Statistica Sinica **27** (2017), 733-760
doi:http://dx.doi.org/10.5705/ss.202015.0292

MIXTURE OF MULTIVARIATE t LINEAR MIXED MODELS FOR MULTI-OUTCOME LONGITUDINAL DATA WITH HETEROGENEITY

Wan-Lun Wang

Feng Chia University

Abstract: The issues of model-based clustering and classification of longitudinal data have received increasing attention in recent years. In this paper, we propose a finite mixture of multivariate t linear mixed-effects model (FM-MtLMM) for analyzing longitudinally measured multi-outcome data arisen from more than one heterogeneous sub-population. The motivation behind this work comes from a cohort study of patients with primary biliary cirrhosis, where the interest is in classifying new patients into two or more prognostic groups on the basis of their longitudinally observed bilirubin and albumin levels. The proposed FM-MtLMM offers robustness and flexibility to accommodate fat tails or atypical observations contained in one or several of the groups. An efficient alternating expectation conditional maximization (AECM) algorithm is employed for the computation of maximum likelihood estimates of parameters. The calculation of standard errors is effected by an information-based method. Practical techniques for clustering of multivariate longitudinal data, estimation of random effects, and classification of future patients are also provided. The methodology is illustrated by analyzing Mayo Clinic Primary Biliary Cirrhosis sequential (PBCseq) data and a simulation study.

Key words and phrases: AECM algorithm, clustering multiple longitudinal profiles, heavy-tailed distribution, maximum likelihood estimation, mixture modeling.

1. Introduction

Finite mixtures of linear mixed models (FM-LMM; Verbeke and Lesaffre (1996)), combining the potency of (univariate) linear mixed models (LMM; Laird and Ware (1982)) and finite normal mixture models (McLachlan and Peel (2000)), have emerged as one of the most effective tools for clustering grouped longitudinal data in which a single continuous outcome is observed and class memberships may not be known a priori. De la Cruz-Mesía, Quintana and Marshall (2008) proposed a nonlinear formulation of FM-LMM for classification of hormone trajectories with nonlinear profiles. The clustering method for discrete longitudinal data can be carried out in a similar way by replacing LMMs with generalized linear mixed models (GLMM; Molenberghs and Verbeke (2005)), where a normal mixture framework is assumed for the random effects (Spiessens, Verbeke and Komárek (2002)); Komárek and Lesaffre (2008)). Earlier developments of FM-LMM can be found in Spiessens, Verbeke and Komárek (2002), Gaffney and Smyth (2003), Pfeifer (2004), Celeux, Martin and Lavergne (2005), Ng et al. (2006), and Booth, Casella and Hobert (2008), among others. In recent years, a number of authors, for example, Pinheiro, Liu and Wu (2001), Rosa, Gianola and Padovani (2004), Lin and Lee (2006), Lin and Lee (2007), and Song, Zhang and Qu (2007), have put forward robust generalizations of LMM using the multivariate t distribution (Kotz and Nadarajah (2004)), known as the t linear mixed model (tLMM). Bai, Chen and Yao (2016) presented a finite mixture of tLMMs (FM-tLMM) to accommodate heterogeneity among repeated measures. However, the application of FM-tLMM approach is limited to single-outcome longitudinal data.

In many biomedical studies or clinical trials, it is common to have data with more than one response variables on the same subject measured repeatedly over time leading to multivariate longitudinal data. For handling such data, Shah, Laird and Schoenfeld (1997) pioneered the introduction of a multivariate generalization of LMM by exploiting a correlation structure across responses: the multivariate linear mixed-effects model (MLMM). They developed an iterative EM algorithm (Dempster, Laird and Rubin (1977)) to estimate the parameters of the model. Villarroel, Marshall and Barón (2009) performed cluster analysis using the multivariate nonlinear mixed-effects model (MNLMM) proposed by Marshall et al. (2006), who described an EM algorithm for parameter estimation based on the first-order Taylor approximation. In the case where group labels are predefined, Marshall et al. (2009) developed a discrimination procedure using the MNLMM for predicting the class membership of future subjects. Komárek and Komárková (2013) investigated the problem of clustering for multivariate mixed-type longitudinal data using an appropriate Bayesian Markov chain Monte Carlo (MCMC) scheme. For robust inference against potential outliers in multi-outcome longitudinal data, Wang and Fan (2011) have extended the tLMM to a multivariate version, called the multivariate t linear mixed model (MtLMM). Further developments along this line can be found in Wang (2013) and Wang and Lin (2014). In spite of having robustness against non-normality due to outliers or atypical (influential) observations, the MtLMM is still limited to its practical use for clustering and classification of grouped multivariate longitudinal data.

The objective of this paper is to establish a framework for providing extra flexibility, called finite mixtures of multivariate t linear mixed-effects model (FM-MtLMM), constructed by imposing mixtures of multivariate t distributions (Peel and McLachlan (2000)) for the random effects and within-subject errors jointly.

In this model, the number of components is fixed but possibly unknown and can be determined using the penalized likelihood-based information selection criteria. Notably, the proposed FM-MtLMM allows practitioners to simultaneously cluster/classify multiple longitudinal profiles into several internally homogeneous sub-populations, to describe association-of-the-evolutions and evolution-of-the-association (Wang (2013)) for multi-outcome repeated measures, and to capture the fat-tailed phenomena existing in the data.

Maximum likelihood (ML) estimation of the FM-MtLMM is considerably more complicated than that of single-component MtLMM because its mixture framework does not offer explicit analytical solutions for the ML estimators of model parameters. To cope with the computational difficulty, we develop an efficient alternating expectation conditional maximization (AECM) algorithm (Meng and van Dyk (1997)) on the basis of three convenient hierarchical representations. Once the parameter estimates have been obtained, it is more meaningful in practice to cluster subjects into a pre-specified number of groups even without a priori known class memberships and further to discriminate new subjects based on the results of training data. Therefore, the problems of model-based clustering of multi-outcome longitudinal profiles and discrimination of external subjects are also investigated.

The outline for the rest of this paper is as follows. Section 2 describes a motivating example concerning a preliminary analysis of Primary Biliary Cirrhosis sequential (PBCseq) data. Section 3 introduces the proposed FM-MtLMM and presents some relevant properties. Section 4 presents an efficient AECM algorithm for parameter estimation and an approximation method for standard-errors calculation. Practical issues on the clustering of longitudinal profiles, estimation of random effects and discriminant analysis for new subjects are also provided. The proposed methodology is illustrated in Section 5 with the analysis of PBCseq data and in Section 6 with a simulation study. Section 7 delivers summaries of the paper and some directions for future research. Technical details and additional computing results are sketched in the supplementary document.

2. Motivating Example: Primary Biliary Cirrhosis Sequential Data

In a follow-up study of primary biliary cirrhosis, participants were repeatedly measured for their serum bilirubin and serum albumin, as well as other fractions of blood and plasma. An extremely higher level than the standard that bilirubin is excreted in bile and urine can indicate certain diseases. Serum albumin may be harmful to humans having too high or too low circulating serum albumin levels. Typically, it is believed that there exist some relationships between serum bilirubin and serum albumin levels, and thus a joint analysis of the longitudinally collected bilirubin and albumin has received increasing emphasis in diagnosing liver diseases. The issues of how the bilirubin/albumin levels evolve over time, how the evolution of bilirubin is related to the evolution of albumin, and how the association between bilirubin and albumin evolves over time are natural ones.

2.1. Description of the PBCseq data

In a subset of the data from the Primary Biliary Cirrhosis sequential (PBCseq) cohort study, 312 patients were recruited from the Mayo Clinic between January 1974 and May 1984, and participated in either of two double-blind, placebo-controlled, randomized trials with D-penicillamine for treating primary biliary cirrhosis until April 1988. A clinical laboratory database was established on each patient who was collected repeatedly and prospectively at yearly intervals under standardized forms, definitions, and study protocols. The collected variables comprised ID number; five time variables, including *age* and total number of follow-up days; eight categorical variables, including *sex*, *drug* and *status*; two censoring indicators for events; and seven continuous measurement variables, including the natural logarithm scale of *bili* and *albumin*. A total of 1945 visit rows and 38 variables on the 312 randomized patients are freely available from the R package mixAK (Komárek and Komárková (2014)) and electronically at `http://lib.stat.cmu.edu/datasets/pbcseq`.

Among these patients (36 males and 276 females), the total number of follow-up days ranges from 41 to 5,225 days, and the age at entry ranges from 26 to 79 years. At the endpoint of this cohort study, 140 of the patients had died, Group 1, while 172 were known to be alive, Group 0. A comprehensive description of the clinical background can be found in Dickson et al. (1989), Markus et al. (1989), and Fleming and Harrington (1991); the data have been previously analyzed by Murtaugh et al. (1994).

Biomedical research indicates that serum bilirubin and serum albumin are two of primary indicators to help evaluate and track the absence of liver diseases. Orthotopic liver transplantation can be treated as potentially life-saving alternative for patients with advanced or end-stage primary biliary cirrhosis. As a consequence, we concentrate on modeling the dependence of the longitudinal profiles of two markers, say natural logarithm of serum bilirubin (*lbili*) and the natural logarithm of serum albumin (*lalbumin*), on time (visited years) and other covariates of interest (e.g., sex, drug, age). Investigating how to cluster or classify the bivariate longitudinal markers can raise many new statistical interests and challenges.

2.2. Preliminary analysis

Figure 1 displays the trajectories for exploring the evolution of lbili and lalbumin markers in Groups 0 and 1. It can be observed that the trend of population mean profiles vary over time, and patients differ in their initial levels and time trends between the two groups. Supplementary Table S1 presents the observed correlations between lbili and lalbumin levels across time (in diagonal) along with the observed autocorrelations of each response at lags 1–16 in terms of scheduled years (lbili in upper-triangular entries and lalbumin in lower-triangular entries). Because the response variables are measured irregularly, observations that are not measured on schedule are treated as at nearest regularly scheduled years in the computation of correlations. From Table S1, the two markers appear to be negatively correlated across time. Meanwhile, the magnitude of autocorrelation on each outcome may decay across time lag. Thus, an uncorrelated structure or a compound symmetry assumption on within-patient variability of each outcome across time may be inappropriate.

Let y_{i1k} and y_{i2k} be the levels of lbili and lalbumin markers, respectively, for the ith patient measured at the kth occasion. Assuming a linear trend in time for the population average and patient-specific intercepts and slopes for the random effects, we preliminarily fit LMMs for y_{i1k} and y_{i2k} separately by using the lme (Pinheiro et al. (2014)) R package:

$$y_{ijk} = \beta_{j0} + \beta_{j1} t_{ik} + b_{ij0} + b_{ij1} t_{ik} + e_{ijk}, \quad \text{for } j = 1, 2,$$

where $(\beta_{j0},\ \beta_{j1})$ are the regression coefficients of fixed effects for marker j, (b_{ij0}, b_{ij1}) are multivariate normally distributed random effects, and independent of $(e_{ij1}, \ldots, e_{ijs_i})$, which follow a multivariate normal distribution with zero mean and variance-covariance matrix of a continuous-type autoregressive process of order 1, and $t_{ik} = \text{month}_{ik}/12$ is the kth visit years for patient i.

The upper panel of Figure 2 displays scatter plots along with the 95% confidence ellipses and summary histograms of empirical Bayes estimates of random intercepts and slopes obtained after fitting the LMMs to the two markers. Different colors and symbols represent different groups. In the lower panel of this figure, scatter plots of fitted values against residuals, along with their 95% confidence ellipses and boxplots of residuals, for lbili are shown in the two graphs on the left-hand side, while those for lalbumin are shown in the two graphs on the right-hand side. The scatter plots exhibit a difference in the variations of random effects and within-subject errors between the two groups, suggesting that the homogeneity assumption for the underlying distributions of random effects and within-subject errors might not be realistic. The boxplots exhibit the heavy-tailed phenomenon for residuals, especially for those patients in Group 1, revealing that some atypical observations or outliers might exist in the data. From these findings, the routine homogeneity and normality assumptions for the random effects and errors appear inappropriate for modeling this data set.

