International Statistical Review

International Statistical Review (2024) doi: 10.1111/insr.12577

Joint Robust Variable Selection of Mean and Covariance Model via Shrinkage Methods

Yeşim Güney[1], **Fulya Gokalp Yavuz**[2] and **Olcay Arslan**[1]

[1]*Department of Statistics, Ankara University, Ankara, Turkey*
[2]*Department of Statistics, Middle East Technical University, Ankara, Turkey*
Correspondence *Yeşim Güney, Department of Statistics, Ankara University. Email: ydone@ankara.edu.tr*

Summary

A valuable and robust extension of the traditional joint mean and the covariance models when data subject to outliers and/or heavy-tailed outcomes can be achieved using the joint modelling of location and scatter matrix of the multivariate t-distribution. This model encompasses three models in itself, and the number of unknown parameters in the covariance model increases quadratically with the matrix size. As a result, selecting the important variables becomes a crucial aspect to consider. In this context, the variable selection combined with the parameter estimation is considered under the normality assumption. However, because of the non-robustness of the normal distribution, the resulting estimators will be sensitive to outliers and/or heavy taildness in the data. This paper has two objectives to overcome these problems. The first is to obtain the maximum likelihood estimates of the parameters and propose an expectation-maximisation type algorithm as an alternative to the Fisher scoring algorithm in the literature. We also consider simultaneous parameter estimation and variable selection in the multivariate t-joint location and scatter matrix models. The consistency and oracle properties of the regularised estimators are also established. Simulation studies and real data analysis are provided to assess the performance of the proposed methods.

Key words: Bridge; joint mean-covariance model; LASSO; penalised estimation; SCAD; t-distribution.

1 INTRODUCTION

In biomedical, sociological, and economic studies, it is common to investigate a problem by observing multiple outcomes over time for the same subject repeatedly. Therefore, longitudinal data arise more frequently in various scientific domains involving extensive research. Unlike other types of multivariate data, the assumption of independence between different subjects and dependency within each subject often poses a fundamental challenge for statistical modelling. A popular approach to analysing the longitudinal data is to use the joint location and scatter matrix model (JLSM) defined as follows.

Suppose that we have m subjects, with response from each subject $i \in \{1, \ldots, m\}$ measured n_i times. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$ be the n_i repeated measurements at time point $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})^T$ of the i-th subject. Assume that for each $i \in \{1, \ldots, m\}$, the responses are

independent. Let $\mathbf{Y}_i$'s come from any distribution with location vector $\boldsymbol{\mu}_i$, scale $\boldsymbol{\Sigma}_i$ and the JLSM is defined as

$$\boldsymbol{Y}_i \sim .(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

where $\boldsymbol{\Sigma}_i = \left[\sigma_{ij}\right]$ is the $n_i \times n_i$ positive definite scatter matrix. Here, the most common way to model the location is to use a linear model. In JLSMs, the estimation of a covariance matrix is an important problem. The current literature on modelling covariance matrices for multivariate longitudinal data relies on decompositions of the covariance matrices. The sample covariance matrix is known to be a positive-definite and unbiased estimator of a covariance matrix. However, it is pretty unstable when the dimension of the covariance matrix is large (Huang *et al.*, 2006; Lin, 1985; Wong *et al.*, 2003). It is also difficult to keep the estimated matrix positive definite. To handle this difficulty, Pourahmadi (1999, 2000) has developed the following modified Cholesky decomposition (MCD) and guaranteed the positive-definiteness of the estimated covariance at no additional computational cost. For any $n_i \times n_i$ covariance matrix $\boldsymbol{\Sigma}_i$, let define

$$\mathbf{L}_i \boldsymbol{\Sigma}_i \mathbf{L}_i^T = \mathbf{D}_i, \tag{2}$$

where $\mathbf{L}_i$ is a lower triangular matrix with ones on the diagonal, $\mathbf{D}_i$ is a diagonal matrix, and the elements below diagonal in the i-th row of $\mathbf{L}_i$ can be interpreted as regression coefficients of the i-th component on its predecessors; the elements of $\mathbf{D}_i$ give the corresponding prediction variances. The elements of $\mathbf{D}_i$ and $\mathbf{L}_i$ can be modelled using linear models.

The overall model consists of three regression sub-models. These are the models for the location of the response vector, the elements of the generalised autoregressive matrices, and the innovation variances. The JLSMs are based on the MCD or its extensions have been considered by several authors (for more details, see Pourahmadi, 2013, and references therein) and many estimation methods have been proposed by Pourahmadi (2000), Fan & Zhang (2000), Wu & Pourahmadi (2003), Pan & Mackenzie (2003), Fan *et al.* (2007) and Fan & Wu (2008) to estimate the parameters of JLSMs under the normality assumption. Further, the generalised estimating equations (GEE) method has also been used to estimate the JLSMs parameters (see Liang & Zeger, 1986, Pan & Ye, 2004, Leng *et al.*, 2010). However, it is well known that these approaches may be sensitive to outliers due to the normality assumption or using non-robust GEE. To handle the outlier problem, some robustification has been done in the literature. One can see Cantoni (2004), He *et al.* (2005), Wang *et al.* (2005), Qin & Zhu (2007), Qin *et al.* (2009) and (Croux *et al.*, 2012) to see robust approaches to JLSMs. However, their methods do not deal with irregular observed measurements. When the assumption of normality is questionable like when unusual points exist or the underlying data exhibit heavy tails, then other heavy-tailed distributions might be reasonable alternatives. In this context, Lin & Wang (2009) have proposed a JLSM of t-distribution (t-JLSM), and Guney *et al.* (2022) have proposed JLSM of multivariate Laplace distribution. However, in both studies, the variable selection has not been considered.

Numerous possible variables can be added to the model at the initial stage of the model construction to reduce potential biases in the model. This causes the number of variables to be high-dimensional. In addition, the number of unknown parameters in the covariance model grows quadratically with the matrix size. The parameter estimation seems computationally intensive in this case. Therefore, variable selection is as important as parameter estimation in JLSM. Since all possible subset searches are time-consuming and not practically useful, when the number of covariates is large, the traditional information-based model selection criteria are not preferable for JLSMs. Another way for variable selection in JLSM is to use penalised likelihood or

penalised estimating equations. Commonly used penalties include the Bridge penalty introduced by Frank & Friedman (1993), least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996), hard thresholding penalty defined by Antoniadis (1997), the smoothly clipped absolute deviation (SCAD) penalty defined by Fan & Li (2001), and adaptive LASSO (ALASSO) in Zou (2006). For example, Huang *et al.* (2006) have proposed a maximum penalised likelihood estimator to select the significant variables and estimate the parameters in JLSM of normal distribution; see also Huang *et al.* (2007) and Levina *et al.* (2008) for some improvements. Kou & Pan (2009) proposed a penalised maximum likelihood method by penalising the normal likelihood using the SCAD penalty. Xu *et al.* (2013) have proposed covariance selection and estimation via penalised normal likelihood using SCAD and established the consistency and asymptotic normality of the penalised maximum likelihood estimators of parameters under certain regularity conditions. Jhong *et al.* (2017) have developed a novel updating-based method for penalised estimators for the mean vector and the covariance matrix. Kou & Pan (2020) have used the LASSO, SCAD and hard thresholding penalty to penalise the likelihood function. The constraint of these studies is that they consider the normal distribution assumption. In the analysis of longitudinal data, such classical modelling approaches can be challenged by heavy-tailed errors and outliers, model misspecification, and others. These challenges demand the development of robust methods that can be insensitive to model specifications and outliers. Nevertheless, the discussion on robust variable selection methods has been limited. For example, Zheng *et al.* (2014) have developed a penalised robust estimating equations-based method to select important variables. Their method is a robustified version of the generalised estimation method. In this paper, we proposed a robust variable selection method in JLSMs using the t-distribution, which can deal with the issue of heavy-tailed and/or noisy data. We also developed an EM-type algorithm for numerical computations using the stochastic representation of the multivariate t-distribution.

In this article, following Lin & Wang (2009), we consider the t-JLSM. This paper has two goals. First, we adapt the EM algorithm to obtain the ML estimates of the parameters using the scale-mixture representation of the multivariate t-distribution. The second goal of this article is to develop a penalised likelihood method for t-JLSM to select the important variables that make a significant contribution to the JLSMs. We consider a penalised likelihood method that can simultaneously perform parameter estimation and variable selection in the JLSMs. We use LASSO, SCAD, and Bridge penalties. In addition, we propose an EM-type algorithm to obtain the penalised likelihood estimates. The asymptotic properties of the resulting estimators are also considered.

The rest of the article is organised as follows: Section 2 describes the model in detail. A Fisher scoring algorithm for the implementation of ML estimation is given in this section as well. Section 3 introduces the EM-type algorithm for computing the ML estimates of the parameters. The penalised estimator is proposed, and an EM-type algorithm is devoloped for parameter estimation and variable selection in Section 4. Section 4 also provides some theoretical justifications. In Section 5, we present results from two simulation studies. The first one investigates how the variable selection method, based on different penalties, concentrates around the true covariates in location and scale models, while the second one investigates the advantages that one may have when using the proposed method for modelling the multivariate longitudinal data with contamination. Section 6 applies the methods to a data set that is the Framingham Cholesterol data set (qrLMM package in R, (Galarza & Lachos, 2017)) to illustrate the proposed method. Finally, Section 7 concludes the paper with a brief discussion. Theoretical proofs of the theorems that summarise the asymptotic results are presented in Appendix A.

2 JLSM OF MULTIVARIATE T-DISTRIBUTION

By assuming responses follow a multivariate t-distribution rather than a normal distribution in JLSMs, we can conduct data analysis for repeated or clustered measurement data with tails that extend beyond the normal distribution. The degrees of freedom parameters of the t-distributed responses provide a convenient way for achieving a flexible trade-off between robustness and efficiency. Given the aforementioned issues, we consider the t-JLSM for the heavy-tailed repeated or clustered measurement data.

In this section, we summarise the fundamental results concerning the multivariate t-distribution, introduce t-JLSM, and derive the 'complete-data' likelihood equations, associated ML estimates, and the Newton–Raphson algorithm to compute the estimates.

2.1 Multivariate t-Distribution

An n-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is said to have a multivariate t-distribution ($\mathbf{Y} \sim t_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, v)$) with parameters, $\boldsymbol{\mu} \in \boldsymbol{R}^n$ is the location vector, $\boldsymbol{\Sigma}$ is the positive definite scatter matrix and $v \in (0, \infty)$ degrees of freedom if its density function is as follows:

$$f(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, v) = \frac{\Gamma\left(\frac{v+n}{2}\right)|\boldsymbol{\Sigma}|^{-1/2}}{(\pi v)^{n/2}\Gamma\left(\frac{v}{2}\right)}\left[1 + \frac{1}{v}(\mathbf{y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right]^{-\frac{v+n}{2}} \tag{3}$$

where $\Gamma(\cdot)$ is the gamma function. The expectation and the variance of $\mathbf{Y}$ are $E(\mathbf{Y}) = \boldsymbol{\mu}$ and $Var(\mathbf{Y}) = \frac{v}{v-2}\boldsymbol{\Sigma}$ for $v > 2$.

As a special case $v = 1$, the distribution becomes a multivariate Cauchy distribution, and as $v \to \infty$, the distribution rolls back to the multivariate normal. Thus, the family of t-distributions provides a heavy-tailed alternative to the normal family.

The multivariate t-distribution can be defined as a scale mixture of $n-$ variate normal and the Chi-square distribution. Let $\mathbf{U} \sim N_n(\mathbf{0}, \mathbf{I})$ and $V \sim \chi^2_v$ be independent random variables, then $\mathbf{Y} = \boldsymbol{\mu} + (v\boldsymbol{\Sigma})^{1/2}\mathbf{U}/\sqrt{V}$ will have $t_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, v)$. The conditional distribution of $\mathbf{Y}$ given V is $N_n\left(\boldsymbol{\mu}, V^{-1}v\boldsymbol{\Sigma}\right)$. Then the joint pdf of $\mathbf{Y}$ and V will be

$$f(\mathbf{y}, v) = \frac{|\boldsymbol{\Sigma}|^{-1/2}}{(\pi v)^{n/2}2^{\frac{n+v}{2}}\Gamma\left(\frac{v}{2}\right)}v^{\frac{n+v}{2}-1}\exp\left\{-\frac{v}{2}\left(1 + \frac{1}{v}\Delta\right)\right\}, \tag{4}$$

where $\Delta = (\mathbf{y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ denotes the Mahalanobis squared distance between $\mathbf{y}$ and $\boldsymbol{\mu}$ (with $\boldsymbol{\Sigma}$ as the scatter matrix). Using the joint pdf and the pdf of $\mathbf{Y}$, the conditional density function of V given $\mathbf{Y}$=$\mathbf{y}$ is $Gamma\left(\frac{n+v}{2}, \frac{2}{1+\frac{1}{v}\Delta}\right)$. The following conditional expectations can be easily obtained:

$$\psi = E(V|y) = \frac{n+v}{1+\frac{1}{v}\Delta}, \tag{5}$$

$$E(\log V|y) = DG\left(\frac{n+v}{2}\right) + \ln\left(\frac{2}{1+\frac{1}{v}A}\right), \tag{6}$$

where $DG(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$. The above results will be useful in the EM-type algorithm to obtain the ML estimates of the model parameters.

2.2 t-Joint Location and Scatter Matrix Model

Suppose there are m independent subjects and the i-th subject has n_i repeated measurements. Specifically, denote the response vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$ for the i-th subject, $i \in \{1, \ldots, m\}$, which are observed at time $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})^T$. Assume that for each i, the response vector follows a multivariate t-distribution. Lin & Wang (2009) have defined t-JLSM as follows.

$$\mathbf{Y}_i \sim t_{n_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, v), \tag{7}$$

where $\boldsymbol{\mu}_i$ is the location vector, $\boldsymbol{\Sigma}_i = [\sigma_{ij}]$ is the $n_i \times n_i$ positive definite scatter matrix and v is the degrees of freedom. One classic way of modelling the location of the data is to use a linear model such as

$$\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^T = \mathbf{X}_i\boldsymbol{\beta}, \tag{8}$$

where $\mathbf{X}_i = [\mathbf{x}_{ij}]_{\substack{i=1,\ldots,m,\\ j=1,2,\ldots,n_i}}$ represents the design matrix of each subject, with size $n_i \times p$, and could have a column of 1's if an intercept term is desired. $\mathbf{X}_i$ is known and assumed to be of full column rank. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of regression coefficients.

