第50卷第3期 中国科学技术大学学报 Vol. 50, No. 3
2020年3月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Mar. 2020

文章编号:0253-2778(2020)03-0317-11

A robust joint modeling approach for longitudinal data

TAN Jiaxin, ZHANG Weiping

(*Department of Statistics and Finance University of Science and Technology of China*)

Abstract: A robust method is proposed for analyzing longitudinal continuous responses with potential outliers by using the multivariate t distribution. Unlike the existing approaches which mainly focus on the inference of regression mean, our approach aims to reveal the dynamics in the location function, marginal scale function and association by joint parsimoniously modeling the location and dependence structure. An ECME-based algorithm is applied to speed up the computation associated with the EM algorithm for maximum likelihood estimation. The resulting estimators are shown to be consistent and asymptotic normality. Numerical studies demonstrate the effectiveness of the proposed approach.

Key words: longitudinal data; robust estimation; em algorithm; joint modeling

CLC number: O212.4 **Document code**: A doi: 10.3969/j.issn.0253-2778.2020.03.009

引用格式: TAN Jiaxin, ZHANG Weiping. A robust joint modeling approach for longitudinal data[J]. Journal of University of Science and Technology of China, 2020,50(3):317-327.
檀佳欣,张伟平. 纵向数据的一种稳健同时建模方法[J]. 中国科学技术大学学报,2020,50(3):317-327.

纵向数据的一种稳健同时建模方法

檀佳欣,张伟平

(中国科学技术大学统计与金融系,安徽合肥 230026)

摘要:使用多元 t 分布,提出了一种分析带有异常值的连续纵向数据的同时建模方法.不同于已有主要推断回归均值的稳健方法,本文旨在通过稳健同时参数化建模来揭示位置参数,边际尺度参数和相依参数的动态变化机制.为了加速极大似然估计过程中EM算法的速度,采用一种基于ECME的最大似然估计求解算法,所得到的估计量被证明具有相合性和渐近正态性.数据分析表明所提方法是有效的.

关键词:纵向数据;稳健估计;EM 算法;同时建模

0 Introduction

A typical characteristic of longitudinal studies is that study subjects are measured over repeated time intervals. Thus, observations for the same subject are intrinsically correlated. It is fundamentally important to account for within-subject correlation in analyzing such data.

Received: 2019-05-02; **Revised**: 2019-06-14

Foundation item: Supported by National Key Research & Development Plan(2016YFC0800104), National Natural Science Foundation of China(11671374, 71771203,71631006).

Biography: TAN Jiaxin, female, born in 1993,Research field: Longitudinal data analysis. E-mail: tjxggsd@163.com

Corresponding author: ZHANG Weiping, PhD/professor. E-mail: zwp@ustc.edu.cn

Regression models on the mean and variance functions for understanding longitudinal data have been extensively studied in the literature. Ref. [1] gave an excellent overview of various approaches in this filed. To understand the dynamics in the mean function and covariance structure, a class of mean-variance-correlation modeling framework has been explored; see Refs. [2-9]. These methods perform well under certain assumptions, but are not resistant to outliers. In this paper, one proposes a joint modeling method of t distribution with missing data.

Compared with the robust methods for regression mean, the study on robust approaches for jointly parsimoniously modeling with longitudinal data received limited attention although valuable. Ref. [10] considered robustification on the mean and covariance where they set up estimating equations for both the mean and the dispersion parameter. The constraint of their approach is that they assumed an inflexible covariance structure determined by two parameters. Refs. [11-12] developed robust estimation for the mean and covariance jointly for the regression model of longitudinal data within the framework of generalized estimating equations[13]. As an alternative, the t-distribution is widely used for longitudinal complete data. Ref. [14] discussed the robust statistical modeling using the t-distribution in a general framework. Ref. [15] proposed a multivariate regression model with its mean and scale covariance modeled jointly based on modified Cholesky decomposition for the analysis of longitudinal data. Ref. [16] obtained robust estimation of the correlation matrix of longitudinal data based on alternative Cholesky decomposition. Ref. [17] used t-distribution to carry out Bayesian inference in longitudinal data. These results show that the t-distribution perform well to obtain robust estimation.

Our reasons for revisiting this topic are threefold. Firstly, the existing literature in joint modeling the regression mean and covariance structure frequently assume normality, while such assumption is routinely made for mathematical convenience. However, such assumption is not always realistic because of the presence of atypical observations and the existing joint modeling approaches are sensitive to outliers, contamination, or heavy-tailed distributions. To remedy this weakness, we considered the use of the multivariate t distribution for robust estimation of regression models, since inference based on parsimonious modeling under t-distribution combines conceptual simplicity with generality. The degree of freedom parameter of the t-distribution provides a convenient dimension for achieving robust statistical inference. Secondly, the existing robust joint modeling approaches can be viewed as indirectly robust modeling the variances and covariances of the longitudinal measurements. More specifically, due to the modified Cholesky decomposition, the resulting variance functions of the aforementioned approaches cannot be directly interpreted as those of the repeated measurements. Moreover, the same interpretation issue also arises for the covariance and correlation structures when these approaches are applied. Therefore, for practical applications, additional effort and extra care are necessary for interpreting the resulting variance and covariance functions. We therefore propose to directly model the regression mean, and the dependence structure simultaneously. Thirdly, the parameter estimates under standard maximum likelihood procedure can be of little practical interest by themselves because they can critically influence the behavior of iterative numerical optimization algorithm especially for small or unknown degrees of freedom. We apply an ECME method[18] to speed up the Monte Carlo implementation of the EM algorithm.

The rest of the paper is organized as follows. In Section 1 we give some general notations for the t-distribution and introduce the joint modeling approach for the mean and covariance structure. In

Section 2, we introduce the likelihood and ECME-based algorithm. In Section 3, we carry out numerical studies to investigate the finite sample properties and demonstrate the effectiveness of the proposed method. We conclude this paper by summarizing the main findings and outlining future research in Section 4.

