


Alternative modified Cholesky decomposition of the precision matrix of longitudinal data

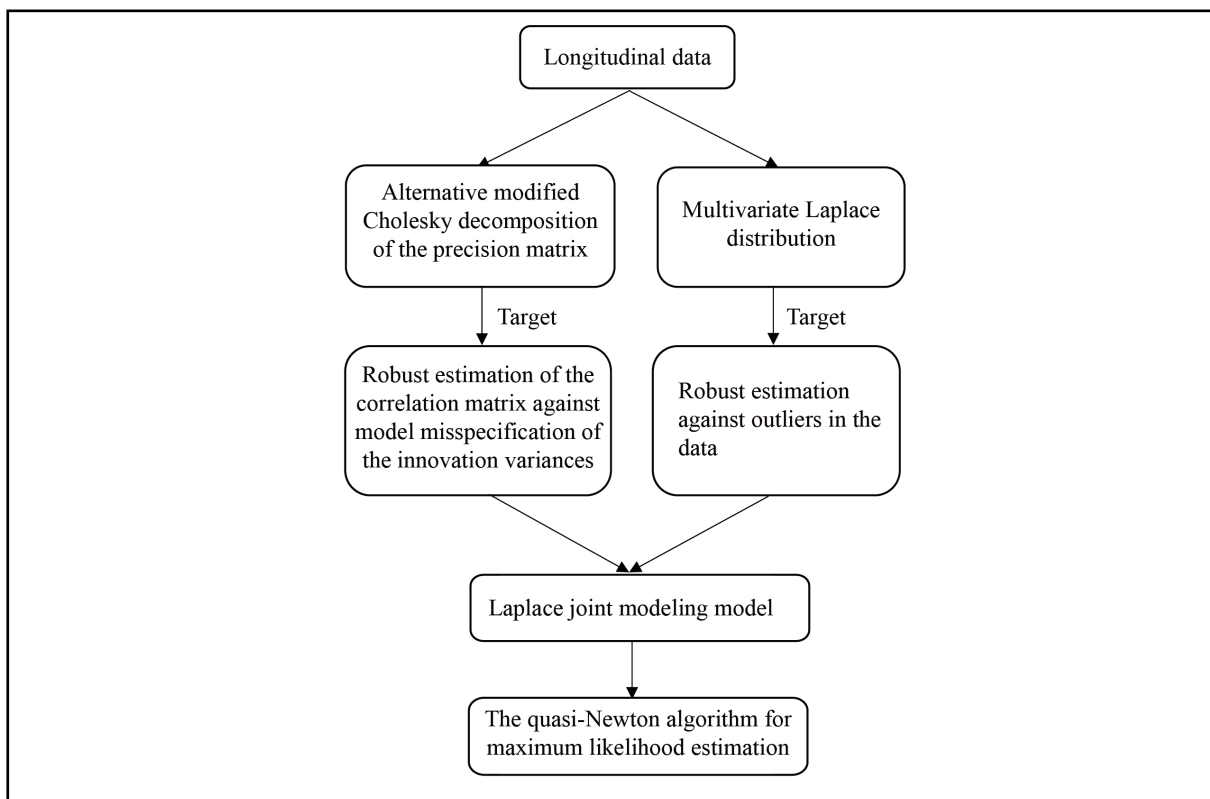
Fei Lu , and Yuting Zeng

College of Science, Zhejiang Sci-Tech University, Hangzhou 310018, China

✉ Correspondence: Fei Lu, E-mail: lufeiby@163.com

© 2024 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Graphical abstract




The framework of the double robust Laplace joint modeling model for longitudinal data.

Public summary

- We propose an alternative modified Cholesky decomposition (AMCD) of the precision matrix of longitudinal data, which results in robust estimation of the correlation matrix against model misspecification of the innovation variances.
- A joint mean-covariance model with multivariate normal distribution and AMCD is established, the quasi-Fisher scoring algorithm is developed, and the maximum likelihood estimators are proved to be consistent and asymptotically normally distributed.
- A double-robust joint modeling approach with multivariate Laplace distribution and AMCD is established, and the quasi-Newton algorithm for maximum likelihood estimation is developed.

Alternative modified Cholesky decomposition of the precision matrix of longitudinal data

Fei Lu , and Yuting Zeng

College of Science, Zhejiang Sci-Tech University, Hangzhou 310018, China

✉ Correspondence: Fei Lu, E-mail: lufeiby@163.com

© 2024 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



Cite This: *JUSTC*, 2024, 54(3): 0306 (16pp)



Read Online

Abstract: The correlation matrix might be of scientific interest for longitudinal data. However, few studies have focused on both robust estimation of the correlation matrix against model misspecification and robustness to outliers in the data, when the precision matrix possesses a typical structure. In this paper, we propose an alternative modified Cholesky decomposition (AMCD) for the precision matrix of longitudinal data, which results in robust estimation of the correlation matrix against model misspecification of the innovation variances. A joint mean-covariance model with multivariate normal distribution and AMCD is established, the quasi-Fisher scoring algorithm is developed, and the maximum likelihood estimators are proven to be consistent and asymptotically normally distributed. Furthermore, a double-robust joint modeling approach with multivariate Laplace distribution and AMCD is established, and the quasi-Newton algorithm for maximum likelihood estimation is developed. The simulation studies and real data analysis demonstrate the effectiveness of the proposed AMCD method.

Keywords: Cholesky decomposition; precision matrix; correlation matrix; multivariate Laplace distribution; robustness

CLC number: O212.4

Document code: A

2020 Mathematics Subject Classification: 62H20

1 Introduction

Longitudinal data arises frequently in the economics, social sciences, epidemiology and biological research. The observations of each subject are measured repeatedly over time and thus are intrinsically correlated^[1]. It is crucial to correctly address the within-subject covariation structure since ignoring this structure may lead to inefficient estimators of the mean parameter. Moreover, the covariation structure itself may be of scientific interest^[2], while covariance and precision matrices suffer from positive definiteness constraints and high-dimensional parameters, and the correlation matrix should satisfy the constraint that its diagonals must be 1's in addition to the aforementioned two issues. Thus, there is a frequently-used strategy to estimate the correlation matrix, i.e., first estimating the covariance or precision matrix and then calculating the corresponding correlation matrix, which motivates us to develop the decomposition method proposed in this paper.

Now it is vital to decide to decompose and model whether the covariance matrix or the precision matrix. In fact, the covariance or precision matrix may possess typical structures, and the matrix inversion operation does not protect the original structure, such as sparsity, regression relationship, etc., then decomposing and modeling the inverse of the covariance matrix or the inverse of the precision matrix may result in efficiency loss. For the precision matrix, Pourahmadi^[3] introduced the modified Cholesky decomposition (MCD),

which leads to an unconstrained parameterization for the precision matrix that automatically guarantees its positive definiteness. The entries in MCD can be interpreted as generalized autoregressive parameters and innovation variances, which can be estimated by fitting a class of unconstrained joint mean-covariance models via maximum likelihood estimation. For the covariance matrix, Zhang and Leng^[4] analyzed within-subject covariation by decomposing the covariance matrix itself rather than its inverse by using the moving average Cholesky decomposition (MACD). The entries in this decomposition are moving average parameters and innovation variances. Based on MACD, the corresponding correlation expression depends on the innovation variances, and thus is not necessarily robust with respect to their model misspecification. In contrast, Chen and Dunson^[5] proposed an alternative Cholesky decomposition (ACD), which directly models the covariance matrix but in a way that the estimates of the correlation matrix do not depend on the quality of modeling and estimating the innovation variances; that is, estimation of the correlation matrix is robust with respect to model misspecification of the innovation variances, the components shared by both MACD and ACD. ACD is closely related to the moving average model of “standardized” measurements on a longitudinal subject^[6,7]. By directly considering the correlation matrix, Zhang et al.^[8] developed the hyperspherical parameterization of the Cholesky factor (HPC), which is very appealing since the resulting parameters are unconstrained on the support and directly interpretable in regard to the correlations.

HPC can always lead to a more parsimonious model, however, at the cost of intensive computations. To the best of our knowledge, when the precision matrix has a typical structure, robust estimation of the correlation matrix against model misspecification of innovation variances has been much less investigated.

Moreover, longitudinal data may suffer from outliers, and several heavy distributed joint modeling approaches have been investigated. Lin and Wang^[9] proposed a joint mean and scale covariance model based on multivariate t distribution and MCD. Maadooliat et al.^[7] further investigated the joint model based on the t distribution and ACD. However, this kind of joint t -regression approach is computationally intensive since joint regression parameters together with the degree of freedom need to be estimated. As an alternative, Guney et al.^[10] adopted a joint model with multivariate Laplace distribution and MCD. The Laplace distribution is heavy-tailed and robust to outliers. It does not have the degree of freedom parameter; thus, its computation is more convenient than that of the t distribution. However, this joint modeling approach could not lead to the aforementioned robust estimation of the correlation matrix.

In this paper, motivated by ACD, we propose an alternative modified Cholesky decomposition (AMCD) of the precision matrix for longitudinal data. Specifically, the inverse of the diagonal matrix of innovation standard deviations is placed outside the two triangular matrices, which alone determines the correlations, and thus results in the aforementioned robust estimation of the correlation matrix. Then, we establish a joint mean-covariance model with multivariate normal distribution and AMCD, and the quasi-Fisher scoring algorithm, and show that the maximum likelihood estimators are consistent and asymptotically normally distributed. Furthermore, we present the double-robust joint modeling approach with multivariate Laplace distribution and AMCD, and the quasi-Newton algorithm for maximum likelihood estimation. We carry out the simulation studies, cattle data analysis, and sleep dose-response data analysis, which indicated that the proposed AMCD method performed well.

The outline of this paper is organized as follows. In Section 2, we review MACD, ACD and MCD, propose AMCD, and discuss their close relationships. In Section 3, we establish the joint normal mean covariance model with AMCD and give the iterative algorithm and asymptotic properties. In Section 4, we investigate the joint model with multivariate Laplace distribution and AMCD, and develop the quasi-Newton algorithm for maximum likelihood estimation. In Section 5, we carry out the simulations. In Section 6, real data analysis is conducted. In Section 7, the paper is summarized. All the calculations and proofs are in the Appendix.

2 Existing and proposed Cholesky-type decompositions of the covariance structure

Assume that y_{ij} is the measurement at the j th time point t_{ij} for the i th subject ($i = 1, \dots, n, j = 1, \dots, m_i$). Let $y_i = (y_{i1}, \dots, y_{im_i})^T$, and $t_i = (t_{i1}, \dots, t_{im_i})^T$. Moreover, suppose

that y_i follows a normal distribution with $E(y_i) = \mu_i = (\mu_{i1}, \dots, \mu_{im_i})^T$ and $\text{var}(y_i) = \Sigma_i$. Then let $y_i = \mu_i + r_i$, where $r_i = (r_{i1}, \dots, r_{im_i})^T$ is the normal random error, with $E(r_i) = 0$ and $\text{var}(r_i) = \Sigma_i$.

In some cases, the covariance matrix Σ_i (or its inverse, i.e., the precision matrix) is of scientific interest and possesses a typical structure. Then, a specific Cholesky-type decomposition can be established to adapt to the corresponding typical structure. Furthermore, the correlation matrix might be of particular interest; therefore, it is worthwhile to develop a robust estimation procedure based on the specific decomposition process, of course, with some suitable modifications. Along this line, several existing Cholesky-type decompositions are discussed, the proposed AMCD is introduced, and their close connections are investigated.

