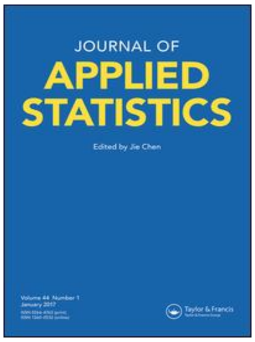

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Variable selection in elliptical linear mixed model

Fulya Gokalp Yavuz & Olcay Arslan

To cite this article: Fulya Gokalp Yavuz & Olcay Arslan (2019): Variable selection in elliptical linear mixed model, Journal of Applied Statistics, DOI: 10.1080/02664763.2019.1702928

To link to this article: https://doi.org/10.1080/02664763.2019.1702928

Published online: 18 Dec 2019.

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2019.1702928

Variable selection in elliptical linear mixed model

Fulya Gokalp Yavuz [a] and Olcay Arslan [b]

[a]Department of Statistics, Middle East Technical University, Ankara, Turkey; [b]Department of Statistics, Ankara University, Ankara, Turkey

ABSTRACT

Variable selection in elliptical Linear Mixed Models (LMMs) with a shrinkage penalty function (SPF) is the main scope of this study. SPFs are applied for parameter estimation and variable selection simultaneously. The smoothly clipped absolute deviation penalty (SCAD) is one of the SPFs and it is adapted into the elliptical LMM in this study. The proposed idea is highly applicable to a variety of models which are set up with different distributions such as normal, student-*t*, Pearson VII, power exponential and so on. Simulation studies and real data example with one of the elliptical distributions show that if the variable selection is also a concern, it is worthwhile to carry on the variable selection and the parameter estimation simultaneously in the elliptical LMM.

ARTICLE HISTORY
Received 7 August 2017
Accepted 5 December 2019

KEYWORDS
Elliptical distributions; mixed models; robust; shrinkage functions; variable selection

1. Introduction

The early applications of the linear mixed models (LMMs) appeared in animal breeding studies [6,14] and there has been a wide variety of applications in the field ranging from genetics to finance since then. Repeated measurements and clustered data are under the scope of these models. Random effects and error terms are assumed to be normally distributed in a classical LMM and parameter estimation is implemented under this assumption. However, the normality assumption does not accommodate for heavy tailedness in the data. Also, the maximum likelihood estimators (MLEs) under normality assumption are typically sensitive to outliers which may not be detected in advance for multidimensional data sets. These obstacles naturally bring some alternative definitions for LMMs.

The robust definitions of the LMM with elliptical distributions are implemented in several studies [21,23,29,30,36] to overcome heavy tailedness and/or outliers in the data. In the study of Manghi *et al.* [23], multilevel models are defined with elliptical distributions to allow light and heavy tailed error distributions. Yavuz and Arslan [36] propose a robust LMM with Laplace distribution, which is a member of the elliptical distribution family, and they carry out model definitions and parameter estimations under this assumption. Similarly, Lindsey [21], Pinheiro *et al.* [29] and Savalli *et al.* [30] confirm the

CONTACT Fulya Gokalp Yavuz fulyagokalp@gmail.com Department of Statistics, Middle East Technical University, Ankara, Turkey

superiority of the robust LMM over counterparts in the presence of heavy tailedness and/or outliers.

Similar to the robust estimation, variable selection is one of the severe topics of the LMM. Recently, shrinkage methods have emerged as efficient model selection procedures. For example, ridge regression is one of the common shrinkage methods and it outperforms the subset regression in terms of accuracy and variance reduction. However, the ridge regression has its own drawbacks [3] and the subset selection procedures are unstable where a small change in the data may result in different selected models [4]. Shrinkage methods that perform variable selection such as the least absolute shrinkage and selection operator (LASSO) overcome these obstacles. Tibshirani [32] proposes LASSO by combining the good features of subset selection and the ridge regression. However, LASSO produces biased estimators for large coefficients. As an alternative, the smoothly clipped absolute deviation penalty (SCAD) is introduced by Fan and Li [7] and it provides the Oracle property.

Robust variable selection has a limited place in the LMM literature. So far, [9] consider a penalized robustified log-likelihood to carry on robust estimation and variable selection in LMM. Bondell [2] uses Cholesky decomposition [22,28] for joint variable selection by reparameterization with adaptive LASSO in LMM. However, their study is based on the normality assumptions for both random effects and error terms. The same decomposition methodology of Bondell [2] is used for the independent cluster model by Müller *et al.* [24] in a classical LMM and by Ibrahim *et al.* [16] in a more general class of LMM. Some further comments about these methods and their drawbacks are mentioned at the fourth section of Müller *et al.* [24].

LASSO and SCAD are used in classical LMM by Cui [5]. Also, Lan [19] extend the variable selection in the LMM with the SCAD throughout a new approach with two different algorithms. In our study, there are two main objectives. First, we extend the [23]'s study for variable selection with a shrinkage penalty in LMM with the elliptically distributed random effects and error terms. We aim to estimate parameters and select important variables simultaneously and to gain some robustness in parameter estimation with elliptical distributions. Second, we would like to consider the variable selection in the LMM under the t-distributed error and the random effect terms. Note that the LMM with t-distribution has already been considered by [29] as a robust alternative to the LMM based on the normal distribution assumption, but the variable selection has not been covered in that paper. Therefore, by doing this, we achieve variable selection and robust estimation in LMM, simultaneously. We think that this will be an additional improvement to the remarkable paper of Pinheiro *et al.* [29].

The rest of the paper is organized as follows. The next section comprises the LMM definitions very briefly. The elliptical distributions and their adaptation into the LMM are covered at the third and fourth sections, respectively. The fifth section includes a brief explanation of the SCAD penalty function and the sixth section explains its adaptation into the LMM. Variable selection is implemented for LMM with the multivariate t-distribution at the seventh section. In Section 8, we provide a simulation study to compare the variable selection performance of the LMM under the multivariate t-distribution versus the multivariate normal distribution assumptions. Real data example is given in the ninth section. The final section of the study includes discussion and conclusion.

2. Linear mixed models

We consider the following LMM for a continuous response variable:

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{e}_i, \quad i = 1, \dots, n, \\ \boldsymbol{u}_i &\sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{D}), \\ \mathbf{e}_i &\sim \mathrm{N}_{n_i}(\mathbf{0}, \boldsymbol{R}_i), \end{aligned} \tag{1}$$

where $\boldsymbol{y}_i$ denotes an $(n_i \times 1)$ vector of continuous responses for the ith subject, $\boldsymbol{\beta}$ denotes a $(p \times 1)$ vector of unknown population parameters, $\boldsymbol{X}_i$ is a known $(n_i \times p)$ design matrix, $\boldsymbol{u}_i$ denotes a $(q \times 1)$ vector of unknown individual effects, $\boldsymbol{Z}_i$ is a known $(n_i \times q)$ design matrix and $\boldsymbol{e}_i$ denotes an $(n_i \times 1)$ vector of residual errors assumed to be independent of $\boldsymbol{u}_i$ [18]. It is often assumed that $\boldsymbol{R}_i = \sigma^2\boldsymbol{I}_{n_i}$ for simplicity. The joint distribution of Equation (1) for $(\boldsymbol{y}_i^T, \boldsymbol{u}_i^T)^{\mathrm{T}}$ is

$$\begin{bmatrix} \boldsymbol{y}_i \\ \boldsymbol{u}_i \end{bmatrix} \sim \mathrm{N}_{n_i+q}\left(\begin{bmatrix} \boldsymbol{X}_i\boldsymbol{\beta} \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{\mathrm{T}} + \sigma^2\boldsymbol{I}_{n_i} & \boldsymbol{Z}_i\boldsymbol{D} \\ \boldsymbol{D}\boldsymbol{Z}_i^T & \boldsymbol{D} \end{bmatrix}\right). \tag{2}$$

The marginal distribution of the response variable is obtained as

$$\boldsymbol{y}_i \sim \mathrm{N}_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i),$$

where $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{\mathrm{T}} + \sigma^2\boldsymbol{I}_{n_i}$. Differentiating the log-likelihood function of model (2) with respect to parameters of interest and solving the resulting equations yield MLEs. However, LMM requires numerical algorithms such as expectation-maximization (EM) [17] or Newton–Raphson (NR) [22], since each parameter estimation includes some of the other unknown parameters. The random effects ($\boldsymbol{u}_i$) are unobserved and required to be *predicted* from the data. An EM-type algorithm is used to meet our objectives for parameter estimation.