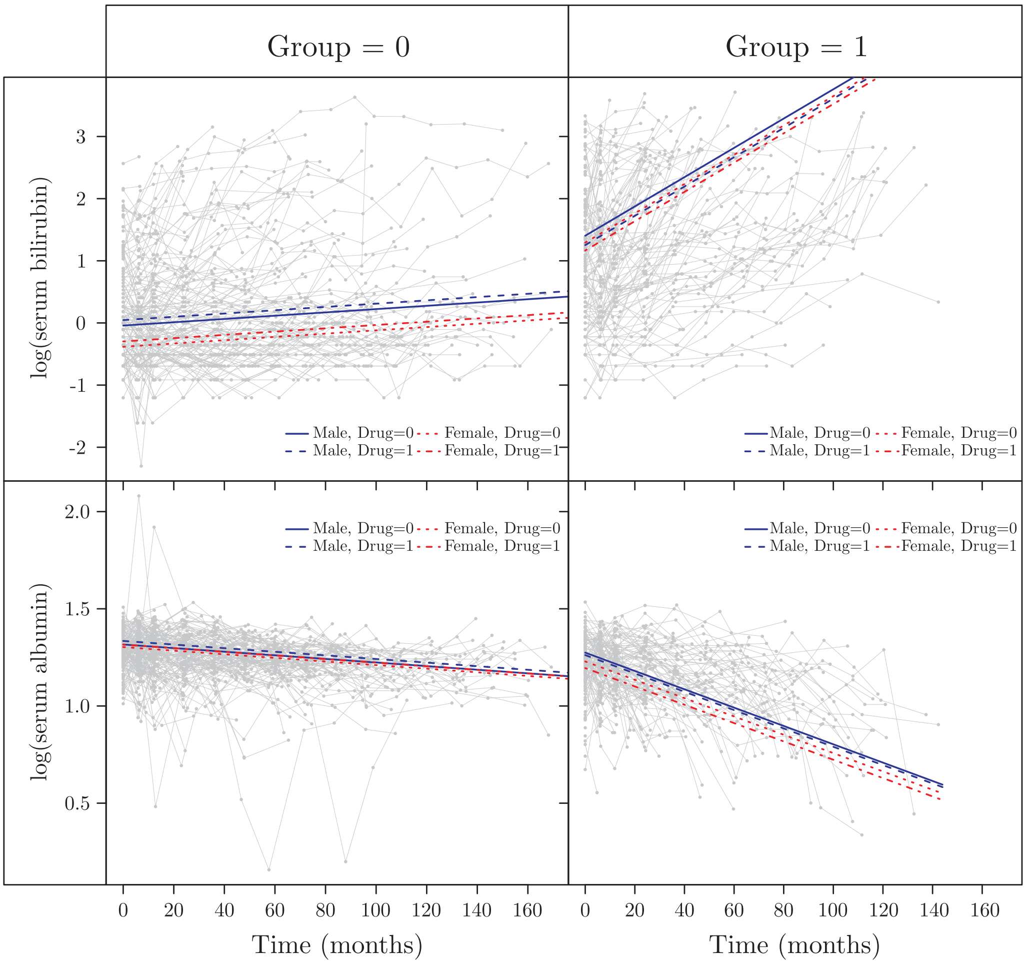

Figure 1. Trajectories plots for the PBCseq data. Observed evolution (in gray) of lbili and lalbumin markers for 312 patients. Solid and dashed (dot and dot-dashed) lines show the fitted mean profiles of male (female) patients who were treated with placebo (Drug=0) and D-penicillamine (Drug=1), respectively, with mean random effects in Group 0 and Group 1 under the fitted FM-MtNLMM with RIS and DEC errors.

Indeed, a misspecified distribution for random quantities in the model can seriously influence parameter estimates as well as their standard errors, subsequently leading to invalid statistical inferences. Besides, a separate analysis of the two markers can lose important information about evolutional relationships among multiple responses across time. This motivates us to establish a more robust and flexible model.

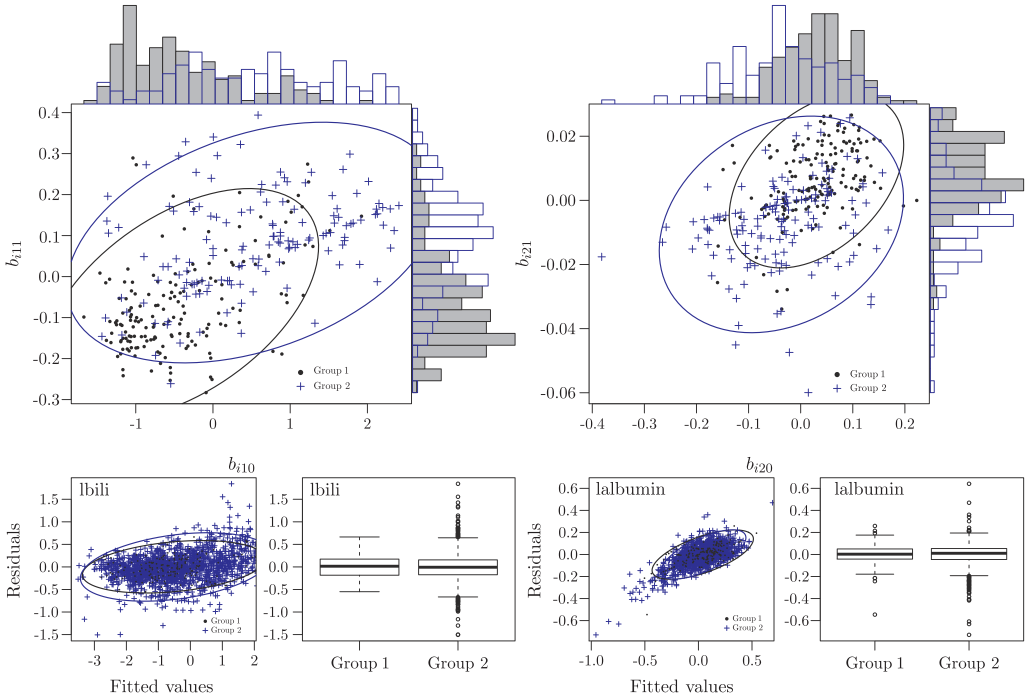

Figure 2. (Upper panel) Scatter plots along with the 95% confidence ellipses and histogram of empirical Bayes estimates for random effects; (Lower left/right panel) Scatter plots along with the 95% confidence ellipses and boxplots of residuals for lbili and lalbumin under the fitted LMMs for lbili and lalbumin markers, separately.

3. Model Formulation

Suppose that there are n subjects from G heterogeneous groups in a study, where the ith subject has r outcome variables measured repeatedly over s_i time points. Let $\mathbf{Y}_i = [\mathbf{y}_{i1} : \cdots : \mathbf{y}_{ir}]$ be a $s_i \times r$ matrix of responses for subject i $(i = 1, \ldots, n)$, where each $\mathbf{y}_{ij} = (y_{ij,1}, \ldots, y_{ij,s_i})^{\mathrm{T}}$ is a $s_i \times 1$ response vector for outcome j $(j = 1, \ldots, r)$. Let $\mathbf{X}_i = \mathrm{diag}\{\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ir}\}$ and $\mathbf{Z}_i = \mathrm{diag}\{\mathbf{Z}_{i1}, \ldots, \mathbf{Z}_{ir}\}$, where $\mathbf{X}_{ij}$ is a $s_i \times p_j$ full-rank design matrix for fixed effects associated with $\mathbf{y}_{ij}$, and $\mathbf{Z}_{ij}$, formed usually by a subset of $\mathbf{X}_{ij}$, is a $s_i \times q_j$ design matrix for random effects. The block-diagonal structures of $\mathbf{X}_i$ and $\mathbf{Z}_i$ allow the analysts to link the grand and subject-specific relationships between covariates and each response, which is collected repeatedly at unequally spaced occasions for each subject, via distinct design matrices for each response. Let $\mathbf{E}_{ig} = [\mathbf{e}_{i1,g} : \ldots : \mathbf{e}_{ir,g}] = [\mathbf{e}^{\mathrm{T}}_{i\cdot 1,g} : \ldots : \mathbf{e}^{\mathrm{T}}_{i\cdot s_i,g}]^{\mathrm{T}}$ be the $s_i \times r$ matrix of within-subject errors of the gth component corresponding to $\mathbf{Y}_i$, where $\mathbf{e}_{ij,g}$ is a $s_i \times 1$ error vector of component g for outcome j over s_i occasions, and $\mathbf{e}_{i\cdot k,g}$ is a $1\times r$ error vector of component g for all outcomes at the same occasion ($k = 1, \ldots, s_i$). For notational convenience, we write $\mathbf{y}_i = \text{vec}(\mathbf{Y}_i)$ and $\boldsymbol{\varepsilon}_{ig} = \text{vec}(\mathbf{E}_{ig})$, where $\text{vec}(\cdot)$ is the vectorization operator. Finally, we let $n_i = s_i r$, and $p = \sum_{j=1}^{r} p_j$ and $q = \sum_{j=1}^{r} q_j$ be the total dimensions of fixed effects and random effects, respectively.

The FM-MtLMM for the ith subject can be formulated as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_g + \mathbf{Z}_i \mathbf{b}_{ig} + \boldsymbol{\varepsilon}_{ig} \quad \text{with mixing probability } w_g, \tag{3.1}$$

for $g = 1, \ldots, G$ subject to $\sum_{g=1}^{G} w_g = 1$, along with the assumption that

$$\begin{bmatrix} \mathbf{b}_{ig} \\ \boldsymbol{\varepsilon}_{ig} \end{bmatrix} \sim \mathcal{T}_{q+n_i}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{D}_g & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{ig} \end{bmatrix}, \nu_g \right), \tag{3.2}$$

where $\mathcal{T}_d(\boldsymbol{\mu}, \boldsymbol{\Omega}, \nu)$ denotes the multivariate t distribution with dimension d, location vector $\boldsymbol{\mu}$, scale-covariance matrix $\boldsymbol{\Omega}$, and degrees of freedom (DOF) ν. Here, $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{g1}^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_{gr}^{\mathrm{T}})^{\mathrm{T}}$ is a $p \times 1$ vector of fixed effects of the gth component with each $p_j \times 1$ sub-vector $\boldsymbol{\beta}_{gj}$ used to describe the component mean profile of outcome j, $\mathbf{b}_{ig} = (\mathbf{b}_{ig,1}^{\mathrm{T}}, \ldots, \mathbf{b}_{ig,r}^{\mathrm{T}})^{\mathrm{T}}$ is a $q \times 1$ vector of (unobservable) component random effects with each $q_j \times 1$ sub-vector $\mathbf{b}_{ig,j}$ corresponding to subject-specific features on $\mathbf{y}_{ij}$, $\mathbf{D}_g$ and $\mathbf{R}_{ig}$ are scale-covariance matrices for component random effects and component within-subject errors, respectively, and ν_g is the component DOF.