1 Models

Let $y_i = (y_{i1}, \cdots, y_{im_i})^{\mathrm{T}}$ be the m_i longitudinal measurements for the ith subject, where the response y_{ij} is observed at time t_{ij}. Let $t_i = (t_{i1}, \cdots, t_{im_i})^{\mathrm{T}}$, and we denote $x_{ij} \in \mathbb{R}^p$ as the covariate for the jth measurement of subject i. To accommodate the presence of atypical observations, we assume that y_i follows a multivariate t distribution, denoted by $t_{m_i}(\mu_i, \Omega_i, \upsilon)$, with density

$$f(y_i;\mu_i,\Omega_i,\upsilon) = \frac{\Gamma((\upsilon+m_i)/2)\,|\,\Omega_i\,|^{-1/2}}{\Gamma(\upsilon/2)(\pi\upsilon)^{m_i/2}} \cdot \left(1+\frac{(y_i-\mu_i)^{\mathrm{T}}\Omega_i^{-1}(y_i-\mu_i)}{\upsilon}\right)^{-(\upsilon+m_i)/2}, \ y_i \in \mathbb{R}^{m_i} \tag{1}$$

where the location parameter $\mu_i = (\mu_{i1}, \cdots, \mu_{im_i})^{\mathrm{T}}$ and scale matrix Ω_i is an $m_i \times m_i$ positive definite matrix. The degree of freedom υ, which controls the thickness of the tails of the distribution, is directly related to the degree of robustness of inference, and smaller υ yields higher robustness. The following lemma shows that a multivariate t distribution $t_m(\mu, \Omega, \upsilon)$ can be seen as the mixture of m-variate normal and Gamma distribution variables.

Lemma 1.1 Let $z \sim N_m(0, \Omega)$, $\tau \sim \gamma(\upsilon/2, \upsilon/2)$ be independent, then $z/\sqrt{\tau} + \mu \sim t_m(\mu, \Omega, \upsilon)$, where the density function of $\gamma(\alpha, \beta)$ is $\beta^{\alpha}\tau^{\alpha-1}\exp\{-\beta\tau\}/\gamma(\alpha)$ with $\tau>0, \alpha>0, \beta>0$.

Proof This well-known representation of multivariate t distribution can be easily found in Refs. [18-20] etc.

For $\upsilon>1$, the mean vector of y_i is defined to be μ_i; for $\upsilon>2$, the covariance matrix y_i is $\dfrac{\upsilon}{\upsilon-2}\Omega_i$. We believe that inference based on a parametric model such as model (1) combines conceptual simplicity with generality, since it can be applied in a wide range of settings. A detailed discussion of mathematical properties and estimation methods for this distribution with complete data can be found in Ref. [20].

With the parametric model (1), it is well-known that modeling covariance (and correlation) matrix is a challenging problem due to the large dimensionality and positive-definiteness constraint. Therefore, with so many parameters in the scale matrix $\{\Omega_i\}$ $(i=1, \cdots, n)$ associated with the heteroscedasticity in longitudinal data, we decompose Ω_i as $\Omega_i = D_i R_i D_i$, where D_i is a diagonal matrix whose diagonal elements $(\sigma_{i1}, \cdots, \sigma_{im_i})$ are the square root of the diagonal element of Ω_i and can be seen as the marginal scale. R_i is a correlation matrix of mixture component z, which is also the correlation matrix of y_i if it exists. Clearly, $\ln(\sigma_{ij})'$ s are unconstrained and parameterized via regression techniques. To parsimoniously model the dynamics in R_i, we follow the idea of Ref. [9] to parameterize them via hyperspherical co-ordinates by the decomposition $R_i = T_i T_i^{\mathrm{T}}$, where T_i is a lower triangular matrix given by

$$T_i = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ c_{i21} & s_{i21} & 0 & \cdots & 0 \\ c_{i31} & c_{i32}s_{i21} & s_{i32}s_{i31} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{im_i1} & c_{im_i2}s_{im_i1} & \cdots & \cdots & \prod_{l=1}^{m_i-1} s_{im_il} \end{pmatrix} \tag{2}$$

with $c_{\mathrm{ijk}} = \cos(\phi_{\mathrm{ijk}})$ and $s_{\mathrm{ijk}} = \sin(\phi_{\mathrm{ijk}})$ being trigonometric functions of angles ϕ_{ijk}. In other words, the non-zero entries in the lower diagonal matrix T_i are given by $T_{i11}=1, T_{ij1}=\cos(\phi_{ij1})$ for $j=2,\cdots,m_i$, and

$$T_{\mathrm{ijk}} = \begin{cases} \cos(\phi_{\mathrm{ijk}})\prod_{l=1}^{k-1}\sin(\phi_{ijl}),\ 2 \leqslant k < j \leqslant m_i \\ \prod_{l=1}^{k-1} sin(\phi_{ijl}),\ k=j, j=1,\cdots,m_i \end{cases} \tag{3}$$

Here the total number of angles ϕ_{ijk} $(1 \leqslant k <$

$j \leqslant m_i$) in expressions (2) and (3) is $m_i(m_i-1)/2$, which is the same as that of the free parameters in an unconstrained correlation matrix. As pointed in Ref. [9], such decomposition automatically leads to positive definite correlation matrices, and the parameters in it are related to well-founded statistical concepts.

Motivated by the above considerations, we propose a joint regression model for the location, the marginal scale and the correlations as

$$\left.\begin{aligned} g(\mu_{ij}) &= x_{ij}^{\mathrm{T}}\beta \\ \ln(\sigma_{ij}) &= z_{ij}^{\mathrm{T}}\lambda \\ \phi_{\mathrm{ijk}} &= w_{\mathrm{ijk}}^{\mathrm{T}}\gamma \end{aligned}\right\} \tag{4}$$

where $g(\cdot)$ is a known link function, which is usually taken as an identity function as in linear models, also β, γ and λ are unknown parameters for parameterizing the location, the marginal scale and the correlation. x_{ij}, z_{ij} and w_{ijk} are $p\times 1, q\times 1$ and $d\times 1$ vectors of covariates available, and z_{ij} does not include intercept for identifiability concern. In practice, natural candidates for w_{ijk} include $(t_{ij}, t_{ik})^{\mathrm{T}}$ and its higher order terms, or more simply a polynomial of the time lag $(t_{ik}-t_{ij})$ such that the resulting correlation is stationary. Further discussion of these covariates can be found in Refs. [6,9]. Remarkably, the angle ϕ_{ijk} can also be transformed by arctan to ensure that it falls in $[0,\pi)$. The model (4) can be generalized easily to nonparametric and semiparametric models, although the focus of the paper is on parametric models as in model(4).