Comprehensive work on parameter estimation for LMM is found in Wu [34] and Searle *et al.* [31], among others. We illustrate the classical definition of LMM in this section. The next two sections cover elliptical distributions, albeit very briefly, and their adaptation into the LMM.

3. Elliptical distributions

If a random vector $\boldsymbol{y}_i = (y_{1i}, y_{2i}, \dots, y_{n_i i})^{\mathrm{T}}$ follows n_i-dimensional elliptical distribution with mean $\boldsymbol{\mu}_i$ and variance proportional to $\boldsymbol{\Sigma}_i$, its density function is defined as

$$f(\boldsymbol{y}_i) = |\boldsymbol{\Sigma}_i|^{-1/2} g(t_i), i = 1, \dots, n, \tag{3}$$

where t_i is the Mahalanobis distance defined as $t_i = (\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)$ and $g(.) : \Re \to [0, \infty)$ is the density generating function such that $\int_0^\infty t^{(n_i/2)-1} g(t)\, \mathrm{d}t < \infty$ [10]. Corresponding variance–covariance matrices are $Var(\boldsymbol{y}_i) = \upsilon_i/(\upsilon_i - 2)\boldsymbol{\Sigma}_i$ for t-distribution [20] and $Var(\boldsymbol{y}_i) = 2^{(1+k_i)}\Gamma[(n_i + 2)(1 + k_i)/2]/n_i\Gamma[n_i(1 + k_i)/2]\boldsymbol{\Sigma}_i$ for power exponential distribution [12], where k_i denotes the shape parameter while $\upsilon_i(\upsilon_i > 2)$ denotes the degrees of freedom.

Even though both a normal distribution and an elliptical distribution have elliptical contours, an elliptical distribution may have heavier or lighter tails than the normal. There are several situations one may need to take into account heavy/light tailedness and also outliers in the data. Elliptical distributions aid to overcome these types of situations as an alternative to the normal, which is also one of the scopes of this study detailed at the following section.

4. Linear mixed model with elliptical distributions

The MLEs of the model (2) depend on the assumption of having normally distributed random effects and error terms, therefore MLEs present the lack of robustness against outliers. The generalization of this definition with elliptical distributed random effects and error terms is used to deal with this deficiency. Savalli *et al.* [30] define the model (2) as follows:

$$\begin{bmatrix} \boldsymbol{y}_i \\ \boldsymbol{u}_i \end{bmatrix} \sim \mathrm{El}_{n_i+q}\left(\begin{bmatrix} \boldsymbol{X}_i\boldsymbol{\beta} \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{V}_i & \boldsymbol{Z}_i\boldsymbol{D} \\ \boldsymbol{D}\boldsymbol{Z}_i^T & \boldsymbol{D} \end{bmatrix} \right), \tag{4}$$

where $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{\mathrm{T}} + \sigma^2\boldsymbol{I}_{n_i}$. $\boldsymbol{V}_i$ and $\boldsymbol{D}$ are proportional by a quantity $\alpha_i > 0$ to the $Var(\boldsymbol{y}_i)$ and $Var(\boldsymbol{u}_i)$, obtained from the characteristics function [10]. $\boldsymbol{u}_i$ and $\boldsymbol{e}_i$ are uncorrelated. The joint distribution of two elliptical vectors, $(\boldsymbol{y}_i^T, \boldsymbol{u}_i^T)^{\mathrm{T}}$, is not necessarily elliptical. Therefore, it is harder to use the hierarchical representation of LMM for inferences. However, the hierarchical representation of LMM with multivariate t-distribution does not have this difficulty and LMM can be defined with a hierarchical form with multivariate t-distribution. It is detailed more in Section 7. Concerning the general elliptical distributions, the log-likelihood function can be written as

$$l(\boldsymbol{\beta}; \boldsymbol{D}, \boldsymbol{R}_i, \boldsymbol{Y}) = -\frac{1}{2}\sum_{i=1}^{n} \ln|\boldsymbol{V}_i| + \sum_{i=1}^{n} \ln g\left[(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right], \tag{5}$$

where $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{\mathrm{T}} + \sigma^2\boldsymbol{I}_{n_i}$ and $g(.)$ is the density generating function. Taking the partial derivatives of the log-likelihood function with respect to the parameters of interest and setting them to zero gives the ML estimators for the parameters of a general class of elliptical distributions. For further details, see [23]. The following iterative procedure for the parameter estimators is given by Savalli *et al.* [30] and Manghi *et al.* [23]

$$\boldsymbol{\beta}^{(k+1)} = \left[\sum_{i=1}^{n} \upsilon(t_i)^k \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1(k)}\boldsymbol{X}_i\right]^{-1} \left[\sum_{i=1}^{n} \upsilon(t_i)^k \boldsymbol{X}_i^T\boldsymbol{V}_i^{-1(k)}\boldsymbol{y}_i\right], \tag{6}$$

$$\tau^{(k+1)} = \mathrm{argmax}(l(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\tau})),$$

where $\upsilon(t_i) = g'(t_i)/g(t_i)$, $g'(t_i) = \mathrm{d}g(t_i)/\mathrm{d}t_i$ and $g(t_i)$ is a positive decreasing function. $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_d)^{\mathrm{T}}$ is the vector of distinct parameters in $\boldsymbol{V}_i$ [30]. $\upsilon(t_i)$ may be interpreted as a weight in (6) and in particular, $\upsilon(t_i) = (\nu_i + n_i)/(\nu_i + t_i)$ for t-distribution, where ν_i is the degrees of freedom. Since t_i denotes the Mahalanobis distance, the smaller weights are given for larger observations. The score functions and the Fisher information matrices are found in Savalli *et al.* [30] for the elliptical LMM.

We use the penalized form of the log-likelihood function given above to simultaneously carry on variable selection and parameter estimation in elliptical LMM. The next section includes the brief definition of aforementioned penalty function.

5. Smoothly clipped absolute deviation penalty (SCAD)

Shrinkage penalty functions have an important role in variable selection for a general class of regression models. LASSO, introduced by Tibshirani [32], is used for parameter estimation and variable selection simultaneously by minimizing the residual sum of squares (RSS) subject to the sum of the absolute values of the coefficients with a bound. It shrinks small components of $\boldsymbol{\beta}$ to zero and so corresponding components are deleted from the model. However, LASSO leads to biased estimators when the values of true parameters are large [24]. Also, the optimization procedure of LASSO is challenging because of its non-differentiable property at zero.

Fan and Li [7] proposed SCAD penalty function as an alternative shrinkage function for variable selection by improving both the hard threshold penalty function and the LASSO penalty. They also extend SCAD to semi-parametric analysis for longitudinal data.

For a classical regression, the penalized least square estimator of $\boldsymbol{\beta}$ is obtained by minimizing the following penalized least square objective function:

$$\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+n\sum_{j=1}^{d}p_{\lambda}(|\beta_j|), \tag{7}$$

where $\boldsymbol{Y}$ is an $(n\times 1)$ vector, $\boldsymbol{X}=(x_{ij})$ is the standardized $(n\times d)$ design matrix, $p_{\lambda}(.)$ is the non-negative penalty function and λ is a non-negative regularization (tuning) parameter. The first derivative of SCAD penalty function is

$$p'_{\lambda}(\beta_j)=\lambda\left\{I(\beta_j\leq\lambda)+\frac{(a\lambda-\beta_j)_+}{(a-1)\lambda}I(\beta_j>\lambda)\right\}, \tag{8}$$

where $\beta_j>0, j=1,\ldots,d$, I is the indicator function, $a>2$ is generally taken as 3.7, $(s)_+=s$, for $s>0$ and zero otherwise. This penalty function is symmetric and it is continuously differentiated on the real line except the origin. Also, its derivatives are zero outside the range $-\lambda a$ and $+\lambda a$. This property of the SCAD penalty function yields the Oracle property of the resulting estimators, which is not satisfied by LASSO [7]. On the other hand, SCAD does not have the continuous second-order derivatives. Therefore, a local quadratic approximation algorithm can be used to locally approximate the penalty function to overcome the computational difficulty. Following [7], the local quadratic approximation applied to SCAD is

$$p_{\lambda}(|\beta_j|)\approx p_{\lambda}(|\beta_{j0}|)+\frac{1}{2}\left\{\frac{p'_{\lambda}(|\beta_{j0}|)}{|\beta_{j0}|}\right\}(\beta_j^2-\beta_{j0}^2), \tag{9}$$

where β_{j0} is an initial value for β. Similar to the pioneers [7,8] and others [25,27,35] in the literature of the variable selection via shrinkage methods for partially linear models area, we preferred to use linear quadratic approximation (LQA) algorithm. Using the quadratic approximation for SCAD penalty and combining this with the quadratic form of the objective function with t-distribution, which is obtained from the scale mixture representation, make the maximization procedure tractable enough to analytically obtain the maximum point at each step. Further, the LQA algorithm can be easily embedded in the EM algorithm that is used to obtain the estimates for the parameters of interest. Please refer [15] for details. In the following section, LQA approximation is adapted into ECM algorithm to shrink coefficients for LMM.

