A Student $t$-mixture autoregressive model with applications to heavy-tailed financial data

BY C. S. WONG, W. S. CHAN
Department of Finance, The Chinese University of Hong Kong, Shatin, Hong Kong
albertw@baf.msmail.cuhk.edu.hk, chanws@cuhk.edu.hk

AND P. L. KAM
Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong
h102934@graduate.hku.hk

SUMMARY

In this paper we introduce the class of Student $t$-mixture autoregressive models which is a promising candidate for financial time series modelling. The model is able to capture serial correlations; time-varying means and volatilities; and the shape of the conditional distributions can be time varied from short-tailed to long-tailed, or from unimodal to multi-modal. The use of $t$-distributed errors in each component of the model allows for conditional leptokurtic distribution which could account for the commonly observed unconditional kurtosis in the financial data. Methods of parameter estimation and model selection are derived. Finally, the proposed modelling procedure is illustrated through a real example.

Some key words: EM algorithm; Interest rate; Mixture distribution; Non-linear time-series model.

1. INTRODUCTION

Recently, Wong & Li (2000) introduced a mixture autoregressive model which is potentially useful in modelling financial returns. However, financial returns usually exhibit excess kurtosis. Since their model allies with the assumption that return series in each of the autoregressive component is conditionally Gaussian, it may underestimate the occurrence of extreme financial events, such as market crashes.

In this paper, we extend the mixture autoregressive model by replacing the Gaussian assumption with the Student $t$ distribution. It consists of a mixture of $g$ autoregressive components. This new model is called the Student $t$-mixture autoregressive model. We consider the Student $t$ distribution because it provides heavier tails than normal distribution and would be able to accommodate some aberrant returns occasionally observed in financial markets. Extreme observations often cause unsatisfactory estimation results for normal mixture type model in EM iterations, see McLachlan & Peel (1998). On the other hand, the use of Student $t$ assumption gives less extreme estimates of the posterior probabilities of the component membership of the mixture model, and hence gives a more robust estimation in our model.

The degrees of freedom in the Student $t$ distributions are parameters in our model. They can be pre-fixed in advance according to some prior information or empirically determined from data. In this paper, we assume that the degrees of freedom are unknown. As the normal distribution is
a limiting case of the Student $t$ distribution, the mixture autoregressive model is a limiting case of our model.

A standardized Student $t$ distribution with unit variance is used. Without standardization, the variance of Student $t$ distribution depends on its degrees of freedom. This will interfere the estimation of the component variance. By using a standardized Student $t$ distribution, the estimation of component variances and degrees of freedom will be separated which corresponds to the modelling of variability and shapes of conditional distribution, respectively.

The stationarity properties and the fourth order moment are examined. The parameter estimation can be carried out via the EM algorithm (Dempster et al., 1977). The standard errors of the parameter estimates can be computed by the missing information principle (Louis, 1982). The main difficulty in estimation is that two sets of missing random variables need to be handled. The first set corresponds to the component identities of the observations and the second set corresponds to the mixing random variables in the Student $t$ distribution. For model selection, the Bayesian information criterion is adopted.

Since the tails of the component distributions can be adjusted, the flexibility of our model is much greater than that of the mixture autoregressive model. Furthermore, our model enjoys some nice statistical properties. First, a mixture of a non-stationary autoregressive component and a stationary autoregressive component results in overall stationary process. Secondly, the conditional distribution of the time series can be multi-modal. Thirdly, our model can capture conditional heteroscedasticity (Engle, 1982) which is a stylized fact in financial time series.

2. THE STUDENT $t$-MIXTURE AUTOREGRESSIVE MODEL

We propose a $g$-component Student $t$-mixture autoregressive model which is defined by

$$F(y_t | F_{t-1}) = \sum_{k=1}^{g} \alpha_k F_{\nu_k} \left( \frac{y_t - \phi_k y_{t-1} - \cdots - \phi_k y_{t-p_k}}{\sigma_k} \right),$$

(1)

where $\alpha_1 + \cdots + \alpha_g = 1$, $\alpha_k > 0$ and $\nu_k > 2$ for $k = 1, \ldots, g$. This model is denoted by $\text{TMAR}(g; \nu_1, \ldots, \nu_g)$. Here, $F(y_t | F_{t-1})$ is the conditional cumulative distribution function of $Y_t$ given the past information, evaluated at $y_t$; $F_t$ is the information set up to time $t$; $F_{\nu_k} (\cdot)$ is the cumulative distribution function of the standardized Student $t$ distribution with $\nu_k$ degrees of freedom. The probability distribution function of a standardized Student $t$ distribution with unit variance and $\nu$ degrees of freedom is

$$f_{\nu}(x) = \frac{\Gamma \left( \frac{\nu+1}{2} \right)}{\sqrt{\pi (\nu - 2)}} \Gamma \left( \frac{\nu}{2} \right) \left( 1 + \frac{x^2}{\nu - 2} \right)^{-\frac{\nu+1}{2}},$$

(2)

where $2 < \nu < \infty$, and $\Gamma (\cdot)$ is the gamma function.

