Journal of Multivariate Analysis 174 (2019) 104533

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

An innovative strategy on the construction of multivariate multimodal linear mixed-effects models

Zahra Mahdiyeh, Iraj Kazemi [*]

Department of Statistics, University of Isfahan, Isfahan 81746, Iran

ARTICLE INFO

Article history:
Received 6 June 2018
Received in revised form 11 July 2019
Accepted 11 July 2019
Available online 16 July 2019

AMS 2010 subject classifications:
62H12
62J99
60E05

Keywords:
Clustered random effects
ECM algorithm
Low-back pain
MCMC
Multimodality
Multiple longitudinal data

ABSTRACT

This paper presents an attractive extension of multivariate mixed-effects models to allow the modeling of correlated responses. By initiating a new multivariate multimodal distribution, the proposed strategy takes multimodality and the asymmetric structure into account in a flexible way. It can also accommodate clustered random effects on multiple longitudinal responses when data comprise various hidden sub-populations that are not directly identifiable. We introduce an explicit stochastic hierarchical representation of the proposed model to render its theoretical properties straightforward and to carry out estimation processes easily. A fully Bayesian approach is proposed to compute posterior distributions using MCMC techniques in modeling multivariate longitudinal data. Moreover, we present an EM-based maximum likelihood estimation procedure. To facilitate Bayesian computation, the estimation process of mixed models utilizes a data augmentation scheme. We analyze two real-life data on the low-back pain study and the height of school-girls to illustrate the usefulness of our proposed model in practical applications.

1. Introduction

An important topic in statistical modeling is devoted to longitudinal data wherein a set of subjects are repeatedly measured on specified conditions or periods. In the analysis of longitudinal data, any working model should allow measurements of the same response for each subject to be correlated, while observations from different subjects are independent. Furthermore, the correlation between measurements of different response variables should be considered for each subject. This is often performed by using joint regression models with postulating random effects as shared-parameter. A key disadvantage of shared parameter models is that they often involve very strong, sometimes unrealistic, assumptions about the association between multiple responses. This restriction can be relaxed by amending a separate mixed-effects model [20] for each response and combining them by allowing the random effects of assorted responses to be correlated [11].

A routine assumption in mixed-effects modeling is the normality of random effects which makes the individual response vector follow a multivariate normal. This conventional assumption can often be violated [35] mainly when certain latent sub-populations exist in the data generating process [29], or, when some important categorical covariates are omitted from the fixed part of the model. Hence, the implementation of suitable joint mixed-effects models is challenged. In this case, to avoid misleading inference and to guarantee its robustness, a model-based clustering is typically suggested. Consequently, the collected measurements can be classified based on the adoption of a multimodal distribution for

[*] Corresponding author.
E-mail address: i.kazemi@stat.ui.ac.ir (I. Kazemi).

https://doi.org/10.1016/j.jmva.2019.104533
0047-259X/© 2019 Elsevier Inc. All rights reserved.

random effects. A simple idea to set up a multimodal structure is according to fitting a mixture of multiple unimodal distributions. Several cases of mixture distributions have been proposed in the literature for certain purposes. In particular, the finite mixture distribution with normal components is widely used [1,31]. The application of normal mixtures as the random-effects distribution in the linear mixed-effects (LME) models was proposed by [32] and described further by [33] and [34].

As is documented, any continuous distribution may be well approximated by a finite mixture distribution [19]. However, in practical applications, there are several issues on using mixture distributions in fitting multivariate LME models. One is the identifiability issue [14] as the unstructured form of covariance matrices generates a large number of unknown parameters. It can make the use of basic estimation procedures complicated. Any mixture distribution also needs convincing prior information on the choice of true number of components to perform effectively. Moreover, fixing the number of components to avoid overfitting and the normality assumption of components about each cluster are rather strong. It requires setting up some extra prior information.

These typical challenges motivated us to investigate alternative techniques. Consequently, we introduce a new family of multivariate multimodal distributions which offers great flexibility in jointly modeling of responses with the multimodal structure. This is originated from a mixing strategy addressed in the univariate case by [15] together with our initiation of a multimodal extension of the multivariate normal distribution (see also [12,16,23]). While a mixture distribution tends to impose additional components and parameters to capture more peaks, our proposed multivariate multimodal normal ($\mathcal{MMN}$) distribution is able to cover most of the peaks through a limited number of parameters, without any requirement of prior information.