To guarantee the positive definiteness of $\boldsymbol{\Sigma}_i$, we re-parameterise $\boldsymbol{\Sigma}_i$ via the MCD as given in (2). Here, $\mathbf{L}_i = [l_{jk}]$ is unit lower triangular matrices with 1's as diagonal entries and (j, k)-th entry being $-\phi_{jk}$, $1 \le k \le j - 1$, and $\mathbf{D}_i = diag\{\sigma_j^2\}_{j=1}^{n_i}$. The ϕ_{jk}'s are called generalised autoregressive parameters and the negatives of the coefficients of $\hat{Y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1}\phi_{jk}(Y_{ik} - \mu_{ik})$. In other words, the ϕ_{jk}'s are the linear least-squares predictor of Y_{ij} based on its predecessors $Y_{i1}, \ldots, Y_{i(j-1)}$. The diagonal elements of $\mathbf{D}_i$, σ_j^2's, are called innovation variances of $\boldsymbol{\Sigma}_i$ (the prediction error variances) in the form of $\sigma_j^2 = \frac{v}{v-2}Var(Y_{ij} - \hat{Y}_{ij})$ for $1 \le i \le m$, $1 \le j \le n_i$. From these definitions, it is clear that $\boldsymbol{\Sigma}_i^{-1} = \mathbf{L}_i^T\mathbf{D}_i^{-1}\mathbf{L}_i$.

The unconstrained parameters ϕ_{jk} and σ_j^2 can be modelled as follows:

$$\phi_{jk} = \mathbf{z}_{jk}^T\boldsymbol{\gamma}, \tag{9}$$

$$\log\,\sigma_j^2 = \mathbf{w}_j^T\boldsymbol{\lambda}. \tag{10}$$

Here, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)^T$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)^T$ are d and q-dimensional vectors of parameters, $\mathbf{z}_{jk}$ and $\mathbf{w}_j$ are d and q-dimensional covariate vectors (Pan & Mackenzie, 2003). $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are assumed to be common for all $\boldsymbol{\Sigma}_i$'s for exhibiting the same covariance structure.

2.3 Maximum Likelihood Estimation of the Parameters

Given a sample $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ from model (7), the log-likelihood function will be as follows:

$$logL(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \sum_{i=1}^{m}\left(\log\Gamma\left(\frac{v+n_i}{2}\right) - \log\Gamma\left(\frac{v}{2}\right) - \frac{n_i}{2}\log(\pi v)\right) - \frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{m}(v+n_i)\log\left[1+\frac{1}{v}(\mathbf{Y}_i - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i)\right] \tag{11}$$

Let $\mathbf{r}_i = \mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta} = \{r_{ij}\}_{j=1}^{n_i}$ be the vector of residuals, $\hat{\mathbf{r}}_i = \{\hat{r}_{ij}\}_{j=1}^{n_i} = \sum_{k=1}^{j-1} r_{ik}\mathbf{z}_{jk}^T\boldsymbol{\gamma}$ be the predictor of $\mathbf{r}_i$, and $n = \sum_{i=1}^{m} n_i$ be the total number of observations. Using each of the sub-models in Equations (8), (9), and (10) and the following result, $\mathbf{L}_i\mathbf{r}_i = \mathbf{r}_i - \hat{\mathbf{r}}_i$, given in Pourahmadi (1999), the log-likelihood function has the following representation (Lin & Wang, 2009).

$$logL(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = \sum_{i=1}^{m}\left(\log\Gamma\left(\frac{v+n_i}{2}\right) - \log\Gamma\left(\frac{v}{2}\right) - \frac{n_i}{2}\log(\pi v)\right) - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{w}_j^T\boldsymbol{\lambda} - \frac{1}{2}\sum_{i=1}^{m}(v+n_i)\log\left(1+\frac{(\mathbf{r}_i - \mathbf{Z}_i\boldsymbol{\gamma})^T\mathbf{D}_i^{-1}(\mathbf{r}_i - \mathbf{Z}_i\boldsymbol{\gamma})}{v}\right) \tag{12}$$

Here, $\mathbf{Z}_i$ is an $n_i \times d$ matrix defined by

$$\mathbf{Z}_i = [\mathbf{z}(i,1), \mathbf{z}(i,2), \ldots, \mathbf{z}(i,n_i)]^T,\ \mathbf{z}(i,j) = \sum_{k=1}^{j-1} r_{ik}\mathbf{z}_{jk}^T. \tag{13}$$

Taking the derivatives of this log-likelihood function with respect to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\lambda}$, and v setting them to zero we obtain the following estimating equations:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m}\hat{\psi}_i\mathbf{X}_i^T\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{m}\hat{\psi}_i\mathbf{X}_i^T\hat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{Y}_i\right), \tag{14}$$

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{m}\hat{\psi}_i\hat{\mathbf{Z}}_i^T\hat{\mathbf{D}}_i^{-1}\hat{\mathbf{Z}}_i\right)^{-1}\left(\sum_{i=1}^{m}\hat{\psi}_i\hat{\mathbf{Z}}_i^T\hat{\mathbf{D}}_i^{-1}\tilde{\mathbf{r}}_i\right), \tag{15}$$

$$\hat{\boldsymbol{\lambda}} = \left(\sum_{i=1}^{m}\hat{\psi}_i\sum_{j=1}^{n_i}\mathbf{w}_j\mathbf{w}_j^T\right)^{-1}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{\psi}_i\mathbf{w}_j\hat{\mathbf{D}}_i^{-1}\left(\hat{\boldsymbol{\epsilon}}_i^2 - \hat{\Omega}_i^2 - \hat{\mathbf{D}}_i\log\hat{\Omega}_i^2\right)\right), \tag{16}$$

and

$$\hat{v} = \left\{\frac{1}{n}\left(\sum_{i=1}^{m}\left[DG\left(\frac{\hat{v}+n_i}{2}\right) - DG\left(\frac{\hat{v}}{2}\right)\right] - \sum_{i=1}^{m}\log\left(1+\frac{1}{\hat{v}}\hat{\Delta}_i\right) + \frac{1}{\hat{v}^2}\sum_{i=1}^{m}\hat{\psi}_i\hat{\Delta}_i\right)\right\}^{-1}, \tag{17}$$

where $\hat{\psi}_i = \dfrac{\hat{v}+n_i}{1+\frac{1}{\hat{v}}\hat{\Delta}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}})}$, $\hat{\Delta}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)^T\hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)$, $\tilde{\mathbf{r}}_i = \mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i$, $\hat{\boldsymbol{\epsilon}}_i^2 = \left((\tilde{r}_{i1} - \hat{r}_{i1})^2, \ldots, (\tilde{r}_{in_i} - \hat{r}_{in_i})^2\right)$, and $\hat{\Omega}_i^2 = \left(\hat{\sigma}_{i1}^2, \hat{\sigma}_{i2}^2, \ldots, \hat{\sigma}_{in_i}^2\right)^T$. Note that the estimating equations for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are given before in literature (see Lin & Wang, 2009).

Here, ψ_i can be seen as a weight function, which is a decreasing function of Δ_i. Using this weight function, data points with large residuals receive small weights and hence will be

down-weighted. In this study, we consider the degrees of freedom v is known and set to some small values for the sake of robustness (see Lange *et al.*, 1989; Arslan & Genç, 2003; Arslan & Genc, 2009).

The ML estimates of the parameters can be obtained by maximising the log-likelihood function given in the Equation (11). One way to get the estimates is to use the Newton–Raphson or Fisher scoring algorithms as in (Lin & Wang, 2009). The estimates can be obtained using the following Newton–Raphson updating equations.

$$\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}} - \mathbf{H}^{-1}\left(\tilde{\boldsymbol{\theta}}\right)\mathbf{U}\left(\tilde{\boldsymbol{\theta}}\right). \tag{18}$$

Here, $\mathbf{H}$ is Hessian matrix and $\mathbf{U}$ is the score vector. By replacing the Hessian matrix in Equation (18) with the expected Fisher information matrix $\mathbf{F}(\boldsymbol{\theta}) = -E(\mathbf{H}(\boldsymbol{\theta}))$, which can be found in (Lin & Wang, 2009), the ML estimates of the parameters can also be obtained via the iterative Fisher scoring method:

$$\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}} + \mathbf{F}^{-1}\left(\tilde{\boldsymbol{\theta}}\right)\mathbf{U}\left(\tilde{\boldsymbol{\theta}}\right). \tag{19}$$

The steps of Fisher scoring algorithm are as follows.

Step 1. Let $\boldsymbol{\theta}^{(0)} = \left(\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)}, \boldsymbol{\lambda}^{(0)}\right)$ is be the initial parameter vector and form the $\mathbf{L}_i^{(0)}$ and $\mathbf{D}_i^{(0)}$ matrices using the models given in (9)–(10) and $\boldsymbol{\Sigma}_i^{(0)}$.

Step 2. For $h = 0, 1, \ldots$, using $\boldsymbol{\theta}^{(h)} = \left(\boldsymbol{\beta}^{(h)}, \boldsymbol{\gamma}^{(h)}, \boldsymbol{\lambda}^{(h)}\right)$ compute the value $\boldsymbol{\gamma}^{(h+1)}$ using the Equation (15). To calculate $\boldsymbol{\lambda}^{(h+1)}$ either use the updating Equation (16) or the following Fisher scoring equation.

$$\boldsymbol{\lambda}^{(h+1)} = \boldsymbol{\lambda}^{(h)} + \left(\mathbf{F}_{\lambda\lambda}^{(h)}\right)^{-1}\mathbf{U}_{\lambda}^{(h)} \tag{20}$$

Step 3. Compute the inverse of $\boldsymbol{\Sigma}_i^{(h+1)}$

$$\left(\boldsymbol{\Sigma}_i^{(h+1)}\right)^{-1} = \mathbf{L}_i^T\left(\boldsymbol{\gamma}^{(h+1)}\right)\mathbf{D}_i^{-1}\left(\boldsymbol{\lambda}^{(h+1)}\right)\mathbf{L}_i^T\left(\boldsymbol{\gamma}^{(h+1)}\right)$$

Step 4. Use $\boldsymbol{\beta}^{(h)}$ and $\boldsymbol{\Sigma}_i^{(h+1)}$ update $\psi_i^{(h+1)}$, calculate $\boldsymbol{\beta}^{(h+1)}$ using the Equation (14).

Step 5. Repeat Steps 2 to 4 until a pre-specified criterion is met.

Remark: Concerning the estimation of the degrees of freedom v, one can use the updating equation given in (17) before Step 5.

Note that these algorithms require a Hessian matrix, which increases the computational costs of each iteration. On the other hand, given the scale mixture representation of the t-distribution, the EM algorithm emerges as a prominent inference tool, renowned for its numerical stability and straightforward implementation, enabling the computation of parameter estimates. Comparing the EM algorithm with previous algorithms by only looking at the number of iterations, those algorithms are faster. Yet as we have already pointed out the computational costs of each iteration for those algorithms are higher compared to the EM algorithm due to the calculation of the Hessian matrix (Jørgensen & Petersen, 2012). Further, it is shown that the log-likelihood function including the penalty term used in the EM algorithm is monotone increasing that will lead to at least a local maximum.

3 EM-TYPE ALGORITHM TO COMPUTE THE ML ESTIMATES

Let $\mathbf{Y}_i$ and V_i be observed and missing data, respectively and for $i = 1, 2, \ldots, m$, $(\mathbf{Y}_i, V_i)$, be the complete data. Let $\mathbf{Y}_i \sim t_{n_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu)$ and $\mathbf{Y}_i = \boldsymbol{\mu}_i + (\nu\boldsymbol{\Sigma}_i)^{1/2}\mathbf{U}_i/\sqrt{V_i}$. Using the joint density function of $(\mathbf{Y}_i, V_i)$ given in (4), we can get the following complete data log-likelihood function.

$$logL(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = -\frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{m} V_i\left(1 + \frac{1}{\nu}\Delta_i\right). \tag{21}$$

E-step: Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$. Finding the conditional expectation of the complete data log-likelihood function yields the following objective function:

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(h)}\right) = E[logL(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i|\mathbf{y}_i)] = -\frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{m}\psi_i\Delta_i, \tag{22}$$

where ψ_i given in Equation (5).

M-step: In the M-step, maximise Q with respect to $\boldsymbol{\theta}$ to find a new estimate $\boldsymbol{\theta}^{(h+1)}$.

These two steps can be implemented with the following iteratively re-weighting algorithm (IRA).

Algorithm Steps:

Step 1. Take the initial estimates $\boldsymbol{\theta}^{(0)}$ and determine a stopping rule ϵ.

Step 2. (E Step) Set $h = 0, 1, 2, \ldots$. Compute $\mathbf{r}_i^{(h)} = \mathbf{Y}_i - \boldsymbol{X}\boldsymbol{\beta}^{(h)}$, $\sigma_j^{2(h)}$, $\phi_{jk}^{(h)}$ using Equations (9) and (10), respectively. Also compute $\hat{r}_{ij}^{(h)} = \sum_{k=1}^{j-1} r_{ik}^{(h)}\phi_{jk}^{(h)}$ and $\mathbf{Z}_i^{(h)}$ using Equation (13). Form the following matrices $\mathbf{D}_i^{(h)} = diag\left\{\sigma_j^{2(h)}\right\}_{j=1}^{n_i}$ and $\mathbf{L}_i^{(h)} = \left[l_{jk}^{(h)}\right]$ with $(j, k) -$ the entry being $-\phi_{jk}^{(h)}$, $1 \le k \le j - 1$ and then compute $\boldsymbol{\Sigma}_i^{-1(h)} = \mathbf{L}_i^{(h)T}\mathbf{D}_i^{(h)-1}\mathbf{L}_i^{(h)}$ and $\Delta_i^{(h)} = \left(\mathbf{y}_i - \boldsymbol{X}\boldsymbol{\beta}^{(h)}\right)^T\boldsymbol{\Sigma}_i^{-1(h)}\left(\mathbf{y}_i - \boldsymbol{X}\boldsymbol{\beta}^{(h)}\right)$. Then compute the current value of conditional expectation $\psi_i^{(h)}$ given in Equation (5).