The conditional expectation of $y_t$ is $E(y_t | F_{t-1}) = \sum_{k=1}^{g} \alpha_k \mu_{k,t}$ where $\mu_{k,t} = \phi_{k0} + \phi_{k1} y_{t-1} + \cdots + \phi_{kp_k} y_{t-p_k}$. The shapes of the conditional distributions may change from unimodal to multi-modal. Therefore, the conditional expectation of $y_t$ may not be the best predictor of the future values. Another important property is the ability to model changing conditional variance. The conditional variance of $y_t$ is $\text{var}(y_t | F_{t-1}) = \sum_{k=1}^{g} \alpha_k \sigma_k^2 + \sum_{k=1}^{g} \alpha_k \mu_{k,t}^2 - (\sum_{k=1}^{g} \alpha_k \mu_{k,t})^2$. The first term summarizes the conditional variance on the past errors. The second and third terms characterize the change of the conditional variance resulting from the differences in the conditional means of the components.

The first-order and second-order stationarity conditions for our model (1) are exactly the same as that for the mixture autoregressive model in Wong & Li (2000). Theorems 1 and 2 give the
conditions of first-order stationarity for a TMAR($g; p_1, \ldots, p_g$) model and second-order station-
arity for a TMAR($g; 1, \ldots, 1$) model. From the theorems, both the first-order and second-order
stationarity conditions involve the mixing proportions, $\alpha_k$, of each component in the model. It is
possible that a mixture of an explosive autoregressive component and a stationary autoregressive
component results in an overall stationary process. The proofs of all theorems are given in a
longer version of this paper which is available from the authors. Let $p = \max(p_1, \ldots, p_g)$.

**Theorem 1:** A necessary and sufficient condition for the process $Y_t$ to be first-order stationary is
that the roots $z_1, \ldots, z_p$ of the equation

$$1 - \sum_{i=1}^{p} \left( \sum_{k=1}^{g} \alpha_k \phi_{ki} \right) z^{-i} = 0$$

all lie inside the unit circle, where $\phi_{ki} = 0$ for $i > p_k$.

**Theorem 1.** Suppose that the process $Y_t$ following a TMAR($g; 1, \ldots, 1$) model is first-order
stationary. A necessary and sufficient condition for the process to be second-order stationary is

$$\alpha_1 \phi_{11}^2 + \cdots + \alpha_g \phi_{g1}^2 < 1.$$  

The major advantage of our model (1) is its ability to model the tails of the conditional dis-
tribution. The shape and tails of the conditional distribution can be effectively characterized by
the combination of various values of degrees of freedom in all components. A smaller degrees
of freedom in a component will contribute more to the heaviness of the tails of the overall con-
ditional distribution.

Kurtosis of $Y_t$, which is used to describe the degree of heavy-tailedness of the distribution, can
be computed. First, we derive the condition for the existence of the fourth-order moment. Since
the derivation is extremely tedious for the general case, for illustrative purpose, we only consider
the TMAR($g; 0, \ldots, 0$) model with $\phi_{ki} = 0$ ($k = 1, \ldots, g$).

**Theorem 2.** The fourth-order moment of a stationary TMAR($g; 0, \ldots, 0$) model with $\phi_{ki} = 0$
($k = 1, \ldots, g$), exists if and only if $\nu_k > 4$.

The existence condition of the fourth-order moment for our model (1) is the same as those for the Student $t$ distribution. If $\nu_k > 4$, the expression for the fourth-order moment is given
by $3 \sum_{k=1}^{g} \alpha_k \sigma_k^4 (\nu_k - 2)/(\nu_k - 4)$. It can be shown that the kurtosis of $Y_t$ is generally greater
than 3. Hence, our model cater for heavier conditional distributions as compared to Gaussian
processes. It should be noted that the expression of the fourth-order moment for our model is
similar to that for the mixture autoregressive model with the additional terms $(\nu_k - 2)/(\nu_k - 4)$
while the expressions of the variance are the same for both models. Since the additional terms
are larger than 1 unless $\nu_k \to \infty$, the marginal distribution from our model can be much more
heavy-tailed than that from the mixture autoregressive model.

The autocorrelations for our model are similar to that for the mixture autoregressive model.
They satisfy a system of equations similar to the Yule–Walker equations. For a second-order
stationary process $Y_t$ that follows a TMAR($g; p_1, \ldots, p_g$) model, it is easy to show that

$$\rho_j = \sum_{i=1}^{p} \left( \sum_{k=1}^{g} \alpha_k \phi_{ki} \right) \rho_{j-i}, \text{ for } j = 1, \ldots, p,$$

where $\rho_j$ is the lag $j$ autocorrelation.