2 Likelihood and estimation

2.1 Maximum likelihood estimation

Under the sample $y_1, \cdots, y_n$ and model (4), the log-likelihood function of the multivariate t-distribution (1) ignoring constant is given by

$$l(\beta,\gamma,\lambda,\upsilon) = \sum_{i=1}^{n}\Big[\ln\Gamma\Big(\frac{\upsilon+m_i}{2}\Big) - \ln\Gamma\Big(\frac{\upsilon}{2}\Big) - \frac{1}{2}\ln|\Omega_i| + \frac{\upsilon}{2}\ln\upsilon - \frac{\upsilon+m_i}{2}\ln(\upsilon + \|\delta_i\|^2)\Big] \tag{5}$$

where $\delta_i = \Omega_i^{-1/2}(y_i-\mu_i)$. The score equations for β, $\eta=(\gamma^{\mathrm{T}},\lambda^{\mathrm{T}})^{\mathrm{T}}$ and υ can be derived from the *log*-likelihood function (5), namely

$$U(\beta) = \sum_{i=1}^{n}\frac{\upsilon+m_i}{\upsilon+\|\delta_i\|^2}\frac{\partial\mu_i^{\mathrm{T}}}{\partial\beta}\Omega_i^{-1}(y_i-\mu_i) = 0 \tag{6}$$

$$U(\eta) = \frac{\partial l}{\partial\eta} = -\frac{1}{2}\sum_{i=1}^{n}\Big[\frac{\partial\ln|\Omega_i|}{\partial\eta} + \frac{\upsilon+m_i}{\upsilon+\|\delta_i\|^2}\frac{\partial\Omega_i^{-1}}{\partial\eta^{\mathrm{T}}}\mathrm{vec}((y_i-\mu_i)(y_i-\mu_i))\Big] = 0 \tag{7}$$

$$U(\upsilon) = \frac{1}{2}\sum_{i=1}^{n}\Big[\psi\Big(\frac{\upsilon+m_i}{2}\Big) - \psi\Big(\frac{\upsilon}{2}\Big) + \ln(\upsilon) + 1 - \ln(\upsilon+\|\delta_i\|^2) - \frac{\upsilon+m_i}{\upsilon+\|\delta_i\|^2}\Big] = 0 \tag{8}$$

where $\psi(x) = d\ln\Gamma(x)/dx$ is the digamma function. The derivative of a matrix A with respect to vector $u=(u_1,\cdots,u_d)^{\mathrm{T}}$ is defined as $\frac{\partial A}{\partial u} = (\mathrm{vec}(\frac{\partial A}{\partial u_1})^{\mathrm{T}},\cdots,\mathrm{vec}(\frac{\partial A}{\partial u_d}))$.

$$E\left[1+\frac{\|\delta_i\|^2}{\upsilon}\right]^{-k} = \frac{(\upsilon/2+k-1)\cdots(\upsilon/2)}{((\upsilon+m_i)/2+k-1)\cdots((\upsilon+m_i)/2)} \tag{9}$$

and conditioning on $\|\delta_i\|=r$, δ_i is uniformly distributed on the sphere $\|\delta_i\|=r$, therefore the block elements of expected Hessian matrix J_n with respect to $\theta=(\beta^{\mathrm{T}}, \lambda^{\mathrm{T}}, \gamma^{\mathrm{T}}, \upsilon)^{\mathrm{T}}$ can be obtained as follows:

$$J_{n11}(\theta) = \sum_{i=1}^{n}\frac{\upsilon+m_i}{\upsilon+m_i+2}\frac{\partial\mu_i^{\mathrm{T}}}{\partial\beta}\Omega_i^{-1}\frac{\partial\mu_i}{\partial\beta^{\mathrm{T}}} \tag{10}$$

$$J_{n22}(\theta) = \sum_{i=1}^{n}\Big[\frac{\upsilon+m_i}{\upsilon+m_i+2}Z_i^{\mathrm{T}}(I_{m_i}+R_i^{-1}\circ R_i)Z_i - \frac{2}{\upsilon+m_i+2}Z_i^{\mathrm{T}}11^{\mathrm{T}}Z_i\Big] \tag{11}$$

$$J_{n23}(\theta) = \sum_{i=1}^{n}\Big[\frac{\upsilon+m_i}{\upsilon+m_i+2}(Z_i^{\mathrm{T}}R_i^{-1})\otimes e_{1m_i}^{\mathrm{T}}\frac{\partial R_i}{\partial\gamma} - \frac{1}{2(\upsilon+m_i+2)}Z_i^{\mathrm{T}}1\frac{\partial R_i}{\partial\gamma^{\mathrm{T}}}\mathrm{vec}(R_i^{-1})\Big] \tag{12}$$

$$J_{n24}(\theta) = -\sum_{i=1}^{n}\frac{2}{(\upsilon+m_i+2)(\mu+m_i)}Z_i^{\mathrm{T}}1 \tag{13}$$

$$J_{n33}(\theta) = \sum_{i=1}^{n}\Big[\frac{\upsilon+m_i}{2(\upsilon+m_i+2)}\frac{\partial R_i}{\partial\gamma^{\mathrm{T}}}(R_i^{-1}\otimes R_i^{-1})\frac{\partial R_i}{\partial\gamma} -$$

$$\frac{1}{2(\upsilon+m_i+2)}\frac{\partial R_i}{\partial\gamma^{\mathrm{T}}}\mathrm{vec}(R_i^{-1})\mathrm{vec}^{\mathrm{T}}(R_i^{-1})\frac{\partial R_i}{\partial\gamma}] \tag{14}$$

$$J_{n34}(\theta)= -\sum_{i=1}^{n}\frac{2}{(\upsilon+m_i+2)(\mu+m_i)}\frac{\partial R_i}{\partial\gamma^{\mathrm{T}}}\mathrm{vec}(R_i^{-1})) \tag{15}$$

$$J_{n44}(\theta)= \sum_{i=1}^{n}[\frac{1}{4}\psi_1(\frac{\upsilon}{2})+\frac{1}{2(\upsilon+m_i)}-\frac{1}{4}\psi_1(\frac{\upsilon+m_i}{2})- \frac{m_i}{2\upsilon(\upsilon+m_i)}-\frac{\upsilon+2}{2\upsilon(\upsilon+m_i+2)}] \tag{16}$$

where $Z_i^{\mathrm{T}}=(z_{i1},\cdots,z_{im_i})$, the derivative $\frac{\partial R_i}{\partial\gamma_k}$ can be easily obtained by Ref. [9]. $e_{1m_i}=(1,0,\cdots,0)^{\mathrm{T}}$ is the first canonical basis of $\mathbb{R}^{m_i}$. And $A\circ B$ denotes the Hadamard product of matrix A and B, $A\otimes B$ is the Kronecker product of A and B. Finally, $\psi_1(x)=\mathrm{d}\psi(x)/\mathrm{d}x$.