2.1 MACD of the covariance matrix

If the covariance matrix itself possesses a typical structure, it is reasonable to decompose the symmetric and positive definite covariance matrix as $\Sigma_i = C_i C_i^T$ based on the standard Cholesky decomposition. Here C_i is a unique lower triangular matrix with positive diagonal elements. Define $\text{diag}(C_i) = (c_{i11}, \dots, c_{im_i m_i})^T$, and $\Lambda_i = \text{diag}(c_{i11}, \dots, c_{im_i m_i})$. For simplification, let $d_{ij} = c_{ijj}, j = 1, \dots, m_i$; then, $\Lambda_i = \text{diag}(d_{i1}, \dots, d_{im_i})$. In Ref. [4], the standard Cholesky decomposition was transformed into MACD; that is, the matrix Λ_i is located inside two triangular matrices:

$$\Sigma_i = C_i \Lambda_i^{-1} \Lambda_i^2 \Lambda_i^{-1} C_i^T = L_i \Lambda_i^2 L_i^T, \tag{1}$$

where $L_i = C_i \Lambda_i^{-1}$ results in a standardized C_i , namely dividing each column of C_i by its diagonal elements. Let $L_i = (l_{ijk})_{m_i \times m_i}$ and $D_i = \Lambda_i^2 = \text{diag}(d_{i1}^2, \dots, d_{im_i}^2)$. Denote $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{im_i})^T = L_i^{-1} r_i$; then, $\text{var}(\varepsilon_i) = \Lambda_i^2$ and $r_i = L_i \varepsilon_i$. Note that L_i is a lower triangular matrix with diagonals being 1's; then, a moving average representation for the residual r_{ij} can be expressed as follows:

$$r_{ij} = \sum_{t=1}^{j-1} l_{ijt} \varepsilon_{it} + \varepsilon_{ij}, j = 1, \dots, m_i, \tag{2}$$

where $\sum_{t=1}^0 l_{ijt} \varepsilon_{it}$ is denoted as 0. Then, from (2), for any $1 \leq j, k \leq m_i$, with $j \wedge k = \min\{j, k\}$, it follows that

$$\text{cov}(r_{ij}, r_{ik}) = \sum_{t=1}^{j \wedge k} l_{ijt} l_{ikt} d_{it}^2,$$

and thus, the correlation between r_{ij} and r_{ik} is given by

$$\text{corr}(r_{ij}, r_{ik}) = \frac{\sum_{t=1}^{j \wedge k} l_{ijt} l_{ikt} d_{it}^2}{\sqrt{\sum_{t=1}^j l_{ijt}^2 d_{it}^2 \sum_{t=1}^k l_{ikt}^2 d_{it}^2}}, \tag{3}$$

which is determined by the L_i and Λ_i matrices. Then it might not be robust against misspecification of the model for the innovation variances $d_{ij}^2, j = 1, \dots, m_i$.

2.2 ACD of the covariance matrix

In contrast, the standard Cholesky decomposition can be transformed into ACD^[7]; that is, the matrix Λ_i is located outside the two triangular matrices:

$$\Sigma_i = \Lambda_i \Lambda_i^{-1} C_i C_i^T \Lambda_i^{-1} \Lambda_i = \Lambda_i A_i A_i^T \Lambda_i,$$

where $A_i = \Lambda_i^{-1} C_i$ leads to a standardized C_i , namely dividing each row of C_i by its diagonal elements. Denote $A_i = (a_{ijk})_{m_i \times m_i}$. It is obvious that $\Lambda_i^{-1} r_i$ has covariance matrix $A_i A_i^T$, within which Λ_i disappears. Specifically, let $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{im_i})^T = (\Lambda_i A_i)^{-1} r_i$; then, $\text{var}(\varepsilon_i) = I_{m_i}$ and $\Lambda_i^{-1} r_i = A_i \varepsilon_i$. In fact, A_i is also a lower triangular matrix with diagonals being 1's; then, the "standardized" residual r_{ij}/d_{ij} can be represented as

$$\frac{r_{ij}}{d_{ij}} = \sum_{t=1}^{j-1} a_{ijt} \varepsilon_{it} + \varepsilon_{ij}, j = 1, \dots, m_i. \tag{4}$$

From (4), it follows that

$$\text{cov}(r_{ij}, r_{ik}) = d_{ij} d_{ik} \sum_{t=1}^{j \wedge k} a_{ijt} a_{ikt},$$

and thus

$$\text{corr}(r_{ij}, r_{ik}) = \frac{\sum_{t=1}^{j \wedge k} a_{ijt} a_{ikt}}{\sqrt{\sum_{t=1}^j a_{ijt}^2 \sum_{t=1}^k a_{ikt}^2}},$$

which is determined by the A_i matrix alone. Modeling a correlation matrix via ACD is highly appealing since it is robust with respect to misspecification of the model for d_{ij}^2 , $j = 1, \dots, m_i$.

2.3 MCD of the precision matrix

However, the precision matrix might possess a typical structure and should be decomposed appropriately. Based on the standard Cholesky decomposition, $\Sigma_i^{-1} = C_i^{-T} C_i^{-1} = B_i^T B_i$, where $B_i = C_i^{-1}$ is a unique lower triangular matrix with positive diagonals. Similar to (1),

$$\Sigma_i^{-1} = B_i^T \Lambda_i \Lambda_i^{-2} \Lambda_i B_i = F_i^T \Lambda_i^{-2} F_i, \tag{5}$$

where $F_i^T = B_i^T \Lambda_i$ leads to a standardized B_i^T , namely dividing each column of B_i^T by its diagonal elements d_{ij}^{-2} . In fact, (5) can be transformed into $F_i \Sigma_i F_i^T = \Lambda_i^2$, i.e., MCD^[6,9]. Notice that F_i can be expressed as a lower triangular matrix, with the diagonal entries being 1's and the (j, k) th nonzero entries being $-f_{ijk}, k < j$. Then let $\varepsilon_i = F_i r_i$, and $\text{var}(\varepsilon_i) = \Lambda_i^2$, from which the autoregressive representation for the residuals r_{ij} can be expressed as follows:

$$r_{ij} = \sum_{t=1}^{j-1} f_{ijt} r_{it} + \varepsilon_{ij}, j = 1, \dots, m_i.$$

Then, by simple calculation, the correlation between r_{ij} and r_{ik} is determined by the F_i and Λ_i matrices, and thus is not robust against misspecification of the model for innovation variances.

2.4 AMCD of the precision matrix

Conversely, the proposed AMCD causes the matrix Λ_i^{-1} to be outside the two triangular matrices:

$$\Sigma_i^{-1} = \Lambda_i^{-1} \Lambda_i B_i^T B_i \Lambda_i \Lambda_i^{-1} = \Lambda_i^{-1} T_i^T T_i \Lambda_i^{-1}, \tag{6}$$

where $T_i^T = \Lambda_i B_i^T$ results in a standardized B_i^T , namely dividing each row of B_i^T by its diagonals d_{ij}^{-2} . In fact, (6) is equivalent to $T_i \Lambda_i^{-1} \Sigma_i \Lambda_i^{-1} T_i^T = I_{m_i}$. Apparently, the covariance matrix of $\Lambda_i^{-1} r_i$ is $T_i^{-1} T_i^{-T}$, where Λ_i disappears. Let $\varepsilon_i = T_i \Lambda_i^{-1} r_i$; then, $\text{var}(\varepsilon_i) = I_{m_i}$. Note that

$$T_i = \begin{pmatrix} 1 & 0 & \dots & 0 \\ -\phi_{i21} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\phi_{im_1} & -\phi_{im_2} & \dots & 1 \end{pmatrix}.$$

Then, the "standardized" residual r_{ij}/d_{ij} can be modeled by

$$\frac{r_{ij}}{d_{ij}} = \sum_{t=1}^{j-1} \phi_{ijt} \frac{r_{it}}{d_{it}} + \varepsilon_{ij}, j = 1, \dots, m_i, \tag{7}$$

where the generalized autoregressive parameter ϕ_{ijt} 's are unconstrained, the "standardized" factor d_{ij} 's are restricted to be positive, the innovation variances $\text{var}(\varepsilon_{ij}) = 1$, and the innovation covariances $\text{cov}(\varepsilon_{ij}, \varepsilon_{ik}) = 0, j \neq k$, implying their uncorrelation in general, more specifically, their independence when assuming normality. Note that d_{ij}^2 's are the innovation variances in MCD. For ease of comparison, they are also referred to as innovation variances rather than squared "standardized" factors, in AMCD.

Obviously, $\text{corr}(r_{ij}, r_{ik})$ is determined by the T_i matrix alone, and thus is robust against misspecification of the model for $d_{ij}^2, j = 1, \dots, m_i$. This advantage makes the proposed AMCD attractive for modeling the correlation matrix. Moreover, it is more suitable to adopt AMCD than ACD when the precision matrix, rather than the covariance matrix, possesses a typical structure.

3 Normal joint modeling approach

3.1 Joint mean-covariance model

Based on the proposed AMCD in (6), the estimation of Σ_i^{-1} is equivalent to that of T_i and Λ_i . Define $\log(\Lambda_i^2) = \text{diag}\{\log(d_{i1}^2), \dots, \log(d_{im_i}^2)\}$. Following the general approach in Ref. [11], the unconstrained nonredundant entries of $\log(\Lambda_i^2)$ and T_i can be modeled together with μ_i by generalized linear regression models

$$g(\mu_{ij}) = x_{ij}^T \beta, \quad \log(d_{ij}^2) = z_{ij}^T \lambda, \quad \phi_{ijk} = w_{ijk}^T \gamma,$$

where $g(\cdot)$ is assumed to be a monotone and differentiable known link function; x_{ij}, z_{ij} and w_{ijk} are the vectors of covariates; and β, λ and γ are the $p \times 1, s \times 1$ and $q \times 1$ vectors, respectively, of the corresponding associated parameters. The covariates x_{ij}, z_{ij} and w_{ijk} may consist of the baseline covariates, polynomials in time (for x_{ij} and z_{ij}) or time lag (for w_{ijk}), and even their interactions. For instance, when the entries of $\mu_i, \log(\Lambda_i^2)$ and L_i are modeled by polynomials in time and time lag, the covariates may be constructed as

$$x_{ij} = (1, t_{ij}, \dots, t_{ij}^{p-1})^T,$$

$$z_{ij} = (1, t_{ij}, \dots, t_{ij}^{q-1})^T,$$

$$w_{ijk} = (1, (t_{ij} - t_{ik}), \dots, (t_{ij} - t_{ik})^{q-1})^T,$$

assuming that the correlation between r_{ij} and r_{ik} relies only on the time lag between t_{ij} and t_{ik} ($1 \leq k < j \leq m_i$).