For the sake of parsimony, we can assume that $\mathbf{e}_{ij,g} \sim \mathcal{T}_{s_i}(\mathbf{0}, \sigma_{jj,g}\mathbf{C}_{ig}, \nu_g)$ for $j = 1, \ldots, r$, and $\mathbf{e}_{i \cdot k, g} \sim \mathcal{T}_r(\mathbf{0}, \boldsymbol{\Sigma}_g, \nu_g)$ for $k = 1, \ldots, s_i$, where $\boldsymbol{\Sigma}_g = [\sigma_{jj',g}] \in \mathbb{R}^{r \times r}$ is used to describe the variances and covariances among r outcome variables, and $\mathbf{C}_{ig} \in [-1, 1]^{s_i \times s_i}$ is a time-dependence correlation matrix used to address possibly serial correlation among s_i irregularly observed occasions. Accordingly, the within-subject error matrix $\mathbf{E}_{ig}$ follows the matrix-t distribution (Kibria (2006)), and thereby the stacked $n_i \times 1$ vector $\boldsymbol{\varepsilon}_{ig}$ follows the multivariate t distribution with the DOF ν_g, location vector zero, and scale-covariance matrix of having a Kronecker product (KP) structure, written as $\mathbf{R}_{ig} = \boldsymbol{\Sigma}_g \otimes \mathbf{C}_{ig}$, which helps us to estimate $\mathbf{R}_{ig}$ more accurately. As suggested by Galecki (1994), to avoid the non-identifiability problem resulting from non-unique solutions of $\boldsymbol{\Sigma}_g$ and $\mathbf{C}_{ig}$ in estimating $\mathbf{R}_{ig}$ with a KP structure (Lee et al. (2013)), we need to specify $\mathbf{C}_{ig}$ as a correlation matrix rather than a covariance matrix. To make estimation of $\mathbf{C}_{ig}$ more precise, we could choose a parsimonious structure on this correlation matrix, which can be a function of parameters ρ_g as well as time points $\mathbf{t}_i$, denoted by $\mathbf{C}_{ig} = \mathbf{C}_i(\rho_g)$, based on the characteristics of the data at hand.

Under (3.1) and (3.2), the marginal density of $\mathbf{y}_i$ is

$$f(\mathbf{y}_i) = \sum_{g=1}^{G} w_g t_{n_i}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_g, \boldsymbol{\Lambda}_{ig}, \nu_g), \tag{3.3}$$

where $\boldsymbol{\Lambda}_{ig} = \mathbf{Z}_i\mathbf{D}_g\mathbf{Z}_i^{\mathrm{T}} + \boldsymbol{\Sigma}_g \otimes \mathbf{C}_{ig}$, and $t_d(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ is the probability density function (pdf) of d-variate t distribution with location vector $\boldsymbol{\mu}$, scale-covariance matrix $\boldsymbol{\Sigma}$, and DOF ν. Summing over the natural logarithm of the marginal densities of $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^{n}$ gives the log-likelihood function of the full model parameter $\boldsymbol{\Theta} = \{w_g, \boldsymbol{\beta}_g, \mathbf{D}_g, \boldsymbol{\Sigma}_g, \rho_g, \nu_g\}_{g=1}^{G}$, denoted by $\ell(\boldsymbol{\Theta}|\mathbf{y}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i)$. As there is no explicit analytical solution for the ML estimator $\hat{\boldsymbol{\Theta}}$ through a direct maximization of $\ell(\boldsymbol{\Theta}|\mathbf{y})$, we utilize the AECM algorithm described in Section 4.1.

To develop the AECM algorithm, we present several hierarchies of the FM-MtLMM by introducing the allocation indicator vector $\mathbf{u}_i = (u_{i1}, \ldots, u_{iG})$, in which the entry $u_{ig} = 1$ if $\mathbf{y}_i$ belongs to the gth group and $u_{ig} = 0$ otherwise. Accordingly, the vector $\mathbf{u}_i$ independently follows a multinomial distribution with one trial and cell probabilities $(w_1, \ldots, w_G)$ subject to $\sum_{g=1}^{G} w_g = 1$. As such, a two-level hierarchy of the FM-MtLMM takes the form of

$$\begin{aligned} \mathbf{y}_i|(u_{ig} = 1) &\sim \mathcal{T}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}_g, \boldsymbol{\Lambda}_{ig}, \nu_g), \\ \mathbf{u}_i &\sim \mathcal{M}(1, w_1, \ldots, w_G). \end{aligned} \tag{3.4}$$

Using the definition of the multivariate t distribution in conjunction with the marginal distribution of $\mathbf{u}_i$ in (3.4), we obtain a flexible three-level hierarchy:

$$\begin{aligned} \mathbf{y}_i|(\tau_i, u_{ig} = 1) &\sim \mathcal{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}_g, \tau_i^{-1}\boldsymbol{\Lambda}_{ig}), \\ \tau_i|(u_{ig} = 1) &\sim \mathrm{Gamma}\Big(\frac{\nu_g}{2}, \frac{\nu_g}{2}\Big), \end{aligned} \tag{3.5}$$

where the τ_i's are independent and identically distributed (*i.i.d.*) latent inverse variances that follow the gamma distribution with shape $\nu_g/2$ and rate $\nu_g/2$ given $u_{ig} = 1$. Combining the conditional distribution of τ_i given $u_{ig} = 1$ in (3.5) and the marginal distribution of $\mathbf{u}_i$ in (3.4) leads to the four-level hierarchy:

$$\begin{aligned} \mathbf{y}_i|(\mathbf{b}_{ig}, \tau_i, u_{ig} = 1) &\sim \mathcal{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}_g + \mathbf{Z}_i\mathbf{b}_{ig}, \tau_i^{-1}\mathbf{R}_{ig}), \\ \mathbf{b}_{ig}|(\tau_i, u_{ig} = 1) &\sim \mathcal{N}_q(\mathbf{0}, \tau_i^{-1}\mathbf{D}_g). \end{aligned} \tag{3.6}$$

4. Computation Methodology

4.1. Parameter estimation via the AECM algorithm

The AECM algorithm (Meng and van Dyk (1997)), a variant of the EM algorithm (Dempster, Laird and Rubin (1977)), uses different complete data of the model in order to obtain simple closed-form expressions of updating estimators and achieve acceleration of the algorithm. To employ the AECM algorithm for the fitting of FM-MtLMM, we partition the set of unknown parameters $\mathbf{\Theta}$ into the subsets $\mathbf{\Theta}_1 = \{w_g, \rho_g, \nu_g\}_{g=1}^G$, $\mathbf{\Theta}_2 = \{\boldsymbol{\beta}_g\}_{g=1}^G$ and $\mathbf{\Theta}_3 = \{\mathbf{D}_g, \mathbf{\Sigma}_g\}_{g=1}^G$. In each iteration, the AECM algorithm consists of three cycles with each cycle updating different subsets of parameters based on three hierarchical forms of the FM-MtLMM, say (3.4), (3.5) and (3.6). The following result is useful for the evaluation of required conditional expectations involved in $Q^{[1]}$ and $Q^{[2]}$ functions described in detail in Supplementary Material.

Proposition 1. *From* (3.4) *to* (3.6), *the conditional probability of the allocation indicator* u_{ig} *given* $\mathbf{y}_i$, *and the conditional distributions of random inverse variance* τ_i *and random effects* $\mathbf{b}_{ig}$ *given* $\mathbf{y}_i$ *and* $u_{ig} = 1$ *are*

$$p_{ig} = P(u_{ig} = 1|\mathbf{y}_i) = \frac{w_g t_{n_i}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_g, \mathbf{\Lambda}_{ig}, \nu_g)}{\sum_{l=1}^G w_l t_{n_i}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}_l, \mathbf{\Lambda}_{il}, \nu_l)}, \tag{4.1}$$

$$\tau_i|(\mathbf{y}_i, u_{ig} = 1) \sim Gamma\Big(\frac{\nu_g + n_i}{2}, \frac{\nu_g + \Delta_{ig}}{2}\Big),$$

$$\mathbf{b}_{ig}|(\mathbf{y}_i, u_{ig} = 1) \sim \mathcal{T}_q\Big(\mathbf{D}_g\mathbf{Z}_i^{\mathrm{T}}\mathbf{\Lambda}_{ig}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_g), \Big(\frac{\nu_g + \Delta_{ig}}{\nu_g + n_i}\Big)\mathbf{V}_{\boldsymbol{b}_{ig}}, \nu_g + n_i\Big),$$

where $\Delta_{ig} = (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_g)^{\mathrm{T}}\mathbf{\Lambda}_{ig}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_g)$ *and* $\mathbf{V}_{\boldsymbol{b}_{ig}} = (\mathbf{D}_g^{-1} + \mathbf{Z}_i^{\mathrm{T}}\mathbf{R}_{ig}^{-1}\mathbf{Z}_i)^{-1}$.

Proof: The proof of the proposition is straightforward based on the Bayes' rule and standard matrix factorizations (Anderson (2003)).

Given an appropriate initial value of the model parameter $\hat{\mathbf{\Theta}}^{(0)}$, the AECM algorithm proceeds as follows.

The 1st cycle

E-step: Evaluate the conditional expectation of complete-data log-likelihood function (S.1) given the observed data $\mathbf{y}$ and current values $\hat{\mathbf{\Theta}}^{(h)} = (\hat{\mathbf{\Theta}}_1^{(h)}, \hat{\mathbf{\Theta}}_2^{(h)}, \hat{\mathbf{\Theta}}_3^{(h)})$, which gives the $Q^{[1]}$ function.

CM-step: Update $\hat{w}_g^{(h)}$ by maximizing the $Q^{[1]}$ function, yielding $\hat{w}_g^{(h+1)} = \sum_{i=1}^n \hat{u}_{ig}^{(h)}/n$.

CML-step: Update $\hat{\rho}_g^{(h)}$ and $\hat{\nu}_g^{(h)}$ by maximizing the constrained actual log-likelihood functions evaluated at $w_g = \hat{w}_g^{(h)}$, $\boldsymbol{\beta}_g = \hat{\boldsymbol{\beta}}_g^{(h)}$, $\mathbf{D}_g = \hat{\mathbf{D}}_g^{(h)}$ and $\mathbf{\Sigma}_g = \hat{\mathbf{\Sigma}}_g^{(h)}$.

This can be done by carrying out the default-install R `optim` function (R Development Core Team. (2014)) subject to a two-dimensional box constraint. The `optim` command is a general-purpose optimization routine based on Nelder-Mead, quasi-Newton and conjugate-gradient algorithms.

The 2nd cycle

E-step: Given $\mathbf{y}$ and $\hat{\boldsymbol{\Theta}}^{(h+1/3)} = (\hat{\boldsymbol{\Theta}}_1^{(h+1)}, \hat{\boldsymbol{\Theta}}_2^{(h)}, \hat{\boldsymbol{\Theta}}_3^{(h)})$, evaluate the conditional expectation of complete-data log-likelihood function (S.3), which gives the $Q^{[2]}$ function.

CM-step: Update $\hat{\boldsymbol{\beta}}_g^{(h)}$ by maximizing the $Q^{[2]}$ function, yielding

$$\hat{\boldsymbol{\beta}}_g^{(h+1)} = \Big(\sum_{i=1}^{n} \hat{u}_{ig}^{(h)} \hat{\tau}_{ig}^{(h)} \mathbf{X}_i^{\mathrm{T}} \hat{\boldsymbol{\Lambda}}_{ig}^{(h)^{-1}} \mathbf{X}_i\Big)^{-1} \Big(\sum_{i=1}^{n} \hat{u}_{ig}^{(h)} \hat{\tau}_{ig}^{(h)} \mathbf{X}_i^{\mathrm{T}} \hat{\boldsymbol{\Lambda}}_{ig}^{(h)^{-1}} \mathbf{y}_i\Big), \tag{4.2}$$

where $\hat{u}_{ig}^{(h)}$ is calculated by (4.1) evaluated at $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{(h+1/3)}$, and $\hat{\tau}_{ig}^{(h)} = (\hat{\nu}_g^{(h)} + n_i)\big/(\hat{\nu}_g^{(h)} + \hat{\Delta}_{ig}^{(h)})$ with $\hat{\Delta}_{ig}^{(h)}$ being Δ_{ig} evaluated at $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{(h+1/3)}$.