The maximum likelihood estimators $\hat\beta,\hat\lambda,\hat\gamma,\hat\upsilon$ can be shown to be consistent and asymptotic normal distributed. Assuming the following regularity conditions:

Condition 2.1 The dimensions p, q and d of covariates x_{ij}, z_{ij} and w_{ijk} are fixed; $\max_{1\leqslant i\leqslant n}m_i$ is bounded.

Condition 2.2 The parameter space Θ of $\theta=(\beta^{\mathrm{T}},\lambda^{\mathrm{T}},\gamma^{\mathrm{T}},\upsilon)^{\mathrm{T}}$ is a bounded compact set in $\mathbb{R}^{p+q+d}\times\mathbb{R}^{+}$, and the true value $\theta_0=(\beta_0{}^{\mathrm{T}},\lambda_0{}^{\mathrm{T}},\gamma_0,\upsilon_0)^{\mathrm{T}}$ is in the interior of Θ.

Condition 2.3 As $n\to\infty$, the average negative expected Hessian matrix converges to a positive definite matrix $I(\theta_0)$, i. e., $\lim_{n\to\infty}\frac{1}{n}J_n(\theta_0)=I(\theta_0)$.

Condition 2.1 is routinely made for longitudinal data from the practical perspective. Condition 2.2 is a conventional assumption for theoretical analysis of the maximum likelihood approach. Condition 2.3 is a natural requirement for the regression analysis in unbalanced longitudinal data modeling.

Theorem 2.1 Under the distribution (1) and regularity conditions 1-3, Let $\hat\theta_n=(\hat\beta^{\mathrm{T}},\hat\lambda^{\mathrm{T}},\hat\gamma^{\mathrm{T}},\hat\upsilon)^{\mathrm{T}}$ be the maximum likelihood estimator of the true parameter value θ_0 in Eq. (5). then (a) $\hat\theta_n$ is strongly consistent for the true value θ_0. and (b) $\hat\theta_n$ is asymptotically normally distributed, that is

$$\sqrt{n}(\hat\theta_n-\theta_0)\rightsquigarrow N[0,I^{-1}(\theta_0)]$$

in distribution.

Proof We only need to verify the regular conditions for maximum likelihood estimation. Let $l_i=\ln f_i(y_i,\theta)$, $(i=1,\cdots,n)$. Then ignoring the constant $\frac{1}{2}m_i\ln(\pi)$, we obtain that

$$l_i=\ln\Gamma(\frac{\upsilon+m_i}{2})-\ln\Gamma(\frac{\upsilon}{2})+\frac{\upsilon}{2}\ln\upsilon- \frac{1}{2}\ln|\Omega_i|-\frac{\upsilon+m_i}{2}\ln(\upsilon+\|\delta_i\|^2).$$

Notice that $\ln(\upsilon+\|\delta_i\|^2)\leqslant\ln\upsilon+\|\delta_i\|/\upsilon$ and boundedness of $E_0\|\delta_i\|^2$, therefore by Kolmogorov's strong law of large numbers we have that

$$\frac{1}{n}\sum_{i=1}^{n}l_i-\frac{1}{n}\sum_{i=1}^{n}E_0(l_i)\to 0,\ \text{a. s.}.$$

where the expectation E_0 is taken under the distribution of y_i with true parameters θ_0. It can be shown that $\frac{1}{n}\sum_{i=1}^{n}E_0(l_i(\theta))$ is equicontinuous in θ, then following the proof of Theorem 2.1 in Ref. [21], it is easy to show the consistency of $\hat\theta$.

The proof of asymptotic normality of $\hat\theta_n$ is essentially the same as that of Theorem 2.2 in Ref. [21].

Since $\hat\theta$ is consistent estimator for θ_0, the Fisher information matrix $I(\theta_0)$ can be consistently estimated by a matrix $\frac{1}{n}J_n(\hat\theta_n)$. From Theorem 2.1, $\hat\beta$ is asymptotically independent of $\hat\gamma$, $\hat\lambda$ and $\hat\upsilon$. This is not surprising for statistical inferences of elliptical distributed data, because $\hat\beta$ concerns the location function, and $\hat\gamma$, $\hat\lambda$ and $\hat\upsilon$ are the estimators for parameters of the scale matrix. Therefore, the optimal efficiency of estimating β is assured whenever Ω_i's or the models for σ_{ij}^2 and ϕ_{ijk} are correctly specified. If the model for Ω_i is misspecified, $\hat{\beta}$ is still consistent and asymptotically normal by a result in Ref. [13], although the asymptotic variance of $\hat{\beta}$ would take a sandwich form. On the other hand, the two covariation parameters $\hat{\gamma}$ and $\hat{\lambda}$ are not asymptotically independent in general.

When the probability model (1) is not correctly specified, let population parameter vector $\theta_* = (\beta_*^{\mathrm{T}}, \gamma_*^{\mathrm{T}}, \lambda_*^{\mathrm{T}})^{\mathrm{T}}$ be the unique minimizer of the Kullback-Leibler divergence $KL(f \mid f_0) = E_{f_0} \log f_0 / f$ between a true model with continuous density f_0 and a working model f defined by (1), and denote by $K = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} E_{f_0} U(\theta) U(\theta)^{\mathrm{T}}$ with $U(\theta) = (U^{\mathrm{T}}(\beta), U^{\mathrm{T}}(\beta), U^{\mathrm{T}}(\upsilon))^{\mathrm{T}}$ given in Eqs. (6)-(8). Then we have the following result.

Corollary 2.1 Under regularity conditions 2.1 and 2.2 and existence of $E\log f_0(y_i)$ with θ_0 replaced by θ_*, as $n\to\infty$, we have that the maximum likelihood estimator $\hat{\theta}$ is strongly consistent for θ_*; If additionally condition 2.3 holds, the matrix $K(\theta_*)$ and its inverse are nonsingular, then $\sqrt{n}(\hat{\theta}_n - \theta_*) \rightsquigarrow N[0, I^{-1}(\theta_*) K(\theta_*) I^{-1}(\theta_*)]$.