3. Estimation

3-1. The standardized Student $t$ distribution

The Student $t$ density function can be obtained from a normal scale mixture model
(Peel & McLachlan, 2000). Suppose that the observation $X$ is generated from the model,

$$f(x; \mu, \sigma^2/w)h(w)dw,$$

where $f(x; \mu, \sigma^2/w)$ denotes the density function of the normal dis-
tribution with mean $\mu$ and variance $\sigma^2/w$, and $h(\cdot)$ is a density function of any distribution. The
function $h(\cdot)$ is employed to control the magnitude of the variance in the normal density func-
tion. Assume that we choose the gamma density function for $h(\cdot)$, then the random variable $W$
is distributed as \( W \sim \text{gamma} \{ \nu/2, (\nu - 2)/2 \} \). After integrating out \( w \), we get a standardized
Student \( t \) distribution with location parameter \( \mu \), scale parameter \( \sigma \) and degrees of freedom \( \nu \),
with a density function of \( f(x) = f_{\nu}(x)/\sigma \). Here \( \delta^2 = \{(x - \mu)/\sigma \}^2 \) denotes the Mahalanobis
squared distance between \( x \) and \( \mu \), and \( f_{\nu}(x) \) is given in (2).

The standardized Student \( t \) distribution is symmetric about \( \mu \), and \( \sigma^2 \) is its variance. If \( \nu > 4 \),
the fourth moment exists and is equal to \( E(x^4) = 3\sigma^4(\nu - 2)/(\nu - 4) + 6\sigma^2\mu^2 + \mu^4 \). It can be
shown that the \( t \)-distribution provides heavier tails as compared to the normal distribution when
\( \nu > 0 \). As \( \nu \to \infty \), it approaches the normal distribution. In practice, the curve of the Student
\( t \) density function with degrees of freedom of 30 or more is almost indistinguishable with the
normal curve. A small degrees of freedom implies fat tailed distribution and results in relatively
large amount of atypical observations in the time series. We define \( \nu \leq 5 \) as small in this paper.

### 3.2. Parameter estimation

The \( \text{EM} \) algorithm (Dempster et al., 1977, and McLachlan & Krishnan, 1997) is adopted for
the estimation of the parameters of our model (1). Here, the log-likelihood is constructed by
the normal scale mixture model. Assume that we have observations \( Y = (y_1, \ldots, y_n) \) generated
from the model. Let \( Z = (Z_1, \ldots, Z_n) \) be a \( g \times n \) unobservable random matrix, where \( Z_t =
(Z_{kt}) \) \( (t = 1, \ldots, n) \), is a \( g \)-dimensional column indicator vector showing the origin of the \( k \)
observation, that is, \( Z_{kt} = 1 \), if \( y_t \) is generated from the \( k \)th component of the model and \( Z_{kt} = 0 \)
otherwise.

Analogous to the formulation of \( Z \), we consider another missing random matrix, \( W =
(W_1, \ldots, W_n) \), where \( W_t = (W_{kt}) \) \( (t = 1, \ldots, n) \) is also a \( g \)-dimensional vector. Given \( z_{kt} =
1 \), the conditional distribution of \( W_{kt} \) is \( W_{kt} | z_{kt} = 1 \sim \text{gamma} \{ \nu_k/2, (\nu_k - 2)/2 \} \), and
\( W_1, \ldots, W_n \) are distributed independently. Let \( \alpha = (\alpha_1, \ldots, \alpha_{g-1})^T \), \( \nu = (\nu_1, \ldots, \nu_g)^T \), \( \theta =
(\theta_1^T, \ldots, \theta_g^T)^T \) with \( \theta_k = (\phi_k, 0, \ldots, \phi_{kp_k}, \sigma_k)^T \) \( (k = 1, \ldots, g) \), and the vector of parameters is
defined as \( \Psi = (\alpha^T, \theta^T, \nu^T)^T \). Here, transpose of a vector or a matrix is denoted by the super-
script \( T \).

The conditional log-likelihood is

\[
\ell = \ell_1(\alpha) + \ell_2(\nu) + \ell_3(\theta)
\]

(3)

where

\[
\ell_1(\alpha) = \sum_{k=1}^g \sum_{t=p+1}^n Z_{kt} \log(\alpha_k),
\]

\[
\ell_2(\nu) = \sum_{k=1}^g \sum_{t=p+1}^n Z_{kt} \left[ -\log \left( \Gamma \left( \frac{1}{2} \nu_k \right) \right) + \left( \frac{1}{2} \nu_k \right) \log \left( \frac{\nu_k - 2}{2} \right) \right]
\]

\[
+ \left( \frac{1}{2} \nu_k \right) (\log W_{kt} - W_{kt}) - \log(W_{kt} + W_{kt}) \right],
\]

and

\[
\ell_3(\theta) = \sum_{k=1}^g \sum_{t=p+1}^n Z_{kt} \left[ -\frac{1}{2} \left\{ \log(2\pi) + \log \sigma_k^2 - \log W_{kt} \right\} - \frac{\phi_k^2 W_{kt}}{2\sigma_k^2} \right].
\]

Here, \( e_{kt} = y_t - \phi_k y_t - \ldots - \phi_{kp_k} y_{t-p_k} \).