6. SCAD in LMM

Fan and Li [8] and Cui [5] show that the SCAD estimates are mainly less biased and the percentage of the correctly fitted model is higher than the LASSO [32] and the adaptive-LASSO [37] in longitudinal studies. Ibrahim *et al.* [16] indicate that SCAD performs well for the regression coefficients, but the adaptive-LASSO works better for the variance parameters in their modified mixed models compared to the SCAD.

The main idea of variable selection in the LMM with a penalty term is subtracting this term from the log-likelihood and maximizing the penalized log-likelihood. Therefore, the penalized likelihood of LMM defined in Equation (2) becomes as

$$l(\boldsymbol{\beta}; \boldsymbol{D}, \boldsymbol{R}_i, \boldsymbol{Y}) = -\frac{1}{2}\ln|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) - n\sum_{j=1}^{d} p_{\lambda}(|\beta_j|), \tag{10}$$

where $\boldsymbol{X}$ is the vertically stacked matrix of $\boldsymbol{X}_i$, $\boldsymbol{Y}$ is the vertically stacked vectors of $\boldsymbol{y}_i$, $\boldsymbol{V}$ is the block-diagonal matrix with blocks on the diagonal by $\boldsymbol{V}_i$ matrices.

For the estimation of random terms $\boldsymbol{u}$, which is the vertically stacked vectors of $\boldsymbol{u}_i$, the following penalized joint log-likelihood function up to some constant will be maximized:

$$\begin{aligned} l_{\text{joint}}(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{D}, \boldsymbol{R}, \boldsymbol{Y}) = &-\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) \\ &- \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} - n\sum_{j=1}^{d} p_{\lambda}(|\beta_j|). \end{aligned} \tag{11}$$

For the ease of computation, $\boldsymbol{R}$ can be taken as $\sigma^2\boldsymbol{I}_n$. From this log-likelihood function of the ML estimator of $\boldsymbol{u}$ is obtained as

$$\hat{\boldsymbol{u}} = \boldsymbol{D}\boldsymbol{Z}^T\hat{\boldsymbol{V}}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{12}$$

This prediction is the same as the best linear unbiased prediction of $\boldsymbol{u}$ given in Harville [13]. For the case where $\boldsymbol{D}$ and $\boldsymbol{R}$ are given, the above maximization problem will be equivalent to minimize the following objective function to obtain estimator for $\boldsymbol{\beta}$ and select the significant variables in fixed terms:

$$G(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_{\lambda}(|\beta_j|). \tag{13}$$

EM algorithm is used in LMM estimation procedure by taking the response variable as observed and random terms as unobserved (missing) data [17,18]. Further, since the SCAD is singular at the origin, some additional effort is required to adapt the SCAD penalty into the LMM for the EM algorithm. For this, the local quadratic approximation of the SCAD penalty given in Equation (9) is used when $\beta_j \neq 0$, otherwise set $\beta_j = 0$ if β_{j0} is very close to 0. Hence, using the following simplification of the penalty function,

$$\sum_{j=1}^{d} p_\lambda(|\beta_j|) \approx \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta} + \text{constant}, \tag{14}$$

where $\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0) = \mathrm{diag}\{p'_\lambda(|\beta_{10}|)/|\beta_{10}|, \ldots, p'_\lambda(|\beta_{d0}|)/|\beta_{d0}|\}$ and 'constant' refers the terms not involving $\boldsymbol{\beta}$, the objective function given in Equation (13) will be as follows:

$$G(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta}) + \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}n\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}. \tag{15}$$

Taking the first derivative of (15) with respect to $\boldsymbol{\beta}$ and setting it to 0 yields the following SCAD estimator of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{\mathrm{SCAD}} = [\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X} + n\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)]^{-1}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}\boldsymbol{Y}. \tag{16}$$

Further, we can adapt this estimation procedure into EM for elliptical LMM by subtracting the penalty term from the log likelihood function defined in (5). In this manner, maximizing the resulting function gives $\hat{\boldsymbol{\beta}}_{\mathrm{SCAD}}$ parameter estimators similar to (16).

The EM algorithm is implemented for a grid of λ_m and the solution is obtained for each λ_m. Minimizing a criterion such as AIC, BIC or generalized cross validation (GCV), a λ_m value is chosen among candidates for the final solution. BIC mainly performs the best amongst other approaches [19]. We use BIC given by

$$\mathrm{BIC} = -2l_{SCAD} + q\log(n), \tag{17}$$

where q is the number of the non-zero variables after selection and l_{SCAD} is obtained with the SCAD penalized estimates.

7. Variable selection in *t*-distributed linear mixed models with ECM algorithm

In this part of the study, we give an example of one of the elliptical distributions in LMM for variable selection. The hierarchical representation of the multivariate t-distribution and the joint distribution of response variable and random effects make the analytic calculations possible for variable selection with shrinkage methods. Pinheiro *et al.* [29] propose a robust hierarchical linear mixed model in which both random effects and error terms are multivariate t-distributed. The joint distribution of response variable and random effects in model (1) proceeds with multivariate t-distribution as follows:

$$\begin{bmatrix} \boldsymbol{y}_i \\ \boldsymbol{u}_i \end{bmatrix} \sim \mathrm{t}_{n_i+q}\left(\begin{bmatrix} \boldsymbol{X}_i\boldsymbol{\beta} \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{\mathrm{T}} + \boldsymbol{R}_i & \boldsymbol{Z}_i\boldsymbol{D} \\ \boldsymbol{D}\boldsymbol{Z}_i^{\mathrm{T}} & \boldsymbol{D} \end{bmatrix}, \nu_i\right) \tag{18}$$

[20,29].

Using the scale mixture representation of t-distributed random variable, the conditional distribution of $[\boldsymbol{y}_i^T, \boldsymbol{u}_i^T]^{\mathrm{T}}$ given τ_i will be

$$\begin{bmatrix} \boldsymbol{y}_i \\ \boldsymbol{u}_i \end{bmatrix} \Bigg| \tau_i \sim \mathrm{N}_{n_i+q}\left(\begin{bmatrix} \boldsymbol{X}_i\boldsymbol{\beta} \\ 0 \end{bmatrix}, \frac{1}{\tau_i}\begin{bmatrix} \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{\mathrm{T}} + \boldsymbol{R}_i & \boldsymbol{Z}_i\boldsymbol{D} \\ \boldsymbol{D}\boldsymbol{Z}_i^{\mathrm{T}} & \boldsymbol{D} \end{bmatrix}\right), \tag{19}$$

$$\tau_i \sim \mathrm{Gamma}\left(\frac{\nu_i}{2}, \frac{\nu_i}{2}\right), i = 1, \ldots, n.$$

The LMM with multivariate t-distributed errors and random effects can be written as follows [29]:

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{e}_i, i = 1, 2 \ldots n, \\ \boldsymbol{u}_i &\sim \mathrm{t}_q(\boldsymbol{0}, \boldsymbol{D}, \nu_i), \\ \mathbf{e}_i &\sim \mathrm{t}_{n_i}(\boldsymbol{0}, \boldsymbol{R}_i, \nu_i), \end{aligned} \tag{20}$$

where $\boldsymbol{u}_i$ and $\boldsymbol{e}_i$ are uncorrelated. This model can accommodate both e- and b-outliers, which occur at the within-subject error, $\boldsymbol{e}_i$-level and at the random effects, $\boldsymbol{u}_i$-level, respectively. The degrees of freedom of the multivariate t-distribution (ν_i) are taken as fixed for the sake of the computation. We define $\boldsymbol{R}_i = \sigma_i^2\boldsymbol{I}_{n_i}$. Now, we can implement the EM algorithm to estimate the parameters as well as variable selection. The objective function of the LMM with multivariate t-distributed errors and random effects given in (20) with the stacked vectors and matrices is

$$G(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zu})^{\mathrm{T}}(\boldsymbol{\tau}_{\mathrm{diag}}\sigma^2\boldsymbol{I}_N)^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zu}), \tag{21}$$

where $\boldsymbol{X}$ is the vertically stacked matrix of $\boldsymbol{X}_i$, $\boldsymbol{Y}$ is the vertically stacked vectors of $\boldsymbol{y}_i$, $\mathbf{Z}$ is the block diagonal matrix of $\boldsymbol{Z}_i$, $\boldsymbol{u}$ is the vertically stacked vectors of $\boldsymbol{u}_i$, $\boldsymbol{\tau}_{\mathrm{diag}}$ is a diagonal matrix with its diagonal entries are τ_i values and $N = \sum_{i=1}^n n_i$.