Step 3. (E Step) Form the following objective function using the current estimates:

$$Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(h)}\right) = -\frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{m}\psi_i^{(h)}\Delta_i. \tag{23}$$

Step 4. (M Step) Maximise the $Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(h)}\right)$ with respect to $\boldsymbol{\theta}$ to get the $(h+1)$-th parameter estimates for the parameters. This maximisation yields the following updating equation for $\boldsymbol{\gamma}$

$$\boldsymbol{\gamma}^{(h+1)} = \left(\sum_{i=1}^{m}\psi_i^{(h)}\mathbf{Z}_i^{(h)}T\left(\mathbf{D}_i^{(h)}\right)^{-1}\mathbf{Z}_i^{(h)}\right)^{-1}\left(\sum_{i=1}^{m}\psi_i^{(h)}\mathbf{Z}_i^{(h)}T\left(\mathbf{D}_i^{(h)}\right)^{-1}\tilde{\mathbf{r}}_i^{(h)}\right). \tag{24}$$

Calculate $\boldsymbol{\lambda}^{(h+1)}$ using either the updating Equation (16) or the Fisher scoring Equation (20). Using $\boldsymbol{\beta}^{(h)}$, $\boldsymbol{\gamma}^{(h+1)}$, $\boldsymbol{\lambda}^{(h+1)}$, update $\boldsymbol{\Sigma}_i^{(h+1)}$ and $\psi_i^{(h+1)}$. Then calculate $\boldsymbol{\beta}^{(h+1)}$ by using the updating Equation (14).

Step 5. Repeat E and M steps until the convergence rule $\boldsymbol{\theta}^{(h+1)} - \boldsymbol{\theta}^{(h)} < \epsilon$ is satisfied.

Remark. Note that the estimates of the degrees of freedom ν can be also calculated using Equation (17); however, in our study, we will not do so for the sake of robustness as we have already pointed out in Section 2.3.

4 VARIABLE SELECTION VIA PENALISED LIKELIHOOD METHOD

Since the JLSM includes three sub-models for each parameter, using more variables than required can be harmful to detecting suitable modelling. At this point, variable selection is as

important as parameter estimation in this model. The classical methods considered in the literature are based on the normality assumption, which is unrealistic. Parameter estimation and inference can fail in the presence of outliers. For this reason, we consider the penalised likelihood approach for t-JLSM. The penalised likelihood estimator of the unknown parameter vector $\boldsymbol{\theta}=(\theta_1, \ldots, \theta_s) = \left(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T\right)^T$ with $s = p + d + q$ is defined as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} S(\boldsymbol{\theta}), \tag{25}$$

where the objective function is

$$S(\boldsymbol{\theta}) = logL(\boldsymbol{\theta}) - m \sum_{k=1}^{s} p_{\tau_m}(|\theta_k|). \tag{26}$$

Here, $logL(\boldsymbol{\theta})$ is the log-likelihood function given in Equation (11) and $p_{\tau_m}(|\theta_k|)$ is the penalty term with the tuning parameter τ_m. Note that different penalty functions can be used for the parameter vectors of each sub-models but we prefer using the same penalty function for all the regression coefficients. With appropriate penalty functions, maximising $S(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ leads to certain parameter estimators vanishing from the initial models so that the unnecessary explanatory variables are automatically removed.

The idea behind the family of variable selection methods based on shrinkage methods is to add a penalty function to the negative log-likelihood and then minimise it (or maximise its negative). Such penalties have the property that small components of the parameter vector are completely minimised to zero. These methods differ from traditional subset selection approaches in that they delete unimportant variables if their coefficients are estimated to be zero under the chosen tuning parameters. Thus, selecting significant variables and estimating coefficients are carried out simultaneously. There is a variety of shrinkage methods available, which include but are not limited to the Bridge (Frank & Friedman, 1993), the LASSO (Tibshirani, 1996), and the SCAD (Fan & Li, 2001). Frank and Friedman (1993) suggested using the L_q penalty $p_\tau(|t|) = \tau|t|^q$, which leads to Bridge method. A special case of the Bridge penalty is the LASSO ($q = 1$) proposed by Tibshirani (1996). Fan and Li (2001) suggested using the SCAD penalty function, which is defined by

$$p_\tau(|t|) = \begin{cases} \tau|t| & \text{if } 0 \le |t| < \tau \\ -\left(|t|^2 - 2\alpha\tau|t| + \tau^2\right)/\{2(\alpha - 1)\} & \text{if } \tau \le |t| < \alpha\tau \\ (\alpha+1)\tau^2/2 & \text{if } |t| \ge \alpha\tau \end{cases} \tag{27}$$

where $\alpha > 2$ and $\tau > 0$ are tuning parameters. Note that the SCAD penalty function is symmetric, non-convex on $[0, \infty)$, and singular at the origin. In practice, one could search the best pair (τ, α) over the two-dimensional grids using some criteria, such as cross-validation (CV), generalised cross validation (GCV), Akaike information criteria (AIC), and Bayesian information criteria (BIC). In the simulation study and the real data example, we consider LASSO, SCAD and Bridge penalties. Fan and Li (2001) pointed out that choosing $\alpha = 3.7$ works reasonably well for SCAD penalty, so we follow their suggestion and use BIC to choose the other tuning parameters.

4.1 EM-type Algorithm for Parameter Estimation and Variable Selection

The penalised likelihood functions become non-differentiable at the origin and non-concave with respect to the parameters. These make it difficult to maximise the penalised likelihood

functions. Therefore, Fan & Li (2001) proposed the following local quadratic approximation (LQA) to approximate the penalty function at an initial value $\boldsymbol{\theta}^{(0)}=\left(\theta_1^{(0)}, \ldots, \theta_s^{(0)}\right)$ that is close to the true value of $\boldsymbol{\theta}$. Using this approximation, the maximisation problem given in (26) can be rewritten as

$$\hat{\boldsymbol{\theta}} = \arg\max S^*(\boldsymbol{\theta}) \tag{28}$$

with the objective function

$$S^*(\boldsymbol{\theta}) = logL(\boldsymbol{\theta}) - \frac{m}{2}\boldsymbol{\theta}^T\mathbf{W}\left(\boldsymbol{\theta}^{(0)}\right)\boldsymbol{\theta} \tag{29}$$

where $\mathbf{W}\left(\boldsymbol{\theta}^{(0)}\right) = diag\left\{\frac{p'_{\tau_m}\left(\left|\theta_k^{(0)}\right|\right)}{\left|\theta_k^{(0)}\right|}\right\}_{k=1}^{s}$. Steps 1 and 2 will be the same as given in Section 3.

Step 1. (E Step) Reform the objective function given in (23)

$$Q^*\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(h)}\right) = -\frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{m}\psi_i^{(h)}\Delta_i - \frac{m}{2}\boldsymbol{\theta}^T\mathbf{W}(\boldsymbol{\theta})\boldsymbol{\theta}. \tag{30}$$

Step 2. (M Step) Maximise the $Q^*\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(h)}\right)$ with respect to $\boldsymbol{\theta}$ to get the $(h+1)$-th estimates for the parameters, which yields the following updating equation for $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$:

$$\boldsymbol{\gamma}^{(h+1)} = \left(\sum_{i=1}^{m}\psi_i^{(h)}\mathbf{Z}_i^{(h)T}\left(\mathbf{D}_i^{(h)}\right)^{-1}\mathbf{Z}_i^{(h)} + \tau_m\mathbf{W}_{\boldsymbol{\gamma}}^{(h)}\right)^{-1}\left(\sum_{i=1}^{m}\psi_i^{(h)}\mathbf{Z}_i^{(h)T}\left(\mathbf{D}_i^{(h)}\right)^{-1}\tilde{\mathbf{r}}_i^{(h)}\right), \tag{31}$$

$$\boldsymbol{\lambda}^{(h+1)} = \left(\sum_{i=1}^{m}\psi_i^{(h)}\sum_{j=1}^{n_i}\mathbf{w}_j\mathbf{w}_j^T + 2\tau_m\mathbf{W}_{\boldsymbol{\lambda}}^{(h)}\right)^{-1}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}\psi_i^{(h)}\mathbf{w}_j\hat{\mathbf{D}}_i^{-1}\left(\boldsymbol{\epsilon}_i^{2(h)} - \Omega_i^2(h) - \mathbf{D}_i^{(h)}\log\Omega_i^2(h)\right)\right), \tag{32}$$

where $\tilde{\mathbf{r}}_i^{(h)} = \mathbf{Y}_i - \boldsymbol{\mu}_i^{(h)}$, $\boldsymbol{\epsilon}_i^{2(h)} = \left(\left(\tilde{r}_{i1}^{(h)} - r_{i1}^{(h)}\right)^2, \ldots, \left(\tilde{r}_{in_i}^{(h)} - r_{in_i}^{(h)}\right)^2\right)$ and $\Omega_i^2(h) = \left(\sigma_{i1}^{2(h)}, \ldots, \sigma_{in_i}^{2(h)}\right)^T$.

The second way to compute $\boldsymbol{\gamma}$ is to use the scoring procedure given in (20).

Using $\boldsymbol{\beta}^{(h)}$, $\boldsymbol{\gamma}^{(h+1)}$, $\boldsymbol{\lambda}^{(h+1)}$, update $\boldsymbol{\Sigma}_i^{(h+1)}$ and $\psi_i^{(h+1)}$. Then calculate $\boldsymbol{\beta}^{(h+1)}$ by using

$$\boldsymbol{\beta}^{(h+1)} = \left(\sum_{i=1}^{m}\psi_i^{(h)}\mathbf{X}_i^T\left(\boldsymbol{\Sigma}_i^{(h)}\right)^{-1}\mathbf{X}_i + \tau_m\mathbf{W}_{\boldsymbol{\beta}}^{(h)}\right)^{-1}\left(\sum_{i=1}^{m}\psi_i^{(h)}\mathbf{X}_i^T\left(\boldsymbol{\Sigma}_i^{(h)}\right)^{-1}\mathbf{Y}_i\right), \tag{33}$$

where $\mathbf{W}_{\boldsymbol{\beta}}^{(h)} = \mathbf{W}\left(\boldsymbol{\beta}^{(0)}\right)$, $\mathbf{W}_{\boldsymbol{\gamma}}^{(h)} = \mathbf{W}\left(\boldsymbol{\gamma}^{(0)}\right)$, and $\mathbf{W}_{\boldsymbol{\lambda}}^{(h)} = \mathbf{W}\left(\boldsymbol{\lambda}^{(0)}\right)$.

Step 3. Repeat E and M steps until the convergence rule $\boldsymbol{\theta}^{(h+1)} - \boldsymbol{\theta}^{(h)} < \boldsymbol{\epsilon}$ is satisfied.

Choosing the tuning parameters

In variable selection, choosing the tuning parameter is a crucial problem. In literature, there are many methods such as CV and GCV (Fan & Li, 2001; Tibshirani, 1996) to choose the tuning parameters. In our study, we determine the optimal value of tuning parameter by minimising BIC with the following formula as suggested by Wang *et al.* (2007):

$$BIC(\tau) = -\frac{2}{m}logL\left(\hat{\boldsymbol{\theta}}\right) + df_{\tau}\frac{\log(m)}{m} \tag{34}$$

where $\hat{\boldsymbol{\theta}}$ is the estimate of $\boldsymbol{\theta}$, df_τ $(0 < df_\tau < p+d+q)$ denotes the number of non-zero components of $\hat{\boldsymbol{\theta}}$, and $logL\left(\hat{\boldsymbol{\theta}}\right)$ is defined in Equation (11).

4.2 Theoretical Results

In this subsection, we provide some theoretical justifications. We first prove that the penalised log-likelihood function is increasing in each iteration using the EM-type algorithm. The consistency, sparsity and asymptotic normality of the penalised estimator $\hat{\boldsymbol{\theta}}$ will be established in our context. We only state the main results here and relegate the proofs to Appendix A.

Theorem 1. *Let $p_{\tau_m}(\cdot)$ is a differentiable concave penalty function on $[0, \infty)$, then the penalised log-likelihood function is increasing at each iteration of the EM algorithm.*

Let $\boldsymbol{\theta}_0$ be the true parameter vector. Partitation $\boldsymbol{\theta}_0$ as $\left(\left(\boldsymbol{\theta}_0^{(1)}\right)^T, \left(\boldsymbol{\theta}_0^{(2)}\right)^T\right)^T$ where $\boldsymbol{\theta}_0^{(1)}$ with the dimension s_1 is the vector of all non-zero components and $\boldsymbol{\theta}_0^{(2)}$ with the dimension s_2 is the vector of all zero components. Let $a_m = \max_{1 \le j \le s}\left\{p'_{\tau_m}\left(\left|\theta_{0j}\right|\right) : \theta_{0j} \neq 0\right\}$ and $b_m = \max_{1 \le j \le s}\left\{p''_{\tau_m}\left(\left|\theta_{0j}\right|\right) : \theta_{0j} \neq 0\right\}$. The $p'_{\tau_m}(\boldsymbol{\theta})$ and $p''_{\tau_m}(\boldsymbol{\theta})$ are the first and second derivatives of the function $p_{\tau_m}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. We have the following conditions on the penalty function

C1 For all m and τ_m, $p_{\tau_m}(0) = 0$, and $p_{\tau_m}(\boldsymbol{\theta})$ is symmetric, non-negative, non-decreasing and twice differentiable for all $\boldsymbol{\theta}$ in $(0, \infty)$ with at most a few exceptions.