The 3rd cycle

E-step: Evaluating the conditional expectation of the complete-data log-likelihood function (S.7), given $\mathbf{y}$ and $\hat{\boldsymbol{\Theta}}^{(h+2/3)} = (\hat{\boldsymbol{\Theta}}_1^{(h+1)}, \hat{\boldsymbol{\Theta}}_2^{(h+1)}, \hat{\boldsymbol{\Theta}}_3^{(h)})$, leads to the $Q^{[3]}$ function.

CM-step: Updating $\hat{\mathbf{D}}_g^{(h)}$ and $\hat{\boldsymbol{\Sigma}}_g^{(h)} = [\hat{\sigma}_{g,ls}^{(h)}]$ by maximizing the $Q^{[3]}$ function gives

$$\hat{\mathbf{D}}_g^{(h+1)} = \sum_{i=1}^{n} \frac{\hat{u}_{ig}^{(h)} \widehat{\tau \mathbf{B}}_{ig}^{(h)}}{\sum_{i=1}^{n} \hat{u}_{ig}^{(h)}}, \tag{4.3}$$

$$\hat{\sigma}_{g,ls}^{(h+1)} = \begin{cases} \big(\sum_{i=1}^{n} s_i \hat{u}_{ig}^{(h)}\big)^{-1} \sum_{i=1}^{n} \hat{u}_{ig}^{(h)} \mathrm{tr}\big(\hat{\mathbf{C}}_i^{-1}(\hat{\rho}_g^{(h)}) \hat{\boldsymbol{\psi}}_{ig,ls}^{(h)}(\hat{\boldsymbol{\beta}}_g^{(h+1)})\big), & l = s, \\ \big(2\sum_{i=1}^{n} s_i \hat{u}_{ig}^{(h)}\big)^{-1} \sum_{i=1}^{n} \hat{u}_{ig}^{(h)} \mathrm{tr}\big(\hat{\mathbf{C}}_i^{-1}(\hat{\rho}_g^{(h)}) \big[\hat{\boldsymbol{\psi}}_{ig,ls}^{(h)}(\hat{\boldsymbol{\beta}}_g^{(h+1)}) \\ \qquad + \hat{\boldsymbol{\psi}}_{ig,sl}^{(h)}(\hat{\boldsymbol{\beta}}_g^{(h+1)})\big]\big), & l \neq s, \end{cases} \tag{4.4}$$

for $l, s = 1, \ldots, r$, where $\widehat{\tau \mathbf{B}}_{ig}^{(h)} = \hat{\tau}_{ig}^{(h)} \hat{\mathbf{b}}_{ig}^{(h)} \hat{\mathbf{b}}_{ig}^{(h)^{\mathrm{T}}} + \hat{\mathbf{V}}_{\boldsymbol{b}_{ig}}^{(h)}$, and $\hat{\boldsymbol{\psi}}_{ig,ls}^{(h)}(\boldsymbol{\beta}_g) = \hat{\tau}_{ig}^{(h)} \hat{\mathbf{e}}_{ig,l}^{(h)} \hat{\mathbf{e}}_{ig,s}^{(h)\mathrm{T}} + \mathbf{Z}_{il} \hat{\mathbf{V}}_{\boldsymbol{b}_{ig,ls}}^{(h)} \mathbf{Z}_{is}^{\mathrm{T}}$ is a $s_i \times s_i$ square submatrix of $\widehat{\tau \mathbf{E}}_{ig}^{(h)}(\boldsymbol{\beta}_g)$ given in (S.9) with $\hat{\mathbf{b}}_{ig}^{(h)} = \hat{\mathbf{D}}_g^{(h)} \mathbf{Z}_i^{\mathrm{T}} \hat{\boldsymbol{\Lambda}}_{ig}^{(h)^{-1}} (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}_g^{(h)})$, $\hat{\mathbf{V}}_{\boldsymbol{b}_{ig}}^{(h)}$ being $\mathbf{V}_{\boldsymbol{b}_{ig}}$ defined in Proposition 1 and evaluated at $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}^{(h+2/3)}$, $\hat{\mathbf{e}}_{ig,l}^{(h)} = \mathbf{y}_{il} - \mathbf{X}_{il} \boldsymbol{\beta}_{gl} - \mathbf{Z}_{il} \hat{\mathbf{b}}_{ig,l}^{(h)}$, and $\hat{\mathbf{V}}_{\boldsymbol{b}_{ig,ls}}^{(h)}$ a $q_l \times q_s$ submatrix consisting of the $(\sum_{j=1}^{l-1} q_j + 1)$th to $(\sum_{j=1}^{l} q_j)$th rows and the $(\sum_{j=1}^{s-1} q_j + 1)$th to $(\sum_{j=1}^{s} q_j)$th columns of $\hat{\mathbf{V}}_{\boldsymbol{b}_{ig}}^{(h)}$.

The E-steps and CM/CML-steps within each cycle continue until a user's specified tolerance or the default maximum number of iterations is met. Upon convergence, we obtain the ML estimates, denoted by $\hat{\boldsymbol{\Theta}} = \{\hat{w}_g, \hat{\boldsymbol{\beta}}_g, \hat{\mathbf{D}}_g, \hat{\boldsymbol{\Sigma}}_g, \hat{\rho}_g, \hat{\nu}_g\}_{g=1}^G$. Additional details on the implementation of AECM algorithm for the proposed FM-MtLMM are sketched in Supplementary S.1. In most cases, the EM-type algorithm is not guaranteed to find the global optimum. A poor convergence criterion might lead to premature convergence and subsequently trap one in a spurious solution. These typically are present when the artifactual components are needed to add to cover very small-scale groups of extremely outlying observations (McLachlan and Peel (2000)). Following the strategy adopted in Wang, Ng and McLachlan (2009), we have made a slight modification of the algorithm by adding a simple singularity handling procedure to rule out the occurrence of spurious maxima. To assess the convergence of the AECM algorithm in a strict manner, we make use of Aitken's acceleration method (Aitken (1926); and McLachlan and Krishnan (2008)) for alleviating a premature termination. Letting $\ell^{(h)}$ be the likelihood value evaluated at $\hat{\boldsymbol{\theta}}^{(h)}$, the Aitken accelerated estimate of the log-likelihood at iteration h is calculated as

$$\ell_\infty^{(h+1)} = \ell^{(h)} + \frac{\ell^{(h+1)} - \ell^{(h)}}{1 - a^{(h)}},$$

where $a^{(h)} = (\ell^{(h+1)} - \ell^{(h)})/(\ell^{(h)} - \ell^{(h-1)})$ is the Aitken acceleration factor. The algorithm is stopped as soon as $\ell_\infty^{(h)} - \ell^{(h)} < \epsilon$, where $\epsilon = 10^{-5}$ was used in our numerical experiments.

Rewrite $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_G)$, where $\boldsymbol{\theta}_g = (w_g, \boldsymbol{\beta}_g, \boldsymbol{\alpha}_g, \nu_g)$ represents the parameter vector involved in the gth component with $\boldsymbol{\alpha}_g = (\text{vech}(\mathbf{D}_g), \text{vech}(\boldsymbol{\Sigma}_g), \rho_g)$, for $g = 1, \cdots, G$. The natural logarithm of the multiplication of the pdfs of hierarchy (3.5) for all subjects leads to the complete-data log-likelihood function of parameters $\boldsymbol{\Theta}$. Taking the first and second derivatives of which with respect to each entry of parameters, we have the score vector and Hessian matrix for $\boldsymbol{\Theta}$,

$$\mathbf{s}(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{u}) = \left(\mathbf{s}_{\boldsymbol{\theta}_1}^{\mathrm{T}}, \ldots, \mathbf{s}_{\boldsymbol{\theta}_G}^{\mathrm{T}}\right)^{\mathrm{T}} \quad \text{and} \quad \mathbf{H}(\boldsymbol{\Theta}; \mathbf{y}, \mathbf{u}) = \text{diag}\Big(\{\mathbf{H}_{\boldsymbol{\theta}_g\boldsymbol{\theta}_g}\}_{g=1}^G\Big),$$

where the sub-entry $\mathbf{s}_{\boldsymbol{\theta}_g}$ and the block-diagonal submatrix $\mathbf{H}_{\boldsymbol{\theta}_g\boldsymbol{\theta}_g}$, for $g = 1, \cdots, G$, can be expressed by $\mathbf{s}_{\boldsymbol{\theta}_g} = \sum_{i=1}^n u_{ig}\mathbf{s}_{\boldsymbol{\theta}}^{(i)} = \sum_{i=1}^n u_{ig}\big[s_{w_g}^{(i)}, \mathbf{s}_{\boldsymbol{\beta}_g}^{(i)\mathrm{T}}, \mathbf{s}_{\boldsymbol{\alpha}_g}^{(i)\mathrm{T}}, s_{\nu_g}^{(i)}\big]^{\mathrm{T}}$, and $\mathbf{H}_{\boldsymbol{\theta}_g\boldsymbol{\theta}_g} = -\sum_{i=1}^n u_{ig}\mathbf{H}_{\boldsymbol{\theta}_g\boldsymbol{\theta}_g}^{(i)}$, respectively. The detailed expressions of the individual first-two order derivatives, say $\mathbf{s}_{\boldsymbol{\theta}_g}^{(i)}$ and $\mathbf{H}_{\boldsymbol{\theta}_g\boldsymbol{\theta}_g}^{(i)}$, are given in Supplementary S.2. The notation $\text{diag}(\{\mathbf{H}_g\}_{g=1}^G)$ expresses a $aG \times aG$ block-diagonal matrix composed of G sub-matrices with each $\mathbf{H}_g$ having dimension $a \times a$, where a is the number of unknown parameters in component g. Using the method proposed by Guo and Thompson (1994), the inverse of the asymptotic variance-covariance matrix of ML estimates can be approximated by

$$\mathrm{Var}(\boldsymbol{\Theta};\mathbf{y})^{-1} = \mathrm{diag}\Big(\big\{\sum_{i=1}^{n} p_{ig}\mathbf{H}^{(i)}_{\boldsymbol{\theta}_g\boldsymbol{\theta}_g}\big\}_{g=1}^{G}\Big) - \sum_{i=1}^{n}\mathrm{Cov}(\mathbf{s}^{(i)}(\boldsymbol{\Theta};\mathbf{y}_i,\mathbf{u}_i)|\mathbf{y}_i), \quad (4.5)$$

where p_{ig} is given in (4.1) and the individual covariance $\mathrm{Cov}\big(\mathbf{s}^{(i)}(\boldsymbol{\Theta};\mathbf{y}_i,\mathbf{u}_i)|\mathbf{y}_i\big)$ can be calculated as

$$\begin{bmatrix} p_{i1}(1-p_{i1})\mathbf{s}^{(i)}_{\boldsymbol{\theta}_1}\mathbf{s}^{(i)\mathrm{T}}_{\boldsymbol{\theta}_1} & -p_{i1}p_{i2}\mathbf{s}^{(i)}_{\boldsymbol{\theta}_1}\mathbf{s}^{(i)\mathrm{T}}_{\boldsymbol{\theta}_2} & \cdots & -p_{i1}p_{iG}\mathbf{s}^{(i)}_{\boldsymbol{\theta}_1}\mathbf{s}^{(i)\mathrm{T}}_{\boldsymbol{\theta}_G} \\ -p_{i2}p_{i1}\mathbf{s}^{(i)}_{\boldsymbol{\theta}_2}\mathbf{s}^{(i)\mathrm{T}}_{\boldsymbol{\theta}_1} & p_{i2}(1-p_{i2})\mathbf{s}^{(i)}_{\boldsymbol{\theta}_2}\mathbf{s}^{(i)\mathrm{T}}_{\boldsymbol{\theta}_2} & \cdots & -p_{i2}p_{iG}\mathbf{s}^{(i)}_{\boldsymbol{\theta}_2}\mathbf{s}^{(i)\mathrm{T}}_{\boldsymbol{\theta}_G} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{iG}p_{i1}\mathbf{s}^{(i)}_{\boldsymbol{\theta}_G}\mathbf{s}^{(i)\mathrm{T}}_{\boldsymbol{\theta}_1} & -p_{iG}p_{i2}\mathbf{s}^{(i)}_{\boldsymbol{\theta}_g}\mathbf{s}^{(i)\mathrm{T}}_{\boldsymbol{\theta}_2} & \cdots & p_{iG}(1-p_{iG})\mathbf{s}^{(i)}_{\boldsymbol{\theta}_G}\mathbf{s}^{(i)\mathrm{T}}_{\boldsymbol{\theta}_G} \end{bmatrix}.$$

Under certain regularity conditions, the construction of confidence intervals or the hypothesis tests for parameters can be explicitly established by the asymptotic theory of ML estimators. The asymptotic standard errors are obtained as the square root of the diagonal entries of $\mathrm{Var}(\boldsymbol{\Theta};\mathbf{y})$ in (4.5), with $\boldsymbol{\Theta}$ replaced by its ML estimates.