We assume that the number of random effects can be specified and we try to select the variables associated with the fixed effects. For this purpose, the SCAD penalty term is added to the related part with unknown parameter $\boldsymbol{\beta}$. We treat $\boldsymbol{u}_i$ and $\boldsymbol{e}_i$ as missing values and use the hierarchical definition of t-LMM given in (19) with the fixed degrees of freedom. The local quadratic approximation of the SCAD given in (14) is added to the objective function similar to the one given in (15). So, the following objective function including penalty term is used to find $\hat{\boldsymbol{\beta}}$.

$$\begin{aligned} G(\boldsymbol{\beta}) &= \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zu})^{\mathrm{T}}(\boldsymbol{\tau}_{\mathrm{diag}}\sigma^2\boldsymbol{I}_N)^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zu}) + n\sum_{j=1}^d p_\lambda(|\beta_j|) \\ &\approx \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zu})^{\mathrm{T}}(\boldsymbol{\tau}_{\mathrm{diag}}\sigma^2\boldsymbol{I}_N)^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zu}) + \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}n\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}, \end{aligned} \tag{22}$$

where $\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0) = \mathrm{diag}\{p'_\lambda(|\beta_{10}|)/|\beta_{10}|, \ldots, p'_\lambda(|\beta_{d0}|)/|\beta_{d0}|\}$. Since we use EM algorithm for variable selection and parameter estimation, it is required to find complete data likelihood including the response variable as observed and random terms as unobserved data

[18]. The log-likelihood of the complete data in the multivariate t-LMM is as follows:

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(\boldsymbol{y}_i, \boldsymbol{u}_i, \tau_i; \boldsymbol{\beta}, \boldsymbol{D}, \sigma_i^2). \tag{23}$$

The hierarchical representation of t-LMM allows us to describe the joint distribution function as $f(\boldsymbol{y}_i|\boldsymbol{u}_i, \tau_i)f(\boldsymbol{u}_i|\tau_i)f(\tau_i)$. Hence, the log-likelihood function is as follows:

$$\ln L(\boldsymbol{\beta}, \boldsymbol{D}, \sigma^2) = \ln L_1(\beta, \sigma^2|\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\tau}) + \ln L_2(\boldsymbol{D}|\boldsymbol{u}, \boldsymbol{\tau}) + \text{constant}, \tag{24}$$

where

$$\begin{aligned} \ln L_1(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\tau}) = \sum_{i=1}^{n} \Big[-\frac{n_i}{2}\ln\sigma_i^2 - \tau_i(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i - \boldsymbol{Z}_i\boldsymbol{u}_i)^{\mathrm{T}} \\ \times (\sigma_i^2 \boldsymbol{I}_{n_i})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{u}_i) \Big], \end{aligned} \tag{25}$$

$$\ln L_2(\boldsymbol{D}|\boldsymbol{u}, \boldsymbol{\tau}) = -\frac{n}{2}\ln|\boldsymbol{D}| - \frac{1}{2}\text{trace}\left(\boldsymbol{D}^{-1}\sum_{i=1}^{n}\tau_i\boldsymbol{u}_i\boldsymbol{u}_i^{\mathrm{T}}\right). \tag{26}$$

Since we aim to carry out the variable selection in fixed term, the SCAD penalty related to $\boldsymbol{\beta}$ is subtracted only from the conditional expectation of $\ln L_1$. Notice that the last term of Equation (27) at below is the quadratic approximation of the SCAD penalty function and the variable selection is carried out in the steps of the EM algorithm with the help of this function. At the E-step of EM algorithm, the conditional expectations of the log-likelihoods are described as follows:

$$\begin{aligned} E[\ln L_1(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{u}, \tau)|\boldsymbol{y}, \hat{\boldsymbol{\theta}}] = \sum_{i=1}^{n} -\frac{n_i}{2}\ln\sigma_i^2 - \sum_{i=1}^{n}\frac{1}{2\sigma_i^2}\text{trace}[\hat{\tau}_i((\boldsymbol{y}_i - \boldsymbol{Z}_i\hat{\boldsymbol{u}}_i) \\ \times (\boldsymbol{y}_i - \boldsymbol{Z}_i\hat{\boldsymbol{u}}_i)^{\mathrm{T}} + \boldsymbol{Z}_i\hat{\boldsymbol{\Omega}}_i\boldsymbol{Z}_i^{\mathrm{T}})] + \sum_{i=1}^{n}\frac{\hat{\tau}_i}{\sigma_i^2}\boldsymbol{\beta}^T\boldsymbol{X}_i^T(\boldsymbol{y}_i - \boldsymbol{Z}_i\hat{\boldsymbol{u}}_i) \\ - \sum_{i=1}^{n}\frac{\hat{\tau}_i}{2\sigma_i^2}\boldsymbol{\beta}^T\boldsymbol{X}_i^T\boldsymbol{X}_i\boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}^T n\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}, \end{aligned} \tag{27}$$

$$E[\ln L_2(\boldsymbol{D}|\boldsymbol{u}, \boldsymbol{\tau})|\boldsymbol{y}, \hat{\boldsymbol{\theta}}] = -\frac{n}{2}\ln|\boldsymbol{D}| - \frac{1}{2}\text{trace}\left(\boldsymbol{D}^{-1}\sum_{i=1}^{n}(\hat{\tau}_i\hat{\boldsymbol{u}}_i\hat{\boldsymbol{u}}_i{}^T + \hat{\boldsymbol{\Omega}}_i)\right), \tag{28}$$

where

$$\begin{aligned} \hat{\boldsymbol{u}}_i &= E(\boldsymbol{u}_i|\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \boldsymbol{y}) = \hat{\boldsymbol{D}}\boldsymbol{Z}_i^{\mathrm{T}}\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) \\ &= \hat{\boldsymbol{D}}\boldsymbol{Z}_i^{\mathrm{T}}(\boldsymbol{Z}_i\hat{\boldsymbol{D}}\boldsymbol{Z}_i^{\mathrm{T}} + \hat{\sigma}_i^2\boldsymbol{I}_{n_i})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}), \\ \hat{\boldsymbol{\Omega}}_i &= \tau_i\text{cov}(\boldsymbol{u}_i|\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \boldsymbol{y}) = \hat{\boldsymbol{D}} - \hat{\boldsymbol{D}}\boldsymbol{Z}_i^{\mathrm{T}}(\boldsymbol{Z}_i\hat{\boldsymbol{D}}\boldsymbol{Z}_i^{\mathrm{T}} + \hat{\sigma}_i^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Z}_i\hat{\boldsymbol{D}}, \end{aligned}$$

$$\hat{\tau}_i = E(\tau_i|\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \boldsymbol{y}) = \frac{\nu + n_i}{\nu + \delta_i^2(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{D}}, \hat{\sigma}_i^2)},$$

$$\begin{aligned}\delta_i^2(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{D}}, \hat{\sigma}_i^2) &= (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{Z}_i\hat{\boldsymbol{D}}\boldsymbol{Z}_i^{\mathrm{T}} + \hat{\sigma}_i^2\boldsymbol{I}_{n_i})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})\\ &= (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \boldsymbol{Z}_i\hat{\boldsymbol{u}}_i)^{\mathrm{T}}(\hat{\sigma}_i^2\boldsymbol{I}_{n_i})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \boldsymbol{Z}_i\hat{\boldsymbol{u}}_i).\end{aligned}$$

Taking the first derivative of Equation (27) with respect to σ_i^2, and $\boldsymbol{\beta}$, and setting these derivatives to 0 gives $\hat{\sigma}_i$, and $\hat{\boldsymbol{\beta}}$, respectively. Similarly, the maximization of the conditional expectation of (28) is used to find $\hat{\boldsymbol{D}}$.