Proof The corollary follows directly from Theorem 2.2 and 3.2 of Ref. [22].

2.2 ECME algorithm

The maximum likelihood estimates of parameters could be found by directly solving the score functions (6)-(8) using various optimization algorithms, however, care must be used with the standard maximum likelihood method under t distribution with unknown degree of freedom. Since the score functions $U(\beta)$ and $U(\eta)$ are bounded while $U(\upsilon)$ is unbounded when $\| \delta_i \|^2$ goes to infinity, it can be inferred that areas of likelihood unboundedness are most likely to occur as $\upsilon \to 0$[23]. That is to say, the likelihood function can be arbitrary large with reasonable parameter values when the degree of freedom is small or unknown. Therefore, the parameter estimates under standard maximum likelihood procedure can be of little practical interest by themselves even though they are formally local or even global maxima because they can critically influence the behavior of iterative simulation algorithms designed to summarize the likelihood function[18]. Therefore, we applied the ECME algorithm developed by Ref. [18] to find the maximum likelihood estimates.

Following Lemma 1.1, the multivariate t distribution (1) can be seen as the mixture of m_i-variate Normal and Gamma distribution variable. Therefore we can use EM type algorithm, which is commonly used to calculate the maximum likelihood estimates[24]. Let $\theta = (\beta^{\mathrm{T}}, \lambda^{\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}}$ and the complete-data log-likelihood function of the ith subject be

$$l_{i,\mathrm{full}}(\theta, \upsilon) = \ln f(y_i, \tau_i \mid \theta, \upsilon) = \ln f(\tau_i \mid \upsilon) + \ln f(y_i \mid \tau_i, \theta) \tag{17}$$

$$= l_{i1,\mathrm{full}}(\upsilon) + l_{i2,\mathrm{full}}(\theta) \tag{18}$$

In the E-step, we calculate the expectation of complete-data log-likelihood given the observed data and current values of the parameters. Thus, the E-step for the ith subject at the $(t+1)$st iteration based on Eq. (17) is

$$Q_i(\theta, \upsilon \mid \theta_t, \upsilon_t) = E[l_{i,\mathrm{full}}(\theta, \upsilon) \mid y_i, \theta_t, \upsilon_t] = E[l_{i1,\mathrm{full}}(\upsilon) \mid y_i, \theta_t, \upsilon_t] + E[l_{i2,\mathrm{full}}(\theta) \mid y_i, \theta_t, \upsilon_t] \tag{19}$$

$$= Q_{i1}(\upsilon \mid \upsilon_t) + Q_{i2}(\theta \mid \theta_t) \tag{20}$$

where the expectation E is taken under the conditional distribution of τ_i given y_i and parameters θ_t, υ_t. Direct computation leads to

$$E[l_{i1,\mathrm{full}}(\upsilon) \mid y_i, \theta_t, \upsilon_t] = E[\ln f(\tau_i \mid \upsilon, \theta) \mid y_i, \theta_t, \upsilon_t] = \frac{\upsilon}{2}\ln\frac{\upsilon}{2} + \frac{\upsilon}{2} E[(\ln\tau_i - \tau_i) \mid y_i, \theta_t, \upsilon_t] - \ln\gamma(\frac{\upsilon}{2}) - E[\ln\tau_i \mid y_i, \theta_t, \upsilon_t] \tag{21}$$

$$E[l_{i2,\mathrm{full}}(\theta) \mid y_i, \theta_t, \upsilon_t] = E[\log f(y_i \mid \tau_i, \theta) \mid y_i, \theta_t, \upsilon_t] = -\frac{m_i}{2}\ln 2\pi - \frac{1}{2}\ln |\Omega_i| + \frac{m_i}{2}\zeta_{i(t)}(y_i - \mu_i)'\Omega_i^{-1}(y_i - \mu_i)\hat{\omega}_{i(t)} = E[h_i(\theta, y_i) \mid y_i, \theta_t, \upsilon_t] \tag{22}$$

where

$$h_i(\theta, y_i) = -\frac{m_i}{2}\ln 2\pi - \frac{1}{2}\ln |\Omega_i| + \frac{m_i}{2}\ln \tau_i - \frac{\tau_i}{2}(y_i - \mu_i)'\Omega_i^{-1}(y_i - \mu_i) \quad (23)$$

$$\zeta_{i(t)} = E(\ln\tau_i \mid y_i, \theta_t, \upsilon_t) = \psi(\frac{m_i + \upsilon_t}{2}) + \ln\frac{\upsilon_t + (y_i - \mu_{i(t)})'\Omega_{i(t)}{}^{-1}(y_i - \mu_{i(t)})}{2} \quad (24)$$

and

$$\hat{\omega}_{i(t)} = E(\tau_i \mid y_i, \theta_t, \upsilon_t) = \frac{\upsilon_t + m_i}{\upsilon_t + (y_i - \mu_{i(t)})'\Omega_{i(t)}{}^{-1}(y_i - \mu_{i(t)})} \quad (25)$$

In the M-step, $\theta^{(t+1)}$ and $\upsilon^{(t+1)}$ are chosen to maximize the Q-function

$$Q(\theta, \upsilon \mid \theta_t, \upsilon_t) = \sum_{i=1}^{n} Q_i(\theta, \upsilon \mid \theta_t, \upsilon_t) \quad (26)$$

Especially, $\upsilon^{(t+1)}$ can be obtained by finding the solution to the equation:

$$-\psi(\upsilon/2) + \ln(\upsilon/2) + \frac{1}{n}\sum_{i=1}^{n}[\ln\hat{\omega}_{i(t)} - \hat{\omega}_{i(t)}] + 1 + \frac{1}{n}\sum_{i=1}^{n}\left\{\psi(\frac{\upsilon_t + m_i}{2}) - \ln(\frac{\upsilon_t + m_i}{2})\right\} = 0 \quad (27)$$

where

$$\hat{\omega}_{i(t)} = \frac{\upsilon_t + m_i}{\upsilon_t + (y_i - \mu_{i(t)})'\Omega_{i(t)}{}^{-1}(y_i - \mu_{i(t)})}.$$

Because the last term on the left side of equation (27) is non-positive and $-\psi(\upsilon/2)+\ln(\upsilon/2)$ is decreasing in $(0, \infty)$. A one-dimensional search, such as Half-interval method can be used to solve equation (27).