4.2. Clustering

Once the FM-MtLMM has been fitted, it is of interest to determine to which group a subject should belong on the basis of the optimal Bayes' rule. With the estimated model parameters, a probabilistic clustering of the data into G clusters is performed by comparing estimated p_{ig}'s in (4.1) evaluated at $\hat{\boldsymbol{\Theta}}$, denoted by $\hat{u}_{ig}$. Thus $\mathbf{y}_i$ is assigned to the group s if $\max_g\{\hat{u}_{ig}\}$ occurs at the sth component.

When the group labels of subjects are predefined, the evaluation of classification accuracy can be treated as an alternative measure of fitness of the data. To assess the agreement between a clustering of the data and their true group labels, we adopt two commonly used indices: the correct classification rate (CCR; Lee, Chen and Hsieh (2003)) and the adjusted Rand index (ARI; Hubert and Arabie (1985)). The CCR value, ranging between zero and one, is measured as one minus the lowest classification error among all permutations of predicted cluster memberships against the predefined (true) group labels. The ARI takes into account the effect of agreement due to chance. Loosely speaking, the larger the values of CCR and ARI, the higher the quality of classification: a value of close to 0 indicates a poor classification, a value of near 1 signifies an ideal agreement between two clusterings.

4.3. Estimation for random effects and fitted responses

In addition to the estimation of fixed effects, it is important to estimate the random effects as they are useful for evaluating such subject-specific quantities of interest as individually changed intercepts and slopes. The empirical Bayes estimates of random effects $\mathbf{b}_i$ are derived from the Bayesian specification of the model (Laird and Ware (1982)). Specifically,

$$\hat{\mathbf{b}}_i = E(\mathbf{b}_i \mid \mathbf{y}_i, \hat{\boldsymbol{\Theta}}) = \sum_{g=1}^{G} \hat{u}_{ig}\hat{\mathbf{b}}_{ig}, \quad (i = 1, \ldots, n), \tag{4.6}$$

where $\hat{\mathbf{b}}_{ig}$ is $\hat{\mathbf{b}}_{ig}^{(h)}$ in (S.10) with $\hat{\boldsymbol{\Theta}}^{(h)}$ replaced by $\hat{\boldsymbol{\Theta}}$. Consequently, the resulting fitted response for $\mathbf{y}_i$ is calculated as

$$\hat{\mathbf{y}}_i = \sum_{g=1}^{G} \hat{u}_{ig}(\mathbf{X}_i\hat{\boldsymbol{\beta}}_g + \mathbf{Z}_i\hat{\mathbf{b}}_{ig}). \tag{4.7}$$

Alternative estimates of (4.6) and (4.7) relying on the *classification ML* approach (McLachlan and Peel (2000)) are defined by replacing $\hat{u}_{ig}$ with $\tilde{u}_{ig}$, where $\tilde{u}_{ig} = 1$ if $\hat{u}_{ig} \geq \hat{u}_{sg}$ for $s \neq g$ and $\tilde{u}_{ig} = 0$ otherwise.

4.4. Discriminant analysis

Discriminant analysis (Fisher (1936)) is a classical technique that explores a rule for classifying new individuals into one of the predefined groups. Previous extensions of the traditional discriminant analysis to multivariate repeated measure data have been investigated in a number of papers (Albert (1983); Tomasko, Helms and Snapinn (1999); Morrell et al. (2005); Roy (2006); Roy and Leiva (2007)). Recently, Komárek et al. (2010) developed a Bayesian approach for classification of multiple longitudinal markers using the MLMM with a normal mixture for the random effects. Marshall et al. (2009) described a discrimination procedure based on the MNLMM with possible missing values.

Suppose that we have a training data set where the memberships of the involved subjects to prognostic groups ($g = 1, \ldots, G$) are known. Given a priori probabilities $\pi_1, \ldots, \pi_G$, each prognostic group is characterized by a single component FM-MtLMM, written as

$$\mathbf{y}_i = \mathbf{X}_i^g\boldsymbol{\beta}^g + \mathbf{Z}_i^g\mathbf{b}_i^g + \boldsymbol{\varepsilon}_i^g, \qquad \text{with} \quad \begin{bmatrix}\mathbf{b}_i^g\\ \boldsymbol{\varepsilon}_i^g\end{bmatrix} \sim \mathcal{T}_{q+n_i}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{D}^g & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_i^g\end{bmatrix}, \nu^g\right). \tag{4.8}$$

Let $\boldsymbol{\theta}^g = (\boldsymbol{\beta}^g, \mathbf{D}^g, \mathbf{R}_i^g, \nu^g)$ consist of unknown parameters for group g. Before creating the classification rule, the ML estimates of $\boldsymbol{\theta}^g$ must be obtained by fitting model (4.8) to observations from the respective training samples.

Now, let $\mathbf{y}_{\text{new}} = \text{vec}([\mathbf{y}_{\text{new}1} : \cdots : \mathbf{y}_{\text{new}r}])$ be a $s_{\text{new}}r \times 1$ response vector for a new subject. Without loss of generality, we assume that all of the s_{new} occasions for $\mathbf{y}_{\text{new}}$ are not larger than the maximized history time of the training data. Then, the strength of the allocation of $\mathbf{y}_{\text{new}}$ to the gth group is characterized by a marginal predictive density $p(\mathbf{y}_{\text{new}}|\boldsymbol{\theta}_g)$, estimated as

$$\hat{p}(\mathbf{y}_{\text{new}}|\hat{\boldsymbol{\theta}}_g) = t_{s_{\text{new}}r}(\mathbf{y}_{\text{new}} \mid \mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}^g,\ \mathbf{Z}_{\text{new}}\hat{\mathbf{D}}^g\mathbf{Z}_{\text{new}}^{\text{T}} + \hat{\mathbf{R}}_i^g, \hat{\nu}^g).$$

Under the zero-one loss function which minimizes the expected error rate (Hastie, Tibshirani and Friedman (2001)), the estimated posterior probability of allocating $\mathbf{y}_{\text{new}}$ to group g is given by

$$\hat{\mathcal{P}}_{g,\text{new}} = \frac{\pi_g \hat{p}(\mathbf{y}_{\text{new}}|\hat{\boldsymbol{\theta}}_g)}{\sum_{l=1}^{G} \pi_l \hat{p}(\mathbf{y}_{\text{new}}|\hat{\boldsymbol{\theta}}_l)}, \quad (g = 1, \ldots, G).$$

If $\hat{\mathcal{P}}_{s,\text{new}} > \hat{\mathcal{P}}_{g,\text{new}}$ for $s \neq g$, $g = 1 \ldots, G$, then $\mathbf{y}_{\text{new}}$ is classified to group s.

5. Analysis of PBCseq Data

We applied the proposed FM-MtLMM approach to analyzing the PBCseq data described in Section 2. Let $\mathbf{y}_i = (\mathbf{y}_{i1}, \mathbf{y}_{i2})$ be the response vector for the ith patient, where $\mathbf{y}_{i1}$ and $\mathbf{y}_{i2}$ represent lbili and lalbumin levels, respectively. Apart from the time effect, it is particular of interest to take into account the relationship between the longitudinal evolutions of the two markers and the covariates of interest, including gender, drug treatment, and age. Thus, the design matrix for fixed effects is

$$\mathbf{X}_i = \mathbf{I}_2 \otimes [\mathbf{1}_{s_i} : \mathbf{t}_i : \text{sex}_i\mathbf{1}_{s_i} : \text{drug}_i\mathbf{1}_{s_i} : \text{age}_i\mathbf{1}_{s_i}],$$

where $\mathbf{1}_{s_i}$ is a $s_i \times 1$ vector of ones, $\mathbf{t}_i = (t_{i1}, \ldots, t_{is_i})$ with $t_{ik} = \text{month}_{ik}/12$ (years), sex_i is a gender indicator (0 = male and 1 = female), drug_i is a drug treatment indicator (0 = patient treated with placebo, and 1 = patient treated with D-penicillamine); and age_i is the age of patient i at entry in years. The design matrices for random effects are considered to be random intercept (RI), $\mathbf{Z}_i = \mathbf{I}_2 \otimes \mathbf{1}_{s_i}$, and random intercept plus slope (RIS), $\mathbf{Z}_i = \mathbf{I}_2 \otimes [\mathbf{1}_{s_i} : \mathbf{t}_i]$. To address the possible serial correlation among irregularly observed occasions, we considered the uncorrelated (UNC), the continuous-type autoregressive order 1 (CAR(1)), and the damped exponential correlation (DEC) structures for $\mathbf{C}_{ig}$. The DEC structure (Muñoz et al. (1992)) is defined as $\mathbf{C}_{ig} = \big[\phi_g^{|t_{ik} - t_{ik'}|^{\gamma_g}}\big]$, where the component autoregressive coefficient ϕ_g ranges between 0 to 1, and the component damping parameter γ_g is a nonnegative value.

Table 1. Summary of model selection criteria of 12 candidate models.

$\mathbf{Z}_i$	$\mathbf{C}_{ig}$	m		$-2\ell_{\max}$		AIC		BIC	
		MN	MT	MN	MT	MN	MT	MN	MT
RI	UNC	33	35	-16.440	-225.942	49.560	-155.942	173.079	-24.937
	CAR(1)	35	37	-242.734	-435.686	-172.735	-361.686	-41.729	-223.195
	DEC	37	39	-391.422	-557.750	-317.422	-479.750	-178.931	-333.773
RIS	UNC	47	49	-440.686	-561.638	-346.686	-463.638	-170.765	-280.231
	CAR(1)	49	51	-477.186	-588.346	-379.186	-486.346	-195.778	-295.453
	DEC	51	53	-534.166	-666.360	-432.166	-560.360	-241.273	-361.981

MN: FM-MLMM; MT: FM-MtLMM; RI: random intercept; RIS: random intercept plus slope; m: number of model parameters.

For comparison purposes, the fitting results of finite mixtures of multivariate linear mixed-effects models (FM-MLMM), which can be treated as the limiting case of FM-MtLMM when the component DOFs are infinity, are also presented. We implemented the AECM algorithm presented in Section 4.1 for fitting the twelve candidate models with 10 random starts for initial clustering. Afterward, we fit the MLMM (Shah, Laird and Schoenfeld (1997)) to each partitioned samples across $G = 2$ different groups and took the resulting ML solutions as the stating values. Moreover, the initial w_gs were taken as the sample proportions and the initial values for ν_g were given as relatively large values, say $\nu_g = 50$, corresponding to an assumption of near-normality. If there existed multiple modes, the global ML solution was chosen as the one providing the highest likelihood. For model selection, we adopted the Akaike Information Criterion (AIC$= 2m - 2\ell_{\max}$; Akaike (1973)) and Bayesian Information Criterion (BIC$= m \log n - 2\ell_{\max}$; Schwarz (1978)), where m is the number of model parameters, and $\ell_{\max}$ is the maximized log-likelihood value. Accordingly, the smaller value of AIC or BIC indicates a better fit of the model. The ML estimation results of twelve candidate models, including the values of $-2\ell_{\max}$, AIC and BIC together with number of parameters m, are listed in Table 1. According to the AIC or BIC values, the FM-MtLMMs generally provide a better fitting performance than their normal counterparts. Among models considered, the scenario of RIS plus DEC errors gives the best fit under both FM-MLMM and FM-MtLMM.