To model joint responses with multimodal structures, we let random effects follow the $\mathcal{MMN}$ distribution. It is shown that both joint and marginal distributions of responses belong to the class of multimodal distributions. Several main advantages of this model include (i) presenting great flexibility to model correlations which often exist within subjects and between multiple responses, (ii) covering various types of responses with multimodal and asymmetric behaviors, (iii) acknowledging for highly unbalanced data, (iv) being available to model more than two responses, (v) being simplifying the interpretation of parameter estimates, (vi) non-increasing the dimension of integration in the marginal distribution of response vector and non-involvement the additional computational complexity, (vii) being useful of its stochastic hierarchical representation for implementing easy estimation processes, (viii) dealing with the effect of hidden sub-populations with different behavior in terms of peaks that cannot be directly observed through the value of responses, (ix) being effective when a knowledge discovery of clusters is not available or choosing the number of clusters is not visible, (x) avoiding misleading inference when the classification of responses occurs due to the omission of important categorical covariates and violation of the normality assumption of error terms or random effects. To make the inference of model parameters using the maximum likelihood approach, the corresponding likelihood function appears in a non-closed form of known distributions, due to complicated integrals. Consequently, some advanced numerical integration techniques, stochastic methods or analytically approximations must be implemented. In this paper, we utilize two estimation approaches that are commonly used in practice. First, we focus on the Expectation–Conditional–Maximization (ECM) algorithm [8,25]. Here, as is underlined in the related contexts, the unobserved random effects are treated as missing values which makes the estimation process highly computationally intensive. However, in construction of the E-step, deriving some associated conditional expectations of unobserved quantities, given the observed data, requires numerical methods to approximate such terms.

A general Bayesian computational technique that is particularly designed for complex models that include the high-dimensional integrals, is the Markov chain Monte Carlo (MCMC) approach. It has been developed extensively in situations where the evaluation of the marginal posterior likelihood is computationally expensive. Specifically, we include data augmentation strategies into the MCMC scheme to contribute random components in the joint likelihood function along with gaining our proposed stochastic hierarchical representation for the $\mathcal{MMN}$ distribution to provide efficient MCMC. Then, a combination of the Gibbs sampling and Metropolis–Hastings algorithms is used to facilitate the generation of samples from specified full conditional posterior distributions of all unknown quantities. With these specifications, underlying multivariate mixed-effects models with the $\mathcal{MMN}$ distribution can be readily fitted in the freely available software packages, such as OpenBugs[1] [21] , or JAGS in R[2] [9,27]. Consequently, Bayesian inference of parameters is performed using summarized MCMC outputs.

The rest of the article is organized as follows. In Section 2, we introduce univariate and multivariate multimodal normal distribution and report their main properties. Section 3 presents in short, the specification of multivariate mixed effects models. We also extend modeling strategies for analyzing multivariate multimodal responses. Section 4 demonstrates how the parameters estimation process can be performed using Bayesian computing techniques. Section 5 conducts a simulation study to evaluate the performance of our proposed model. Finally, Section 6 is devoted to highlighting the usefulness of our methodological findings in real-life data analyses of the low-back pain measurements and the height of school-girls magnitudes.

[1] Release 3.2.3 of OpenBugs is freely available at http://www.openbugs.net/w/Downloads.

[2] Release 4.3.0 of JAGS in R is available for free at https://mcmc-jags.sourceforge.net.

2. A multimodal extension of the normal distribution

2.1. The univariate multimodal normal distribution

Let W be a univariate random variable defined on the real line with the probability density function (pdf)

$$f_W(w) = 2h(w)\,G(w)\,, \tag{1}$$

where $h(\cdot)$ is a symmetric pdf and $G(\cdot)$ is a Lebesgue measurable function, satisfying conditions $0 \leq G(w) \leq 1$ and $G(w) + G(-w) = 1$, almost everywhere [3]. The multimodal normal distribution is a special case of (1) that is addressed by [6] and defined in the following definition.

Definition 1. The random variable W follows the multimodal normal distribution, denoted by $W \sim \mathcal{MN}(\mu, \alpha, \lambda, \sigma)$, if its pdf is of the form

$$f_W(w|\mu, \alpha, \lambda, \sigma) = \phi\left(w|\mu, \sigma^2\right)\left\{1 + \frac{1}{\alpha}\sin(\lambda z)\right\}, \tag{2}$$

where $\phi\left(w|\mu, \sigma^2\right)$ denotes the normal pdf with mean μ and variance σ^2, $z = (w - \mu)/\sigma$, the location parameter $\mu \in \mathbb{R}$, the shape parameter $\alpha \geq 1$, the dispersion parameter $\sigma > 0$ and the peak parameter $\lambda \in \mathbb{R}$.

If $\lambda = 0$ or $\alpha \to \infty$ then the normal distribution is retrieved. The density plots of $\mathcal{MN}(0, \alpha, \lambda, \sigma)$ (not shown) show that as λ increases, the number of peaks increases and for any λ between $p - 1$ and p, for $p \in \mathbb{N}$, the number of visible peaks is p. The peaks are obvious for α close to 1, but as α increases, peaks tend to flatten. It is clear that a range of shapes are created by various values of σ.