3.2 Maximum likelihood estimation

According to the proposed AMCD, inverting (6), it can be obtained that

$$\Sigma_i = \Lambda_i T_i^{-1} T_i^{-T} \Lambda_i. \tag{8}$$

Let $\theta = (\beta^T, \lambda^T, \gamma^T)^T$; then, from (6) and (8), twice the negative log-likelihood function of the multivariate normal distribution, except for a constant, can be written as

$$\begin{aligned} -2l(\theta) &= \sum_{i=1}^n \log |\Sigma_i| + \sum_{i=1}^n r_i^T \Sigma_i^{-1} r_i = \\ &= \sum_{i=1}^n \log |\Lambda_i T_i^{-1} T_i^{-T} \Lambda_i| + \sum_{i=1}^n r_i^T \Lambda_i^{-1} T_i^T T_i \Lambda_i^{-1} r_i, \end{aligned}$$

where $r_i = y_i - \mu_i$. In this case, the joint mean-covariance model can be referred to as the normal joint modeling model (NJMM). By taking partial derivatives of $l(\theta)$ with respect to β , λ and γ , respectively, the maximum likelihood estimating equations for these parameters become

$$\begin{aligned} U_1(\beta; \gamma, \lambda) &= \sum_{i=1}^n \frac{\partial \mu_i^T}{\partial \beta} \Sigma_i^{-1} (y_i - \mu_i) = 0, \\ U_2(\lambda; \beta, \gamma) &= \frac{1}{2} \sum_{i=1}^n Z_i^T (h_i - 1_{m_i}) = 0, \\ U_3(\gamma; \beta, \lambda) &= \sum_{i=1}^n (\epsilon_i^T \otimes I_q) \frac{\partial T_i^{-T}}{\partial \gamma} T_i^T \epsilon_i = 0. \end{aligned} \tag{9}$$

Here $\frac{\partial \mu_i^T}{\partial \beta}$ is the $p \times m_i$ matrix with j th column $\frac{\partial \mu_{ij}}{\partial \beta} = \dot{g}^{-1}(x_{ij}^T \beta) x_{ij}$, $\dot{g}^{-1}(\cdot)$ is the derivative of the inverse function $g^{-1}(\cdot)$, and we denote $\mu(\cdot) = g^{-1}(\cdot)$; $Z_i = (z_{i1}^T, \dots, z_{im_i}^T)^T$, $h_i = \text{diag}(T_i^T T_i \Lambda_i^{-1} r_i r_i^T \Lambda_i^{-1})$, and 1_{m_i} is a $m_i \times 1$ vector of 1's; $\frac{\partial T_i^{-1}}{\partial \gamma_j} = -T_i^{-1} \frac{\partial T_i}{\partial \gamma_j} T_i^{-1}$ with

$$\frac{\partial T_i}{\partial \gamma_j} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ -w_{i21,j} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -w_{im_1,j} & -w_{im_2,j} & \dots & 0 \end{pmatrix},$$

for $j = 1, \dots, q$, and $w_{ijk} = (w_{ijk,1}, \dots, w_{ijk,q})^T$.

Since the solutions satisfy the equations (9), the parameters β , λ and γ can be solved iteratively with the others kept fixed. More specifically, the numerical solutions for these parameters can be calculated by the quasi-Fisher scoring algorithm.

First, the expectations of the Hessian matrices are listed below and discussed explicitly in Appendix A. That is,

$$I(\theta) = -E \left(\frac{\partial l}{\partial \theta \partial \theta^T} \right) = \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) & I_{13}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) & I_{23}(\theta) \\ I_{31}(\theta) & I_{32}(\theta) & I_{33}(\theta) \end{pmatrix},$$

where

$$I_{11}(\theta) = \sum_{i=1}^n \frac{\partial \mu_i^T}{\partial \beta} \Sigma_i^{-1} \frac{\partial \mu_i}{\partial \beta^T},$$

$$I_{22}(\theta) = \frac{1}{4} Z_i^T \{I_{m_i} + (T_i^T T_i) \circ (T_i^{-1} T_i^{-T})\} Z_i,$$

$$I_{33}(\theta) = \sum_{i=1}^n \sum_{j=1}^{m_i} \sum_{k=1}^{j-1} v_{ijk} v_{ijk}^T,$$

$$I_{12}(\theta) = I_{21}^T(\theta) = 0,$$

$$I_{13}(\theta) = I_{31}^T(\theta) = 0,$$

$$I_{32}(\theta) = I_{23}^T(\theta) = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^{m_i} \sum_{k=1}^{j-1} v_{ijk} \phi_{ijk} \left(z_{ik}^T + \sum_{t=k+1}^j a_{itk} z_{it}^T \right)$$

with $A \circ B$ denoting the Hadamard product of matrix A and B , $v_{ijk} = -\sum_{t=k+1}^{j-1} \frac{\partial a_{itk}}{\partial \gamma} \phi_{ij,t}$, and a_{itk} being the (t, k) th element of T_i^{-1} .

Then the iterative algorithm is as follows.

Step 1. Initialize the starting values $\beta^{(0)}$, $\lambda^{(0)}$ and $\gamma^{(0)}$. Set $k = 0$.

Step 2. Compute Σ_i with given $\lambda^{(k)}$ and $\gamma^{(k)}$. Update β via

$$\beta^{(k+1)} = \beta^{(k)} + I_{11}^{-1}(\theta) U_1(\beta; \lambda, \gamma) \Big|_{\beta=\beta^{(k)}}. \tag{10}$$

Step 3. Given $\beta = \beta^{(k+1)}$, update λ and γ as

$$\begin{pmatrix} \lambda^{(k+1)} \\ \gamma^{(k+1)} \end{pmatrix} = \begin{pmatrix} \lambda^{(k)} \\ \gamma^{(k)} \end{pmatrix} + \left\{ \begin{pmatrix} I_{22}(\theta) & I_{23}(\theta) \\ I_{32}(\theta) & I_{33}(\theta) \end{pmatrix}^{-1} \begin{pmatrix} U_2(\lambda; \beta, \gamma) \\ U_3(\gamma; \beta, \lambda) \end{pmatrix} \right\} \Big|_{\lambda=\lambda^{(k)}, \gamma=\gamma^{(k)}}. \tag{11}$$

Step 4. Set $k \leftarrow k + 1$. Repeat steps 2 and 3 until a prespecified convergence criterion is satisfied.

Note that λ and γ are updated together, which is caused by their asymptotic dependence, as seen from Theorem 1 in Section 3.3. It can only be ensured that this algorithm converges to a local optimum which relies critically on the starting values. It is natural to choose the starting value of β as the least square estimator in equation (10) with Σ_i 's set as identity matrices. Then, the starting value of γ can be determined assuming $T_i = I_{m_i}$, and the starting value of λ can be given as the least square estimator based on the residuals. The starting estimates of β and λ are clearly \sqrt{n} -consistent. Moreover, the negative log-likelihood function is asymptotically convex around a small neighbourhood of true parameters, according to the theoretical results in Theorem 1 in Section 3.3 and the proofs in Appendix B. Then the final estimates via the iterative algorithm are guaranteed to be the global optima and more efficient than the starting values in terms of the asymptotical viewpoint. For simulation studies and real data analysis, con-

vergence was usually achieved within several iterations.

3.3 Asymptotic properties

In this section, we establish the strong consistency and asymptotic normality of the maximum likelihood estimators. The following regularity conditions are imposed for the theoretical analysis.

(C1) The dimensions p , s and q of covariates x_{ij} , z_{ij} and ω_{ijk} are fixed, $n \rightarrow \infty$, and $\max m_i$ is bounded.

(C2) The parameter space Θ of θ is a compact subset of R^{p+s+q} , and the true parameter $\theta_0 = (\beta_0^T, \lambda_0^T, \gamma_0^T)^T$ lies in the interior of Θ .

(C3) When $n \rightarrow \infty$, $I(\theta_0)/n$ converges to a positive definite matrix $I(\theta_0)$.

Condition 1 is standard for practical longitudinal data analysis. Condition 2 is natural in the theoretical study of the maximum likelihood estimation. Condition 3 is conventional in regression analysis for modeling unbalanced longitudinal data.

Theorem 1. If $n \rightarrow \infty$ and regularity conditions (C1)-(C3) hold, then (a) the maximum likelihood estimator $\widehat{\theta} = (\widehat{\beta}^T, \widehat{\lambda}^T, \widehat{\gamma}^T)^T$ is strongly consistent for the true value $\theta_0 = (\beta_0^T, \lambda_0^T, \gamma_0^T)^T$, and (b) $\widehat{\theta} = (\widehat{\beta}^T, \widehat{\lambda}^T, \widehat{\gamma}^T)^T$ is asymptotically normal, that is, $\sqrt{n}(\widehat{\theta} - \theta_0) \rightarrow N\{0, I^{-1}(\theta_0)\}$ in distribution.

It can be easily seen that the Fisher information matrix $I(\theta_0)$ is a block matrix, more specifically,

$$I(\theta_0) = \begin{pmatrix} I_{11}(\theta_0) & 0 & 0 \\ 0 & I_{22}(\theta_0) & I_{23}(\theta_0) \\ 0 & I_{32}(\theta_0) & I_{33}(\theta_0) \end{pmatrix},$$

from which $\widehat{\beta}$ is asymptotically independent of $\widehat{\lambda}$ and $\widehat{\gamma}$, whereas $\widehat{\lambda}$ and $\widehat{\gamma}$ are not asymptotically independent. Since $\widehat{\theta} = (\widehat{\beta}^T, \widehat{\lambda}^T, \widehat{\gamma}^T)^T$ is a consistent estimator for θ_0 , the asymptotic covariance matrix I^{-1} can be consistently estimated by the inverse of a matrix with block components

$$\widehat{I}_{ij} = \frac{I_{ij}(\widehat{\theta})}{n}, i, j = 1, 2, 3.$$

4 Laplace joint modeling approach

4.1 Laplace joint modeling model

In this section, we investigate the joint modeling approach based on the multivariate Laplace distribution, which is useful when the response variable has heavier tails. However, there are several kinds of multivariate Laplace distributions, which can be respectively regarded as a particular multivariate Linnik distribution^[12], a special case of the multivariate power exponential (PE) distribution^[13,14], a Gaussian scale mixture^[15], and so on^[16]. Among them, we adopt the special case of the multivariate PE distribution.

If y_i follows a m_i -dimensional PE distribution, $i = 1, \dots, n$, then its density function is

$$f(y_i; \mu_i, \Sigma_i, \nu) = \kappa |\Sigma_i|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} [(y_i - \mu_i)^T \Sigma_i^{-1} (y_i - \mu_i)]^\nu \right\}, \quad (12)$$

where

$$\kappa = \frac{m_i \Gamma\left(\frac{m_i}{2}\right)}{2^{1+\frac{m_i}{2\nu}} \pi^{\frac{m_i}{2}} \Gamma\left(1 + \frac{m_i}{2\nu}\right)},$$

with the location vector $\mu_i \in R$, the positive definite dispersion matrix Σ_i and ν . When $\nu = 1$, (12) corresponds to a multivariate normal distribution. When $\nu = 0.5$, (12) is the density function of a multivariate Laplace distribution, i.e.,

$$f(y_i; \mu_i, \Sigma_i, \nu) = \frac{\Gamma\left(\frac{m_i}{2}\right)}{2^{1+m_i} \pi^{\frac{m_i}{2}} \Gamma(m_i)} \times |\Sigma_i|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} [(y_i - \mu_i)^T \Sigma_i^{-1} (y_i - \mu_i)]^{\frac{1}{2}} \right\}, \quad (13)$$

then $E(y_i) = \mu_i$ and $\text{var}(y_i) = 4(m_i + 1)\Sigma_i$. Based on the multivariate Laplace distribution (13), the proposed AMCD of Σ_i^{-1} (6) and the joint model in Section 3.1, we can obtain the Laplace joint modeling model (LJMM).