For the EM algorithm, we take $\boldsymbol{u} = [\boldsymbol{u}_1^T, \dots, \boldsymbol{u}_n^T]^{\mathrm{T}}$ and $\boldsymbol{\tau} = [\tau_1, \dots, \tau_n]^{\mathrm{T}}$ as missing data, $\boldsymbol{y} = [\boldsymbol{y}_1^T, \dots, \boldsymbol{y}_n^T]^{\mathrm{T}}$ as observed data, and so $[\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\tau}]$ are complete data.

We extend the ECM algorithm in [29] with the SCAD for multivariate t-distributed LMM and call it ECM-t-SCAD. Initial values of the algorithm are determined from t-distributed LMM.

The ECM algorithm for t-distributed LMM with the SCAD is defined as follows:

E-step: Standardize $\boldsymbol{X}$ values, compute $\hat{\boldsymbol{u}}_i$, $\hat{\boldsymbol{\Omega}}_i$ and $\hat{\tau}_i$ for $i = 1, \dots, n$.

CM-SCAD step: Fix $\sigma_i^2 = \hat{\sigma}_i^2$ for $i = 1, \dots, n$.

Set a range for $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_t)$ and for each level of λ_i, implement the inner steps, estimate $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ and find BIC. The corresponding λ_i of the smallest BIC is selected as optimal λ_i. The final estimates of $\boldsymbol{\beta}$ are determined with this selected λ_i and the significant variable coefficients are determined with the nonzero components of $\hat{\boldsymbol{\beta}}$.

CM-SCAD inner step: Find the final $\hat{\boldsymbol{\beta}}$ with

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(k+1)} &= \left(\sum_{i=1}^{n} \hat{\tau}_i^{(k)}\boldsymbol{X}_i^{\mathrm{T}}(\hat{\sigma}_i^{2(k)}\boldsymbol{I}_{n_i})^{-1}\boldsymbol{X}_i + \boldsymbol{\Sigma}_\lambda(\hat{\boldsymbol{\beta}}^{(k)})\right)^{-1}\\ &\quad\times \sum_{i=1}^{n} \hat{\tau}_i^{(k)}\boldsymbol{X}_i^{\mathrm{T}}(\hat{\sigma}_i^{2(k)}\boldsymbol{I}_{n_i})^{-1}(\boldsymbol{y}_i - \boldsymbol{Z}_i\hat{\boldsymbol{u}}_i^{(k)}). \qquad (29)\end{aligned}$$

The convergency of $\hat{\boldsymbol{\beta}}^{(k+1)}$ for $k = 1, 2, \dots$ in the CM-SCAD inner step is determined with the relative change:

$$\max\frac{|\hat{\boldsymbol{\beta}}_l^{(k+1)} - \hat{\boldsymbol{\beta}}_l^{(k)}|}{|\hat{\boldsymbol{\beta}}_l^{(k)}| + \eta} < \mathrm{tol}, l = 1, \dots, d,$$

where η is a small positive number. Repeat this step until $\hat{\boldsymbol{\beta}}$ converges for each level of λ_i.

CM inner step 1: Fix $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(k+1)}$ and update $\hat{\sigma}_i^2$ for $i = 1, \dots, n$ with

$$\begin{aligned}\hat{\sigma}_i^{2(k+1)} &= \frac{1}{n_i}\sum_{i=1}^{n}[\hat{\tau}_i^{(k)}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(k+1)} - \boldsymbol{Z}_i\hat{\boldsymbol{u}}_i^{(k)})^{\mathrm{T}}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(k+1)} - \boldsymbol{Z}_i\hat{\boldsymbol{u}}_i^{(k)})\\ &\quad+ \mathrm{trace}(\hat{\boldsymbol{\Omega}}_i^{(k)}\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{Z}_i)]. \qquad (30)\end{aligned}$$

CM inner step 2: Update $\hat{\boldsymbol{D}}$ with

$$\hat{\boldsymbol{D}}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}_i^{(k)}\hat{\boldsymbol{u}}_i^{(k)}\hat{\boldsymbol{u}}_i^{\mathrm{T}(k)} + \hat{\boldsymbol{\Omega}}_i^{(k)}). \qquad (31)$$

All steps are repeated until the convergent and the parameter estimation procedure is finalized.

8. Simulation study

We use a similar scenario of [27] for the simulation study. LMM is carried out in the form of

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{e}_i, i = 1, \dots, n,$$

where $\boldsymbol{y}_i = [y_{i1}, \dots, y_{in_i}]$, i represents the subjects/observations. Nine independent variables are generated from a multivariate normal distribution. We add intercept along with 1s to $\boldsymbol{X}$. The first four columns of $\boldsymbol{X}$ are taken to generate $\boldsymbol{Z}$. Simulation studies are implemented for three different scenarios and 100 sets of data are generated for each scenario.

- *Scenario 1*: The random errors $\boldsymbol{e}_i$ are generated from a multivariate normal distribution $\boldsymbol{e}_i \sim \mathrm{N}_{n_i}(0, \sigma_i^2\boldsymbol{I}_{n_i})$ with the model variance $\sigma_i^2 = 1$ for simplicity. Random effects $\boldsymbol{u}_i$ are generated from a multivariate normal distribution $\boldsymbol{u}_i \sim \mathrm{N}_q(0, \boldsymbol{D})$.
- *Scenario 2*: The random errors $\boldsymbol{e}_i$ are generated from a multivariate t-distribution $\boldsymbol{e}_i \sim \mathrm{t}_{n_i}(0, \sigma_i^2\boldsymbol{I}_{n_i}, \nu)$ with the model variance $\sigma_i^2 = 1$. Random effects $\boldsymbol{u}_i$ are generated from a multivariate t-distribution, $\boldsymbol{u}_i \sim \mathrm{t}_q(0, \boldsymbol{D}, \nu)$. ν is taken as 3 for simplicity.
- *Scenario 3*: Randomly chosen three subjects are replaced with $\boldsymbol{X}_i^{\mathrm{T}}(\text{new}) = \boldsymbol{X}_i^{\mathrm{T}} + 10$ (contamination).

The true value of the fixed effects coefficients is $\boldsymbol{\beta} = [0, 1, 1, 0, 0, 0, 0, 0, 0]$ with the intercept $\beta_0 = 1$. The true variance–covariance matrix of random effects with four dimensions is

$$\boldsymbol{D} = \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{bmatrix}.$$

For the CM-SCAD step, λ is taken as a sequence of 100 values from 2^{-14} to $2^{0.75}$ similar to the simulation scenario in Cui [5]. The BIC values are calculated for each λ with Equation (17). The optimal λ associated with the smallest BIC is determined and the rest of the calculations are implemented with this optimal λ for the data. a is taken as 3.7 in the SCAD function, Fan and Li [7] argued that the variable selection is not improved with different a values by using data driven methods.

For all scenarios, four different conditions are tried:

- *Condition 1*: Consider $n = 30$ subjects, $n_i = 5$ repeats per subject.
- *Condition 2*: Consider $n = 60$ subjects, $n_i = 5$ repeats per subject.
- *Condition 3*: Consider $n = 30$ subjects, $n_i = 10$ repeats per subject.
- *Condition 4*: Consider $n = 60$ subjects, $n_i = 10$ repeats per subject.

We conduct simulation studies with two different methods for the comparisons. We use ECM algorithm for t-distributed LMM that is proposed in this study and call it ECM-t-SCAD. Second, we use ECM algorithm for the classical LMM and call it ECM-SCAD.

Five different components are measured in simulation studies to evaluate the model selection method. Two of them are model size and the mean number of nonzero coefficients except β_0. The third one is the median of the mean squared error (MME), which is reported to examine the performances of the parameter estimations (excluding intercept):

$$\hat{\boldsymbol{\beta}}_{\text{MME}} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} E(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

where $E(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})$ is estimated with $\boldsymbol{X}$'s sample covariance, since the mean vector of $\boldsymbol{X}$ is zero. The fourth component is the number of correct model among 100 simulations (correct model ratio, CMR) and the last one is the relative efficiency, which is the ratio of ECM-SCAD MME to ECM-t-SCAD MME.

8.1. Simulation results

All simulations are conducted by generating R codes. Data generating procedures are implemented with *mnormt* [1] and *mvtnorm* [11] libraries defined in R. The comparison results of our proposed method (ECM-t-SCAD) via classical LMM (ECM-SCAD) are given in Tables 1 and 2.

Table 1. Simulation results for all conditions and scenarios.