Table 2 summarizes the ML estimates of model parameters and the standard errors (SE) of the fixed effects under the 'best' fitted FM-MLMM and FM-MtLMM. The SE of the fixed effects were calculated via the approximate observed information matrix $\mathrm{Var}(\boldsymbol{\Theta}; \mathbf{y})^{-1}$ given in (4.5). From this table, we observe that the SE of fixed effects under FM-MtLMM are smaller than those under the FM-MLMM. Focusing on the fitted FM-MtLMM, the patients in Group 0

(Group 1) receiving D-penicillamine have 0.087 higher (0.143 lower) lbili levels than those who were treated with placebo. The patients in Group 0 (Group 1) receiving D-penicillamine have 0.016 higher (0.020 lower) lalbumin levels than those who were treated with placebo. However, the differences of lbili and lalbumin levels between placebo and drug treatments are not highly statistically significant for patients in either Group 0 or Group 1. It can be found from the parameters involving the 'Time' covariate that the increasing trend of lbili levels ($\hat{\beta}_{111} = 0.040$, $\hat{\beta}_{211} = 0.266$) and the decreasing trend of lalbumin levels ($\hat{\beta}_{121} = -0.014$, $\hat{\beta}_{221} = -0.051$) are significantly apparent. The range of changes over time for Group 1 is much larger than that for Group 0. In addition to the parameters involving the 'Time' and 'Drug' covariates, other significant parameters include the intercepts ($\beta_{120}, \beta_{210}, \beta_{220}$), sex ($\beta_{112}, \beta_{222}$) and age ($\beta_{224}$). The results suggest that the baseline levels of lalbumin marker for both groups are significantly different from zero, while only that of the lbili marker for Group 1 differs from zero significantly. In Group 0, the female patients show 0.345 lower lbili levels than the male patients at baseline, while the lalbumin levels of female and male patients show no significant difference. In Group 1, male and female patients have no significantly different lbili levels, while female patients show 0.089 lower lalbumin levels than male patients at baseline. Besides, the lalbumin levels for Group-1 patients decrease 0.005 unit when the age increases one year.

From the estimates of variance components $\mathbf{D}_g$ and $\mathbf{\Sigma}_g$, we found that the between-patient variation for patients in Group 1 is larger than that in Group 0. The estimated correlation coefficients of within-patient errors between the two markers are around $\sigma_{121}/\sqrt{\sigma_{111}\sigma_{122}} = 0.04$ and $\sigma_{221}/\sqrt{\sigma_{211}\sigma_{222}} = -0.15$ for Group 0 and Group 1, respectively. The estimates of autoregressive and damping parameters are 0.391 and 0.387, respectively, confirming the existence of positive serial correlations among occasions for both markers. The estimates of DOFs are small ($\hat{\nu}_1 = 6.500$ and $\hat{\nu}_2 = 11.805$), suggesting that the patient-specific variability for deviating from the mean profiles of both groups exhibits fat-tailed behaviors. We displayed the estimated mean profiles of male (female) patients treated with placebo (Drug=0) and D-penicillamine (Drug=1) in different styles and colors superimposed on the trajectory curves depicted in Figure 1. As can bee seen, the differences in baseline measurements between male and female as well as drug treatments are relatively minor, consistent with the significance of fixed effects for sex and drug covariates shown in Table 2.

Since the group levels for 312 patients are predefined, it is of interest to compare the classification results using the two 'best' fitted models shown in Table 3. The proposed classification method assigns patients to one cluster according to the estimated posterior probabilities of allocation indicators, which can be obtained immediately as a by-product of the AECM algorithm. As can

Table 2. Summary of parameter estimates along with standard errors of fixed effects (in parentheses) under the fitted FM-MLMM (MN) and FM-MtLMM (MT) with RIS and DEC errors for the PBCseq data.

		Fixed effects			Variances for Random effects			Variances for Within-subject errors	
		MN	MT		MN	MT		MN	MT
Group 0	β_{110}(Intercept)	0.345(0.285)	0.072(0.239)	d_{111}	0.217	0.149	σ_{111}	0.043	0.068
	β_{111}(Time)	0.053(0.009)	0.040(0.007)	d_{121}	0.004	0.001	σ_{121}	0.001	0.001
	β_{112}(Sex)	-0.276(0.166)	-0.345(0.129)	d_{122}	0.006	0.003	σ_{122}	0.005	0.009
	β_{113}(Drug)	0.060(0.094)	0.087(0.067)	d_{131}	-0.006	-0.008	ϕ_1	10^{-6}	0.391
	β_{114}(Age)	-0.006(0.005)	0.00040(0.004)	d_{132}	-0.000035	0.00011	γ_1	0.740	0.387
	β_{120}(Intercept)	1.386(0.037)	1.381(0.034)	d_{133}	0.003	0.001	ν_1	—	6.500
	β_{121}(Time)	-0.017(0.002)	-0.014(0.001)	d_{141}	-0.001	-0.00041			
	β_{122}(Sex)	-0.030(0.021)	-0.025(0.020)	d_{142}	-0.001	-0.00044			
	β_{123}(Drug)	0.007(0.012)	0.016(0.012)	d_{143}	-0.00010	0.000026			
	β_{124}(Age)	-0.001(0.001)	-0.001(0.001)	d_{144}	0.00014	0.000064			
Group 1	β_{210}(Intercept)	0.843(0.485)	0.917(0.450)	d_{211}	0.831	0.721	σ_{211}	0.304	0.267
	β_{211}(Time)	0.253(0.021)	0.266(0.018)	d_{221}	0.027	-0.007	σ_{221}	-0.008	-0.009
	β_{212}(Sex)	-0.050(0.233)	-0.061(0.222)	d_{222}	0.020	0.014	σ_{222}	0.023	0.013
	β_{213}(Drug)	-0.192(0.159)	-0.143(0.123)	d_{231}	-0.045	-0.038	ϕ_2	0.398	0.391
	β_{214}(Age)	0.006(0.008)	0.005(0.006)	d_{232}	-0.002	0.00043	γ_2	0.461	0.387
	β_{220}(Intercept)	1.488(0.063)	1.548(0.052)	d_{233}	0.003	0.002	ν_2	—	11.805
	β_{221}(Time)	-0.046(0.004)	-0.051(0.003)	d_{241}	-0.008	-0.006			
	β_{222}(Sex)	-0.060(0.030)	-0.089(0.025)	d_{242}	-0.002	-0.001			
	β_{223}(Drug)	0.007(0.020)	-0.020(0.015)	d_{243}	0.001	0.001			
	β_{224}(Age)	-0.004(0.001)	-0.005(0.001)	d_{244}	0.00040	0.001			

be seen from Table 3, the FM-MtLMM provides more accurate classification performance than the FM-MLMM in terms of the CCR and ARI values. As an alternative way to summarize the classification results, the sensitivity and specificity can be readily evaluated. If $A = \{$the patient is assigned into Cluster 1$\}$, and $B = \{$the patient actually belongs to Group 0$\}$, the sensitivity and specificity can be calculated as $P(A|B)$ and $P(\bar{A}|\bar{B})$, respectively. From the results listed in Table 3, the FM-MLMM gives 63.37% sensitivity and 84.29% specificity, while the FM-MtLMM produces 69.77% sensitivity and 82.14% specificity. Figure 3 shows the "smoothed" receiver operating characteristic (ROC) curves (Fawcett (2006); Robin et al. (2011)) and the area under the ROC curve (AUC) for the FM-MLMM and FM-MtLMM. The value of AUC measure ranges from zero to one, has an expected value of 0.5 under random classification, and takes the value one for perfect classification. From Figure 3, the ROC curves reveal

Table 3. Agreements and differences between the clinical and model classifications using the FM-MLMM (RIS-DEC) and FM-MtLMM (RIS-DEC) scenarios.

		FM-MLMM		FM-MtLMM		Total
Classify to:		1	2	1	2	
True	Group 0	**109**	63	**120**	52	172
	Group 1	22	**118**	25	**115**	140
	Total	131	181	145	167	312
	CCR	0.728		0.753		
	ARI	0.204		0.254		

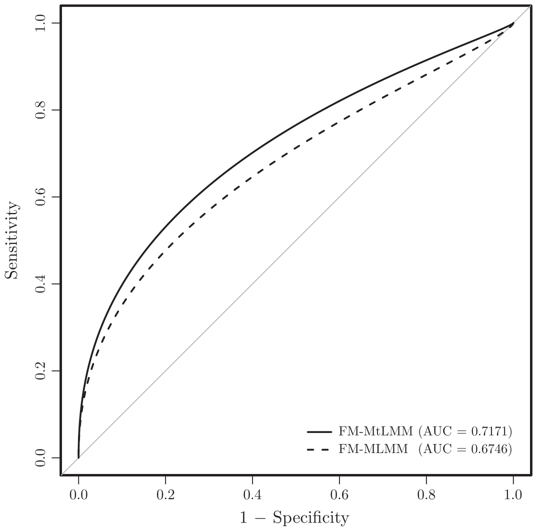

Figure 3. ROC curves with areas for FM-MtLMM with RIS-DEC and FM-MLMM with RIS-DEC scenarios fitted to lbili and lalbumin markers.

that both models provide predictive ability for determining alive or dead patients in this population because their AUC values are larger than 0.5.

6. Simulation Study

We conducted a small simulation study to compare the proposed FM-MtLMM with the existing FM-tLMM approach (Bai, Chen and Yao (2016)) in terms of parameter estimation and clustering performance. The two issues to be explored were how badly the FM-tLMM could perform if the outcome variables of longitudinal data are intrinsically correlated with each other, and whether the proposed FM-MtLMM can produce similar or even better performances than the FM-tLMM when any two outcome variables are uncorrelated.