4.2 The quasi-Newton algorithm for maximum likelihood estimation

Twice the negative log-likelihood function of the multivariate Laplace distribution, except for a constant, can be written as

$$-2L_L(\theta) = \sum_{i=1}^n \log |\Lambda_i T_i^{-1} T_i^{-T} \Lambda_i| + \sum_{i=1}^n (r_i^T \Lambda_i^{-1} T_i^T T_i \Lambda_i^{-1} r_i)^{\frac{1}{2}}. \quad (14)$$

Let $\Delta_i = r_i^T \Lambda_i^{-1} T_i^T T_i \Lambda_i^{-1} r_i$. By taking partial derivatives of $L_L(\theta)$ with respect to β , λ and γ , respectively, the maximum likelihood estimating equations are

$$\begin{aligned} S_1(\beta; \gamma, \lambda) &= \frac{1}{2} \sum_{i=1}^n \Delta_i^{-\frac{1}{2}} \frac{\partial \mu_i^T}{\partial \beta} \Sigma_i^{-1} (y_i - \mu_i) = 0, \\ S_2(\lambda; \beta, \gamma) &= \frac{1}{4} \sum_{i=1}^n \Delta_i^{-\frac{1}{2}} Z_i^T h_i - \frac{1}{2} \sum_{i=1}^n Z_i^T 1_{m_i} = 0, \\ S_3(\gamma; \beta, \lambda) &= \frac{1}{2} \sum_{i=1}^n \Delta_i^{-\frac{1}{2}} (\epsilon_i^T \otimes I_q) \frac{\partial T_i^{-T}}{\partial \gamma} T_i^T \epsilon_i = 0. \end{aligned} \quad (15)$$

where $\Delta_i^{-\frac{1}{2}}$ can be regarded as a weight, providing robustness against outliers in the data.

We then estimate θ by minimizing expression (14) via the quasi-Newton algorithm. More specifically, the parameters in θ can be split into $\theta_1 = \beta$ and $\theta_{23} = (\lambda^T, \gamma^T)^T$, and solved sequentially with the other parameters held fixed. The detailed algorithm is as follows.

Step 1. Initialize the parameters $\beta^{(0)}$, $\lambda^{(0)}$ and $\gamma^{(0)}$ as in Section 3.2, the inverse Hessian matrix $H_1^{(0)}$ for θ_1 and $H_{23}^{(0)}$ for θ_{23} as identity matrix. Set $k = 0$.

Step 2. For $\theta_1 = \beta$, compute the score function

$$S_1^{(k)} = S_1(\theta_1^{(k)}; \theta_{23}^{(k)}) = S_1(\beta^{(k)}; \gamma^{(k)}, \lambda^{(k)}).$$

Step 3. Update the search direction

$$p_1^{(k)} = -H_1^{(k)} S_1^{(k)},$$

and compute the step size by an approximate line minimization

$$\alpha_1^{(k)} = \arg \min_{0 < \alpha_1 \leq 1} \left\{ -2L_1(\theta_1^{(k)} + \alpha_1 p_1^{(k)}, \theta_{23}^{(k)}) \right\}.$$

Step 4. Update θ_1 and the score function as

$$\theta_1^{(k+1)} = \theta_1^{(k)} + \alpha_1^{(k)} p_1^{(k)}, S_1^{(k+1)} = S_1(\theta_1^{(k+1)}; \theta_{23}^{(k)}) = S_1(\theta_1^{(k+1)}; \gamma^{(k)}, \lambda^{(k)}).$$

Step 5. Update the inverse Hessian matrix via the BFGS formula

$$H_1^{(k+1)} = H_1^{(k)} + \frac{(\theta_1^{(k+1)} - \theta_1^{(k)})(\theta_1^{(k+1)} - \theta_1^{(k)})^T}{(\theta_1^{(k+1)} - \theta_1^{(k)})^T (S_1^{(k+1)} - S_1^{(k)})} - \frac{\{H_1^{(k)}(S_1^{(k+1)} - S_1^{(k)})\} \{H_1^{(k)}(S_1^{(k+1)} - S_1^{(k)})\}^T}{(S_1^{(k+1)} - S_1^{(k)})^T H_1^{(k)} (S_1^{(k+1)} - S_1^{(k)})} + (S_1^{(k+1)} - S_1^{(k)})^T H_1^{(k)} (S_1^{(k+1)} - S_1^{(k)}) u u^T,$$

where

$$u = \frac{\theta_1^{(k+1)} - \theta_1^{(k)}}{(\theta_1^{(k+1)} - \theta_1^{(k)})^T (S_1^{(k+1)} - S_1^{(k)})} - \frac{H_1^{(k)}(S_1^{(k+1)} - S_1^{(k)})}{(S_1^{(k+1)} - S_1^{(k)})^T H_1^{(k)} (S_1^{(k+1)} - S_1^{(k)})}.$$

Step 6. For $\theta_{23} = (\lambda^T, \gamma^T)^T$, compute the score function

$$S_{23}^{(k)} = S_{23}(\theta_{23}^{(k)}; \theta_1^{(k+1)}) = (S_2(\lambda^{(k)}; \beta^{(k+1)}, \gamma^{(k)})^T, S_3(\gamma^{(k)}; \beta^{(k+1)}, \lambda^{(k)})^T)^T.$$

Step 7. Update the search direction

$$p_{23}^{(k)} = -H_{23}^{(k)} S_{23}^{(k)},$$

and compute the step size by an approximate line minimization:

$$\alpha_{23}^{(k)} = \arg \min_{0 < \alpha_{23} \leq 1} \left\{ -2L_1(\theta_1^{(k+1)}, \theta_{23}^{(k)} + \alpha_{23} p_{23}^{(k)}) \right\}.$$

Step 8. Update θ_{23} and the score function as

$$\theta_{23}^{(k+1)} = \theta_{23}^{(k)} + \alpha_{23}^{(k)} p_{23}^{(k)},$$

$$S_{23}^{(k+1)} = S_{23}(\theta_{23}^{(k+1)}; \theta_1^{(k+1)}) = (S_2(\lambda^{(k+1)}; \beta^{(k+1)}, \gamma^{(k+1)})^T, S_3(\gamma^{(k+1)}; \beta^{(k+1)}, \lambda^{(k+1)})^T)^T.$$

Step 9. Update the inverse Hessian matrix via the BFGS formula as in Step 5, with the subscript “1” replaced by “23”.

Step 10. Set $k = k + 1$ and repeat Steps 2 to 9 until a pre-specified convergence criterion is met.

Note that the BFGS algorithm is implemented in this section since it is demonstrated to be one of the best quasi-Newton algorithms for solving unconstrained smooth optimization problems and performs very well here. Alternative quasi-Newton algorithms are also deserved to be investigated in the future. For more details, see Refs. [17, 18].

5 Simulation studies

In this section, we investigate the finite-sample performance of the proposed AMCD, compare the robustness of AMCD,

MCD and ACD for modeling correlation matrices based on various typical covariation structures, and compare the modeling capacities of NJMM and LJMM with AMCD based on different distributions.

Study 1. The 1000 replications of the datasets were generated from the following NJMM:

$$y_{ij} = \beta_0 + x_{ij1}\beta_1 + x_{ij2}\beta_2 + x_{ij3}\beta_3 + r_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, m_i,$$

where r_{ij} follows the proposed AMCD, i.e., expression (7), with

$$\log(d_{ij}^2) = \lambda_0 + z_{ij1}\lambda_1 + z_{ij2}\lambda_2 + z_{ij3}\lambda_3,$$

$$\phi_{ijk} = \gamma_0 + w_{ijk1}\gamma_1 + w_{ijk2}\gamma_2.$$

Here, $n = 50, 100$, or 200 , $m_i - 1 \sim \text{binomial}(6, 0.8)$, and t_{ij} 's were generated from the uniform distribution in the unit interval. The covariates $x_{ij} = (x_{ij1}, x_{ij2}, x_{ij3})^T$ were generated from the multivariate normal distribution with mean zero, marginal variance 1 and correlation 0.5. In addition, we take $z_{ij} = x_{ij}$ and $w_{ijk} = \{1, t_{ij} - t_{ik}, (t_{ij} - t_{ik})^2\}$. In Table 1, “Bias” denotes the average bias of the parameter estimates, and “SD” represents the sample standard deviation of the estimates for the parameters, which can be regarded as the true standard deviation of the resulting estimates. “SE” denotes the sample average of the estimated standard errors by the formula, and “Std” represents the standard deviation of these standard errors. Table 1 indicates that the proposed AMCD approach results in unbiased parameter estimates, and the standard error formula works well, especially when n is large.

Study 2. We compare the performances of the proposed AMCD with those of MCD and ACD based on different datasets generated from various types of covariation structures in terms of the estimation robustness of the correlation matrices against model misspecification of the innovation variances. To assess the estimation accuracy of the correlation matrices, we define the entropy loss function $\Delta_1(R_i, \widehat{R}_i) = \text{trace}(R_i^{-1} \widehat{R}_i) - \log |R_i^{-1}| - m_i$ and the quadratic loss function $\Delta_2(R_i, \widehat{R}_i) = \text{trace}(R_i^{-1} \widehat{R}_i - I_{m_i})^2$, where R_i is the true correlation matrix and \widehat{R}_i is an estimated correlation matrix. Both of the loss functions are 0 when $R_i = \widehat{R}_i$ and positive when $R_i \neq \widehat{R}_i$. It is obvious that a smaller loss implies better estimates of R_i . Furthermore, the loss functions for all subjects are defined by $L_i = \frac{1}{n} \sum_{i=1}^n \Delta_i(R_i, \widehat{R}_i), i = 1, 2$.

In the following, we generated 1000 replications of the datasets with $n = 100$ from the joint mean-covariance model similar to that in Study 1, except that r_{ij} follows respectively the proposed AMCD, MCD and ACD. Then the estimates and losses of the correlation matrices were computed respectively based on the true model $\log(d_{ij}^2) = \lambda_0 + z_{ij1}\lambda_1 + z_{ij2}\lambda_2 + z_{ij3}\lambda_3$ and the misspecified model $\log(d_{ij}^2) = \lambda_0 + z_{ij1}\lambda_1$ for the innovation variances. For comparison, the losses obtained from fitting the true and misspecified models for the innovation variances are exhibited in Tables 2 and 3, respectively, where ENL and QUL represent respectively the average of the en-

Table 1. Simulation results for Study 1. All the results are multiplied by a factor 10^2 .

	True	$n = 50$			$n = 100$			$n = 200$		
		Bias	SD	SE (Std)	Bias	SD	SE (Std)	Bias	SD	SE (Std)
β_0	1	-0.11	2.00	1.88 (0.19)	0.02	1.38	1.35 (0.09)	-0.01	0.96	0.96 (0.04)
β_1	-0.6	-0.17	3.22	3.07 (0.38)	0.02	2.37	2.19 (0.19)	-0.02	1.53	1.54 (0.10)
β_2	0.6	0.08	3.42	3.12 (0.38)	0.01	2.32	2.22 (0.18)	0.04	1.55	1.57 (0.09)
β_3	0.4	0.09	3.44	3.13 (0.35)	-0.03	2.43	2.25 (0.18)	0.06	1.53	1.58 (0.09)
λ_0	-0.8	-3.88	8.64	7.73 (0.22)	-2.21	6.42	5.50 (0.09)	-1.07	4.22	3.89 (0.05)
λ_1	0.8	0.50	8.91	7.98 (0.51)	0.27	6.06	5.71 (0.26)	0.41	4.23	4.05 (0.13)
λ_2	-0.5	-0.59	8.66	7.99 (0.52)	-0.10	5.85	5.71 (0.26)	-0.09	4.22	4.04 (0.12)
λ_3	0.25	0.15	8.81	8.00 (0.53)	-0.10	6.27	5.72 (0.26)	-0.14	4.27	4.05 (0.12)
γ_0	-0.5	-2.55	9.41	8.96 (0.32)	-0.83	6.63	6.35 (0.16)	-0.50	4.63	4.48 (0.08)
γ_1	0.3	2.61	58.08	52.66 (2.37)	0.32	39.65	37.27 (1.18)	-0.72	26.29	26.27 (0.56)
γ_2	-0.5	-6.35	72.12	64.96 (3.87)	-1.71	49.57	45.92 (1.90)	-0.22	32.53	32.30 (0.93)

Table 2. Simulation results for Study 2. Fitting the true model for $\log(d_{ij}^2)$.