The parameters are estimated by maximizing the log-likelihood (3). The estimation can be
done via the iterative \( \text{EM} \) procedure which consists of an \( E \) step and a \( M \) step.
E step. Assume that $\Psi$ is known. The missing data $Z$, $W$ and $\log W$ in the log-likelihood are
replaced by their expectations, conditional over the parameters and the observed data $Y$. The
E-step equations are

$$
\tau_{kt} = E_\Psi(Z_{kt} | y_t) = \frac{\alpha_k \sigma_k^{-1} f_{\nu_k}(\delta_{kt})}{\sum_{j=1}^{g} \alpha_j \sigma_j^{-1} f_{\nu_j}(\delta_{jt})} \quad (k = 1, \ldots, g),
\tag{4}
$$

$$
\eta_{kt} = E_\Psi(W_{kt} | y_t, z_{kt} = 1) = \frac{\nu_k + 1}{\delta_{kt}^2 + \nu_k - 2} \quad (k = 1, \ldots, g),
\tag{5}
$$

$$
E_\Psi(\log W_{kt} | y_t, z_{kt} = 1) = \log \eta_{kt} + \left\{ \psi\left(\frac{\nu_k + 1}{2}\right) - \log \left(\frac{\nu_k + 1}{2}\right) \right\} \quad (k = 1, \ldots, g),
\tag{6}
$$

where $\delta_{kt}^2 = e_{kt}^2 / \sigma_k^2$ and $\psi(s) = \{d \Gamma(s)/ds\} / \Gamma(s)$ is the Digamma function.

M step. Assume that the missing data are known. By maximizing the log-likelihood function
(3), the estimates of our model (1) are obtained. The estimates of mixing proportions and com-
ponent standard deviations are

$$
\hat{\alpha}_k = \frac{\sum_{t=p+1}^{n} \tau_{kt}}{n - p} \quad (k = 1, \ldots, g), \quad \text{and} \quad \hat{\sigma}_k = \left(\frac{\sum_{t=p+1}^{n} \tau_{kt} \eta_{kt} \delta_{kt}^2}{\sum_{t=p+1}^{n} \tau_{kt}}\right)^{\frac{1}{2}} \quad (k = 1, \ldots, g).
\tag{7}
$$

Here, $\hat{\tau}_{kt} = y_t - \hat{\phi}_{k0} - \hat{\phi}_{k1}y_{t-1} - \cdots - \hat{\phi}_{kp_k}y_{t-p_k}$. The estimates of $\phi_{ki}$’s are obtained by solving
the following system of equations.

$$
\sum_{t=p+1}^{n} \tau_{kt} \eta_{kt} u(y_t, i) = \sum_{j=0}^{p_k} \phi_{kj} \sum_{t=p+1}^{n} \tau_{kt} \eta_{kt} u(y_t, j) u(y_t, i) \quad (k = 1, \ldots, g; i = 0, \ldots, p_k),
\tag{8}
$$

where $u(y_t, i) = 1$ for $i = 0$ and $u(y_t, i) = y_{t-i}$ for $i > 0$. The estimates of degrees of freedom
must satisfy the following equations,

$$
\left(\frac{\nu_k}{\nu_k - 2}\right) + \log \left(\frac{\nu_k - 2}{2}\right) - \psi\left(\frac{\nu_k}{2}\right) + \psi\left(\nu_k^{(m)} / 2\right) - \log \left(\nu_k^{(m)} / 2\right)
+ \frac{1}{\sum_{t=p+1}^{n} \tau_{kt}^{(m)}} \sum_{t=p+1}^{n} \tau_{kt}^{(m)} \left(\log \eta_{kt}^{(m)} - \eta_{kt}^{(m)}\right) = 0 \quad (k = 1, \ldots, g).
\tag{9}
$$

Here, $\nu_k^{(m)}$ is the estimated $\nu_k$ in the $m$th iteration of the EM algorithm. Since equation (9) is not
linear, the numerical solution of $\nu_k$ is obtained by the Newton–Raphson method. Details can be
found in the Appendix. In practice, it is possible that the estimated values of $\nu_k$ are less than 2.
We avoid these cases by restricting the degrees of freedom of the Student $t$ distribution in each
component of the model to be greater than 2 during the EM estimation.

### 3.3. Observed information matrix

We employ the method of missing information principle, which was firstly introduced by
Louis (1982), for calculating standard errors of the parameter estimates. According to this prin-
ciple, the observed information matrix $I$ is computed by following formula, $I = I_c - I_m =
E(\frac{\partial E}{\partial \theta} \frac{\partial E}{\partial \theta} | \Psi, Y) - \text{var}(\frac{\partial E}{\partial \theta} | \Psi, Y)$, where $I_c$ and $I_m$ are the complete information
matrix and missing information matrix respectively. The formulae for $I_c$ and $I_m$ are given in
a longer version of the paper which is available from the authors.
Table 1. Simulation results for Model (A) with \( n=1000 \) and 1000 replications

