Kong, which he joined in 2006. He had previously held several research appointments, namely, at King's College London, London, U.K.; Stanford University, Stanford, CA, USA; Telecommunications Technological Center of Catalonia, Barcelona, Spain; Royal Institute of Technology, Stockholm, Sweden; University of Rome "La Sapienza," Rome, Italy; and Princeton University, Princeton, NJ, USA. His current research interests include applications of convex optimization theory and signal processing in financial systems, and big data analytics.

Dr. Palomar was a Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING 2016 Special Issue on "Financial Signal Processing and Machine Learning for Electronic Trading," an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY and the IEEE TRANSACTIONS ON SIGNAL PROCESSING, a Guest Editor for the IEEE SIGNAL PROCESSING MAGAZINE 2010 Special Issue on "Convex Optimization for Signal Processing," the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2008 Special Issue on "Game Theory in Communication Systems," and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2007 Special Issue on "Optimization of MIMO Transceivers for Realistic Communication Networks." He was the recipient of the 2004/2006 Fulbright Research Fellowship, the 2004 and 2015 (co-author) Young Author Best Paper Awards by the IEEE Signal Processing Society, the 2015–2016 HKUST Excellence Research Award, the 2002/2003 best Ph.D. prize in information technologies and communications by the UPC, the 2002/2003 Rosina Ribalta first prize for the Best Doctoral Thesis in information technologies and communications by the Epson Foundation, and the 2004 prize for the best Doctoral Thesis in advanced mobile communications by the Vodafone Foundation and COIT.