True structure	AMCD		MCD		ACD	
	ENL	QUL	ENL	QUL	ENL	QUL
AMCD	0.0342 (0.0287)	0.1239 (0.1415)	0.3999 (0.0949)	1.6446 (0.7985)	0.1611 (0.0355)	0.4953 (0.3158)
MCD	0.5481 (0.0868)	2.7191 (1.3806)	0.0432 (0.0307)	0.1413 (0.1545)	0.7785 (0.1191)	4.9234 (2.2544)
ACD	0.0807 (0.0300)	0.1274 (0.1133)	0.2208 (0.0444)	0.4502 (0.2350)	0.0359 (0.0299)	0.0701 (0.0852)

Table 3. Simulation results for Study 2. Fitting $\lambda_0 + z_{ij1}\lambda_1$ for $\log(d_{ij}^2)$.

True structure	AMCD		MCD		ACD	
	ENL	QUL	ENL	QUL	ENL	QUL
AMCD	0.0364 (0.0308)	0.1348 (0.1584)	0.4551 (0.0974)	1.9058 (0.8514)	0.1638 (0.0375)	0.5323 (0.3441)
MCD	0.5551 (0.0908)	2.8642 (1.4454)	0.1051 (0.0350)	0.3242 (0.2610)	0.7857 (0.1215)	5.1367 (2.3220)
ACD	0.0819 (0.0310)	0.1362 (0.1220)	0.2411 (0.0448)	0.4912 (0.2397)	0.0364 (0.0301)	0.0728 (0.0898)

tropy loss L_1 and quadratic loss L_2 , with empirical standard errors in parentheses.

In terms of the average losses in Table 2, it is vital to know the true covariation structure; that is, when the true covariation structure follows AMCD, the losses of estimating the correlation matrices increase when the covariation structure is incorrectly decomposed based on MCD or ACD, and vice versa. By comparing Table 2 with Table 3, we can analyze the influence of model misspecification for the innovation variances on estimating the correlation matrices. Specifically, by comparing the left two AMCD columns of Table 2 with those of Table 3, ENL and QUL both vary little, no matter whether the true covariation structure is AMCD, MCD or ACD; the losses in the right two ACD columns of Tables 2 and 3 also exhibit a similar pattern; however, the losses in the middle two MCD columns of Tables 2 and 3 vary significantly.

Thus, the estimation of the correlation matrices is robust with respect to the model misspecification for the innovation variances when fitting the AMCD or ACD structure, but is nevertheless not robust when fitting the MCD structure.

Study 3. We generated 1000 replications of the datasets with $n = 100$ respectively from several distributions, where the joint model was the same as that in Study 1. Specifically, y_i follows respectively the multivariate normal distribution (Scenario 1), Laplace distribution (Scenario 2) and PE distribution with $\nu = 0.7$ (Scenario 3), and we fitted NJMM and LJMM under each scenario to compare their modeling capacities.

To assess the estimation accuracy of the parameters, Bias, SD and mean squared errors (MSE) were calculated, and the simulation results are displayed in Tables 4–6. From Table 4, NJMM yields smaller MSE than does LJMM for all param-

Table 4. Simulation results for Study 3. Fitting NJMM and LJMM based on the data generated from the multivariate normal distribution-Scenario 1.

	True	NJMM			LJMM		
		Bias	SD	MSE	Bias	SD	MSE
β_0	1	0.000293	0.013820	0.000191	0.000389	0.014523	0.000211
β_1	-0.6	0.000956	0.023084	0.000534	0.001186	0.023680	0.000562
β_2	0.6	-0.001044	0.022847	0.000523	-0.001132	0.023518	0.000554
β_3	0.4	-0.000942	0.024125	0.000583	-0.001063	0.025185	0.000635
λ_0	-0.8	-0.018044	0.057580	0.003641	-3.284723	0.059723	10.792970
λ_1	0.8	0.002058	0.061634	0.003803	0.002269	0.064317	0.004142
λ_2	-0.5	-0.001617	0.060490	0.003662	-0.002075	0.064486	0.004163
λ_3	0.25	0.002351	0.057719	0.003337	0.002540	0.061003	0.003728
γ_0	-0.5	-0.010487	0.063410	0.004131	-0.010345	0.064888	0.004318
γ_1	0.3	-0.000795	0.379933	0.144350	0.000034	0.387449	0.150117
γ_2	-0.5	-0.015246	0.473920	0.224833	-0.018321	0.486507	0.237025

Table 5. Simulation results for Study 3. Fitting NJMM and LJMM based on the data generated from the multivariate Laplace distribution-Scenario 2.

	True	NJMM			LJMM		
		Bias	SD	MSE	Bias	SD	MSE
β_0	1	0.000443	0.074431	0.005540	-0.001020	0.070935	0.005033
β_1	-0.6	0.001874	0.125683	0.015800	-0.000377	0.115993	0.013454
β_2	0.6	-0.006187	0.120606	0.014584	-0.004163	0.112968	0.012779
β_3	0.4	0.003073	0.124357	0.015474	0.004704	0.116733	0.013649
λ_0	-0.8	3.325543	0.090872	11.067496	-0.019765	0.085474	0.007696
λ_1	0.8	0.000266	0.068353	0.004672	0.000632	0.063271	0.004004
λ_2	-0.5	-0.000353	0.068978	0.004758	-0.000868	0.065576	0.004301
λ_3	0.25	0.003889	0.067912	0.004627	0.003261	0.063900	0.004094
γ_0	-0.5	-0.009415	0.072748	0.005381	-0.007799	0.069382	0.004875
γ_1	0.3	-0.012116	0.428129	0.183441	-0.016835	0.406756	0.165733
γ_2	-0.5	0.003418	0.527299	0.278056	0.011848	0.507051	0.257241

Table 6. Simulation results for Study 3. Fitting NJMM and LJMM based on the data generated from the multivariate PE distribution with $\nu=0.7$ -Scenario 3.

	True	NJMM			LJMM		
		Bias	SD	MSE	Bias	SD	MSE
β_0	1	0.000279	0.031374	0.000984	-0.000075	0.031226	0.000975
β_1	-0.6	-0.000363	0.053292	0.002840	-0.001130	0.052419	0.002749
β_2	0.6	0.000324	0.053262	0.002837	0.001559	0.053214	0.002834
β_3	0.4	-0.000064	0.054821	0.003005	-0.000298	0.053428	0.002855
λ_0	-0.8	1.697678	0.076425	2.887951	-1.610802	0.075379	2.600366
λ_1	0.8	0.005810	0.065800	0.004363	0.006267	0.063930	0.004126
λ_2	-0.5	-0.001172	0.067431	0.004548	-0.001043	0.066281	0.004394
λ_3	0.25	0.000540	0.064320	0.004137	0.000625	0.063059	0.003977
γ_0	-0.5	-0.007865	0.068495	0.004753	-0.007884	0.068537	0.004759
γ_1	0.3	-0.000655	0.422247	0.178293	0.000933	0.421153	0.177371
γ_2	-0.5	-0.019594	0.521888	0.272751	-0.022436	0.515721	0.266472

ers, when the data follows the multivariate normal distribution (Scenario 1). Moreover, Bias of NJMM and LJMM are generally small, except that Bias for Λ_0 when fitting LJMM are rather large under Scenario 1, which might be due to the influence of distribution misspecification and the difficulty of estimating the constant term in innovation variance models. From Table 5, we can carry out a similar discussion for Scenario 2, and find out that LJMM outperforms NJMM when the data follows the multivariate Laplace distribution. Table 6 shows that LJMM has slightly smaller MSE than NJMM, when the data follows the multivariate PE distribution with $\nu = 0.7$ (Scenario 3), which is between the multivariate normal distribution and the multivariate Laplace distribution.

6 Real data analysis

6.1 The cattle data

We applied the proposed AMCD approach to the balanced cattle data, which was initially introduced in Ref. [19]. This dataset consists of two treatment groups, A and B, and the weights of each cattle were measured 11 times over a period of 133 days. Notably, the 30 animals in group A were analyzed in Refs. [11, 20] and [8]. Pan and Mackenzie^[20] found out that it is reasonable to adopt three polynomials for modeling jointly the mean, the log-innovation variances and the autoregressive coefficients based on MCD, and the optimal triplet of the polynomial orders is determined as (8,3,4) in terms of the Bayesian information criterion (BIC).

From expression (8) based on AMCD, we can calculate the corresponding innovation variances d_{ij}^2 and generalized autoregressive parameters ϕ_{ijk} ($j > k$) of the sample covariance matrix. Fig. 1 displays $\log(d_{ij}^2)$ versus time and ϕ_{ijk} versus time lag, both indicate trends which might be properly characterized by polynomials. Then we adopt three polynomials

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 t_{ij} + \dots + \beta_p t_{ij}^p, \\ \log(d_{ij}^2) &= \lambda_0 + \lambda_1 t_{ij} + \dots + \lambda_s t_{ij}^s, \\ \phi_{ijk} &= \gamma_0 + \gamma_1 (t_{ij} - t_{ik}) + \dots + \gamma_q (t_{ij} - t_{ik})^q \end{aligned} \quad (16)$$

for modelling jointly the mean and the covariation structure based on the proposed AMCD, and select the optimal model in terms of

$$\text{BIC}(p, s, q) = -\frac{2}{n} \widehat{l}_{\max} + (p + s + q + 3) \frac{\log(n)}{n}, \quad (17)$$

where p , s and q are the three polynomial orders, and \widehat{l}_{\max} is the corresponding maximum log-likelihood. Then some model with the smallest BIC is generally judged to be the optimal model, which can capture the dynamics much more precisely and parsimoniously. Due to the asymptotic orthogonality of the mean parameter and the covariation parameters, these two kinds of parameters can be searched separately, which is more computationally efficient. Note that p within the optimal model in Refs. [20] and [8] are both 8, which motivates us to first fix $p = 8$ and search for s and q . The optimal (s, q) is determined to be (3,3) by comparing several candidate models, and the major searching process is displayed in Table 7. Then, we fix $(s, q) = (3, 3)$, search for p , and determine the optimal order for AMCD as $(p, s, q) = (8, 3, 3)$. Fig. 1 shows the optimal AMCD-based fitted curves for $\log(d_{ij}^2)$'s and ϕ_{ijk} 's and their corresponding 95% pointwise confidence intervals obtained using the bootstrapping method; these curves capture the pattern well for the log-innovation variances and rather well for the generalized autoregressive parameters.