<table>
<thead>
<tr>
<th>Model Parameter</th>
<th>( k )</th>
<th>( \nu_k )</th>
<th>( \alpha_k )</th>
<th>( \sigma_k )</th>
<th>( \phi_{k0} )</th>
<th>( \phi_{k1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 True values</td>
<td>1</td>
<td>5.00</td>
<td>0.50</td>
<td>1.00</td>
<td>0.00</td>
<td>-0.50</td>
</tr>
<tr>
<td>Mean of estimates</td>
<td>1</td>
<td>5.00</td>
<td>0.50</td>
<td>1.00</td>
<td>0.00</td>
<td>-0.50</td>
</tr>
<tr>
<td>Empirical standard error</td>
<td>1</td>
<td>3.74</td>
<td>0.02</td>
<td>0.15</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>Theoretical standard error</td>
<td>1</td>
<td>2.88</td>
<td>0.02</td>
<td>0.15</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>2 True values</td>
<td>2</td>
<td>8.00</td>
<td>0.50</td>
<td>2.00</td>
<td>0.00</td>
<td>1.10</td>
</tr>
<tr>
<td>Mean of estimates</td>
<td>2</td>
<td>10.58</td>
<td>0.50</td>
<td>2.00</td>
<td>0.00</td>
<td>1.10</td>
</tr>
<tr>
<td>Empirical standard error</td>
<td>2</td>
<td>7.12</td>
<td>0.02</td>
<td>0.10</td>
<td>0.10</td>
<td>0.03</td>
</tr>
<tr>
<td>Theoretical standard error</td>
<td>2</td>
<td>6.64</td>
<td>0.02</td>
<td>0.11</td>
<td>0.10</td>
<td>0.03</td>
</tr>
</tbody>
</table>

4. Simulation experiments

We have done extensive simulation studies to assess the performance of the EM estimation algorithm and the estimated standard errors based on the Louis’s (1982) missing information principle. Only two of these simulation studies are reported here. In both simulation experiments, 1000 independent sample paths were generated from our model (1) with 1000 data points. Model (A) in the first experiment is a \( t \)-MAR(2;1,1) model, which is a mixture of stationary autoregressive process and an explosive autoregressive process, with parameter values

\[(\alpha_1, \alpha_2, \nu_1, \nu_2, \sigma_1, \phi_{10}, \phi_{11}, \phi_{20}, \phi_{21}) = (0.50, 0.50, 4.00, 8.00, 1.00, 2.00, 0.00, -0.50, 0.00, 1.10).\]

Model (B) in the second experiment is a \( t \)-MAR(3;1,1) model with parameter values

\[(\alpha_1, \alpha_2, \nu_1, \nu_2, \nu_3, \sigma_1, \sigma_2, \phi_{10}, \phi_{11}, \phi_{12}, \phi_{20}, \phi_{21}, \phi_{30}, \phi_{31}, \phi_{32}, \phi_{33}) = (0.30, 0.30, 0.40, 4.00, 6.00, 10.00, 2.00, 1.00, 0.50, 0.50, 0.50, 0.24, -5.00, -0.90, 0.00, 1.50, -0.74, 0.12).\]

The results of the simulation studies are shown in Tables 1 and 2. These studies indicate that the performance of the proposed EM algorithm and the Louis’s method is satisfactory. There may be a bias problem in the estimation of degrees of freedom, especially when the underlying degrees of freedom is large. It might be due to the fact that \( \nu_k \)’s are shape parameters which are relatively difficult to estimate. However, the bias problem in the EM estimation of degrees of freedom should not hinder the use of our model in practice. The characteristics of Student \( t \) distributions with slightly different degrees of freedom are quite similar, especially when \( \nu_k \geq 8 \).

To further study the impact of sample size on the estimation of \( \nu_k \)’s, we repeat the simulation experiments with a larger sample size. The results are summarized in Table 3. When sample size is increased to 2000 and 5000, the relative biases in the estimation of degrees of freedom are reduced significantly. Their empirical standard errors are now much closer to their theoretical standard errors. These studies show that the performance of the EM algorithm and the Louis’s method in the estimation of \( \nu_k \)’s improve when the sample size increases.

5. Model selection

We propose using the Bayesian information criterion (BIC) (Schwarz, 1978) as a tool for model selection of our model (1). Although the use of BIC to choose the number of component, \( g \), is somewhat nonstandard as it corresponds to testing problems with nuisance parameters that do not exist under the null hypothesis (Davies, 1977, 1987), it can serve as a rough guide in model selection. Within the class of Student \( t \)-mixture autoregressive processes, \( \text{BIC} = -2\ell^* + \log(n - p)(4g - 1 + \sum_{k=1}^{g} p_k), \) where \( \ell^* = \sum_{k=p+1}^{n} \log f(y_t | F_{t-1}) = \sum_{k=p+1}^{n} \log \{dF(y_t | F_{t-1})/dy_t\} \) is a maximized log-likelihood calculated from the conditional...
### Student t-mixture autoregressive model