Since the convergence of the EM algorithm with unknown υ can be very slow, Ref. [25] proposed a multi-cycle version of ECM, called the MCECM algorithm, to estimate parameters for multivariate t distribution. Moreover, Ref. [18] proposed an ECME algorithm, and the ECME converges substantially faster than EM, ECM or MCECM. The ECME algorithm is as follows:

(Ⅰ) E-step: Calculate the expected complete-data log likelihood given current estimates of parameters $(\beta_t, \gamma_t, \lambda_t, \upsilon_t)$. The E-step of ECME is the same as EM;

(Ⅱ) CM-step1: Fix $\upsilon = \upsilon_t$, and calculate $\theta_{t+1} = (\beta_{t+1}^{\mathrm{T}}, \gamma_{t+1}^{\mathrm{T}}, \lambda_{t+1}^{\mathrm{T}})^{\mathrm{T}}$ using Eq. (22) with υ replaced by υ_t;

(Ⅲ) CM-step2: Given $\theta_{t+1} = (\beta_{t+1}^{\mathrm{T}}, \gamma_{t+1}^{\mathrm{T}}, \lambda_{t+1}^{\mathrm{T}})^{\mathrm{T}}$, and calculate υ_{t+1} to maximise Eq. (28)

$$-\psi(\upsilon/2) + \& \ln(\upsilon/2) + \frac{1}{n}\sum_{i=1}^{n}[\ln\hat{\omega}_i - \hat{\omega}_i] + 1 + \frac{1}{n}\sum_{i=1}^{n}\left\{\psi(\frac{\upsilon + m_i}{2}) - \ln(\frac{\upsilon + m_i}{2})\right\} = 0 \quad (28)$$

where $\hat{\omega}_i = (\upsilon + m_i)/(\upsilon + \| \delta_i(\theta_{t+1}) \|^2)$.

(Ⅳ) Repeat (Ⅰ)~(Ⅲ) until a pre-specified convergence criterion is met.

3 Numerical studies

3.1 Simulations

In this section the finite sample performance of the proposed approach is investigated through simulations. The continuous longitudinal responses y_i are generated from (1) under the following model:

$$\left.\begin{array}{l} \mu_{ij} = \beta_{\{0\}} + x_{ij1}\beta_{\{1\}} + x_{ij2}\beta_2 \\ \ln(\sigma_{ij}) = z_{ij1}\lambda_1 + z_{ij2}\lambda_2 \\ \phi_{\mathrm{ijk}} = \gamma_0 + w_{ijk1}\gamma_1 + w_{\mathrm{ijk2}}\gamma_2 \\ (i = 1, \cdots, n; j = 1, \cdots, m_i) \end{array}\right\} \quad (29)$$

The covariate $(x_{ij1}, x_{ij2})'$ is generated from a multivariate normal distribution with mean 0, marginal variance 1 and correlation 0.5. We take $(z_{ij1}, z_{ij2}) = (x_{ij1}, x_{ij2})$, and $w_{ijk1} = (t_{ij} - t_{ik})$, $w_{\mathrm{ijk2}} = (t_{ij} - t_{ik})^2$. The parameters are set to be $(\beta_0, \beta_1, \beta_2) = (1, -0.5, 0.5)$, $(\gamma_0, \gamma_1, \gamma_2) = (0.3, -0.2, 0.3)$, $(\lambda_1, \lambda_2) = (0.5, -0.3)$. The measurement times t_{ij} is generated uniformity. The degree of freedom, υ, for the multivariate t distribution is 3 as existing studies have shown that the t-distribution with 3 degrees of freedom has sufficiently long tails and almost covers all extreme outliers[14,26]. Finally, we generate 500 data sets and consider sample sizes for $n=50$, 100 and 400 with $m_i = 5$. The proposed approach under t-distribution and maximum likelihood estimation under normal distribution are used to estimate the parameters respectively.

Tab. 1 Simulation results when the data sets are generated under *t*-distribution

n		50			100			400		
Model	Par	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP
t	β	−0.0010	0.0071	93%	0.0005	0.0042	94%	0.0004	0.0019	94%
		−0.0004	0.0027	95%	0.0002	0.0017	97%	0.0001	0.0007	94%
		0.0003	0.0018	95%	−0.0001	0.0011	97%	− 0.0001	0.0005	93%
	γ	−0.0037	0.0092	97%	−0.0016	0.0061	93%	−0.0010	0.0032	94%
		0.0010	0.0160	95%	0.0007	0.0088	95%	0.0011	0.0055	95%
		−0.0020	0.0171	95%	−0.0012	0.0094	93%	− 0.0013	0.0060	94%
	λ	−0.0001	0.0008	96%	0.0001	0.0005	94%	0.0000	0.0002	92%
		0.0001	0.0009	96%	−0.0001	0.0006	92%	0.0000	0.0003	96%
	υ	−0.1438	0.5971	93%	−0.0166	0.5088	93%	0.0418	0.2957	95%
normal	β	0.0016	0.0223	99%	0.0013	0.0061	95%	−0.0015	0.0043	96%
		0.0009	0.0099	99%	0.0005	0.0023	92%	−0.0028	0.0044	97%
		−0.0005	0.0072	99%	−0.0003	0.0014	94%	0.0019	0.0007	95%
	γ	0.0423	0.0389	87%	0.0486	0.0272	87%	0.0412	0.0150	89%
		−0.0395	0.0895	97%	−0.0328	0.0282	81%	−0.0359	0.0256	99%
		0.0530	0.1108	97%	0.0474	0.0312	96%	0.0541	0.0190	97%
	λ	−0.0004	0.0037	98%	0.0002	0.0012	94%	−0.0048	0.0042	99%
		0.0004	0.0043	99%	−0.0002	0.0014	99%	0.0038	0.0017	98%

When the data sets are generated from multivariate t distribution, Tab. 1 reports the accuracy of the estimated parameters by the ECME algorithm in terms of their mean biases (Bias), standard errors (SE) and the coverage percentage for the 95% confidence interval (CP), where model denotes the model distribution and Par denotes the parameters. It clearly indicates that the proposed method works reasonably well for data with potential outliers. All the biases are small especially when n is large under t distribution. Additionally, to evaluate the inference procedure, we report the coverage percentage of 95% confidence interval which is quite close to the nominal level, especially for large n. This demonstrates the validity of Theorem 2.1.