The optimal order based on AMCD is slightly less than that based on MCD, but identical to that based on ACD, which is also determined by using a similar searching process. To compare the performances of the aforementioned three decompositions, we regard the sample correlation matrix as the true matrix, and still assess the estimation accuracy of the cor-

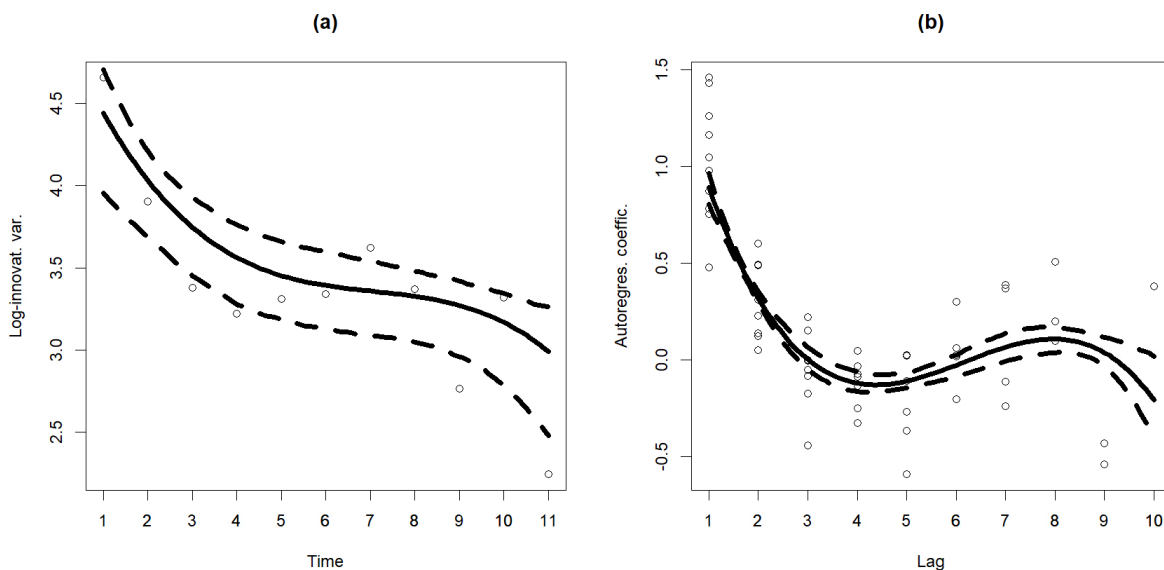


Fig. 1. Cattle data: sample regressograms and fitted curves for (a) log-innovation variances and (b) autoregressive coefficients (solid lines, curves fitted by the proposed AMCD method; dashed lines, 95% pointwise confidence intervals using the bootstrapping method).

Table 7. Cattle data: comparison of various models based on the proposed AMCD approach.

Poly(8,s,q)	Number of parameters	\hat{t}_{\max}	BIC
(8,1,1)	13	-808.07	55.35
(8,1,2)	14	-764.56	52.56
(8,1,3)	15	-750.13	51.71
(8,1,4)	16	-749.86	51.80
(8,2,1)	14	-806.79	55.37
(8,2,2)	15	-763.07	52.57
(8,2,3)	16	-749.16	51.76
(8,2,4)	17	-749.00	51.86
(8,3,1)	15	-806.52	55.47
(8,3,2)	16	-760.96	52.54
(8,3,3)†	17	-746.07	51.67
(8,3,4)	18	-745.82	51.76
(8,4,1)	16	-806.45	55.58
(8,4,2)	17	-759.88	52.59
(8,4,3)	18	-746.06	51.78
(8,4,4)	19	-745.80	51.87

relation matrix by the entropy loss L_1 and quadratic loss L_2 . The (L_1, L_2) values for the optimal models based on AMCD, ACD and MCD are 3.0866 and 35.6300; 2.8660 and 30.8312;

and 1.9237 and 12.9392, respectively. Obviously, the MCD-based model yields more accurate estimates of the correlation matrix, and the AMCD and ACD-based models have similar performances. This indicates that the MCD structure might be more suitable for this dataset than AMCD and ACD, in terms of the estimation accuracy of the correlation matrix. However, when we keep p and q fixed, and take the polynomial order for the innovation variances as $s = 1$ rather than the optimal order $s = 3$, the (L_1, L_2) 's are respectively (3.0743, 31.8634), (2.9381, 30.0376), and (2.8819, 29.2350). The estimation accuracy of the correlation matrix for AMCD and ACD-based models varies slightly, again implying robustness with respect to the model misspecification of the innovation variances; while the losses for the MCD-based model increase significantly which might lead to the opposite conclusion.

6.2 The sleep dose-response data

We apply NJMM and LJMM based on the proposed AMCD to the sleep dose-response data, which was initially investigated by Ref. [21] and subsequently analyzed in Ref. [22] using Bayesian inference for the joint model based on the t distribution and MCD. The days of sleep deprivation and the corresponding average reaction times were recorded for each of the 18 participants in the 3-h group. Fig. 2a shows the trajectories of the average reaction times over an equally spaced 10-day period, together with the mean profile plot and the corresponding ± 1 standard deviations across the period. There seem to be sudden jumps and drops in the trajectories

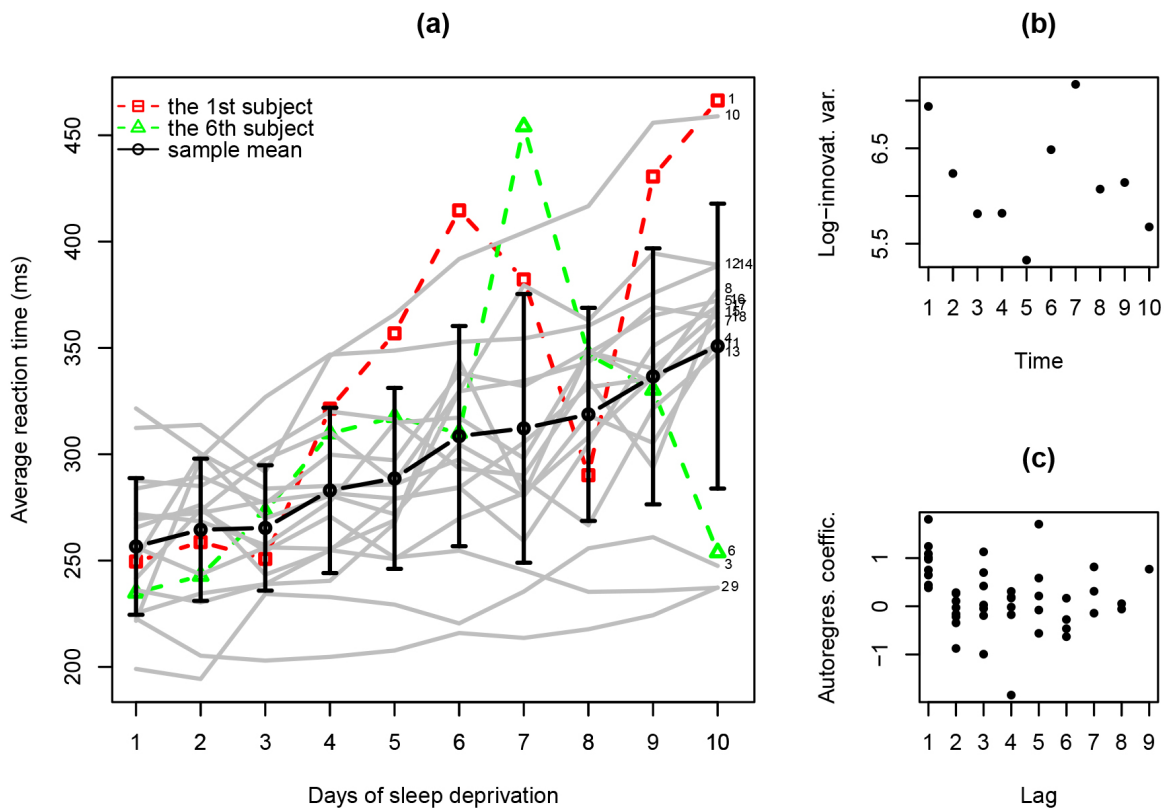


Fig. 2. Sleep dose-response data: (a) trajectories of average reaction time; (b) sample regressograms for log-innovation variances; (c) sample regressograms for autoregressive coefficients.

of the first and sixth subjects, which may be obvious outliers and thus not able to be properly handled using NJMM.

The mean response varies linearly over time, and thus can be modeled by a 1st-degree polynomial in time. Fig. 2b and c show the sample regressograms for log-innovation variances and autoregressive coefficients, implying higher-order polynomial functions in time or time lag. Hence, it is suitable to adopt three polynomials expressed as (16) with $p = 1$ and s, q to be determined. We fitted the Sleep dose-response data using AMCD-based NJMM and LJMM for various degrees of the Poly(1, s, q) models. The numbers of parameters, \widehat{l}_{\max} and BIC(1, s, q) values for the searched Poly(1, s, q)'s are listed in Table 8, where the BIC values are calculated via formula (17). According to the BIC values, Poly(1,3,4) and Poly(1,3,2) are the best for NJMM and LJMM, respectively, and this data is fitted significantly better when using LJMM. Table 9 displays the maximum likelihood estimates and standard errors when fitting respectively NJMM and LJMM with the above two best polynomials. Note that NJMM and LJMM have similar parameter estimates. However, the standard errors of $\widehat{\lambda}$ and $\widehat{\gamma}$ for LJMM are smaller than those of NJMM with Poly(1,3,2); the standard errors of $\widehat{\lambda}$ for LJMM

are slightly larger than those of NJMM with Poly(1,3,4), and the standard errors of $\widehat{\gamma}$ for LJMM are still smaller in this case. This indicates convincingly that parameter estimates of LJMM always possess lower variability for this data.

7 Conclusions

In longitudinal data analysis, we always suffer from various kinds of covariation structure and need to choose the most suitable decomposition among the candidates. If the covariance matrix possesses a typical structure, we might apply the MACD method; otherwise, we might adopt the MCD method for the precision matrix. Furthermore, when we are interested in robust estimation of the correlation matrix against model misspecification of the innovation variances, we can apply the ACD method to the covariance matrix. However, the decomposition of the precision matrix targeting the above robustness has been rather less investigated. In this paper, we have proposed AMCD of the precision matrix and established its role in providing robust estimator for the correlation matrix. In addition, we have investigated the AMCD-based LJMM which may achieve both the aforementioned robust estimation and robustness to outliers in the data.

Table 8. Sleep dose-response data: comparison of various Poly(1, s, q) choices between NJMM and LJMM.

Poly(1, s, q)	Number of parameters	\widehat{l}_{\max}		BIC	
		NJMM	LJMM	NJMM	LJMM
(1,1,1)	6	-710.32	-454.85	79.89	51.50
(1,1,2)	7	-698.82	-441.85	78.77	50.22
(1,1,3)	8	-698.07	-441.25	78.85	50.31
(1,1,4)	9	-696.49	-440.01	78.83	50.34
(1,1,5)	10	-695.85	-439.72	78.92	50.46
(1,2,1)	7	-710.13	-454.47	80.03	51.62
(1,2,2)	8	-698.66	-441.54	78.91	50.34
(1,2,3)	9	-697.86	-440.87	78.99	50.43
(1,2,4)	10	-696.40	-439.79	78.98	50.47
(1,2,5)	11	-695.78	-439.52	79.08	50.60
(1,3,1)	8	-709.45	-454.21	80.11	51.75
(1,3,2)	9	-695.00	-438.03	78.67	50.11
(1,3,3)	10	-694.00	-437.36	78.72	50.20
(1,3,4)	11	-691.02	-435.40	78.55	50.14
(1,3,5)	12	-690.22	-435.04	78.62	50.27
(1,4,1)	9	-708.21	-453.47	80.13	51.83
(1,4,2)	10	-694.34	-437.76	78.76	50.25
(1,4,3)	11	-693.20	-436.96	78.79	50.32
(1,4,4)	12	-690.52	-435.11	78.65	50.27
(1,4,5)	13	-689.79	-434.79	78.73	50.40
(1,5,1)	10	-707.14	-452.99	80.18	51.94
(1,5,2)	11	-693.42	-437.21	78.81	50.35
(1,5,3)	12	-692.06	-436.24	78.82	50.40
(1,5,4)	13	-689.65	-434.64	78.72	50.38
(1,5,5)	14	-688.83	-434.26	78.78	50.50

Table 9. Sleep dose-response data: parameter estimates for NJMM and LJMM with the two best polynomials.