#### Table 2. Simulation results for Model (B) with n=1000 and 1000 replications

<table>
<thead>
<tr>
<th>$k$</th>
<th>Model Parameter</th>
<th>( \nu_k )</th>
<th>( \alpha_k )</th>
<th>( \sigma_k )</th>
<th>( \phi_{k0} )</th>
<th>( \phi_{k1} )</th>
<th>( \phi_{k2} )</th>
<th>( \phi_{k3} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True values</td>
<td>4.00</td>
<td>0.30</td>
<td>2.00</td>
<td>5.00</td>
<td>0.50</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean of estimates</td>
<td>4.57</td>
<td>0.30</td>
<td>2.01</td>
<td>5.00</td>
<td>0.50</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Empirical standard error</td>
<td>2.27</td>
<td>0.02</td>
<td>0.22</td>
<td>0.11</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Theoretical standard error</td>
<td>1.61</td>
<td>0.02</td>
<td>0.23</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>True values</td>
<td>6.00</td>
<td>0.30</td>
<td>1.00</td>
<td>-5.00</td>
<td>-0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean of estimates</td>
<td>7.45</td>
<td>0.30</td>
<td>1.00</td>
<td>-5.00</td>
<td>-0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Empirical standard error</td>
<td>4.89</td>
<td>0.02</td>
<td>0.07</td>
<td>0.06</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Theoretical standard error</td>
<td>4.29</td>
<td>0.02</td>
<td>0.07</td>
<td>0.06</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>True values</td>
<td>10.00</td>
<td>0.40</td>
<td>0.50</td>
<td>0.00</td>
<td>1.50</td>
<td>-0.74</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>Mean of estimates</td>
<td>13.28</td>
<td>0.40</td>
<td>0.50</td>
<td>0.00</td>
<td>1.50</td>
<td>-0.74</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>Empirical standard error</td>
<td>8.91</td>
<td>0.02</td>
<td>0.33</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Theoretical standard error</td>
<td>12.94</td>
<td>0.02</td>
<td>0.33</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

#### Table 3. Estimation results of \( \nu_k \)'s with various sample sizes

<table>
<thead>
<tr>
<th>$k$</th>
<th>( \nu_k )</th>
<th>( \nu_k ) (Model A)</th>
<th>( \nu_k ) (Model B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True values</td>
<td>4.00</td>
<td>4.00</td>
</tr>
<tr>
<td></td>
<td>Mean of estimates</td>
<td>4.40</td>
<td>4.12</td>
</tr>
<tr>
<td></td>
<td>Empirical standard error</td>
<td>1.57</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>Theoretical standard error</td>
<td>1.23</td>
<td>0.63</td>
</tr>
<tr>
<td>2</td>
<td>True values</td>
<td>8.00</td>
<td>8.00</td>
</tr>
<tr>
<td></td>
<td>Mean of estimates</td>
<td>9.34</td>
<td>8.41</td>
</tr>
<tr>
<td></td>
<td>Empirical standard error</td>
<td>4.32</td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td>Theoretical standard error</td>
<td>3.64</td>
<td>1.70</td>
</tr>
<tr>
<td>3</td>
<td>True values</td>
<td>10.00</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td>Mean of estimates</td>
<td>12.49</td>
<td>10.91</td>
</tr>
<tr>
<td></td>
<td>Empirical standard error</td>
<td>7.43</td>
<td>3.37</td>
</tr>
<tr>
<td></td>
<td>Theoretical standard error</td>
<td>7.98</td>
<td>3.29</td>
</tr>
</tbody>
</table>

The number of component and the autoregressive order, \( p_k \), in each component can be selected based on minimum BIC combination. A simple simulation experiment is carried out to evaluate the performance of BIC as a tool to select \( g \) and \( p_k \)'s in the model. The model under consideration is a TMAR(2;2,1) model with parameter values \((\alpha_1, \alpha_2, \nu_1, \nu_2, \sigma_1, \sigma_2, \phi_{10}, \phi_{11}, \phi_{12}, \phi_{20}, \phi_{21}) = (0.50, 0.50, 4.00, 10.00, 1.00, 2.00, 0.00, 0.90, -0.60, 0.00, -0.50)\). The sample sizes are 1000 and 2000 and the two simulation experiments are repeated for 100 times. The BIC values are computed for each order combination of \( \{g \leq 3\} \) and \( \{p_k \leq 5, k = 1, \ldots, g\} \). The selected order is the combination that has the smallest BIC value. The rates of correct identification of true model are 87% and 94% respectively for \( n = 1000 \) and 2000. The simulation results indicate that the performance of BIC seems acceptable.
Table 4. Estimation result of the TMAR(3;0,4,4) model for return series of the 3-year Treasury Constant Maturity Rate

<table>
<thead>
<tr>
<th>k</th>
<th>Estimate</th>
<th>s.e.</th>
<th>Estimate</th>
<th>s.e.</th>
<th>Estimate</th>
<th>s.e.</th>
<th>Estimate</th>
<th>s.e.</th>
<th>Estimate</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32.32</td>
<td>0.30</td>
<td>0.66</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.69</td>
<td>0.33</td>
<td>2.58</td>
<td>0.06</td>
<td>0.32</td>
<td>0.07</td>
<td>-0.04</td>
<td>0.06</td>
<td>-0.47</td>
<td>0.06</td>
</tr>
<tr>
<td>3</td>
<td>2.96</td>
<td>0.37</td>
<td>1.76</td>
<td>0.06</td>
<td>-0.38</td>
<td>0.01</td>
<td>0.45</td>
<td>0.04</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

6. Example: The 3-year Treasury Constant Maturity Rate

The 3-year Treasury Constant Maturity Rate is an interest rate series published by the Federal Reserve Board. Our study period is between January 2, 1996 and December 30, 2005. This data set, which contains 2502 observations, was downloaded from the website of The Federal Reserve Bank of St Louis (www.stls.frb.org/fred). The original series is clearly non-stationary. After taking the first differences of the logarithm of the series, the resulting log return series, rt, seems to be stationary.