Condition	Scenario	ECM	Model size[a]	MME	CMR	Efficiency[b]
1	1	t-SCAD	2.27	0.0862	75	1.9486
		SCAD	2.39	0.1680	66	
	2	t-SCAD	2.47	0.2035	54	2.3339
		SCAD	2.50	0.4750	41	
	3	t-SCAD	2.26	0.5625	75	2.8309
		SCAD	2.33	1.5925	70	
2	1	t-SCAD	2.15	0.0410	86	1.3891
		SCAD	2.21	0.0569	82	
	2	t-SCAD	2.18	0.0905	80	2.5459
		SCAD	2.41	0.2304	57	
	3	t-SCAD	2.08	0.0968	92	3.5234
		SCAD	2.14	0.3411	87	
3	1	t-SCAD	2.10	0.0435	90	2.3601
		SCAD	2.19	0.1027	82	
	2	t-SCAD	2.16	0.1070	77	3.6216
		SCAD	2.28	0.3876	56	
	3	t-SCAD	2.05	0.1463	94	4.5644
		SCAD	2.06	0.6679	84	
4	1	t-SCAD	2.12	0.0243	88	1.9474
		SCAD	2.50	0.0474	85	
	2	t-SCAD	2.05	0.0437	95	4.1394
		SCAD	2.28	0.1809	74	
	3	t-SCAD	2.01	0.0182	98	4.6388
		SCAD	2.07	0.0844	92	

[a] True value is 2.
[b] (MME/MME(t)).

Table 2. The percentage of the number of times the nonzero coefficients obtained for 100 simulations for all conditions and scenarios.

Condition	Scenario	ECM	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
1	1	t-SCAD	13	100	100	3	2	7	0	1	1
		SCAD	21	100	100	3	2	5	2	5	1
	2	t-SCAD	12	98	99	10	6	5	7	6	7
		SCAD	21	92	97	10	7	8	7	7	5
	3	t-SCAD	7	100	100	2	5	3	1	3	5
		SCAD	15	100	100	3	5	2	1	3	4
2	1	t-SCAD	9	100	100	1	0	1	1	3	0
		SCAD	15	100	100	0	0	2	0	2	2
	2	t-SCAD	10	99	98	4	1	2	2	1	2
		SCAD	21	98	96	7	5	3	4	5	3
	3	t-SCAD	2	100	100	0	1	2	0	3	0
		SCAD	8	100	100	1	3	1	0	1	0
3	1	t-SCAD	9	100	100	1	0	0	0	0	0
		SCAD	17	100	100	2	0	0	0	0	0
	2	t-SCAD	17	100	97	1	0	0	1	0	1
		SCAD	25	95	94	2	6	3	3	4	4
	3	t-SCAD	5	100	100	0	0	0	0	0	1
		SCAD	12	96	96	0	1	1	0	0	0
4	1	t-SCAD	12	100	100	0	0	0	0	0	0
		SCAD	15	100	100	0	0	0	0	0	0
	2	t-SCAD	5	100	100	0	0	0	0	0	0
		SCAD	20	100	100	2	1	1	1	1	2
	3	t-SCAD	1	100	100	0	0	1	0	0	0
		SCAD	7	99	100	0	0	0	0	0	0

The true model size is 2 and the model sizes are closer to 2 for ECM-t-SCAD in all scenarios and conditions. This value is getting closer to 2 when the sample size is 60 (Table 1).

The median of MMEs is reported in the second columns in Table 1 and they are smaller for ECM-t-SCAD.

We may conclude that the proposed model is preferred, especially when the sample size is relatively small. The correct model ratios of ECM-t-SCAD are greater than the correct model ratios of ECM-SCAD and they are getting closer to 100 for larger sample sizes (Table 1). On the other hand, even the correct model ratios of ECM-t-SCAD and ECM-SCAD models are similar for larger sample sizes, the median of the means squared errors is still smaller for the proposed model. In particular, the results for contaminated data show that MME of ECM-SCAD is higher than ECM-t-SCAD. Therefore, the proposed method is more robust to contaminated data for all conditions.

The relative efficiency is greater than 1 for almost all conditions and scenarios (Table 1). This value is promising to use ECM-t-SCAD for model selection. The same value gets greater for t-distributed random effects and error terms. Also, the number of correct model is higher for greater N values which is shown in Tables 1 and 2.

9. Real data example

We use 'veneer' data [26] located in *WWGbook* library in R. The researchers aim to see whether different amounts of contour differences (CDA) affect gingival health over time. The data is composed of seven variables which are GCF (gingival crevicular fluid adjacent

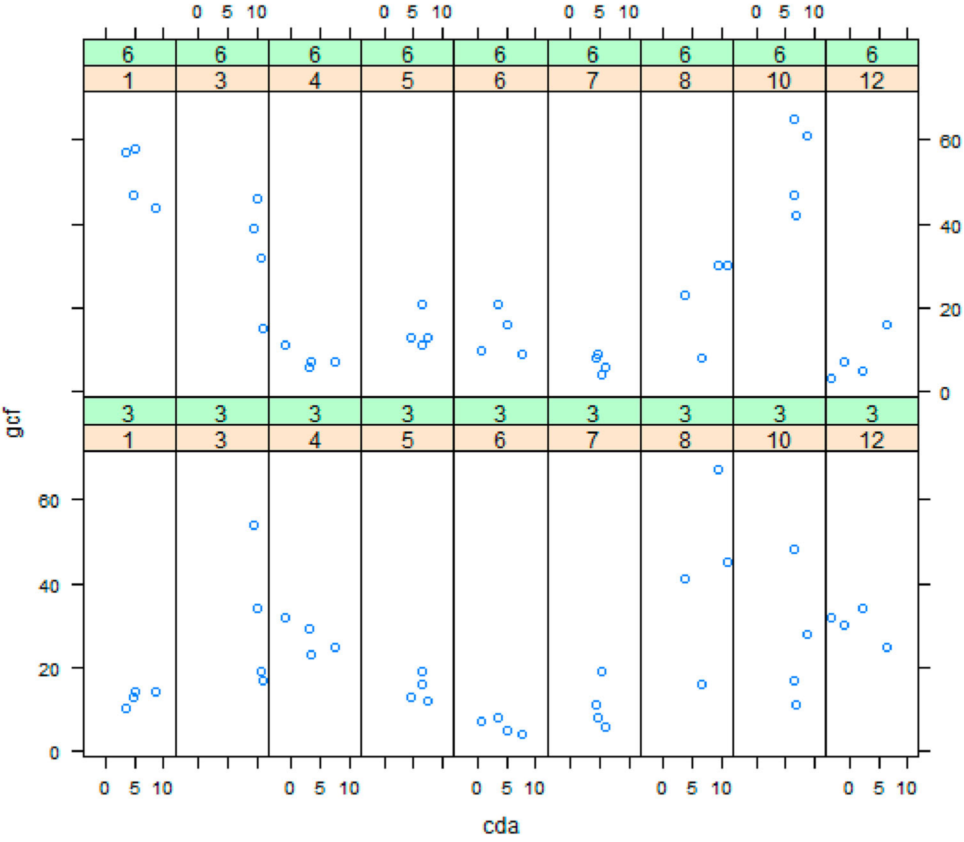

Figure 1. The panels of raw GCF versus CDA for each time points.

to the tooth, dependent variable), patient (patient/subject ID), age (age was recorded when veneer was placed), tooth (tooth number), base-GCF (baseline measure of GCF for the tooth), CDA (average contour difference in the tooth after veneer replacement) and time (measurement times; third and sixth months). Age, base-GCF and CDA are constant for all observations on the same tooth. Nine observations with two repeats at the third and sixth months (post-treatment time points) are used in this study. The panels of raw GCF versus CDA for each time points are located in Figure 1. All analyses and figures are implemented in R.