C2 As $m \to \infty$, $b_m = o(1)$.

C3 For $T_m = \left\{\boldsymbol{\theta};\ 0 < \boldsymbol{\theta} \le m^{-1/2}\log m\right\}$, $\lim_{m\to\infty}\inf_{\boldsymbol{\theta} \in T_m} p'_{\tau_m}(\boldsymbol{\theta})/\sqrt{m} = \infty$.

These conditions guarantee $\sqrt{m}$-consistency of the estimators. The following assumptions are also needed:

A1 $\mathbf{Y}_i \sim t_{n_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, v)$, for each i.

A2 The covariates $\mathbf{x}_{ij}$, $\mathbf{z}_{jk}$, and $\mathbf{w}_j\,(i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, n_i,\ k = 1, 2, \ldots, j-1)$ are fixed and finite. The number of repeated measurements (n_i) are fixed.

A3 The parameter space is compact and the $\boldsymbol{\theta}_0$ is in the interior of the parameter space.

The following theorems states the consistency, sparsity and asymptotic normality of $\hat{\boldsymbol{\theta}}$

Theorem 2. *Assume that $a_m = O_p\left(m^{-1/2}\right)$, $b_m \to 0$, and $\tau_m \to 0$ as $m \to \infty$. Under the conditions (A1)-(A3), with probability tending to 1 there must exist a local maximiser $\hat{\boldsymbol{\theta}}_m$ of the penalised likelihood function $S^*(\boldsymbol{\theta})$ in (29) such that $\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0 = O_p\left(m^{-1/2}\right)$.*

Let $A_m = \mathrm{diag}\left(p''_{\tau_m}\left(\left|\theta_{01}^{(1)}\right|\right), , \ldots, p''_{\tau_m}\left(\left|\theta_{0s_1}^{(1)}\right|\right)\right)$, $c_m = \left(p'_{\tau_m}\left(\left|\theta_{01}^{(1)}\right|\right)sgn\left(\theta_{01}^{(1)}\right), \ldots, p'_{\tau_m}\left(\left|\theta_{0s_1}^{(1)}\right|\right)sgn\left(\theta_{0s_1}^{(1)}\right)\right)^T$, $\theta_{0j}^{(1)}$ is the j-th component of $\boldsymbol{\theta}_0^{(1)}$, and $F_m(\boldsymbol{\theta})$ represents the Fisher information matrix of $\boldsymbol{\theta}$.

Theorem 3. *Assume that the conditions in Theorem 1 are satisfied, and the function $p_{\tau_m}(\boldsymbol{\theta})$ satisfies conditions C1–C3. If the penalty function has* $\liminf_{m\to\infty}\liminf_{t\to 0^+}\frac{p'_{\tau_m}(t)}{\tau_m} > 0$ *when $\tau_m \to 0$ and $\sqrt{m}\tau_m \to \infty$ as $m \to \infty$, then for any $\sqrt{m}$ -consistent estimator $\hat{\boldsymbol{\theta}}_m$ of $\boldsymbol{\theta}$, as $m \to \infty$, we have*

(i) Consistency in the variable selection: $P\left(\hat{\boldsymbol{\theta}}_m^{(2)} = 0\right) \to 1$, (ii) Asymptotic normality:

$$\sqrt{m\left(\mathbf{F}_m^{(1)}\right)}\left(\mathbf{F}_m^{(1)}+A_m\right)\left\{\left(\hat{\boldsymbol{\theta}}_m^{(1)}-\boldsymbol{\theta}_0^{(1)}\right)+\left(\mathbf{F}_m^{(1)}+A_m\right)^{-1}c_m\right\}\underset{\rightarrow}{D}N_{s_1}(\boldsymbol{\theta},\boldsymbol{I}_{s_1}),$$

where "$\underset{\rightarrow}{D}$" stands for the convergence in distribution; $\mathbf{F}^{(1)}$ is the $(s_1 \times s_1)$ submatrix of n corresponding to the non-zero components $\boldsymbol{\theta}_0^{(1)}$ and $\boldsymbol{I}_{s_1}$ is the $(s_1 \times s_1)$ identity matrix.

5 SIMULATION STUDY

In this section, a simulation study is performed to compare the performance of the proposed methods in terms of estimation and variable selection over the ML estimation method. All simulations are conducted using R [?]. We use the same design as in the study of (Xu *et al.*, 2013). The true values of the parameters in the mean, generalised autoregressive parameters, and log-innovation variances are $\boldsymbol{\beta}=[1,\ -0.5, 0,\ 0.5, 0,\ 0, 0]^T$, $\boldsymbol{\gamma}=[-0.3, 0.3, 0,\ 0, 0]^T$ and $\boldsymbol{\lambda}=[0, 0.5, 0.4, 0,\ 0]^T$, respectively. In the models, $x_{ij}=\left(1,\ x_{ijt}\right)$ are generated from a 7-variate multivariate normal distribution with mean zero, marginal variance 1, and all correlations 0.5. We then form the covariates $w_{ij}=\left(x_{ijt}\right)_{t=1}^{5}$ and $z_{ijk}=\left(1,\ t_{ij}-t_{ik},\ (t_{ij}-t_{ik})^2,\ \ldots,\ (t_{ij}-t_{ik})^4\right)$ for the log-innovation variances and the generalised autoregressive parameters with the measurement times t_{ij}, which are generated from the uniform distribution $U(0,2)$. Note that $\mathbf{z}(i,\ 1)=\mathbf{0}$ so that the first row of $\mathbf{Z}_i$ is zero. Using these values, the mean $\boldsymbol{\mu_i}$ and covariance matrix Σ_i are constructed through the MCD. The initial values of parameters for EM are gathered from ordinary least squares as given in (Pan & Pan, 2017).

We carry on the theoretical part of the paper for assuming different n_i for each subject, but we only consider the balanced models ($n_i = n$) in simulation and in our real data for the ease of computation. We simulate 100 data sets for each setting with sample sizes m=100, 200 and 400. We simulate m subjects, each has $n = 12$ observations drawn from the $t_{12}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, v)$. In the simulation study, the degrees of freedom v of the t-distribution are fixed and set to 3, as recommended for the sake of robustness in the literature (see Lange *et al.*, 1989; Arslan & Genç, 2003; Arslan & Genc, 2009). For SCAD, the tuning parameter α is taken as 3.7 as suggested by Fan & Li (2001), and the parameter τ_m is selected by BIC. Figure 1 represents the distribution of y-values without contamination for one of our generated data sets. The fluctuations between and within observations repetitions encourage us to use a model including these variations.

The variable selection performance is assessed by the proportion of times that the correct model is selected (CF). We compared model errors of different estimators by the square root of the median of model error $RMME=\sqrt{Median\left\{\left(\hat{\boldsymbol{\theta}}_M-\boldsymbol{\theta}_0\right)^T\left(\hat{\boldsymbol{\theta}}_M-\boldsymbol{\theta}_0\right)\right\}}$. Here, $\hat{\boldsymbol{\theta}}_M$ is the estimate of the parameter vector obtained from M-th simulated data set of 100 data sets. Table 1 presents the simulation results with the CF and RMME.

Figure 2 depicts the estimates of non-zero β and one of the zero βs with MLE and shrinkage methods for each simulation. It is observed that MLE (green dots) are further away from the actual values (horizontal blue line), which indicates the biases, compared to the shrinkage methods. Figure 3 and Figure 4 show the estimation of all γ and λ values for MLE and shrinkage methods, respectively. We can again notice that the MLE (green dots) are further away from the actual values (horizontal blue line) compared to the shrinkage methods.

We also aimed to investigate how the simulation results varied when estimating parameters using a normal distribution approach suggested by Kou & Pan (2009). To achieve this, we employed $v = 50$ to ensure that the estimation was conducted using a normal distribution rather than a t-distribution with the same simulated data. The findings are presented in Table 2 below.

The analysis reveals that the simulation results obtained from the normal distribution do not outperform those obtained from the t-distribution, as documented in Table 1.

We would further like to explore the robustness of the proposed methods against to the different outlier scenarios. The contamination scenarios are designed as follows:

Scenario 1: 100 added to target Mean values of randomly chosen 10% of y_i,

Scenario 2: 100 added to target Sigma values of randomly chosen 10% of y_i,

Scenario 3: Randomly chosen 10% of the μ_i values are generated from different Betas reported below: $\boldsymbol{\beta}_{new}=[1, 0, 100, 100, 0, 0, 0]^T$.

Scenario 4: 100 added to target mean values of randomly chosen 10% of X.

Figure 5 shows how data changes with each scenario. It can be observed from these graphs that Scenario 1 and 2 deviate significantly from the original data, while 3 and 4 do not cause significant contamination.

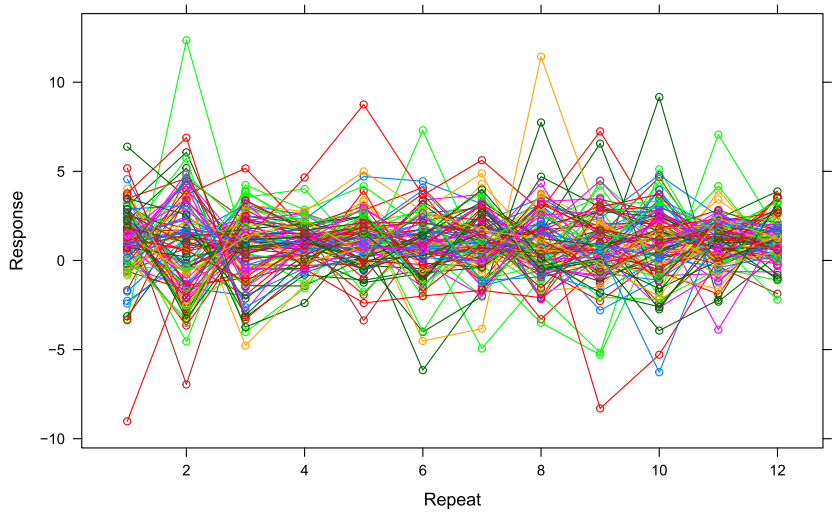

FIGURE 1. *An example of response values without perturbation.*

Table 1. *Simulation results without contamination.*

		LASSO		*SCAD*		*Bridge*	
	n	*CF*	*RMME*	*CF*	*RMME*	*CF*	*RMME*
β	100	96	0.0299	95	0.0315	96	0.0276
	200	100	0.0207	99	0.0180	99	0.0199
	400	100	0.0145	100	0.0134	100	0.0134
γ	100	100	0.0235	100	0.0252	100	0.0235
	200	100	0.0166	100	0.0172	100	0.0168
	400	100	0.0109	100	0.0117	100	0.0106
λ	100	84	0.1186	84	0.1205	84	0.1187
	200	99	0.0854	99	0.0857	99	0.0854
	400	100	0.0526	100	0.0521	100	0.0532

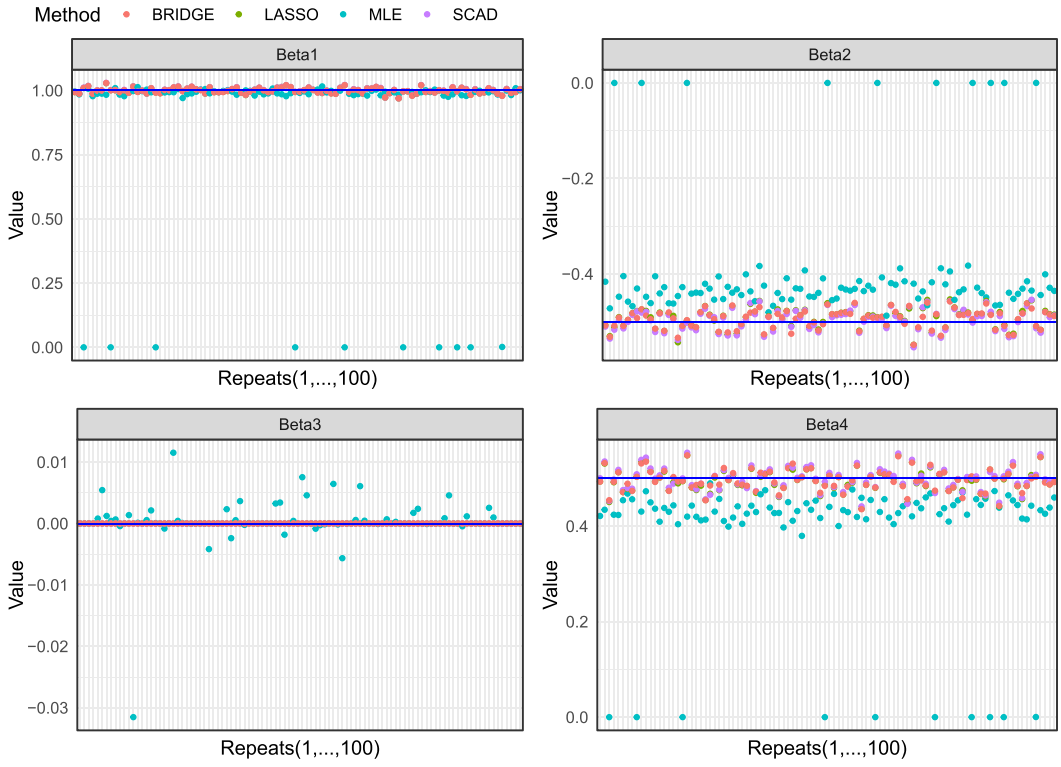

FIGURE 2. *β estimations (m = 100).*

The CF and RMME are reported in Tables 3 and 4 for contamination Scenario 1, 2, 3 and 4.