We generated the $\mathbf{Y}_i$'s with $r = 3$ outcome variables in two cases. In (a), the data were generated from (3.1) with $G = 2$ groups, the design matrices $\mathbf{X}_i$'s including an intercept and scheduled visits of time (1 to 7), and $\mathbf{Z}_i$'s containing an intercept only. The presumed parameters are

$$\begin{aligned}
&\boldsymbol{\beta}_1 = (\beta_{111}, \beta_{112}, \beta_{121}, \beta_{122}, \beta_{131}, \beta_{132})^{\mathrm{T}} = (1, 2, -2, -4, -1, 2)^{\mathrm{T}}, \qquad (6.1)\\
&\boldsymbol{\beta}_2 = (\beta_{211}, \beta_{212}, \beta_{221}, \beta_{222}, \beta_{231}, \beta_{232})^{\mathrm{T}} = (1, 3, -1, -3, -4, 1.5)^{\mathrm{T}},\\
&\mathbf{D}_g = \begin{bmatrix} 2 & 0.25 & 0.25 \\ 0.25 & 2 & 0.25 \\ 0.25 & 0.25 & 2 \end{bmatrix}, \boldsymbol{\Sigma}_g = \begin{bmatrix} 2 & 1.414 & 1.225 \\ 1.414 & 4 & 1.732 \\ 1.225 & 1.732 & 3 \end{bmatrix}, \ \mathbf{C}_{ig} = \mathbf{I}_{s_i},
\end{aligned}$$

and $\nu_g = \nu$, for $g = 1, 2$. Two DOFs were considered for each component: a low value ($\nu = 5$) that yields heavy tails and a high value ($\nu = 50$) that resembles the normal distribution except for a slightly heavier tail. In (b), the three outcomes $\mathbf{y}_{i1}, \mathbf{y}_{i2}$ and $\mathbf{y}_{i3}$ were generated from three 2-component FM-tLMMs, separately, in which the true values for fixed effects, including $\boldsymbol{\beta}_{11} = (\beta_{111}, \beta_{112})$ and $\boldsymbol{\beta}_{21} = (\beta_{211}, \beta_{212})$ for the first outcome $\mathbf{y}_{i1}$, $\boldsymbol{\beta}_{12} = (\beta_{121}, \beta_{122})$ and $\boldsymbol{\beta}_{22} = (\beta_{221}, \beta_{222})$ for the second outcome $\mathbf{y}_{i2}$, and $\boldsymbol{\beta}_{13} = (\beta_{131}, \beta_{132})$ and $\boldsymbol{\beta}_{23} = (\beta_{231}, \beta_{232})$ for the third outcome $\mathbf{y}_{i3}$, are the same as those specified in (6.1). Besides, the assumption for component random effects was $b_{ijg} \sim \mathcal{T}_{q_j}(0, 2, \nu)$, and that for component within-subject errors was $\mathbf{e}_{i1g} \sim \mathcal{T}_{s_i}(\mathbf{0}, 2\mathbf{I}_{s_i}, \nu)$, $\mathbf{e}_{i2g} \sim \mathcal{T}_{s_i}(\mathbf{0}, 4\mathbf{I}_{s_i}, \nu)$, and $\mathbf{e}_{i3g} \sim \mathcal{T}_{s_i}(\mathbf{0}, 3\mathbf{I}_{s_i}, \nu)$, for $i = 1, \ldots, n$, $j = 1, 2, 3$, and $g = 1, 2$. The values of DOF were the same as those considered in case (a). The specification of $\mathbf{D}_g$ and $\boldsymbol{\Sigma}_g$ in (6.1) yields a positive correlation between two responses, while the generation process under (b) results in nearly uncorrelated responses. The sample sizes were small ($n = 20$), moderate ($n = 50$) and relatively large ($n = 100$), and each simulated data set was fitted under the FM-MtLMM and FM-tLMM approaches. Under the FM-tLMM approach, each of the three outcome variables was assumed to be independent and fitted separately one at a time, while the FM-MtLMM could be applied to analyze multiple responses simultaneously. The estimation accuracy for model parameters was evaluated by the mean square errors (MSE) and mean absolute relative errors (MARE), and the classification performance was measured by CCR and ARI. A total of 100 replications were run across each combination of DOFs ν and sample sizes n. In the estimation procedure, the algorithm failed, though seldom, to achieve convergence for a particular data set. To ensure that we were comparing different methods based on the same simulated data, an additional data set would be regenerated if one of the model fittings did not converge. To this end, we handled the error-recovery problem by using the R `try()` function.

For finite mixture models, there is a non-identifiability problem due to possible label switching (McLachlan and Peel (2000)), which means that the likelihood is invariant under a permutation of the class labels. A number of authors have sought distinct strategies to solve this problem from both the likelihood-based and Bayesian perspectives. See, for example, Celeux, Hurn and Robert (2000), Stephens (2000), Yao (2015), and Yao and Lindsay (2009). However, the label switching is not a concern when carrying out ML estimation via the EM-type algorithm with only one replication. Yet the switching of class labels is an important issue when carrying out a series of Monte Carlo simulations or a single trial with different initial values to calculate the component-related quantities such as the SE for parameters or CCR and ARI. As adopted by Lin, McLachlanc and Lee (2016), we considered all permutations of the class labels for the estimated component parameters and chose the one that gave the minimum Euclidean distance to the true parameter values.

For the accuracy in parameter estimation, we focus on the estimates of fixed effects. Web Tables S2 and S3 summarize the averages of MSE and MARE of estimated fixed effects $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ under cases (a) and (b) for a total of 100 replications. For (a), the FM-MtLMM almost gives relatively smaller MSE and MARE for the estimates of fixed effects compared with the FM-tLMMs because it takes into account the correlations among outcome variables. In (b), the FM-MtLMM and FM-tLMMs provide similarly accurate estimates of the fixed effects due to a very weak correlation among the generated responses. In both cases, the MSE and MARE values decrease as the sample size increases, revealing the finite-sample property of ML estimators.

Figure 4 shows the boxplots of CCR and ARI values of 100 trails for different DOFs and sample sizes under the settings of (a) and (b). It is readily observed that the CCR and ARI values given by the FM-MtLMM are higher than the best clustering result given by the FM-tLMMs. Even if the outcome variables are uncorrelated with each other, the clustering performance of FM-MtLMM could perform better than that of FM-tLMMs, especially when the random effects and errors exhibit heavy tails. Although this simulation seems limited, the results apparently reveal that the FM-tLMM may fail to perform satisfactorily in estimating parameters and clustering observations because of a lack of mechanism to take into account the correlation between variables.

7. Conclusion

This paper proposes a robust extension of FM-MLMM using the multivariate t distribution, called the FM-MtLMM, that is expected to provide more reliable clustering structure for multivariate longitudinal data in the presence of population heterogeneity and heavy-tailed noises. A computationally flexible AECM algorithm is developed by borrowing attractive data augmentation schemes for parameter estimation. Techniques for estimation of random effects, clustering,

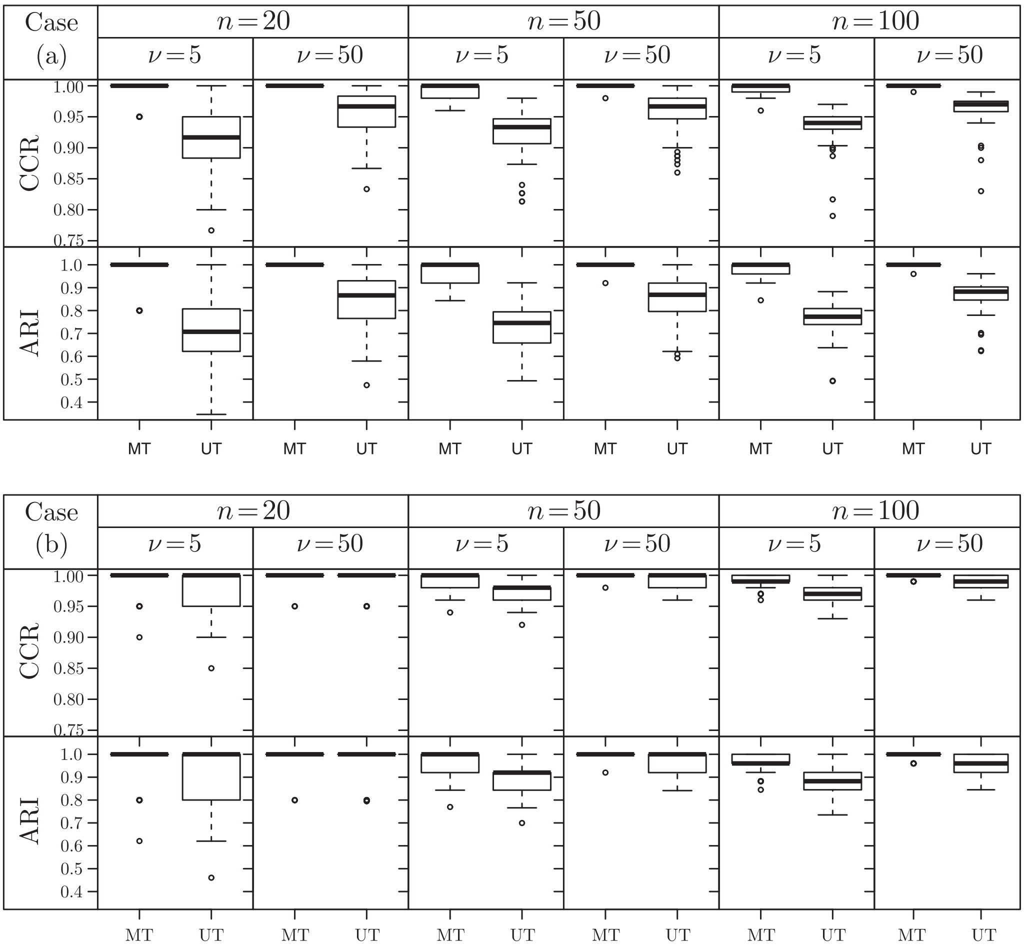

Figure 4. Boxplots for the CCR and ARI values under the fitted FM-MtLMM (MT) and FM-tLMMs (UT) for various DOFs ν and sample sizes n under cases (a) and (b).

and discriminant analysis are also provided and convenient to implement. Numerical results have highlighted the effectiveness of FM-MtLMM on the provision of an improved model fit and classification accuracy.

It is widely acknowledged that the likelihood function in finite mixture models can be unbounded, at least in normal mixtures with unequal variances (Hathaway (1985)), such that the ML solution is not well defined. For multivariate mixture models such as that defined in (3.3), a degraded component due to the unboundedness of the likelihood function might cause *degenerate or spurious* solutions. There have been some strategies proposed to deal with such ill-posed problems by imposing constraints on the parameter space (Hathaway (1986);

Ingrassia (2004)) or using the penalized ML methods (Chen, Tan and Zhang (2008)). Under different constraint conditions, Yao (2010) offered an alternative maximizer based on the profile log-likelihood method. Further investigations along these directions would be helpful in pursuing an improved procedure. The topic is beyond the scope of the present paper and is left for future work.

We foresee several possibilities for future research along these lines in terms of methodology and application. The proposed model can be extended to deal with multivariate longitudinal data with censored observations following the approaches developed by Vaida, Fitzgerald and DeGruttola (2007), Vaida and Liu (2009), Lachos, Bandyopadhyay and Dey (2011), and Wang, Lin and Lachos (2015). Missing data are unavoidable in many longitudinal studies for such reasons as missed visits, withdrawal from a study, loss to follow-up, adverse side effects, and so on. Although the FM-MtLMM has shown its robustness to accommodate heavy tails, inferential procedures can still be dramatically obscured in the presence of highly asymmetric observations (Lin and Lee (2008); Ho and Lin (2010)). Thus, it is of interest to establish a broader framework for fitting the FM-MtLMM that allows one to simultaneously address population heterogeneity, missingness, censored responses, or impacts of skewness and heavy tails. With increased computer power, it would be worthwhile to develop modern MCMC methods (Hastings (1970); Richardson and Green (1997)) coupled with the inverse Bayes formulae (Tan, Tian and Ng (2003, 2006); Wang and Fan (2012); Wang and Lin (2015)) for efficient Bayesian estimation of the FM-MtLMM. When the observed covariates carry group information, incorporating the dependence of mixing weights w_g's on the observed covariates might improve the performance in data clustering. Some authors, for example Fokoué (2005) and Tan and Nott (2014), have considered the logit function with an identifiable constraint to model the relationship between prior classification probabilities and covariates in the frameworks of FMM and FM-LMM. Finally, it would be interesting to extend the current approach by taking into account the effect of covariates on mixing weights.

Supplementary Materials

Some detailed derivations and results for the preliminary analysis of PBCseq data and simulation studies are available online at the *Statistica Sinica* website.

Acknowledgements

The author would like to express her deepest gratitude to the Co-Editor, the Associate Editor, and two anonymous reviewers for their insightful comments and suggestions that greatly improved this paper. This work was partially supported by the Ministry of Science and Technology under Grant no. MOST 103-2118-M-035-001-MY2 of Taiwan.

References

Aitken, A. C. (1926). On Bernoulli's numerical solution of algebraic equations. *Proc. Roy. Soc. Edinb.* **46**, 289-305.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the* 2nd *International Symposium on Information Theory* (Edited by B. N. Petrov and F. Csaki), 267-281. Akademiai Kiado, Budapest.