We find that a TMAR(3;0,4,4) model without intercept is the best fitted model to the return series. It minimizes the BIC over all two- and three-component models with or without intercept. The parameter estimates and standard errors are shown in Table 4. The value of BIC is 9447.25.

This model comprises two stationary fourth-order autoregressive components mixed with a portion of observations generated from white noise process with large degrees of freedom but small standard deviation. As two of the three estimated degrees of freedom νk are less than 4, this implies that the fourth moment may not exist under the fitted model.

For comparison, a mixture autoregressive model (Wong & Li, 2000) is also fitted to the return series. The best model is MAR(3;0,6,7) model without intercept and the value of BIC is 9507.52. The fitted mixture autoregressive model uses a white-noise component with large variance to capture the extreme observations in the return series while the fitted TMAR(3;0,4,4) model uses the two heavy-tailed fourth-order autoregressive components to capture the same phenomenon.

Besides the mixture autoregressive model, we try to compare our model with autoregressive (AR), moving average (MA), random walk (RW), generalized autoregressive conditional heteroscedastic (GARCH) and GARCH with Student t distribution (GARCH–t) models on the basis of their ability to describe the predictive distribution of the return series. The estimation results for these models are not shown here and are available from the authors. The empirical coverages of the in-sample one-step-ahead prediction intervals, based on \( F(r_{t+1} \mid F_t) \), \( t = 9, \ldots, 2501 \), for the return series generated by each model are given in Table 5. If a model describes the conditional distributions of the data well, the empirical coverages of the one-step-ahead prediction intervals generated should be close to the nominal levels. Generally speaking, performance of prediction under Student t-mixture autoregressive model is better than those under other models, except that the GARCH(1,1)–t model has similar performance. The empirical coverages of the prediction intervals based on the TMAR(3;0,4,4) and GARCH(1,1)–t models are closer to the nominal coverages. On the other hand, for the other models, the deviations of empirical coverages from the nominal coverages are increased as the nominal level of prediction intervals is decreased.

We generated the empirical coverages of one-sided lower prediction intervals for the return series. The lower prediction intervals for return series are related to the concept of Value-at-Risk.
Table 5. Empirical coverage of the \((1 - \alpha)100\%\) prediction intervals for the 3-year Treasury Constant Maturity Rate

<table>
<thead>
<tr>
<th>Model</th>
<th>99</th>
<th>98</th>
<th>95</th>
<th>90</th>
<th>80</th>
<th>70</th>
<th>60</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMAR(3;0,4,4)</td>
<td>99.04</td>
<td>97.91</td>
<td>94.82</td>
<td>89.21</td>
<td>79.98</td>
<td>70.75</td>
<td>61.12</td>
<td>51.28</td>
</tr>
<tr>
<td>MAR(3;0,6,7)</td>
<td>98.88</td>
<td>98.48</td>
<td>95.71</td>
<td>90.09</td>
<td>79.17</td>
<td>71.43</td>
<td>61.96</td>
<td>53.61</td>
</tr>
<tr>
<td>AR(2)</td>
<td>97.03</td>
<td>96.19</td>
<td>93.86</td>
<td>91.01</td>
<td>85.71</td>
<td>80.62</td>
<td>74.72</td>
<td>68.02</td>
</tr>
<tr>
<td>MA(1)</td>
<td>97.07</td>
<td>96.19</td>
<td>93.90</td>
<td>91.09</td>
<td>85.79</td>
<td>80.58</td>
<td>75.04</td>
<td>67.74</td>
</tr>
<tr>
<td>RW</td>
<td>97.03</td>
<td>96.11</td>
<td>93.90</td>
<td>90.97</td>
<td>85.55</td>
<td>80.62</td>
<td>75.28</td>
<td>67.94</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>98.03</td>
<td>96.87</td>
<td>94.90</td>
<td>91.21</td>
<td>83.75</td>
<td>75.40</td>
<td>66.05</td>
<td>56.82</td>
</tr>
<tr>
<td>GARCH(1,1)-t</td>
<td>98.88</td>
<td>98.03</td>
<td>95.02</td>
<td>89.77</td>
<td>80.18</td>
<td>69.78</td>
<td>60.39</td>
<td>48.39</td>
</tr>
</tbody>
</table>

Table 6. Empirical coverage of the \((1 - \alpha)100\%\) one sided lower prediction intervals for 3-year Treasury Constant Maturity Rate