When the data $y_i \sim N_{m_i}(\mu_i, \Omega_i)$ under model (29), we generate 500 data sets with sample sizes $n=50$, 100 and 400. In this case, the multivariate t distribution is misspecified while the normal model is correctly specified. Tab. 2 shows that the proposed approach under t distribution performs almost as well as normal distribution. The estimated degree of freedom $\hat{\upsilon} = 3328.34$, 3497.25, 5911.89 for different sample sizes respectively, indicating the essential normality.

Tab. 2 Simulation results when the data sets are generated under normal distribution

n		50			100			400		
Model	Par	Bias	SE	CP	Bias	SE	CP	Bias	SE	CP
normal	β	−0.0010	0.0062	93%	−0.0001	0.0020	93%	−0.0001	0.0019	94%
		−0.0003	0.0025	92%	−0.0001	0.0015	93%	−0.0001	0.0007	93%
		0.0002	0.0016	93%	0.0001	0.0010	96%	0.0000	0.0004	94%
	γ	−0.0024	0.0067	93%	−0.0020	0.0047	92%	−0.0006	0.0020	95%
		0.0012	0.0141	93%	0.0017	0.0107	94%	0.0004	0.0047	93%
		−0.0021	0.0147	95%	−0.0026	0.0107	96%	−0.0005	0.0043	92%
	λ	−0.0004	0.0008	98%	0.0000	0.0005	98%	0.0000	0.0003	98%
		0.0000	0.0007	94%	0.0000	0.0005	97%	0.0000	0.0003	96%
t	β	−0.0010	0.0062	93%	−0.0001	0.0042	93%	−0.0001	0.0019	94%
		−0.0003	0.0025	92%	−0.0001	0.0016	96%	−0.0001	0.0007	92%
		−0.0002	0.0016	93%	0.0001	0.0010	95%	0.0000	0.0004	93%
	γ	−0.0030	0.0067	92%	−0.0024	0.0050	90%	−0.0007	0.0021	92%
		0.0017	0.0143	94%	0.0020	0.0108	94%	0.0004	0.0048	94%
		−0.0027	0.0149	95%	−0.0031	0.0109	95%	−0.0006	0.0044	95%
	λ	0.0000	0.0008	98%	0.0000	0.0004	98%	0.0000	0.0003	96%
		0.0000	0.0007	94%	0.0000	0.0004	97%	0.0000	0.0003	94%

To study the robustness of proposed method, we consider the following contaminated normal distribution

$$y_i \sim (1-\pi)N_{m_i}(\mu_i, \Omega_i) + \pi\, N_{m_i}(\mu_i, \delta_e \Omega_i),$$

where $0 \leqslant \pi \leqslant 1$ corresponds to the percent of contamination, and $\delta_e > 1$ is a parameter that determines the deviation of the wider component. Since the multivariate t distribution and normal distribution are both misspecified, we compare them via the following relative error measurements

$$\mathrm{err}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} \| \hat{\mu}_i - \mu_i \| / \| \mu_i \|,$$

$$\mathrm{err}(\hat{\Omega}) = \frac{1}{n}\sum_{i=1}^{n} \| \hat{\Omega}_i - \Omega_i \| / \| \Omega_i \|.$$

We generate 500 data sets for different sample sizes $n = 50$, 100 and 400 with $m_i = 5$. Tab. 3 reports these two error measurements for different sample sizes and $\delta_e = 4$, 16 respectively. The corresponding estimates of degree of freedom in all the cases range from 2.79 to 7.54, which indicate a long tail of population distribution. Obviously, the proposed approach is more robust than the maximum likelihood estimation under normal distribution.

Tab. 3 Simulation results when the data sets are generated under contaminated normal distribution with $\pi = 5\%$

Model	n	$\delta_e = 4$			$\delta_e = 16$		
		50	100	400	50	100	400
normal	$\mathrm{err}(\hat{\mu}) \times 10^2$	0.61	0.52	0.12	0.62	0.56	0.12
	$\mathrm{err}(\hat{\Omega}) \times 10^2$	1.34	0.87	0.23	1.43	0.89	0.24
t	$\mathrm{err}(\hat{\mu}) \times 10^2$	0.09	0.05	0.02	0.11	0.08	0.06
	$\mathrm{err}(\hat{\Omega}) \times 10^2$	0.07	0.06	0.02	0.09	0.06	0.05

3.2 Analysis of CD4 cell data

We apply the proposed robust joint modelling approach to an unbalanced longitudinal data set, previously studied by Refs. [6, 8-9, 27]. HIV destroys T-lymphocytes called CD4 cells, which play a vital role in immune function. Disease progression can be assessed by measuring the number or percentage of CD4 cells. which on average decrease throughout the disease incubation period. The CD4 cell count of 369 people infected with human immunodeficiency virus with a total of 2 376 values were collected for this study, covering a period of approximately 8.5 years. The data set is observational and these counts were measured at different times for each individual. The number of measurements for each individual varies from 1 to 12 and the time points are not equally spaced. As in Ref. [26], square roots of CD4 cell counts are used. compared to the models in Ref. [9], we use their optimal polynomials for the mean, logarithm of marginal variance and the angles in the correlation matrix. That is,

$$y_{ij}=\beta_0+t_{ij}\beta_1+t_{ij}^2\beta_2+\cdots+t_{ij}{}^8\beta_8+\epsilon_{ij},$$

$$\ln(\sigma_{ij})=t_{ij}\lambda_1,\ (i=1,\cdots,n;j=1,\cdots,m_i)$$

$$\phi_{ijk}=\gamma_0+(t_{ij}-t_{ik})\gamma_1,$$

where $n=369$, y_{ij} is the square roots of the CD4 cell numbers. To address the potential outliers in the data set, we assume $\epsilon_i\sim t_{m_i}(0,\Omega_i,\upsilon)$.

By the proposed approach in section 2.2, the parameter estimates are $\beta_0=29.181(0.284)$, $\beta_1=-3.908(0.252)$, $\beta_2=-1.184(0.238)$, $\beta_3=0.974(0.134)$, $\beta_4=0.208(0.066)$, $\beta_5=-0.153(0.028)$, $\beta_6=-0.005(0.004)$, $\beta_7=0.009(0.002)$ and $\beta_8=-0.001(0.000)$; $\gamma_0=1.066(0.0161)$ and $\gamma_1=0.062(0.008)$; $\lambda_1=0.046(0.008)$ with standard error being given in the parenthesis. The estimated degree of freedom, $\hat{\upsilon}=9.865(1.446)$, indicates possible violation of normality assumption.