Albert, A. (1983). Discriminant analysis based on multivariate response curve: a descriptive approach to dynamic allocation. *Statist. Med.* **2**, 95-106.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis.* 3rd edition. Wiley and Sons, New York.

Bai, X., Chen, K. and Yao, W. (2016). Mixture of linear mixed models using multivariate t distribution. *J. Statist. Comput. Simulation* **86**, 771-787.

Booth, J. G., Casella, G. and Hobert, J. P. (2008). Clustering using objective functions and stochastic search. *J. Roy. Statist. Soc. B* **70**, 119-139.

Celeux, G., Hurn, M. and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95**, 957-970.

Celeux, G., Martin, O. and Lavergne, C. (2005). Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statist. Model.* **5**, 243-267.

Chen, J., Tan, X. and Zhang, R. (2008). Inference for normal mixture in mean and variance. *Statist. Sinica* **18**, 443-465.

De la Cruz-Mesía, R., Quintana, F. A. and Marshall, G. (2008). Model-based clustering for longitudinal data. *Comput. Statist. Data Anal.* **52**, 1441-1457.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D. and Langworthy, A. (1989). Prognosis in primary biliary-cirrhosis — Model for decision-making. *Hepatology* **10**, 1-7.

Fawcett, T. (2006). An introduction to ROC analysis. *Patt. Recog. Lett.* **27**, 861-874.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7**, 179-188.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* Wiley, New York.

Fokoué, E. (2005). Mixtures of factor analyzers: an extension with covariates. *J. Multivariate Anal.* **95**, 370-384.

Gaffney, S. J. and Smyth, P. (2003). Curve clustering with random effects regression mixtures. In*Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics* (Edited by Bishop, C. M., Frey, B. J.) Key West, FL.

Galecki, A. T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Commun. Statist. Theory Methods* **23**, 3105-3119.

Guo, S. W. and Thompson, E. A. (1994). Monte Carlo estimation of mixed models for large complex pedigrees. *Biometrics* **50**, 417-432.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). *Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.* **13**, 795-800.

Hathaway, R. J. (1986). A constrained EM algorithm for univariate mixtures. *J. Statist. Comput. Simulation* **23**, 211-230.

Ho, H. J. and Lin, T. I. (2010). Robust linear mixed models using the skew t distribution with application to schizophrenia data. *Biometrical J.* **52**, 449-469.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *J. Classification* **2**, 193-218.

Ingrassia, S. (2004). A likelihood-based constrained algorithm for multivariate normal mixture models. *Stat. Methods Appl.* **13**, 151-166.

Kibria, B. M. G. (2006). The matrix-t distribution and its applications in predictive inference. *J. Multivariate Anal.* **97**, 785-795.

Komárek, A., Hansen, B. E., Kuiper, E. M. M., van Buurenc, H. R. and Lesaffre, E. (2010). Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statist. Med.* **29**, 3267-3283.

Komárek, A. and Komárková, L. (2013). Clustering for multivariate continuous and discrete longitudinal data. *Ann. Appl. Stat.* **7**, 177-200.

Komárek, A. and Komárková, L. (2014). Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data. *J. Stat. Softw.* **59**, 1-38.

Komárek, A. and Lesaffre, E. (2008). Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. *Comput. Statist. Data Anal.* **52**, 3441-3458

Kotz, S. and Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications.* Cambridge University Press, Cambridge.

Lachos, V. H., Bandyopadhyay, D. and Dey, D. K. (2011). Linear and nonlinear mixed-effects models for censored HIV viral loads using normal/independent distributions. *Biometrics* **67**, 1594-1604.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963-974.

Lee, K., Michael, J., Daniels, M. J. and Joo, Y. (2013). Flexible marginalized models for bivariate longitudinal ordinal data. *Biostatistics* **14**, 462-476.

Lee, W., Chen, Y. and Hsieh, K. (2003). Ultrasonic liver tissues classification by fractal feature vector based on M-band wavelet transform. *IEEE Trans. Med. Imaging* **22**, 382-392.

Lin, T. I. and Lee, J. C. (2006). A robust approach to t linear mixed models applied to multiple sclerosis data. *Statist. Med.* **25**, 1397-1412.

Lin, T. I. and Lee, J. C. (2007). Bayesian analysis of hierarchical linear mixed modeling using the multivariate t distribution. *J. Statist. Plann. Inference* **137**, 484-495.

Lin, T. I. and Lee, J. C. (2008). Estimation and prediction in linear mixed models with skew normal random effects for longitudinal data. *Statist. Med.* **27**, 1490-1507.

Lin, T. I., McLachlanc, G. J. and Lee, S. X. (2016). Extending mixtures of factor models using the restricted multivariate skew-normal distribution. *J. Multivariate Anal.* **143**, 398-413.

Markus, B. H., Dickson, E. R., Grambsch, P. M., Fleming, T. R., Mazzaferro, V., Klintmalm, G. B. G., Wiesner, R. H., Vanthiel, D. H. and Starzl, T. E. (1989). Efficacy of liver-transplantation in patients with primary biliary-cirrhosis. *New Engl. J. Med.* **320**, 1709-1713.

Marshall, G., De la Cruz-Mesia, R., Baron, A. E., Rutledge, J. H. and Zerbe, G. O. (2006). Non-linear random effects model for multivariate responses with missing data. *Statist. Med.* **25**, 2817-2830.

Marshall, G., De la Cruz-Mesia, R., Quintana, F. A. and Baron, A. E. (2009). Discriminant analysis for longitudinal data with multiple continuous responses and possibly missing data. *Biometrics* **65**, 69-80.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions.* 2nd edition. Wiley, New York.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models.* Wiley, New York.

Meng, X. L. and van Dyk, D. (1997). The EM algorithm - an old folk-song sung to a fast new tune. *J. Roy. Statist. Soc. Ser. B* **59**, 511-567.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* Springer, New York.

Morrell, C. H., Brant, L. J., Sheng, S. and Metter, E. J. (2005). Using multivariate mixed-effects models to predict prostate cancer. *Proc. Amer. Statist. Assoc.*, 332-337. American Statistical Association: Alexandria.

Muñoz, A., Carey, V., Schouten, J. P., Segal, M. and Rosner, B. (1992). A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics* **48**, 733-742.

Murtaugh, P. A., Dickson, E. R., Van Dam, G. M., Malinchoc, M., Grambsch, P. M., Langworthy, A. L. and Gips, C. H. (1994). Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits. *Hepatology* **20**, 126-134.

Ng, S. K., McLachlan, G. J., Wang, K., Ben-Tovim, L. and Ng, S. W. (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics* **22**, 1745-1752.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statist. Comput.* **10**, 339-348.

Pfeifer, C. (2004). Classification of longitudinal profiles based on semi-parametric regression with mixed effects. *Statist. Med.* **4**, 314-323.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team. (2014). nlme: Linear and nonlinear mixed effects models. R package version 3.1-117 `http://CRAN.R-project.org/package=nlme.`

Pinheiro, J. C., Liu, C. H. and Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects model using the multivariate t distribution. *J. Comput. Graph. Statist.* **10**, 249-276.

R Development Core Team. (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. `URLhttp://www.R-project.org/`.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 731-792.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez. J. C. and Muller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77.

Rosa, G. J. M., Gianola, D. and Padovani, C. R. (2004). Bayesian longitudinal data analysis with mixed models and thick-tailed distributions using MCMC. *J. Appl. Stat.* **31**, 855-873.

Roy, A. (2006). A new classification rule for incomplete doubly multivariate data using mixed effects model with performance comparisons on the imputed data. *Statist. Med.* **25**, 1715-1728.

Roy, A. and Leiva, R. (2007). Discrimination with jointly equicorrelated multi-level multivariate data. *Adv. Data Anal. Classification* **1**, 175-199.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Shah, A., Laird, N. and Schoenfeld, D. (1997). A random-effects model for multiple characteristics with possibly missing data. *J. Amer. Statist. Assoc.* **92**, 775-779.

Song, P. X. K., Zhang, P. and Qu, A. (2007). Maximum likelihood inference in robust linear mixed-effects models using multivariate t distributions. *Statist. Sinica* **17**, 929-943.

Spiessens, B., Verbeke, G. and Komárek, A. (2002). A SAS-macro for the classification of longitudinal profiles using mixtures of normal distributions in nonlinear and generalised linear mixed models. Technical Report, Biostatistical Center, Catholic Univ. Leuven.

Stephens, M. (2000). Dealing with label switching in mixture models. *J. Roy. Statist. Soc. Ser. B* **62**, 795-809.

Tan, M., Tian, G. L. and Ng, K. W. (2003). A noniterative sampling method for computing posteriors in the structure of EM-type algorithms. *Statist. Sinica* **13**, 625-639.

Tan, M., Tian, G. L. and Ng, K. W. (2006). Hierarchical models for repeated binary data using the IBF sampler. *Comput. Statist. Data Anal.* **50**, 1272-1286.

Tan, S. L. and Nott, D. J. (2014). Variational approximation for mixtures of linear mixed models. *J. Comput. Graph. Statist.* **23(2)**, 564-585.

Tomasko, L., Helms, R. W. and Snapinn, S. M. (1999). A discriminant analysis extension to mixed models. *Statist. Med.* **18**, 1249-1260.

Vaida, F., Fitzgerald, A. and DeGruttola, V. (2007). Efficient hybrid EM for linear and nonlinear mixed effects models with censored response. *Comput. Statist. Data Anal.* **51**, 5718-5730.

Vaida, F. and Liu, L. (2009). Fast implementation for normal mixed effects models with censored response. *J. Comput. Graph. Statist.* **18**, 797-817.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *J. Amer. Statist. Assoc.* **91**, 217-221.

Villarroel, L., Marshall, G. and Barón, A. E. (2009). Cluster analysis using multivariate mixed effects models. *Statist. Med.* **28**, 2552-2565.

Wang, K., Ng, S. K. and McLachlan, G. J. (2009). Multivariate skew t mixture models: applications to fluorescence-activated cell sorting data. *Proceedings of DICTA 2009, IEEE Computer Society*, 526-531.

Wang, W. L. (2013). Multivariate t linear mixed models for irregularly observed multiple repeated measures with missing outcomes. *Biometrical J.* **55**, 554-571.

Wang, W. L. and Fan, T. H. (2011). Estimation in multivariate t linear mixed models for multiple longitudinal data. *Statist. Sinica* **21**, 1857-1880.

Wang, W. L. and Fan, T. H. (2012). Bayesian analysis of multivariate t linear mixed models using a combination of IBF and Gibbs samplers. *J. Multivariate Anal.* **105**, 300-310.

Wang, W. L. and Lin, T. I. (2014). Multivariate t nonlinear mixed-effects models for multi-outcome longitudinal data with missing values. *Statist. Med.* **33**, 3029-3046.

Wang, W. L. and Lin, T. I. (2015). Bayesian analysis of multivariate t linear mixed models with missing responses at random. *J. Statist. Comput. Simulation* **85**, 3594-3612.

Wang, W. L., Lin, T. I. and Lachos, V. H. (2015). Extending multivariate-t linear mixed models for multiple longitudinal data with censored responses and heavy tails. *Statist. Meth. Med. Res.* DOI: 10.1177/0962280215620229.

Yao, W. (2010). A profile likelihood method for normal mixture with unequal variance. *J. Statist. Plann. Inference* **140**, 2089-2098.

Yao, W. (2015). Label switching and its simple solutions for frequentist mixture models. *J. Statist. Comput. Simulation* **85**, 1000-1012.

Yao, W. and Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *J. Amer. Statist. Assoc.* **104**, 758-767.

Department of Statistics, Graduate Institute of Statistics and Actuarial Science, Feng Chia University, Taichung 40724, Taiwan.

E-mail: wlunwang@fcu.edu.tw

(Received August 2015; accepted March 2016)