When considering contamination scenarios, it is observed that the first two scenarios induce greater distortion to the data than its normal appearance. In contamination scenarios 1 and 2, although the accuracy of estimating the $\boldsymbol{\beta}$ parameter does not change significantly compared to the original simulations reported in Table 1, there is a decrease in the estimation of $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ for small sample sizes. In scenarios 3 and 4, as expected, there is no significant change in the estimation. It is crucial to comprehend the modifications in the parameter estimates obtained from the MLEs and shrinkage methods in response to these scenarios. Figures 6 and 7 illustrate the differences in the $\boldsymbol{\beta}$ estimations between MLE and shrinkage methods for Scenario 1 and 2, respectively. The same plots for Scenario 3 and 4 was not added to save some space, since they are similar to the plots of the data without contamination. The results demonstrate that the MLEs deviate from the true values of $\boldsymbol{\beta}$ when the data are contaminated, while the shrinkage methods still provide reliable estimates.

6 REAL DATA ANALYSIS

In this section, we apply the proposed method with LASSO, SCAD and Bridge to Framingham Cholesterol data set (*qrLMM* package in R, Galarza and Lachos, 2017) that was used to analyse the progression of cardiovascular disease and the role of serum cholesterol as a risk factor in 200 randomly selected individuals over a period of time. The data set includes the following variables: *ID* is the subject number in the population, *cholst* is the cholesterol level for each patient that was measured at the beginning of the study and at 2-year intervals for 10 years, *year*

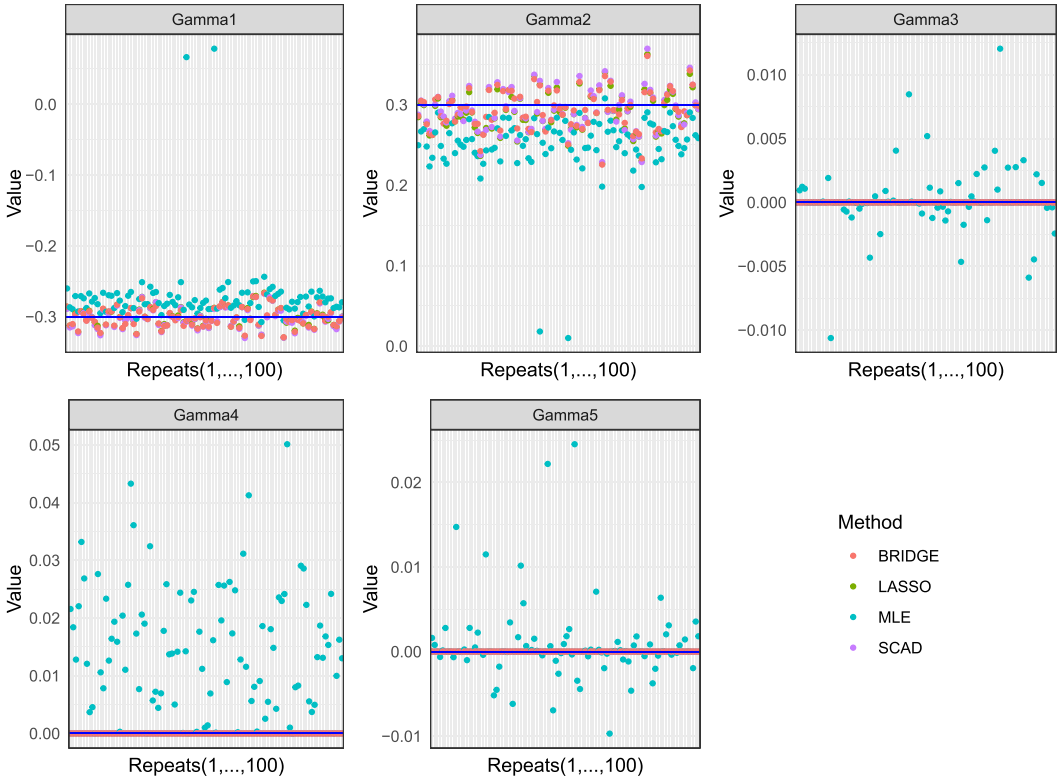

FIGURE 3. *γ estimations (n = 100).*

represents the number of years that have passed since the beginning of the study up to the time of the current measurement, *age* at baseline and *sex*. We included the complete measurements collected over ten years for individuals who remained in the study for the entire duration. In the study of Zhang & Davidian (2001), they also used Cholesterol Data for their proposed method which relaxed the normality assumption for random effects by introducing a semi-nonparametric representation of Gallant & Nychka (1987) in linear mixed models. For these data (Figure 8), we used robust modelling and combined the variable selection with it to carry on robust estimation and variable selection, simultaneously.

To investigate these data preliminarily, we utilised a linear regression model over the years for each individual, and the intercept and slope with their corresponding confidence intervals are shown in Figure 9. Differences in the estimates for intercept and slope in the plot suggest the need to use a model such as mean-covariance that incorporates differences between each of the observations, as opposed to the traditional regression model approach. Furthermore, Figure 9 reveals a grouping by gender, particularly in the intercept graphic on the left. In this case, it will be necessary to add a gender variable and gender-year interaction to the model in addition to the age covariate with all interactions.

When examining the values of standardised residuals in Figure 10, it was observed that there are outliers in both gender groups in the available data, and when Figure 11 is examined, it cannot be concluded that the normality assumption is valid. Therefore, a model that takes into

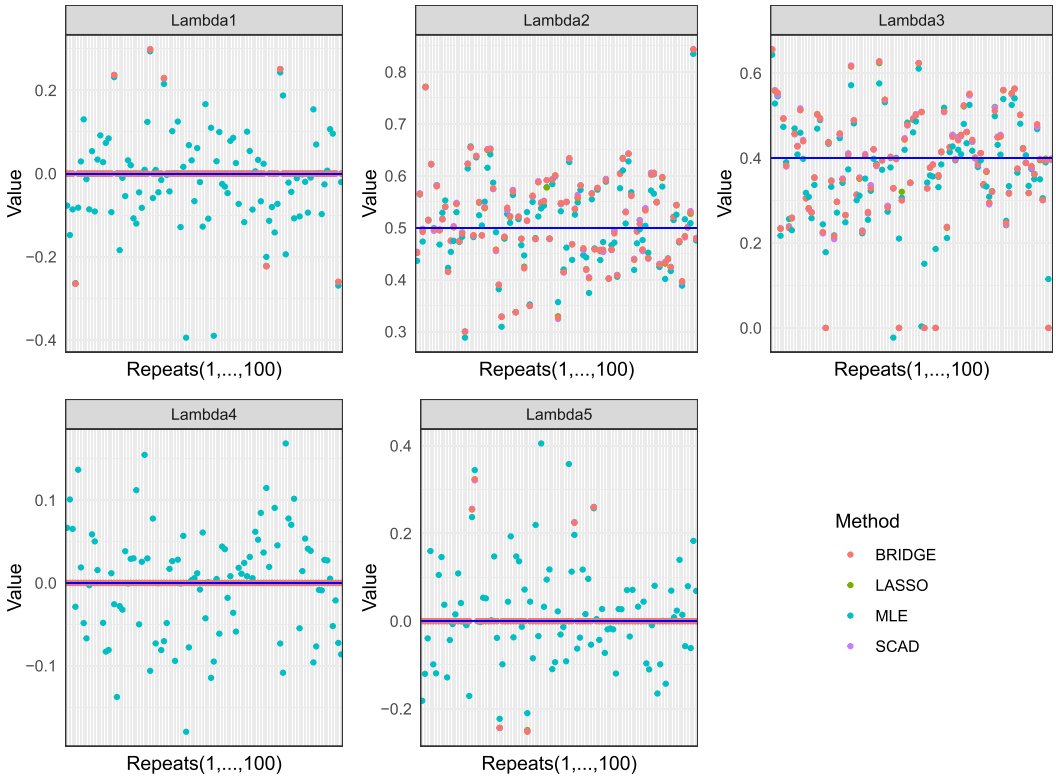

FIGURE 4. *λ estimations (n = 100).*

Table 2. *Simulation results obtained from v = 50 without contamination.*

		LASSO		*SCAD*		*Bridge*	
	n	*CF*	*RMME*	*CF*	*RMME*	*CF*	*RMME*
$\boldsymbol{\beta}$	100	90	0.0326	92	0.0284	85	0.0292
	200	81	0.0279	92	0.0202	88	0.0214
	400	70	0.0373	93	0.0145	91	0.0161
$\boldsymbol{\gamma}$	100	93	0.0253	96	0.0234	87	0.0265
	200	78	0.0242	92	0.0177	90	0.0176
	400	66	0.0284	93	0.0145	89	0.0142
$\boldsymbol{\lambda}$	100	87	0.1165	86	0.1088	81	0.1148
	200	74	0.0863	91	0.0740	88	0.0740
	400	58	0.0943	92	0.0498	89	0.0504

account both within-group and between-group variability while also incorporating outlier values and a heavy-tailed data structure would be appropriate for these data. For this reason, we think that t-JLSM can be used to model this data set. In particular we assume that the cholesterol level $\mathbf{Y_i}$ has a $t_n(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}, v)$ with the following sub-models for $i = 1, \ldots, 174$ and $j = 1, \ldots, 6$:

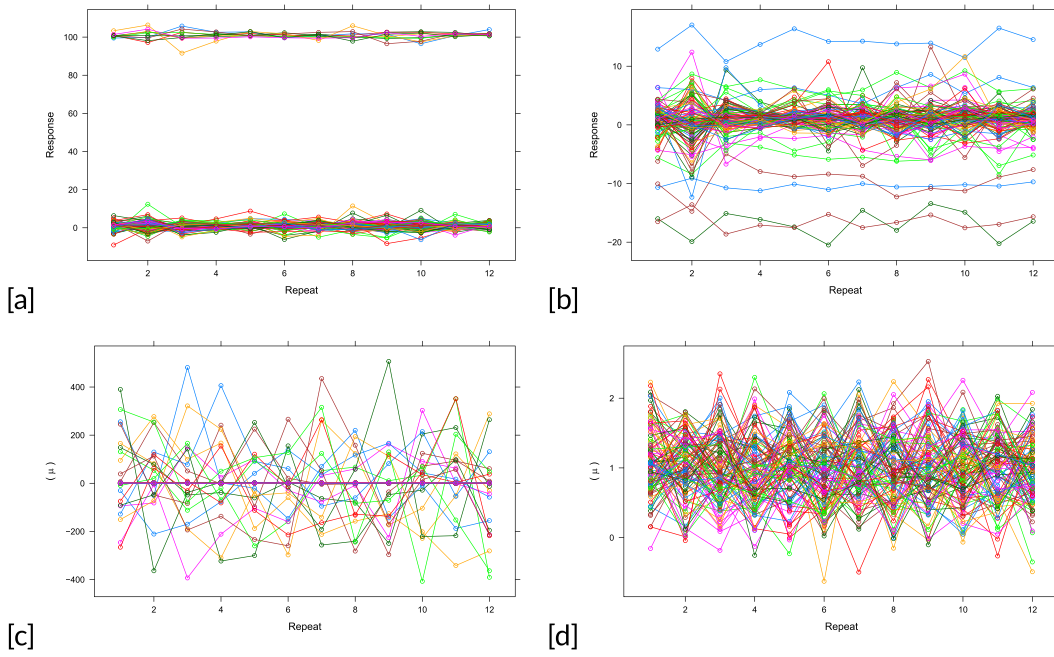

FIGURE 5. *(a) Scenario 1. (b) Scenario 2. (c) Scenario 3. (d) Scenario 4.*

Table 3. *Simulation results for Scenario I and II.*

		Scenario 1						*Scenario 2*					
		LASSO		*SCAD*		*Bridge*		*LASSO*		*SCAD*		*Bridge*	
	n	*CF*	*RMME*	*CF*	*RMME*	*CF*	*RMME*	*CF*	*RMME*	*CF*	*RMME*	*CF*	*RMME*
β	100	92	0.0326	89	0.0327	88	0.0312	95	0.0331	90	0.0366	90	0.0317
	200	97	0.0254	97	0.0240	97	0.0256	99	0.0250	97	0.0232	98	0.0251
	400	100	0.0201	100	0.0167	100	0.0196	100	0.0194	100	0.0165	100	0.0189
γ	100	93	0.1034	94	0.0994	96	0.1014	94	0.0943	68	0.1584	96	0.0918
	200	89	0.1396	91	0.1395	80	0.1401	93	0.1271	93	0.1222	93	0.1236
	400	93	0.0754	94	0.0720	98	0.0726	95	0.0698	95	0.0661	98	0.0660
λ	100	68	0.1558	70	0.1528	68	0.1531	69	0.1524	65	0.1459	69	0.1515
	200	69	0.1708	71	0.1666	63	0.1858	75	0.1574	75	0.1582	75	0.1577
	400	100	0.1339	100	0.1330	100	0.1345	100	0.1285	100	0.1277	100	0.1284

$$\begin{aligned}\mu_{ij} &= \beta_0 + \beta_1 Gender_i + \beta_2 Age_i + \beta_3 t_{ij} + \beta_4 t_{ij}^2 \\ &\quad + \beta_5 Gender_i \times Age_i + \beta_6 Age_i \times t_{ij} + Gender_i \times t_{ij}\end{aligned} \tag{35}$$

Here, z_{jk} in ϕ_{jk} are generated as in the simulation study, and w_j in innovation variance are taken as $w_j = [1 j j^2 ... j^4]$. Similar to in the simulation case, we are taken the degrees of freedom v as 3 (see Lange *et al.*, 1989; Arslan & Genç, 2003; Arslan & Genc, 2009).