In Figure 1, we observe that GCF values versus CDA values may differ for different time points for each subject. This may indicate a time-CDA interaction effect on GCF values and this interaction is added into the model. Figure 2 shows solely GCF values for each subject from third and sixth months. In Figure 2, we observe that GCF values tend to decrease for some subjects while they are decreasing or comparatively constant for others over time. The subjects and tooth nested within subjects are assumed to be random factors. Since the time points are the same for each subject, these random factors are crossed with time. We use the multivariate t-distributed LMM with SCAD for parameter estimation and variable selection simultaneously. The predicted values are gathered after variable selection

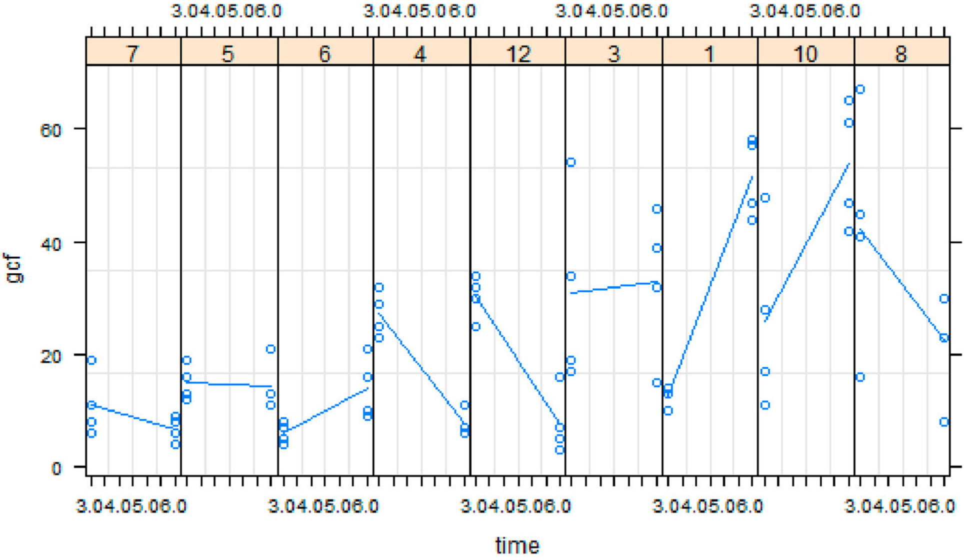

Figure 2. GCF values for each subject from third and sixth months.

procedure for ‘veneer’ data. All independent variables and their interactions with time are included into the model against dependent variable (GCF).

The model specification for GCF$_{tij}$ at time point t on tooth i nested within subject j is as follows:

$$\begin{aligned} GCF_{tij} &= \beta_0 + \beta_1 \times (time_t) + \beta_2 \times (base\_GCF_{ij}) + \beta_3 \times (CDA_{ij}) \\ &\quad + \beta_4 \times (age_j) + \beta_5 \times (time_t \times base\_GCF_{ij}) + \beta_6(time_t \times CDA_{ij}) \\ &\quad + \beta_7(time_t \times age_j) + u_{0j} + u_{1j} \times (time_t) + u_{0i|j} + \epsilon_{tij}, \end{aligned} \tag{32}$$

where $\beta_i (i = 0, \ldots, 7)$ is fixed effect parameters. u_{0j} and u_{1j} are random intercept and random time slope parameters for subjects, respectively. $u_{0i|j}$ is random effects for tooth nested within a subject (cluster) and ϵ_{tij} is the residual term which is random naturally.

The joint distribution of two random effects is

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim t(0, \boldsymbol{D}, \nu),$$

where

$$\boldsymbol{D} = \begin{bmatrix} \sigma^2_{\text{intercept:subject}} & \sigma^2_{\text{intercept,time:subject}} \\ \sigma^2_{\text{intercept,time:subject}} & \sigma^2_{\text{time:subject}} \end{bmatrix}$$

and

$$\epsilon_{tij} \sim t(0, \sigma_2 I_2, \nu).$$

The initial values for the variance matrices of random effects are gathered from the classical LMM model. Model selection procedure is implemented depending on the generated code in R for the proposed model. The SCAD is used as a shrinkage function to select variables for the model (32).

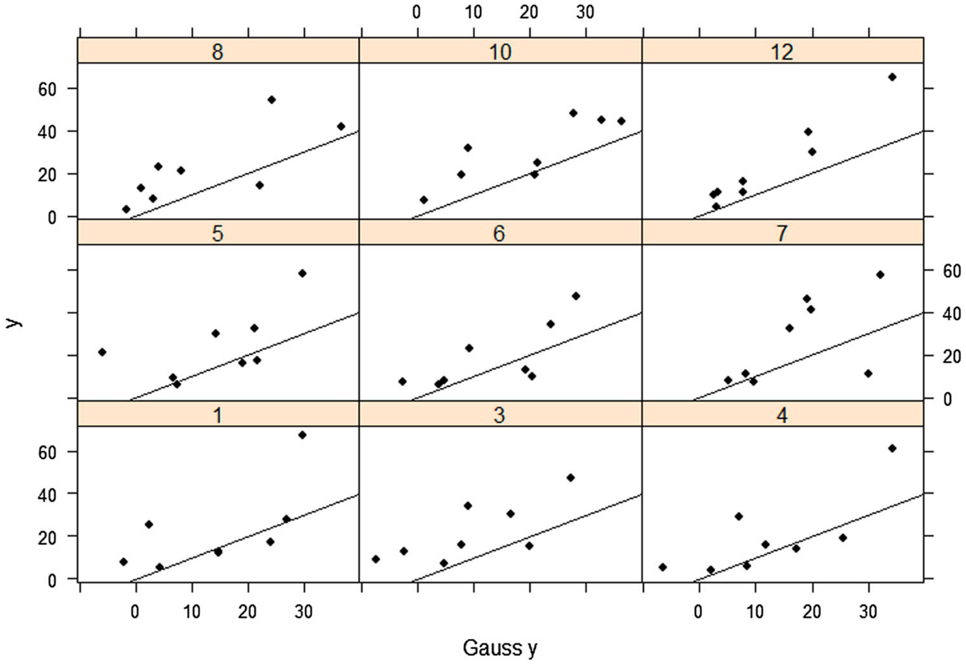

Figure 3. The predicted y values for ECM-SCAD (Gauss y) versus actual y values.

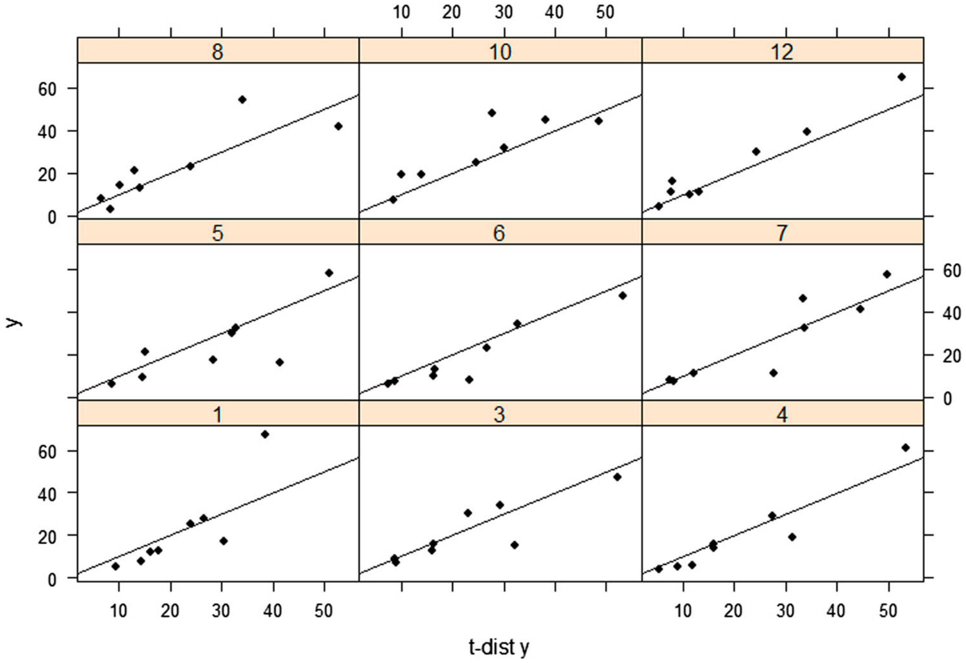

Figure 4. The predicted y values for ECM-t-SCAD (t-dist y) versus actual y values.

Depending on implementations, the SCAD procedure selects two fixed variables, which are CDA and CDA–time interactions with −1.8292 and 0.3052 values, respectively. We fit the model and do model selection with the SCAD for the classical LMM and Figure 3 shows the predicted y values with the selected variables for ECM-SCAD versus actual y values. We also fit the model and do model selection with SCAD for t-LMM and Figure 4 shows the predicted y values for ECM-t-SCAD versus actual y values. Comparing two graphics, Figure 4 is highly promising for the proposed model, especially for this data set in which the number of observations is small.

We included a brief explanation of 'veneer' data analysis results for our method. Readers may refer to the book of West *et al.* [33] for more real data applications of the classical LMM approach.

10. Discussion and conclusion

Variable selection studies for LMM suffer from a lack of robustness to deviations from the normality assumption and the studies accounting for robustness deserve investigations [2]. This study is a contribution to the field of model selection in robust LMM.