Fig. 1(a) shows the fitted curves of the mean with normal distribution (red line) and t distribution (blue dotted line). They coincide with each other except near the boundary. The curve fitted by normal likelihood decreases slightly faster when time goes by than that by multivariate t likelihood, indicating the non-robustness of normality. And Fig. 1 (b) and (c) report the angle in the correlation matrix and log-variance. It is clear that the log-variance under normal likelihood is over-estimated at the beginning and under-estimated at the end of the study. The estimated angle parameters under two methods coincide with each other, indicating the same correlation structure. Therefore, it is useful to assume approximate normality for the distribution of square root of the CD4 cell numbers to study the relationship in the mean, one should be cautious when studying dynamics in the variance.

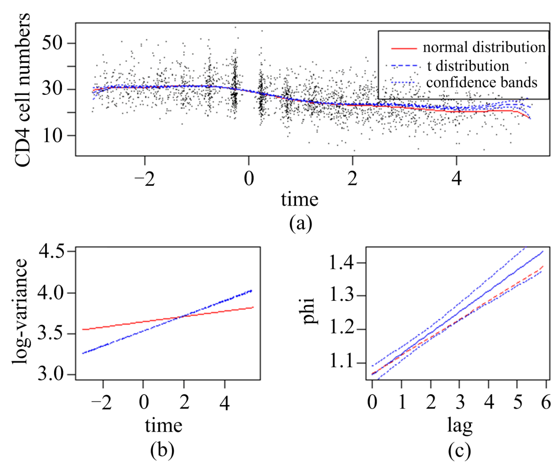

Fig. 1 CD4 cell data: fitted curves of (a) the mean against time, (b) the log-variances against time, (c) the angles against the time lag under normal (red line) and *t* likelihood (blue dotted line) respectively. The dotted lines are asymptotic 95% confidence intervals by *t* likelihood

4 Conclusion

We have proposed a robust parsimoniously joint location-scale modeling approach using t-distribution as an alternative to the classical normality-based approaches in order to provide protection against outliers in the data, and understand the dynamics in the location function, marginal scale function and association. An ECME-based algorithm is applied to speed up the computation associated with the EM algorithm for parameter estimation. Data examples and simulations demonstrate the effectiveness of the proposed approach.

We formulate the problem under the t distribution mainly because of its familiarity and computational simplicity. Other robust distributions can also be used to yield robust estimates, such as the contaminated normal distribution or the exponential power family. Studies comparing these alternative models might be useful, particularly in multivariate settings where previous work appears limited. Another possible extension of the proposed framework would be to longitudinal data with missing response and covariates as well as informative missing.

Acknowledgement

We thank the two referees for their constructive comments and suggestions that have greatly improved the paper.

References

[1] DIGGLE P J. The analysis of longitudinal data[J]. Journal of the American Statistical Association, 2002, 90: 1231-1232.

[2] POURAHMADI M. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation[J]. Biometrika, 1999, 86:677-690.

[3] POURAHMADI M. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix[J]. Biometrika, 2000, 87: 425-435.

[4] POURAHMADI M. Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance-correlation parameters[J]. Biometrika, 2007, 94: 1006-1013.

[5] PAN J, MACKENZIE G. Model selection for joint mean-covariance structures in longitudinal studies[J]. Biometrika, 2003, 90: 239-244.

[6] YE H, PAN J. Modeling covariance structures in generalized estimating equations for longitudinal data [J]. Biometrika, 2006, 93: 927-941.

[7] LENG C, ZHANG W, PAN J. Semiparametric mean-covariance regression analysis for longitudinal data[J]. Journal of the American Statistical Association, 2010, 105: 181-193.

[8] ZHANG W, LENG C. A moving average Cholesky factor model in covariance modeling for longitudinal data[J]. Biometrika, 2012, 99: 141-150.

[9] ZHANG W, LENG C, TANG C Y. A joint modelling approach for longitudinal studies[J]. Journal of the Royal Statistical Society, 2015, 77: 219-238.

[10] CROUX C, GIJBELS I, PROSDOCOMO I. Robust estimation of mean and dispersion functions in extended generalized additive models[J]. Biometrics, 2012, 68: 31-44.

[11] ZHENG X, FUNG W K, ZHU Z. Robust estimation in joint mean-covariance regression model for longitudinal data [J]. Annals of the Institute of Statistical Mathematics, 2013, 65: 617-638.

[12] LV J, GUO C, LI T, et al. Adaptive robust estimation in joint mean-covariance regression model for bivariate longitudinal data[J]. Statistics, 2018, 52 (1):64-83.

[13] LIANG K Y, ZEGER S L. Longitudinal data analysis using generalized linear models[J]. Biometrika, 1986, 73: 13-22.

[14] LANGE K L. Robust statistical modeling using the t distribution [J]. Publications of the American Statistical Association, 1989, 84: 881-896.

[15] LIN T I, WANG Y J. A robust approach to joint modeling of mean and scale covariance for longitudinal data[J]. Journal of Statistical Planning & Inference, 2009, 139(9): 3013-3026.

[16] MAADOOLIAT M, POURAHMADI M, HUANG J Z. Robust estimation of the correlation matrix of longitudinal data[J]. Statistics and Computing, 2013, 23(1): 17-28.

[17] LIN T I, WANG W L. Bayesian inference in joint modeling of location and scale parameters of the t distribution for longitudinal data[J]. Journal of Statistical Planning & Inference, 2011, 141: 1543-1553.

[18] LIU C, RUBIN D B. ML estimation of the t distribution using EM and its extensions, ECM and ECME[J]. Statistic Sinica, 1995, 5(1): 19-39.

[19] LIN P. Some characterizations of the multivariate t distribution [J]. Journal of Multivariate Analysis, 1972, 2: 339-344.

[20] KOTZ S, NADARAJAH S. Multivariate t Distributions and Their Applications[M]. Cambridge, UK: Cambridge University Press, 2004.

[21] CHIU T Y M, LEONARD T, TSUI K W. The matrix-logarithm covariance model [J]. Journal of American Statistical Association, 1996, 91: 198-210.

[22] WHITE H. Maximum likelihood estimation of misspecied models[J]. Econometrica, 1982, 50: 1-25.

(下转第 348 页)