Three main parameters βs, γs and λs are under consideration. It is not anticipated for this data set to have a quadratic relationship with respect to year; however, the inclusion of the

Table 4. *Simulation results for Scenario III and IV.*

		Scenario 3						*Scenario 4*					
		LASSO		*SCAD*		*Bridge*		*LASSO*		*SCAD*		*Bridge*	
	n	*CF*	*RMME*	*CF*	*RMME*	*CF*	*RMME*	*CF*	*RMME*	*CF*	*RMME*	*CF*	*RMME*
β	100	97	0.0315	93	0.0328	93	0.0321	100	0.1119	98	0.1059	98	0.0447
	200	100	0.0225	100	0.0214	100	0.0201	100	0.0786	100	0.0731	100	0.0253
	400	100	0.0173	100	0.0146	100	0.0162	100	0.0591	100	0.0598	100	0.0105
γ	100	100	0.0300	100	0.0299	100	0.0299	100	0.0426	100	0.0430	100	0.0268
	200	100	0.0549	100	0.0497	100	0.0521	100	0.0388	100	0.0398	100	0.0152
	400	100	0.0320	100	0.0251	100	0.0277	100	0.0347	100	0.0336	100	0.0084
λ	100	83	0.1367	89	0.1230	89	0.1235	87	0.1197	86	0.1178	85	0.1121
	200	96	0.1088	96	0.1077	96	0.1081	98	0.0825	98	0.0838	99	0.0811
	400	100	0.0961	100	0.0945	100	0.0957	100	0.0579	100	0.0586	100	0.0547

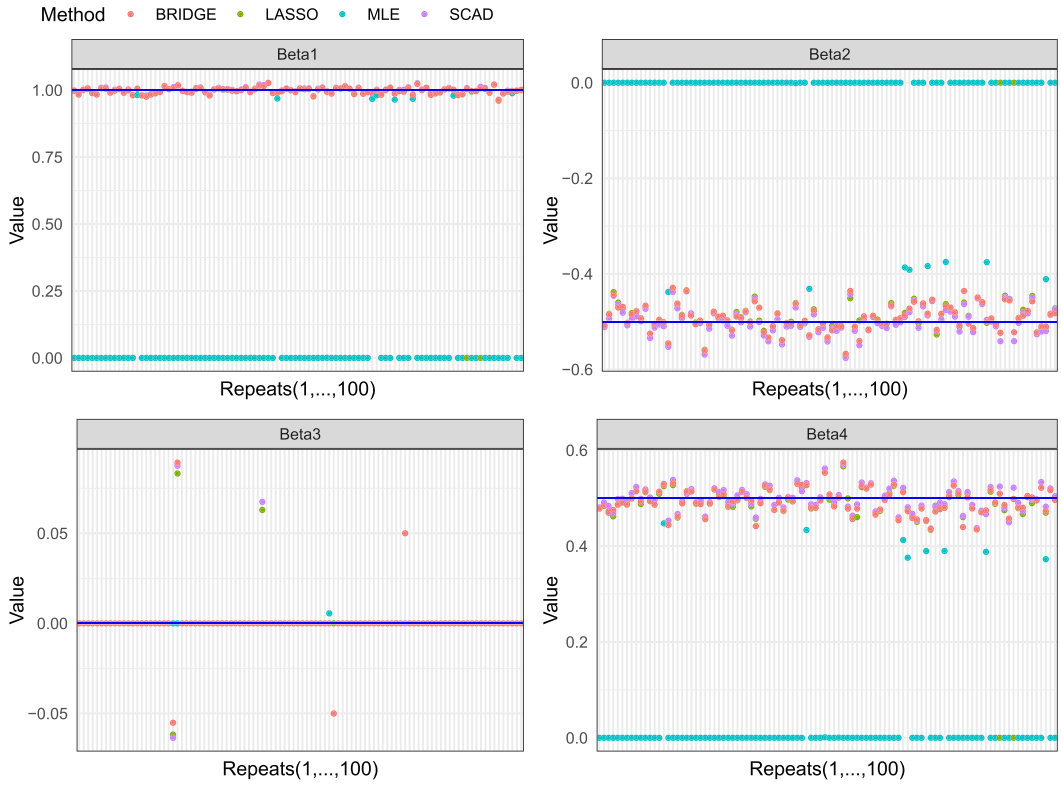

FIGURE 6. *β estimations scenario 1 ($n = 100$).*

corresponding quadratic term and β_4 parameter was deemed necessary to assess the ability of our proposed method in selecting this variable as insignificant. The estimation of parameters and the selection of variables for Cholesterol data were executed using the EM steps as outlined in Section 4.1. Additionally, the mean prediction errors (MPE) are calculated by applying CV,

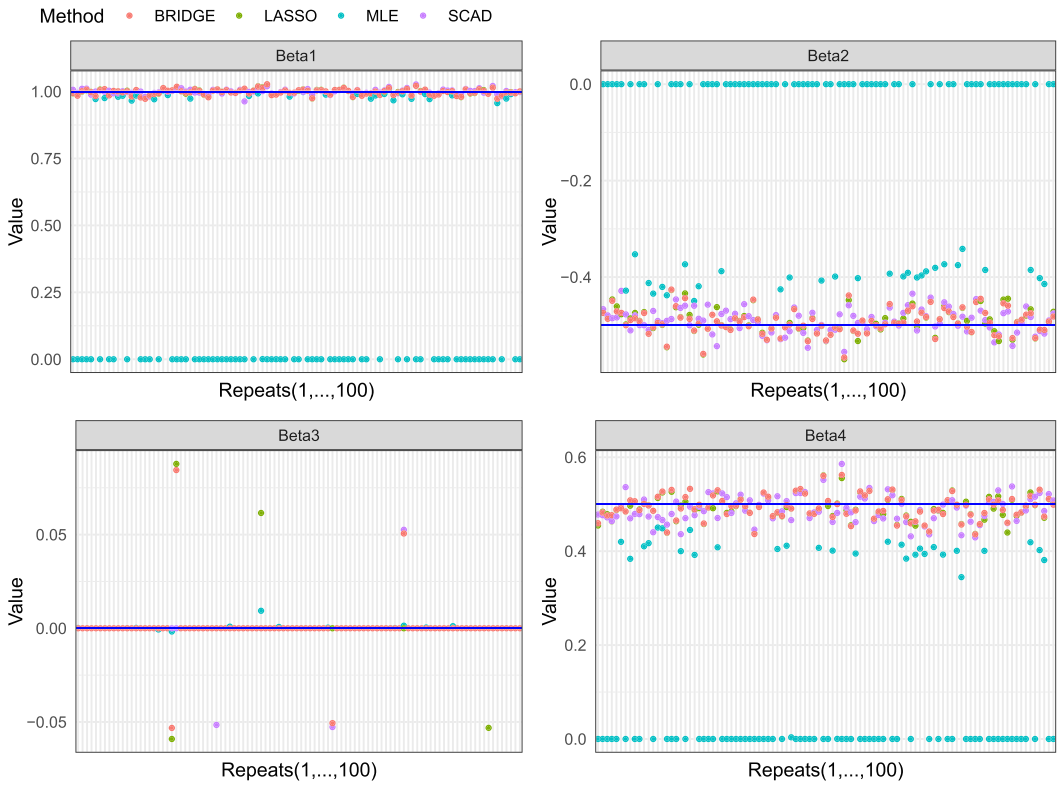

FIGURE 7. *β estimations scenario 2 (n = 100).*

80% of the data being the training data with 10 repeats. The formula of MPE for each CV set is given below

$$MPE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{\Sigma}}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i). \tag{36}$$

The real data results are summarised in Table 5. According to the real data results, the last value of the $\boldsymbol{\lambda}$ is determined as zero from all three shrinkage methods. The model for the innovation variance is chosen as third-order polynomial model. While we thought it would be a fourth-order polynomial model, it came up as a cubic polynomial in time. β_4 is determined as zero that indicates the quadratic term for year is not necessary in the model. SCAD and Bridge methods agree on variable selection, while LASSO forces the quadratic form of the year to be included in the model with a small contribution for this data set. Similar to the simulation results, SCAD also provided the best result followed by Bridge for the real data set in the smallest BIC and MPE.

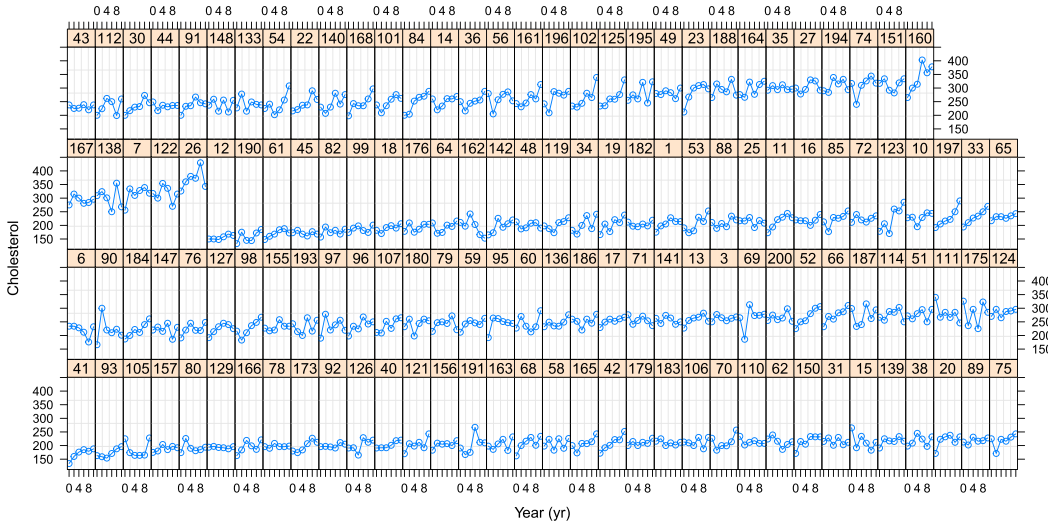

FIGURE 8. *Cholesterol levels over time for each subject in Framingham Cholesterol data set.*

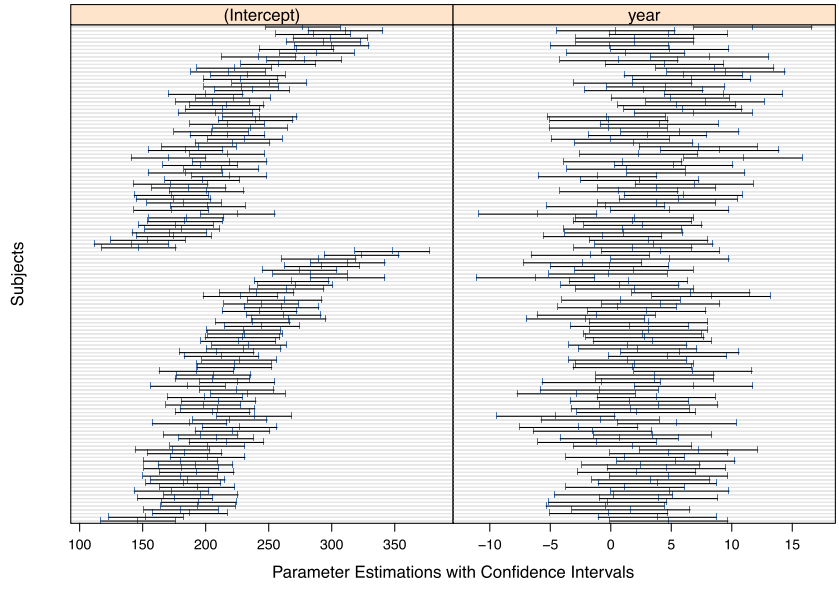

FIGURE 9. *Parameter estimations of each subject for intercept on the left and slope on the right with their confidence intervals.*

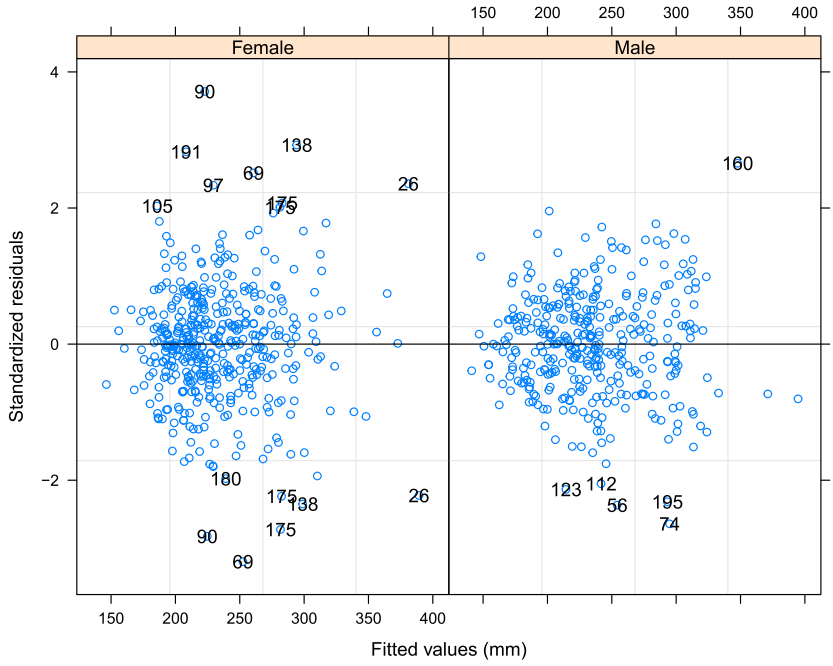

FIGURE 10. *Standardised residuals for each subject by gender groups.*

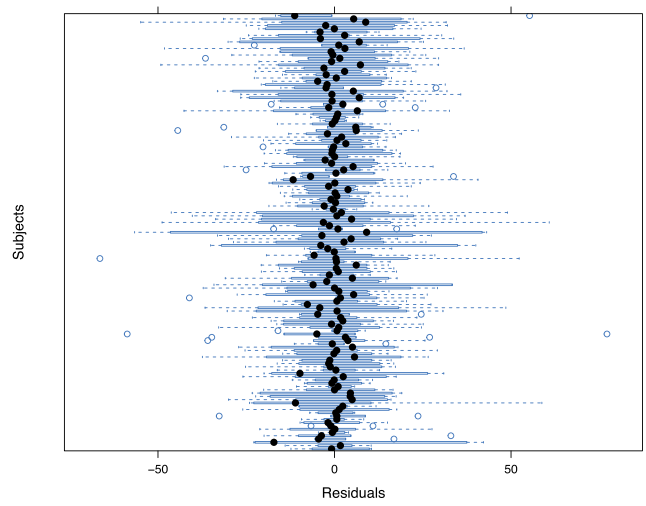

FIGURE 11. *Residuals of LM for each subject.*

7 DISCUSSION

The main contribution of this paper can be summarised as follows. Using the scale mixture representation of the t-distribution, we have proposed an EM-type algorithm to compute the estimates. The simultaneous variable selection and robust parameter estimation have been combined with the EM algorithm for the t-JLSMs. In this EM-type algorithm, the M-step has been

Table 5. *Cholesterol data parameter estimations.*

	MLE	*LASSO*	*SCAD*	*Bridge*
β_0	183.968	177.552	182.978	180.488
β_1	−94.835	−88.989	−94.975	−93.262
β_2	0.768	0.891	0.782	0.822
β_3	5.234	5.342	5.143	5.053
β_4	0.016	0.022	0.000	0.000
β_5	2.072	1.959	2.095	2.045
β_6	−0.078	−0.078	−0.071	−0.072
β_7	1.257	1.241	1.231	1.356
γ_1	0.251	0.196	0.333	0.164
γ_2	0.323	−0.093	−0.475	0.150
γ_3	−0.027	0.036	0.699	1.461
γ_4	−0.113	0.030	3.231	−0.813
γ_5	0.007	0.000	1.812	−2.628
λ_1	5.611	5.423	5.592	5.416
λ_2	2.523	2.508	2.526	2.514
λ_3	−1.608	−1.432	−1.583	−1.414
λ_4	0.352	0.285	0.336	0.277
λ_5	−0.025	0.000	0.000	0.000
BIC	57.461	57.760	57.350	57.693
MPE	11.159	11.193	10.834	10.921

implemented via weighted least squares, with weights computed at the E-step as the expectation of independent Gamma variables. The variable selection has been carried on using the penalised likelihood method based on LASSO, SCAD, and Bridge penalties to choose important variables in t-JLSMs. We have further explored the asymptotic properties of the proposed estimators and shown the consistency, sparsity and the asymptotic normality of the estimator for the parameters of t-JLSMs.