	Poly(1,3,2)				Poly(1,3,4)			
	NJMM		LJMM		NJMM		LJMM	
	MLE	SE	MLE	SE	MLE	SE	MLE	SE
β_0	240.9311	6.7036	241.1218	7.7784	241.6758	6.3496	241.1834	7.7073
β_1	10.1691	1.6517	9.7037	1.6478	10.2647	1.6540	9.8365	1.6582
λ_0	7.5886	0.6562	4.1020	0.6333	7.8139	0.5422	4.1793	0.5746
λ_1	-1.0052	0.4517	-1.0521	0.4055	-1.2803	0.3616	-1.1990	0.3734
λ_2	0.2199	0.0875	0.2107	0.0783	0.2827	0.0753	0.2470	0.0767
λ_3	-0.0126	0.0050	-0.0118	0.0045	-0.0165	0.0044	-0.0141	0.0045
γ_0	0.8835	0.1715	0.9978	0.1456	2.1703	0.5384	2.0808	0.4754
γ_1	-0.3654	0.0978	-0.4222	0.0795	-2.0456	0.7714	-1.8478	0.6405
γ_2	0.0349	0.0112	0.0407	0.0090	0.6838	0.3205	0.5931	0.2577
γ_3	-0.0956	0.0495	-0.0814	0.0393
γ_4	0.0047	0.0025	0.0040	0.0020

Recently, an autoregressive moving average Cholesky decomposition (ARMACD) has been proposed in Ref. [23] ; this decomposition is more general than both MACD and MCD. Therefore, it is worthwhile to develop a robust estimation of the correlation matrix on the basis of ARMACD in the future. Furthermore, these decompositions correspond to various time series models, then various robust time series approaches could also be generalized for robust estimation of the correlation, covariance or precision matrix. In addition, longitudinal data may suffer from missingness, such as informative dropouts^[24]. Then we need to specify a suitable dropout mechanism, establish the complete data log-likelihood, and implement an EM algorithm to calculate the maximum likelihood estimates. Moreover, heterogeneity is also deserved to be investigated, not only for the mean parameters, but also for the covariation structures. The finite mixture model^[25] based on NJMM or LJMM with AMCD may perform well, which is left for future research.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (12101559), the Zhejiang Natural Science Foundation (LQ22A010013), the Science Foundation of Zhejiang Sci-Tech University (21062111-Y), and the Scientific Research Foundation of Zhejiang Sci-Tech University.

Conflict of interest

The authors declare that they have no conflict of interest.

Biography

Fei Lu is currently a lecturer at the College of Science, Zhejiang Sci-Tech University. He received his Ph.D. degree from Beijing University of Technology in 2020. His research mainly focuses on longitudinal data analysis.

References

[1] Diggle P J, Heagerty P J, Liang K Y, et al. Analysis of Longitudinal

Data. Oxford: Oxford University Press, 2002.

- [2] Diggle P J, Verbyla A P. Nonparametric estimation of covariance structure in longitudinal data. *Biometrics*, 1998, 54 (2): 401–415.
- [3] Pourahmadi M. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 1999, 86 (3): 677–690.
- [4] Zhang W, Leng C. A moving average Cholesky factor model in covariance modelling for longitudinal data. *Biometrika*, 2012, 99 (1): 141–150.
- [5] Chen Z, Dunson D B. Random effects selection in linear mixed models. *Biometrics*, 2003, 59 (4): 762–769.
- [6] Pourahmadi M. Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance-correlation parameters. *Biometrika*, 2007, 94 (4): 1006–1013.
- [7] Maadooliat M, Pourahmadi M, Huang J Z. Robust estimation of the correlation matrix of longitudinal data. *Statistics and Computing*, 2013, 23: 17–28.
- [8] Zhang W, Leng C, Tang C Y. A joint modelling approach for longitudinal studies. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2015, 77 (1): 219–238.
- [9] Lin T I, Wang Y J. A robust approach to joint modeling of mean and scale covariance for longitudinal data. *Journal of Statistical Planning and Inference*, 2009, 139 (9): 3013–3026.
- [10] Guney Y, Arslan O, Gokalp-Yavuz F. Robust estimation in multivariate heteroscedastic regression models with autoregressive covariance structures using EM algorithm. *Journal of Multivariate Analysis*, 2022, 191: 105026.
- [11] Pourahmadi M. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 2000, 87 (2): 425–435.
- [12] Anderson D N. A multivariate Linnik distribution. *Statistics & Probability Letters*, 1992, 14 (4): 333–336.
- [13] Ernst M D. A multivariate generalized Laplace distribution. *Computational Statistics*, 1998, 13 (2): 227–232.
- [14] Fernández C, Osiewalski J, Steel M F. Modeling and inference with v -spherical distributions. *Journal of the American Statistical Association*, 1995, 90 (432): 1331–1340.
- [15] Portilla J, Strela V, Wainwright M J, et al. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, 2003, 12 (11): 1338–1351.
- [16] Kotz S, Kozubowski T J, Podgórski K. The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance. Boston, MA: Birkhäuser, 2001.

- [17] Press W H, Teukolsky S A, Vetterling W T, et al. Numerical recipes: The art of scientific computing. 3rd ed. Cambridge: Cambridge University Press, 2007.
- [18] Pan J, Pan Y. jmcM: An R package for joint mean-covariance modeling of longitudinal data. *Journal of Statistical Software*, 2017, 82: 1–29.
- [19] Kenward M G. A method for comparing profiles of repeated measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1987, 36 (3): 296–308.
- [20] Pan J, Mackenzie G. On modelling mean-covariance structures in longitudinal studies. *Biometrika*, 2003, 90 (1): 239–244.
- [21] Belenky G, Wesensten N J, Thorne D R, et al. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. *Journal of Sleep Research*, 2003, 12 (1): 1–12.
- [22] Lin T I, Wang W L. Bayesian inference in joint modelling of location and scale parameters of the t distribution for longitudinal data. *Journal of Statistical Planning and Inference*, 2011, 141 (4): 1543–1553.
- [23] Lee K, Baek C, Daniels M J. ARMA Cholesky factor models for the covariance matrix of linear models. *Computational Statistics & Data Analysis*, 2017, 115: 267–280.
- [24] Zhang W, Xie F, Tan J. A robust joint modeling approach for longitudinal data with informative dropouts. *Computational Statistics*, 2020, 35: 1759–1783.
- [25] Yu J, Nummi T, Pan J. Mixture regression for longitudinal data based on joint mean-covariance model. *Journal of Multivariate Analysis*, 2022, 190: 104956.
- [26] Chiu T Y M, Leonard T, Tsui K W. The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 1996, 91 (433): 198–210.
- [27] Rubin H. Uniform convergence of random functions with applications to statistics. *The Annals of Mathematical Statistics*, 1956, 27 (1): 200–203.
- [28] Royden H L, Fitzpatrick P. Real Analysis. New York:Macmillan, 1968.

Appendix A: The score and expectation of the Hessian matrices

It is trivial to compute $U_1(\beta; \gamma, \lambda)$ and $I_{11}(\theta)$. Since Σ_i relies only on γ and λ , it can be easily obtained that

$$I_{12}(\theta) = -E\left(\frac{\partial^2 l}{\partial \beta \partial \lambda^T}\right) = -E\left\{\sum_{i=1}^n \frac{\partial \mu_i^T}{\partial \beta} \frac{\partial \Sigma_i^{-1}}{\partial \lambda^T} (y_i - \mu_i)\right\} = 0.$$

Similarly, $I_{13}(\theta) = 0$. As $\epsilon_i = T_i \Lambda_i^{-1} r_i$, it is easy to see that

$$-2l(\theta) = \sum_{i=1}^n \log |\Lambda_i^2| + \sum_{i=1}^n \epsilon_i^T \epsilon_i = \sum_{i=1}^n \sum_{j=1}^{m_i} \{\log(d_{ij}^2) + \epsilon_{ij}^2\}.$$

Thus, the partial derivative of $l(\theta)$ with respect to λ can be expressed as

$$U_2(\lambda; \beta, \gamma) = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^{m_i} \left(z_{ij} + 2 \frac{\partial \epsilon_{ij}}{\partial \lambda} \epsilon_{ij} \right). \tag{A.1}$$

With $\epsilon_i = T_i \Lambda_i^{-1} r_i$, it can be easily obtained that $\epsilon_{ij} = \frac{r_{ij}}{d_{ij}} - \sum_{t=1}^{j-1} \phi_{ijt} \frac{r_{it}}{d_{it}}$. Thus,

$$\frac{\partial \epsilon_{ij}}{\partial \lambda} = -\frac{1}{2} \frac{r_{ij}}{d_{ij}^2} z_{ij} + \frac{1}{2} \sum_{k=1}^{j-1} \frac{r_{ik}}{d_{ik}} \phi_{ijk} z_{ik}, \tag{A.2}$$

or equivalently in the matrix form $\frac{\partial \epsilon_i^T}{\partial \lambda} = -\frac{1}{2} Z_i^T \text{diag}(\Lambda_i^{-1} r_i) T_i^T$. Then,

$$U_2(\lambda; \beta, \gamma) = \frac{1}{2} \sum_{i=1}^n \left\{ -\sum_{j=1}^{m_i} z_{ij} + \sum_{k=1}^{m_i} z_{ik} \left(1 - \sum_{j=k+1}^{m_i} \phi_{ijk} \right) \frac{r_{ik}}{d_{ik}} \left(\frac{r_{ij}}{d_{ij}} - \sum_{t=1}^{j-1} \phi_{ijt} \frac{r_{it}}{d_{it}} \right) \right\} = \frac{1}{2} \sum_{i=1}^n Z_i^T (h_i - \mathbf{1}_{m_i}),$$

where the $s \times 1$ vector $h_i = \text{diag}(T_i^T T_i \Lambda_i^{-1} r_i r_i^T \Lambda_i^{-1})$.

From (A.1), we have

$$I_{22}(\theta) = -E\left(\frac{\partial^2 l}{\partial \lambda \partial \lambda^T}\right) =$$

$$\sum_{i=1}^n \sum_{j=1}^{m_i} E\left(\epsilon_{ij} \frac{\partial^2 \epsilon_{ij}}{\partial \lambda \partial \lambda^T} + \frac{\partial \epsilon_{ij}}{\partial \lambda} \frac{\partial \epsilon_{ij}}{\partial \lambda^T}\right) =$$

$$\frac{1}{4} Z_i^T \{I_{m_i} + (T_i^T T_i) \circ (T_i^{-1} T_i^{-T})\} Z_i,$$

where $A \circ B$ denotes the Hadamard product of matrices A and B .