<table>
<thead>
<tr>
<th>Model</th>
<th>99</th>
<th>98</th>
<th>95</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMAR(3;0,4,4)</td>
<td>99.16</td>
<td>97.87</td>
<td>94.86</td>
</tr>
<tr>
<td>MAR(3;0,6,7)</td>
<td>99.48</td>
<td>98.27</td>
<td>95.39</td>
</tr>
<tr>
<td>AR(2)</td>
<td>97.99</td>
<td>97.35</td>
<td>95.79</td>
</tr>
<tr>
<td>MA(1)</td>
<td>98.03</td>
<td>97.51</td>
<td>95.87</td>
</tr>
<tr>
<td>RW</td>
<td>97.95</td>
<td>97.43</td>
<td>95.91</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>98.43</td>
<td>97.75</td>
<td>95.30</td>
</tr>
<tr>
<td>GARCH(1,1)-t</td>
<td>99.16</td>
<td>97.91</td>
<td>94.46</td>
</tr>
</tbody>
</table>

in finance literature. The results are recorded in Table 6. The performance of lower prediction intervals under the TMAR(3;0,4,4) model is better than those under the other models. The relative percentage of absolute deviations of the empirical coverages to nominal coverages are at most 0.16% for TMAR(3;0,4,4)-based lower prediction intervals. In comparison, the corresponding percentage is a bit higher at 0.57% for GARCH(1,1)-t-based 95% lower prediction intervals.

ACKNOWLEDGEMENT

We thank the editor, the associate editor and the referees for comments that led to improvement of the paper. This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region.

APPENDIX

Technical details

The conditional expectation of \(Z_{kt}\) is just the posterior probability that the observation \(y_t\) comes from the \(k\)th component of the mixing distribution, conditional on \(\Psi\) and \(Y\). Hence equation (4) follows.

Since the gamma distribution is the conjugate prior distribution for \(W_{kt}\), it is easy to show that the conditional distribution of \(W_{kt}\), given \(y_t\) and \(z_{kt} = 1\), is a gamma distribution with parameters \((\nu_k + 1)/2\) and \((\delta_k^2 + \nu_k - 2)/2\). Hence, the conditional expectation of \(W_{kt}\) is given in equation (5).
Consider a random variable $X$ distributed as gamma($\alpha, b$). The moment generating function of the random variable $\log X$ is $M_{\log X}(t) = E(e^{t\log X}) = E(X^t) = \Gamma(\alpha + t)/[\Gamma(\alpha) b^t]$. Thus, $E(\log X) = \psi(\alpha) - \log(b)$. The expectation of $\log W_{kt}$, conditional on $y_t$ and $z_{kt} = 1$, is

$$E_\varphi(\log W_{kt} | y_t, z_{kt} = 1) = \psi\left(\frac{\nu_k + 1}{2}\right) - \log \left(\frac{\delta_{kt}^2 + \nu_k - 2}{2}\right).$$

Hence, equation (6) follows.

**M-step equations.** The first derivatives of the log-likelihood (3) with respect to $\Psi$ are

$$\frac{\partial \ell_1}{\partial \alpha_k} = \sum_{t=1}^{n} \left( \frac{Z_{kt}}{\alpha_k} - \frac{Z_{gt}}{\alpha_g} \right) (k = 1, \ldots, g - 1), \quad (A1)$$

$$\frac{\partial \ell_2}{\partial \nu_k} = \sum_{t=1}^{n} Z_{kt} \left[ -\frac{1}{2} \psi\left(\frac{1}{2}\nu_k\right) + \frac{1}{2} \log \left(\frac{\nu_k - 2}{2}\right) \right.\left. + \frac{1}{2} \left(\frac{\nu_k}{\nu_k - 2}\right) \right] (k = 1, \ldots, g), \quad (A2)$$

$$\frac{\partial \ell_3}{\partial \phi_{ki}} = \sum_{t=1}^{n} Z_{kt} W_{kt} u(y_t, i) e_{kt} \left(\frac{\nu_k}{\sigma_k^2} \right) (k = 1, \ldots, g; i = 0, \ldots, p_k), \quad (A3)$$

$$\frac{\partial \ell_3}{\partial \sigma_k^2} = \sum_{t=1}^{n} \left( \frac{W_{kt} e_{kt}^2}{\sigma_k^2} - 1 \right) (k = 1, \ldots, g). \quad (A4)$$

By replacing $Z_{kt}$, $W_{kt}$ and $\log(W_{kt})$ with their conditional expectations into equations (A1) to (A4) and setting these equations to zero, we obtain the estimating equations (7) to (9).

**Newton–Raphson method for estimating $\nu_k$.** Denote the left-hand-side of equation (9) as $f(\nu_k)$, $\nu_k^*\nu_k$ is the value in the previous step and $\nu_k^\nu_k$ is the value in the current step. We can obtain the solution by iterating the following formula, $\nu_k^\nu_k = \nu_k^* - f(\nu_k^*)/f'(\nu_k^*)$, where $f'(\nu_k) = (\nu_k - 4)/(\nu_k - 2)^2 - \frac{1}{2} \psi'(\nu_k/2)$ is the first derivative of $f(\nu_k)$. Here, $\nu_k > 2$ and $\psi(\cdot)$ is the Trigamma function.

**REFERENCES**

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64, 247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 74, 33–43.