The simulation results and real data have yielded results in favour of our proposed method. While simulations have shown accurate results to select for all non-zero parameters, a decrease in the accuracy of selecting $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ parameters has been observed as the data becomes contaminated. However, this degradation disappears as the data size increases. Overall, SCAD provides more accurate results followed by Bridge in variable selection for these types of models for repeated data.

ACKNOWLEDGEMENTS

We thank the editor-in-chief, associate editor, and two anonymous referees for their thorough reading of the former version of the paper. Their helpful comments and numerous suggestions led to a considerable improvement of the manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

Antoniadis, A. (1997). Wavelets in statistics: a review. *J. Italian Stat. Soc.*, **6**(2), 97–130.

Arslan, O. (2004). Convergence behavior of an iterative reweighting algorithm to compute multivariate m-estimates for location and scatter. *J. Stat. Plan. Inference*, **118**(1-2), 115–128.

Arslan, O. & Genc, A.I. (2009). The skew generalized t distribution as the scale mixture of a skew exponential power distribution and its applications in robust estimation. *Statistics*, **43**(5), 481–498.

Arslan, O. & Genç, ⊠.A. (2003). Robust location and scale estimation based on the univariate generalized t (gt) distribution. *Commun. Stat.-Theory Methods*, **32**(8), 1505–1525.

Cantoni, E. (2004). A robust approach to longitudinal data analysis. *Canadian J. Stat.*, **32**(2), 169–180.

Croux, C., Gijbels, I. & Prosdocimi, I. (2012). Robust estimation of mean and dispersion functions in extended generalized additive models. *Biometrics*, **68**(1), 31–44.
Fan, J., Huang, T. & Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Am. Stat. Assoc.*, **102**(478), 632–641.
Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, **96**(456), 1348–1360.
Fan, J. & Wu, Y. (2008). Semiparametric estimation of covariance matrixes for longitudinal data. *J. Am. Stat. Assoc.*, **103**(484), 1520–1533.
Fan, J. & Zhang, J.-T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J. R. Stat. Soc.: Series B (Stat. Methodol.)*, **62**(2), 303–322.
Frank, L.L.E. & Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35** (2), 109–135.
Galarza, C.E. & Lachos, V.H. (2017). qrlmm: Quantile regression for linear mixed-effects models. *R Package Version*, **1**.
Gallant, A.R. & Nychka, D.W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica: J. Econ. Soc.*, **1987**, 363–390.
Guney, Y., Arslan, O. & Yavuz, F.G. (2022). Robust estimation in multivariate heteroscedastic regression models with autoregressive covariance structures using em algorithm. *J. Multivar. Anal.*, **2022**, 105026.
He, X., Fung, W.K. & Zhu, Z. (2005). Robust estimation in generalized partial linear models for clustered data. *J. Am. Stat. Assoc.*, **100**(472), 1176–1184.
Huang, J.Z., Liu, L. & Liu, N. (2007). Estimation of large covariance matrices of longitudinal data with basis function approximations. *J. Comput. Graph. Stat.*, **16**(1), 189–209.
Huang, J.Z., Liu, N., Pourahmadi, M. & Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, **93**(1), 85–98.
Jhong, J.-H., Lee, J., Kim, S. & Koo, J.-Y. (2017). Joint modeling for mean vector and covariance estimation with l 1-penalty. *Quant. Bio-Sci.*, **36**(1), 33–38.
Jørgensen, B. & Petersen, H.C. (2012). Efficient estimation for incomplete multivariate data. *J. Stat. Plan. Inference*, **142**(5), 1215–1224.
Kou, C. & Pan, J. (2009). Variable selection for joint mean and covariance models via penalized likelihood. *Manchester Institute for Mathematical Sciences*.
Kou, C. & Pan, J. 2020. Variable selection in joint mean and covariance models. Recent Developments in Multivariate and Random Matrix Analysis: Festschrift in Honour of Dietrich von Rosen, 219–244.
Lange, K.L., Little, RJA & Taylor, J.eremyM.G. (1989). Robust statistical modeling using the t distribution. *J. Am. Stat. Assoc.*, **84**(408), 881–896.
Leng, C., Zhang, W. & Pan, J. (2010). Semiparametric mean–covariance regression analysis for longitudinal data. *J. Am. Stat. Assoc.*, **105**(489), 181–193.
Levina, E., Rothman, A. & Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *Ann. Appl. Stat.*, **2**(1), 245–263.
Liang, K.-Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.
Lin, S.P. (1985). A monte carlo comparison of four estimators of a covariance matrix. *Multivar. Anal.*, **1985**, 411–429.
Lin, T.-I. & Wang, Y.-J. (2009). A robust approach to joint modeling of mean and scale covariance for longitudinal data. *J. Stat. Plan. Inference*, **139**(9), 3013–3026.
Pan, J. & Mackenzie, G. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika*, **90**(1), 239–244.
Pan, J. & Pan, Y. (2017). jmcm: An r package for joint mean-covariance modeling of longitudinal data. *J. Stat. Softw.*, **82**, 1–29.
Pan, J. & Ye, H. (2004). Modelling covariance structures in generalized estimating equations for longitudinal data. *Model. Covariance Struct. Gener. Estimat. Equ. Longit. Data*, 1000–1005.
Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, **86**(3), 677–690.
Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, **87**(2), 425–435.
Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*, Vol. **882**. John Wiley & Sons.
Qin, G. & Zhu, Z. (2007). Robust estimation in generalized semiparametric mixed models for longitudinal data. *J. Multivariate Anal.*, **98**(8), 1658–1683.

Qin, G., Zhu, Z. & Fung, W.K. (2009). Robust estimation of covariance parameters in partial linear model for longitudinal data. *J. Stat. Plan. Inference*, **139**(2), 558–570.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.: Ser. B (Methodol.)*, **58**(1), 267–288.
Wang, H., Li, R. & Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**(3), 553–568.
Wang, N., Carroll, R.J. & Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Am. Stat. Assoc.*, **100**(469), 147–157.
Wong, F., Carter, C.K. & Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, **90**(4), 809–830.
Wu, W.B. & Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, **90**(4), 831–844.
Xu, D., Zhang, Z. & Wu, L. (2013). Joint variable selection of mean-covariance model for longitudinal data. *Open J. Stat*.
Zhang, D. & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, **57**(3), 795–802.
Zheng, X., Fung, W.K. & Zhu, Z. (2014). Variable selection in robust joint mean and covariance model for longitudinal data analysis. *Stat. Sinica*, **2014**, 515–531.
Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.*, **101**(476), 1418–1429.

APPENDIX A:

In this section, we provide some theoretical justifications. We first prove that the penalised log-likelihood function is increasing in each iteration using the EM-type algorithm. The consistency, sparsity and asymptotic normality of the penalised estimator $\hat{\boldsymbol{\theta}}$ will be established in our context.

Theorem 1 Let $p_{\tau_m}(\cdot)$ is a differentiable concave penalty function on $[0, \infty)$, then the penalised log-likelihood function is increasing at each iteration of the EM algorithm.

Proof of Theorem *The proof of Theorem 1 is very similar to the proof given by Arslan (2004). Here, we will outline the proof.*

Let consider $S(\boldsymbol{\theta})$ *and define*

$$\Phi\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}\right) = -\sum_{i=1}^{m} \log|\boldsymbol{\Sigma}_i| - \sum_{i=1}^{m} \hat{\psi}_i \Delta_i - \sum_{k=1}^{s} p_{\tau_m}\left(\left|\hat{\theta}_k\right|\right) + p'_{\tau_m}\left(\left|\hat{\theta}_k\right|\right)\left(\theta_k - \hat{\theta_k}\right). \tag{A1}$$

Then using estimate $\boldsymbol{\theta}^{(h)}$ *at the* $(h+1)$*-th step of the algorithm, we have*

$$\boldsymbol{\theta}^{(h+1)} = \underset{\boldsymbol{\theta} \in S_p}{\arg\max}\, \Phi\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(h)}\right). \tag{A2}$$

Here, we need to prove that

$$S\left(\boldsymbol{\theta}^{(h+1)}\right) \geq S\left(\boldsymbol{\theta}^{(h)}\right). \tag{A3}$$

Let consider at the h-th step

$$S(\boldsymbol{\theta}) - \Phi\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(h)}\right) = \sum_{k=1}^{s} p_{\tau_m}\left(\left|\theta_k^{(h)}\right|\right) + p'_{\tau_m}\left(\left|\theta_k^{(h)}\right|\right)\left(\theta_k - \theta_k^{(h)}\right) - p_{\tau_m}(|\theta_k|). \tag{A4}$$

Since $p_{\tau_m}(\cdot)$ is concave, then $S(\boldsymbol{\theta}) \geq \Phi\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(h)}\right)$. By considering the h-th step in this result, we have $S\left(\boldsymbol{\theta}^{(\boldsymbol{h})}\right) \geq \Phi\left(\boldsymbol{\theta}^{(\boldsymbol{h})}, \boldsymbol{\theta}^{(h)}\right)$. Then we have

$$S\left(\boldsymbol{\theta}^{(h+1)}\right) \geq \Phi\left(\boldsymbol{\theta}^{(h+1)}, \boldsymbol{\theta}^{(h)}\right) \geq \Phi\left(\boldsymbol{\theta}^{(h)}, \boldsymbol{\theta}^{(h)}\right) = S\left(\boldsymbol{\theta}^{(h)}\right). \tag{A5}$$

Thus, we have proved the monotonicity of our proposed EM-type algorithm.

Let $\boldsymbol{\theta}_0$ be the true parameter vector. Partition $\boldsymbol{\theta}_0$ as $\left(\left(\boldsymbol{\theta}_0^{(1)}\right)^T, \left(\boldsymbol{\theta}_0^{(2)}\right)^T\right)^T$ where $\boldsymbol{\theta}_0^{(1)}$ with the dimension s_1 is the vector of all non-zero components and $\boldsymbol{\theta}_0^{(2)}$ with the dimension s_2 is the vector of all zero components. Let $a_m = \max_{1 \leq j \leq s}\left\{p'_{\tau_m}\left(|\theta_{0j}|\right) : \theta_{0j} \neq 0\right\}$ and $b_m = \max_{1 \leq j \leq s}\left\{p''_{\tau_m}\left(|\theta_{0j}|\right) : \theta_{0j} \neq 0\right\}$. The $p'_{\tau_m}(\boldsymbol{\theta})$ and $p''_{\tau_m}(\boldsymbol{\theta})$ are the first and second derivatives of the function $p_{\tau_m}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. We have the following conditions on the penalty function:

C1 For all m and τ_m, $p_{\tau_m}(0) = 0$, and $p_{\tau_m}(\boldsymbol{\theta})$ is symmetric, non-negative, non-decreasing and twice differentiable for all $\boldsymbol{\theta}$ in $(0, \infty)$ with at most a few exceptions.

C2 As $m \to \infty$, $b_m = o(1)$.

C3 For $T_m = \left\{\boldsymbol{\theta};\ 0 < \boldsymbol{\theta} \leq m^{-1/2}\log m\right\}$, $\lim_{m\to\infty}\inf_{\boldsymbol{\theta} \in T_m} p'_{\tau_m}(\boldsymbol{\theta})/\sqrt{m} = \infty$.

These conditions guarantee $\sqrt{m}$-consistency of the estimators. The following assumptions are also needed:

A1 The observations $\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{V}_i, \boldsymbol{Z}_i$ from t-JLSM are independently distributed each with the conditional student-t density $f(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{V}_i, \boldsymbol{Z}_i, \nu)$.

A2 The covariates $\mathbf{x}_{ij}$, $\mathbf{z}_{jk}$, and $\mathbf{w}_j\,(i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, n_i,\ k = 1, 2, \ldots, j-1)$ are fixed and finite. The number of repeated measurements (n_i) are fixed.

A3 The parameter space is compact and the $\boldsymbol{\theta}_0$ is in the interior of the parameter space.

The following theorems states the consistency, sparsity and asymptotic normality of $\hat{\boldsymbol{\theta}}$. The proofs of Theorems 2 and 3 are very similar to the proofs given by Fan & Li (2001). Here, we will outline the proof.