Similarly, we have

$$U_3(\gamma; \beta, \lambda) = - \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{\partial \epsilon_{ij}}{\partial \gamma} \epsilon_{ij} = \sum_{i=1}^n (\epsilon_i^T \otimes I_q) \frac{\partial T_i^{-T}}{\partial \gamma} T_i^T \epsilon_i, \tag{A.3}$$

where

$$\frac{\partial \epsilon_{ij}}{\partial \gamma} = - \sum_{k=1}^{j-1} v_{ijk} \epsilon_{ik} \tag{A.4}$$

with $v_{ijk} = - \sum_{t=k+1}^{j-1} \frac{\partial a_{itk}}{\partial \gamma} \phi_{ijt}$ and a_{itk} being the (t, k) th element of T_i^{-1} , or in the matrix form $\frac{\partial \epsilon_i^T}{\partial \gamma} = - (\epsilon_i^T \otimes I_q) \frac{\partial T_i^{-T}}{\partial \gamma} T_i^T$.

From (A.3) and (A.4), it is easy to see that

$$\begin{aligned} I_{33}(\theta) &= -E \left(\frac{\partial^2 l}{\partial \gamma \partial \gamma^T} \right) \\ &= \sum_{i=1}^n \sum_{j=1}^{m_i} E \left(\epsilon_{ij} \frac{\partial^2 \epsilon_{ij}}{\partial \gamma \partial \gamma^T} + \frac{\partial \epsilon_{ij}}{\partial \gamma} \frac{\partial \epsilon_{ij}}{\partial \gamma^T} \right) \\ &= \sum_{i=1}^n \sum_{j=1}^{m_i} \sum_{k=1}^{j-1} v_{ijk} v_{ijk}^T. \end{aligned}$$

From (A.2), (A.3), (A.4) and the fact that $E(\epsilon_i r_i^T) = T_i^{-T} \Lambda_i$, i.e.,

$$E(\epsilon_{ij} r_{ik}) = \begin{cases} 0, & k < j, \\ d_{ik}, & k = j, \\ d_{ik} a_{ikj}, & k > j, \end{cases}$$

we have

$$\begin{aligned} I_{32}(\theta) &= -E \left(\frac{\partial^2 l}{\partial \gamma \partial \lambda^T} \right) = \\ &= - \sum_{i=1}^n \sum_{j=1}^{m_i} E \left\{ \sum_{k=1}^{j-1} v_{ijk} \left(\epsilon_{ik} \frac{\partial \epsilon_{ij}}{\partial \lambda^T} + \frac{\partial \epsilon_{ik}}{\partial \lambda^T} \epsilon_{ij} \right) \right\} = \\ &= - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^{m_i} \sum_{k=1}^{j-1} v_{ijk} \phi_{ijk} \left(z_{ik}^T + \sum_{t=k+1}^j a_{itk} z_{it}^T \right). \end{aligned}$$

Appendix B: Proof

Proof of Theorem 1. (a) Let $\theta = (\beta^T, \lambda^T, \gamma^T)^T$ and $l_i(\theta) = \log f_i(y_i, \theta)$, $i = 1, \dots, n$. Then, ignoring the constant $m_i \log(2\pi)$, we obtain that

$$l_i(\theta) = -\frac{1}{2} \log(|\Sigma_i|) - \frac{1}{2} \{y_i - \mu(x_i \beta)\}^T \Sigma_i^{-1} \{y_i - \mu(x_i \beta)\}.$$

Thus, the mean and variance of l_i when $\theta = \theta_0$ are

$$E_0 \{l_i(\theta)\} = -\frac{1}{2} \log(|\Sigma_i|) - \frac{1}{2} \text{tr}(\Sigma_i^{-1} \Sigma_{0i}) - \frac{1}{2} \{\mu(x_i \beta) - \mu(x_i \beta_0)\}^T \Sigma_i^{-1} \{\mu(x_i \beta) - \mu(x_i \beta_0)\},$$

$$\text{var}_0 \{l_i(\theta)\} = \frac{1}{2} \text{tr}(\Sigma_i^{-1} \Sigma_{0i})^2 + \{\mu(x_i \beta) - \mu(x_i \beta_0)\}^T \Sigma_i^{-1} \Sigma_{0i} \Sigma_i^{-1} \{\mu(x_i \beta) - \mu(x_i \beta_0)\},$$

where $\Sigma_i = \Lambda_i T_i^{-1} T_i^{-T} \Lambda_i$ and $\Sigma_{0i} = \Lambda_{0i} T_{0i}^{-1} T_{0i}^{-T} \Lambda_{0i}$. It follows from the compactness of the parameter space and the boundedness of the covariates that $\text{var}_0 \{l_i\} \leq \kappa_0$, for all i , where κ_0 is a constant. Therefore, we obtain that

$$\sum_{i=1}^{\infty} \frac{\text{var}_0 \{l_i(\theta)\}}{i^2} < \infty.$$

Thus, by Kolmogorov’s strong law of large numbers, we obtain that

$$\frac{1}{n} \sum_{i=1}^n l_i(\theta) - \frac{1}{n} \sum_{i=1}^n E_0 \{l_i(\theta)\} \rightarrow 0, \text{ a.s.} \tag{A.5}$$

Then we prove the consistency of $\widehat{\theta}$, which is similar to that of Theorem 1 in Ref. [26]. Note that the above constant κ_0 is independent of θ . By the compactness of Θ and a similar argument as that of Ref. [27], the convergence in (A.5) is uniform in Θ .

Furthermore, it can be shown that $\frac{1}{n} \sum_{i=1}^n E_0 \{l_i(\theta)\}$ is equicontinuous in θ . Since Θ is compact, and by condition (C3), it can be easily seen that $\frac{1}{n} \sum_{i=1}^n E_0 \{l_i(\theta)\}$ converges to a finite limit,

$$K_0(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n E_0 \{l_i(\theta)\}.$$

Then according to Ref. [28], the foregoing convergence is uniform in Θ and the limit $K_0(\theta)$ is continuous in θ . Therefore, by expression (A.5), we obtain that

$$\frac{1}{n} \sum_{i=1}^n l_i(\theta) \rightarrow K_0(\theta), \text{ a.s.}$$

uniformly in Θ . Due to the compactness of Θ , $K_0(\theta)$ is uniformly continuous in Θ .

Since the true parameter θ_0 lies in Θ , we have

$$K_0(\theta) < K_0(\theta_0)$$

for any θ ; that is, $K_0(\theta)$ has a uniform maximum at θ_0 . Since $K_0(\theta)$ is continuous and Θ is compact, $K_0(\theta_0)$ is bounded away from its maximum for any θ bounded away from θ_0 ; that is, for any $\delta > 0$, there exists $\varepsilon > 0$ such that

$$K_0(\theta) \leq K_0(\theta_0) - \varepsilon \tag{A.6}$$

for $|\theta - \theta_0| \geq \delta$.

For a contraction, assume that there is a set of positive probability where $\widehat{\theta}_n(y)$ does not converge to θ_0 . For each y in the set there exists a subsequence $\{m_n\} \subset \{n\}$ and a limit point $\theta_y \neq \theta_0$ in Θ such that $\widehat{\theta}_{m_n}(y) \rightarrow \theta_y \neq \theta_0$. Because $\widehat{\theta}_{m_n}(y)$ produces a maximum for every m_n , we obtain that

$$\frac{1}{m_n} \sum_{i=1}^{m_n} l_i(\widehat{\theta}_{m_n}(y)) \geq \frac{1}{m_n} \sum_{i=1}^{m_n} l_i(\theta_0).$$

Then by uniform convergence and continuity of the limit, for this y we can obtain that

$$K_0(\theta) \geq K_0(\theta_0),$$

but this contradicts (A.6), and thus it can be concluded that $\widehat{\theta}_n(y)$ is strongly consistent for θ_0 .

(b) First, it can be proven that the following necessary conditions for asymptotic normality hold under regular conditions (C1)–(C3).

- (B1) The first and the second derivatives of $f_i(y_i, \theta)$ with respect to θ exist.
- (B2) The expectation of the first derivative of $l_i(\theta)$ with respect to θ equals zero.
- (B3) The information matrices satisfy

$$E \left\{ \frac{\partial l_i(\theta)}{\partial \theta} \frac{\partial l_i(\theta)}{\partial \theta^T} \right\} = -E \left\{ \frac{\partial^2 l_i(\theta)}{\partial \theta \partial \theta^T} \right\},$$

(B4) As $n \rightarrow \infty$, we have

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 l_i(\theta)}{\partial \theta \partial \theta^T} - \frac{1}{n} \sum_{i=1}^n E_0 \left\{ \frac{\partial^2 l_i(\theta)}{\partial \theta \partial \theta^T} \right\} \rightarrow 0, \text{ a.s.}$$

uniformly in Θ .

(B5) The following asymptotic result holds:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial l_i(\theta)}{\partial \theta} \Big|_{\theta=\theta_0} \xrightarrow{d} N\{0, \mathcal{I}(\theta_0)\},$$

where the asymptotic covariance matrix $\mathcal{I}(\theta_0)$ is positive definite.

(B1)–(B3) are straightforward under (C1)–(C2). (B4) can be shown in a similar way to the proof of (A.5). In the following, we show that (B5) holds. In fact, at $\theta = \theta_0$,

$$E_0 \left\{ \left| \psi^T \frac{\partial l_i(\theta)}{\partial \theta} \right|^3 \right\} \leq \kappa$$

for any $\psi \in R^{p+d+q}$, where κ is a positive constant independent of i . Furthermore, at $\theta = \theta_0$, we have

$$\frac{1}{n} \sum_{i=1}^n \text{var}_0 \left\{ \psi^T \frac{\partial l_i(\theta)}{\partial \theta} \right\} = \psi^T \{n^{-1} \mathcal{I}(\theta_0)\} \psi \rightarrow \psi^T \mathcal{I}(\theta_0) \psi > 0,$$

due to the positive definiteness of $\mathcal{I}(\theta_0)$ in (C3). Therefore, (B5) follows from the Liapounov multivariate central limit theorem.

In the following, we prove the asymptotic normality of the maximum likelihood estimator $\widehat{\theta}$ under regular conditions (B1)–(B5). The proof is similar to that of Theorem 2 in Ref. [26] and we only state the key points. Since $\widehat{\theta}_n$ is a consistent sequence of roots to equations (9), with probability 1, we can concentrate on a neighborhood of θ_0 . Define

$$\varphi_i(\zeta) = \frac{\partial l_i(\theta)}{\partial \theta} \Big|_{\theta=\theta_1+\zeta(\theta_2-\theta_1)},$$

where $\zeta \in [0, 1]$, for any θ_1 and θ_2 . Then, by the fundamental theorem of calculus,

$$\varphi_i(1) - \varphi_i(0) = \int_0^1 \dot{\varphi}_i(\zeta) d\zeta.$$

Setting $\theta_1 = \theta_0$ and $\theta_2 = \widehat{\theta}_n$, and summing over i from 1 to n , we can obtain that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial l_i(\theta)}{\partial \theta} \Big|_{\theta=\theta_0} = -\Omega_n \{ \sqrt{n}(\widehat{\theta}_n - \theta_0) \},$$

where

$$\Omega_n = \int_0^1 \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 l_i(\theta)}{\partial \theta \partial \theta^T} \Big|_{\theta=\theta_\zeta} d\zeta$$

with $\theta_\zeta = \theta_0 + \zeta(\widehat{\theta}_n - \theta_0)$, and $\widehat{\theta}_n$ is the root to (9). By (B2)–(B4), it can be shown that

$$-\Omega_n - n^{-1} \mathcal{I}(\theta_0) \xrightarrow{p} 0,$$

where the convergence holds when $\theta = \theta_0$. By (C4), we obtain that when $\theta = \theta_0$,

$$-\Omega_n \xrightarrow{p} \mathcal{I}(\theta_0). \tag{A.7}$$

Thus, conclusion (b) follows from (A.7) and (B5).