Since normal distribution may not be the most appropriate option especially for heavy tailed data, further comparisons are needed for variable selection with different distributions in the LMM. We indicate that expanding the ECM algorithm with the SCAD in elliptical LMM enables us to select fixed effects effectively. Simulation results support that our proposed method outperforms classical ECM-SCAD especially for the data with longer tails. Furthermore, we do variable selection for t-distributed LMM for clustered longitudinal 'veneer' data, in which tooth are nested within clusters (subjects) and repeated measures are gathered over time for each subject. Analysis results show that the proposed method outperforms the classical one comparing the predicted values of the response variable. More applications for high dimensional data will be the natural extension of this study.

The SCAD sets small coefficients to zero, while large coefficients retain as they are because of its singularity at zero property. Thus we obtained sparse set of solutions and more unbiased coefficients with the help of the multivariate t-distributed random effects and error terms in LMM. The proposed method is promising especially for the contaminated data and it is more robust than the classical approach. The selected models are closer to the actual models and the mean squared errors are smaller for the proposed model. Simulation studies show that it is worthwhile to try model selection in the LMM with robust distributed random effects and error terms with the conditional log-likelihoods, that is also suggested in Müller *et al.* [24].

The approach may be extended to both fixed and random effects selection in the LMM defined with the elliptical distributions. The research effort in this direction is an ongoing study and the results will be reported in due course.

Acknowledgements

The authors thank the Editor and anonymous referees for valuable suggestions that greatly improved the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Fulya Gokalp Yavuz http://orcid.org/0000-0002-7750-9767
Olcay Arslan http://orcid.org/0000-0002-7067-4997

References

[1] A. Azzalini and A. Genz, *The R package `mnormt`: The multivariate normal and t distributions* (2016). Available at http://azzalini.stat.unipd.it/SW/Pkg-mnormt.

[2] H.D. Bondell, A. Krishna and S.K. Ghosh, *Joint variable selection for fixed and random effects in linear mixed-effects models*, Biometrics 66 (2010), pp. 1069–1077.

[3] L. Breiman, *Better subset regression using the nonnegative garrote*, Technometrics 37 (1995), pp. 373–384. Available at http://www.jstor.org/stable/1269730.

[4] L. Breiman, *Heuristics of instability and stabilization in model selection*, Ann. Stat. 24 (1996), pp. 2350–2383. Available at http://www.jstor.org/stable/2242688.

[5] R. Cui, *Variable selection in linear mixed models for longitudinal data*, Ph.D. diss., Harvard University, 2011

[6] C. Eisenhart, *The assumptions underlying the analysis of variance*, Biometrics 3 (1947), pp. 1–21. Available at http://www.jstor.org/stable/3001534.

[7] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Am. Stat. Assoc. 96 (2001), pp. 1348–1360.

[8] J. Fan and R. Li, *New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis*, J. Am. Stat. Assoc. 99 (2004), pp. 710–723. Available at http://dx.doi.org/10.1198/016214504000001060.

[9] Y. Fan, G. Qin and Z.Y. Zhu, *Robust variable selection in linear mixed models*, Commun. Stat. Theory Methods 43 (2014), pp. 4566–4581. Available at http://dx.doi.org/10.1080/03610926.2012.724509.

[10] K.W. Fang, K.T. Kotz and S. Ng, *Symmetric Multivariate and Related Distributions*, Chapman & Hall, London, New York, 1990.

[11] A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl and T. Hothorn, *The R package `mvtnorm`: Multivariate Normal and t Distributions* (2017). Available at http://CRAN.R-project.org/package = mvtnorm.

[12] E. Gómez, M.A. Gomez-Villegas and J.M. Marín, *A multivariate generalization of the power exponential family of distributions*, Commun. Stat. Theory Methods 27 (1998), pp. 589–600. Available at http://dx.doi.org/10.1080/03610929808832115.

[13] D. Harville, *Extension of the Gauss–Markov theorem to include the estimation of random effects*, Ann. Stat. 4 (1976), pp. 384–395.

[14] C.R. Henderson, *Estimation of genetic parameters*, Ann. Math. Stat. 21 (1950), pp. 309–310.

[15] D.R. Hunter and R. Li, *Variable selection using mm algorithms*, Ann. Statist. 33 (2005), pp. 1617–1642. Available at https://doi.org/10.1214/009053605000000200.

[16] J.G. Ibrahim, H. Zhu, R.I. Garcia and R. Guo, *Fixed and random effects selection in mixed effects models*, Biometrics 67 (2011), pp. 495–503.

[17] N. Laird, N. Lange and D. Stram, *Maximum likelihood computations with repeated measures: application of the EM algorithm*, J. Am. Stat. Assoc. 82 (1987), pp. 97–105.

[18] N.M. Laird and J.H. Ware, *Random-effects models for longitudinal data*, Biometrics 38 (1982), pp. 963–74. Available at http://www.ncbi.nlm.nih.gov/pubmed/7168798.

[19] L. Lan, *Variable selection in linear mixed model for longitudinal data*, Ph.D. diss., North Carolina State University, 2006.

[20] K.L. Lange, R.J.A. Little and J.M.G. Taylor, *Robust statistical modeling using the t distribution*, J. Am. Stat. Assoc. 84 (1989), pp. 881–896. Available at http://www.jstor.org/stable/2290063.

[21] J.K. Lindsey, *Multivariate elliptically contoured distributions for repeated measurements*, Biometrics 55 (1999), pp. 1277–1280. Available at http://www.jstor.org/stable/2533755.

[22] M. Lindstrom and D. Bates, *Newton–Raphson and EM algorithms for linear models for repeated-measures data*, J. Am. Stat. Assoc. 83 (1988), pp. 1014–1022.

[23] R.F. Manghi, G.A. Paula and F.J.A. Cysneiros, *On elliptical multilevel models*, J. Appl. Stat. 43 (2016), pp. 2150–2171. Available at http://dx.doi.org/10.1080/02664763.2015.1134445.

[24] S. Müller, J.L. Scealy and A.H. Welsh, *Model selection in linear mixed models*, Stat. Sci. 28 (2013), pp. 135–167. Available at http://arxiv.org/abs/1306.2427.

[25] X. Ni, H.H. Zhang and D. Zhang, *Automatic model selection for partially linear models*, J. Multivar. Anal. 100 (2009), pp. 2100–2111. Available at http://www.sciencedirect.com/science/article/pii/S0047259X09001171.

[26] J. Ocampo, *Effect of porcelain laminate contour on gingival inflammation*, Ph.D. diss., University of Michigan, 2005.

[27] H. Peng and Y. Lu, *Model selection in linear mixed effect models*, J. Multivar. Anal. 109 (2012), pp. 109–129. Available at http://www.sciencedirect.com/science/article/pii/S0047259X12000395.

[28] J.C. Pinheiro and D.M. Bates, *Unconstrained parametrizations for variance–covariance matrices*, Stat Comput. 6 (1996), pp. 289–296.

[29] J.C. Pinheiro, C. Liu and Y.N. Wu, *Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution*, J. Comput. Graph. Stat. 10 (2001), pp. 249–276. Available at http://www.tandfonline.com/doi/abs/10.1198/10618600152628059.

[30] C. Savalli, G.A. Paula and F.J.A. Cysneiros, *Assessment of variance components in elliptical linear mixed models*, Stat. Model. 6 (2006), pp. 59–76. Available at http://smj.sagepub.com/content/6/1/59.abstract.

[31] S.R. Searle, G. Casella and C.E. McCulloch, *Variance Components*, Wiley, New York, 1992.

[32] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B 58 (1996), pp. 267–288.

[33] B.T. West, K.B. Welch and A.T. Galeckl, *Linear Mixed Models: A Practical Guide Using Statistical Software*, , CRC Press, Boca Raton, FL, 2007.

[34] L. Wu, *Mixed Effects Models for Complex Data*, CRC Press, Boca Raton, FL, 2010.

[35] P. Wu, X. Luo, P. Xu and L. Zhu, *New variable selection for linear mixed-effects models*, Ann. Inst. Stat. Math. 69 (2017), pp. 627–646. Available at https://doi.org/10.1007/s10463-016-0555-z.

[36] F.G. Yavuz and O. Arslan, *Linear mixed model with Laplace distribution (llmm)*, Stat. Papers 59 (2018), pp. 271–289.Available at https://doi.org/10.1007/s00362-016-0763-x.

[37] H. Zou, *The adaptive lasso and its oracle properties*, J. Am. Stat. Assoc. 101 (2006), pp. 1418–1429.