Theorem 2 Assume that $a_m = O_p\left(m^{-1/2}\right)$, $b_m \to 0$, and $\tau_m \to 0$ as $m \to \infty$. Under the conditions (A1)-(A3), with probability tending to 1 there must exist a local maximiser $\hat{\boldsymbol{\theta}}_m$ of the penalised likelihood function $S^*(\boldsymbol{\theta})$ such that $\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0 = O_p\left(m^{-1/2}\right)$.

Proof of Theorem *Let $\zeta_m = m^{-1/2} + a_m$. We just have to specify that for any given $\varepsilon > 0$, there exists a large constant C such that*

$$\lim_{m\to\infty} P\left(\sup_{u=C} S(\boldsymbol{\theta}_0 + \zeta_m \mathbf{u}) < S(\boldsymbol{\theta}_0)\right) \geq 1 - \varepsilon. \tag{A6}$$

This implies that for large m, with large probability, there is a local maximum in the ball $\{\boldsymbol{\theta}_0 + \zeta_m \mathbf{u};\ \mathbf{u} \leq C\}$. This local maximiser $\hat{\boldsymbol{\theta}}$, satisfies $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = O_p(\zeta_m)$. By the definition of $S(\cdot)$ and $p_{\tau_m}(0) = 0$, we have

$$D_m(u) = S(\boldsymbol{\theta}_0 + \zeta_m \mathbf{u}) - S(\boldsymbol{\theta}_0) \tag{A7}$$

$$= \left[logL(\boldsymbol{\theta}_0 + \zeta_m \mathbf{u}) \; - \; m \sum_{j=1}^{s} p_{\tau_{jm}}\left(\left|\theta_{0j} + \zeta_m u_j\right|\right)\right] \; - \; \left[logL(\boldsymbol{\theta}_0) \; - \; m \sum_{j=1}^{s} p_{\tau_{jm}}\left(\left|\theta_{0j}\right|\right)\right] \tag{A8}$$

$$\leq [logL(\boldsymbol{\theta}_0 + \zeta_m \mathbf{u}) \; - \; logL(\boldsymbol{\theta}_0)] \; - \; m\left[\sum_{j=1}^{s_1} p_{\tau_{jm}}\left(\left|\theta_{0j} + \zeta_m u_j\right|\right) \; - \; \sum_{j=1}^{s_1} p_{\tau_{jm}}\left(\left|\theta_{0j}\right|\right)\right] \tag{A9}$$

where s_1 is the number of non-zero elements of the vector $\boldsymbol{\theta}_0$. By substituting the first-order Taylor's expansions and the triangular inequality, we have

$$D_m(u) \leq \zeta_m [l'(\boldsymbol{\theta}_0)]^T \mathbf{u} + \frac{1}{2}\mathbf{u}^T l''(\boldsymbol{\theta}^*)\mathbf{u} m\zeta_m^2\left(1 + o_p(1)\right) \tag{A10}$$

$$-\left[\sum_{j=1}^{s_1} m\zeta_m p'_{\tau_{jm}}\left(\left|\theta_{0j}\right|\right) sgn\left(\theta_{0j}\right) u_j + \frac{1}{2} m\zeta_m^2 p''_{\tau_{jm}}\left(\left|\theta_{0j}\right|\right) u_j^2\left(1 + O_p(1)\right)\right] \tag{A11}$$

$$= K_1 + K_2 + K_3 \tag{A12}$$

Regularity conditions imply that $logL'(\boldsymbol{\theta}_0) = O_p(\sqrt{m})$. *Thus, the* K_1 *is of the order* $O_p(\sqrt{m}\zeta_m)$. *By choosing a sufficiently large C, the* K_1 *is controlled uniformly by* K_2 *in* $u = C$. *Note that the* K_3 *is bounded by*

$$\left[\sum_{j=1}^{s_1} p_{\tau_{jm}}\left(\left|\theta_{0j} + \zeta_m u_j\right|\right) \; - \; \sum_{j=1}^{s_1} p_{\tau_{jm}}\left(\left|\theta_{0j}\right|\right)\right] = \sqrt{s_1} m\zeta_m a_m u + m\zeta_m^2 b_m u^2. \tag{A13}$$

Since it is assumed that $a_m = O_p\left(m^{-1/2}\right)$ and $b_m \to 0$ as $m \to \infty$, if we choose a sufficiently large C. it is concluded that K_3 is dominated by K_2. Thus, for any given $\varepsilon > 0$, we have $\lim_{m\to\infty} P\left(\sup_{u=C} S\left(\boldsymbol{\theta}_0 + m^{-1/2}\mathbf{u}\right) \; < \; S(\boldsymbol{\theta}_0)\right) \geq 1 \; - \; \varepsilon$ that is (A6).

Let $F_m(\boldsymbol{\theta})$ represents the Fisher information matrix of $\boldsymbol{\theta}$, $\theta_{0j}^{(1)}$ is the j-th component of $\boldsymbol{\theta}_0^{(1)}$, define $A_m = \text{diag}\left(p''_{\tau_m}\left(\left|\theta_{01}^{(1)}\right|\right), , \ldots, p''_{\tau_m}\left(\left|\theta_{0s_1}^{(1)}\right|\right)\right)$ and $c_m = \left(p'_{\tau_m}\left(\left|\theta_{01}^{(1)}\right|\right) sgn\left(\theta_{01}^{(1)}\right), \ldots, p'_{\tau_m}\left(\left|\theta_{0s_1}^{(1)}\right|\right) sgn\left(\theta_{0s_1}^{(1)}\right)\right)^T$.

Theorem 3 Assume that the conditions in Theorem 1 are satisfied, and the function $p_{\tau_m}(\boldsymbol{\theta})$ satisfies conditions C1–C3. If the penalty function has $\liminf_{m\to\infty} \liminf_{t\to 0^+} \frac{p'_{\tau_m}(t)}{\tau_m} > 0$ when $\tau_m \to 0$ and $\sqrt{m}\tau_m \to \infty$ as $m \to \infty$, then for any $\sqrt{m}$-consistent estimator $\hat{\boldsymbol{\theta}}_m$ of $\boldsymbol{\theta}$, as $m \to \infty$, we have

(i) Consistency in the variable selection: $P\left(\hat{\boldsymbol{\theta}}_m^{(2)} = 0\right) \to 1$,

(ii) Asymptotic normality:

$$\sqrt{m\left(\mathbf{F}_m^{(1)}\right)}\left(\mathbf{F}_m^{(1)} + A_m\right)\left\{\left(\hat{\boldsymbol{\theta}}_m^{(1)} \; - \; \boldsymbol{\theta}_0^{(1)}\right) + \left(\mathbf{F}_m^{(1)} + A_m\right)^{-1} c_m\right\} \underset{\rightarrow}{D} N_{s_1}(\boldsymbol{\theta}, \boldsymbol{I}_{s_1}),$$

where "$\underset{\rightarrow}{D}$" stands for the convergence in distribution; $\mathbf{F}^{(1)}$ is the $(s_1 \times s_1)$ submatrix of n corresponding to the non-zero components $\boldsymbol{\theta}_0^{(1)}$ and $\boldsymbol{I}_{s_1}$ is the $(s_1 \times s_1)$ identity matrix.

Proof of Theorem *We first prove part (I). From* $\tau_{\max} \to 0$*, it is easy to show that* $a_m = 0$ *for large* m*. Consider the partition* $\boldsymbol{\theta} = \left(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}\right)$ *for any* $\boldsymbol{\theta}$ *in the neighbourhood* $\boldsymbol{\theta} - \boldsymbol{\theta}_0 = O_p\left(m^{-1/2}\right)$*. Second, we prove that for any given* $\boldsymbol{\theta}^{(1)}$ *satisfying* $\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}_0^{(1)} = O_p\left(m^{-1/2}\right)$ *and any constant* $C > 0$*, we have*

$$S\left(\boldsymbol{\theta}^{(1)}, 0\right) = \max_{\boldsymbol{\theta}^{(1)} \le Cm^{-1/2}} < S\left(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}\right). \tag{A14}$$

In fact, for any θ_j $(j = s_1 + 1, \ldots, s)$*, using the Taylor expansion, we obtain*

$$\frac{\partial S(\boldsymbol{\theta})}{\partial \theta_j} = \frac{\partial logL(\boldsymbol{\theta})}{\partial \theta_j} - mp'_{\tau_{jm}}\left(\left|\theta_j\right|\right) sgn\left(\theta_j\right) \tag{A15}$$

$$= \frac{\partial logL(\boldsymbol{\theta}_0)}{\partial \theta_j} + \sum_{t=1}^{s_1} \frac{\partial^2 logL(\boldsymbol{\theta}^*)}{\partial \theta_j \partial \theta_t}(\theta_t - \theta_{0t}) - mp'_{\tau_{jm}}\left(\left|\theta_j\right|\right) sgn\left(\theta_j\right) \tag{A16}$$

where $\boldsymbol{\theta}^*$ lies between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. In addition, we have

$$\frac{1}{m}\frac{\partial logL(\boldsymbol{\theta}_0)}{\partial \theta_j} = O_p\left(m^{-1/2}\right) \tag{A17}$$

and

$$\frac{1}{m}\left[\frac{\partial^2 logL(\boldsymbol{\theta}_0)}{\partial \theta_j \partial \theta_t} - E\left(\frac{\partial^2 logL(\boldsymbol{\theta}_0)}{\partial \theta_j \partial \theta_t}\right)\right] = O_p(1). \tag{A18}$$

Note that $\left\|\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0\right\| = O_p\left(n^{-1/2}\right)$*, we have*

$$\frac{\partial S(\boldsymbol{\theta})}{\partial \theta_j} = -m\tau_{jm}\left(\tau_{jm}^{-1} p'_{\tau_{jm}}\left(\left|\theta_j\right|\right) sgn\left(\theta_j\right) + O_p\left(m^{-1/2}\tau_{jm}^{-1}\right)\right). \tag{A19}$$

From the assumption given in the theorem, we obtain

$$\liminf \liminf \tau_{jm}^{-1} p'_{\tau_{jm}}\left(\left|\theta_j\right|\right) > 0, \text{ and } \tau_{jm}^{-1} m^{-1/2} \to 0 \tag{A20}$$

So that

$$\frac{\partial S(\boldsymbol{\theta})}{\partial \theta_j} < 0, \text{ for } 0 < \theta_j < Cm^{-1/2}, \tag{A21}$$

$$\frac{\partial S(\boldsymbol{\theta})}{\partial \theta_j} > 0, \text{ for } -Cm^{-1/2} < \theta_j < 0 \tag{A22}$$

Therefore, $S(\boldsymbol{\theta})$ *achieve its maximum at* $\boldsymbol{\theta} = \left(\boldsymbol{\theta}^{(1)T}, \mathbf{0}^T\right)^T$ *and this completes the proof of the first part of theorem.*

Second, we study the asymptotic normality of $\hat{\boldsymbol{\theta}}_m^{(1)}$*. From Theorem 1 and the first part of Theorem 2, there exists a penalised maximum likelihood estimator* $\hat{\boldsymbol{\theta}}_m^{(1)}$ *that is the* $\sqrt{m}-$ *consistent local maximiser of the function* $S\left(\boldsymbol{\theta}^{(1)}, \mathbf{0}\right)$*. The estimator* $\hat{\boldsymbol{\theta}}_m^{(1)}$ *must satisfy*

$$0 = \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_j}\bigg|_{\boldsymbol{\theta}=\left(\boldsymbol{\theta}^{(1)T},\, \mathbf{0}^T\right)^T} - mp'_{\tau_{jm}}\left(\left|\hat{\theta}^{(1)}_{mj}\right|\right) sgn\left(\hat{\theta}^{(1)}_{mj}\right) \tag{A23}$$

$$= \frac{\partial logL(\boldsymbol{\theta}_0)}{\partial \theta_j} + \sum_{t=1}^{s_1}\left[\frac{\partial^2 logL(\boldsymbol{\theta}_0)}{\partial \theta_j \partial \theta_t} + O_p(1)\right]\left(\hat{\theta}^{(1)}_{mt} - \theta^{(1)}_{0t}\right) - mp'_{\tau_{jm}}\left(\left|\theta^{(1)}_{0j}\right|\right) sgn\left(\hat{\theta}^{(1)}_{0j}\right) \tag{A24}$$

$$-m\left[p''_{\tau_{jm}}\left(\left|\theta^{(1)}_{0j}\right|\right) + O_p(1)\right] \times \left(\hat{\theta}^{(1)}_{mj} - \theta^{(1)}_{0j}\right) \tag{A25}$$

In other words, we have

$$\left[\frac{\partial^2 logL(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{(1)} \partial\left(\boldsymbol{\theta}^{(1)}\right)^T} + mA_m + O_p(1)\right]\left(\hat{\boldsymbol{\theta}}^{(1)}_m - \boldsymbol{\theta}^{(1)}_0\right) + c_m = \frac{\partial logL(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{(1)}} \tag{A26}$$

Using the Lyapunov form of the multivariate central limit theorem, we obtain

$$\frac{1}{\sqrt{m}} \frac{\partial logL(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{(1)}} \xrightarrow{D} N\left(\mathbf{0},\, \boldsymbol{I}^{(1)}\right). \tag{A27}$$

Note that

$$\frac{1}{m}\left\{\frac{\partial^2 logL(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{(1)} \partial\left(\boldsymbol{\theta}^{(1)}\right)^T} - E\left[\frac{\partial^2 logL(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{(1)} \partial\left(\boldsymbol{\theta}^{(1)}\right)^T}\right]\right\} = O_p(1) \tag{A28}$$

it follows immediately by using Slustky's theorem that

$$\sqrt{m\left(\mathbf{F}^{(1)}_m\right)}\left(\mathbf{F}^{(1)}_m + A_m\right)\left\{\left(\hat{\boldsymbol{\theta}}^{(1)}_m - \boldsymbol{\theta}^{(1)}_0\right) + \left(\mathbf{F}^{(1)}_m + A_m\right)^{-1} c_m\right\} \xrightarrow{D} N_{s_1}(\boldsymbol{\theta},\, \boldsymbol{I}_{s_1}).$$

[*Received July 2023; accepted April 2024*]