Proposition 1. *If $W \sim \mathcal{MN}(\mu, \alpha, \lambda, \sigma)$, then the expectation, variance and characteristic function of W are*

$$\begin{aligned}
\mathrm{E}(W) &= \mu + \frac{\sigma\lambda}{\alpha}\exp\left(\frac{-\lambda^2}{2}\right), \\
\mathrm{Var}(W) &= \sigma^2\left\{1 - \frac{\lambda^2}{\alpha^2}\exp\left(-\lambda^2\right)\right\}, \\
\psi_W(t) &= \exp\left(-\frac{t^2\sigma^2}{2} + it\mu\right)\left\{1 + \frac{1}{\alpha}\exp\left(\frac{-\lambda^2}{2}\right)\sin(i\lambda\sigma t)\right\}.
\end{aligned}$$

Proof. Using usual statistical methods, these basic properties hold in which any clear proof is omitted. □

2.2. The multivariate multimodal normal distribution

We now introduce a multimodal extension of normal distribution for the p-dimensional random vector $\mathbf{Y}$ in $\mathbb{R}^p$, $p \geq 1$. Denote $\phi_p(\mathbf{y}|\boldsymbol{\mu}, \mathbf{V})$ the p-variate normal pdf with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$. Let $\boldsymbol{\delta}$ be a p dimensional vector with elements $\delta_1, \ldots, \delta_p$ and scalar $\alpha \geq 1$. The multivariate multimodal normal ($\mathcal{MMN}$) distribution is obtained by the following proposition.

Proposition 2. *Let the random vector $\mathbf{Y}|W = w \sim \mathcal{N}_p(\boldsymbol{\mu} + w\boldsymbol{\delta}, \mathbf{V})$ and $W \sim \mathcal{MN}(0, \alpha, \lambda, \sigma)$. The marginal density of $\mathbf{Y}$ is given by*

$$f(\mathbf{y}|\boldsymbol{\delta}, \mathbf{V}, \boldsymbol{\mu}, \alpha, \lambda, \sigma) = \phi_p\left(\mathbf{y}|\boldsymbol{\mu}, \mathbf{V}_y\right)\left[1 + \frac{1}{\alpha_y}\sin\left\{\boldsymbol{\Lambda}_y(\mathbf{y} - \boldsymbol{\mu})\right\}\right] \tag{3}$$

where $\mathbf{V}_y = \mathbf{V} + \sigma^2\boldsymbol{\delta}\boldsymbol{\delta}^\top$, $\alpha_y = \alpha\exp\left\{-\lambda^2\left(2 + 2\sigma^2\boldsymbol{\delta}^\top\mathbf{V}^{-1}\boldsymbol{\delta}\right)^{-1}\right\}$ and $\boldsymbol{\Lambda}_y = \lambda\sigma\boldsymbol{\delta}^\top\mathbf{V}^{-1}\left(1 + \sigma^2\boldsymbol{\delta}^\top\mathbf{V}^{-1}\boldsymbol{\delta}\right)^{-1}$. We denote $\mathbf{Y} \sim \mathcal{MMN}_p(\boldsymbol{\delta}, \mathbf{V}, \boldsymbol{\mu}, \alpha, \lambda, \sigma)$ with the pdf given in (3).

Proof. The joint density function of $\mathbf{Y}$ and W is

$$f(\mathbf{y}, w|\boldsymbol{\delta}, \mathbf{V}, \boldsymbol{\mu}, \alpha, \lambda, \sigma) = \phi_p(\mathbf{y}|\boldsymbol{\mu} + w\boldsymbol{\delta}, \boldsymbol{V})\,\phi\left(w|0, \sigma^2\right)\left\{1 + \frac{1}{\alpha}\sin(\lambda w/\sigma)\right\}.$$

Straightforward algebra yields

$$\begin{aligned}
f(\mathbf{y}|\boldsymbol{\delta}, \mathbf{V}, \boldsymbol{\mu}, \alpha, \lambda, \sigma) &= \int_{\mathbb{R}} f(\mathbf{y}, w|\boldsymbol{\delta}, \mathbf{V}, \boldsymbol{\mu}, \alpha, \lambda, \sigma)\,dw \\
&= \phi_p\left(\mathbf{y}|\boldsymbol{\mu}, \mathbf{V} + \sigma^2\boldsymbol{\delta}\boldsymbol{\delta}^\top\right)\int_{\mathbb{R}} \phi\left(w|\mu_w, \sigma_w^2\right)\left\{1 + \frac{1}{\alpha}\sin(\lambda w/\sigma)\right\}dw,
\end{aligned}$$

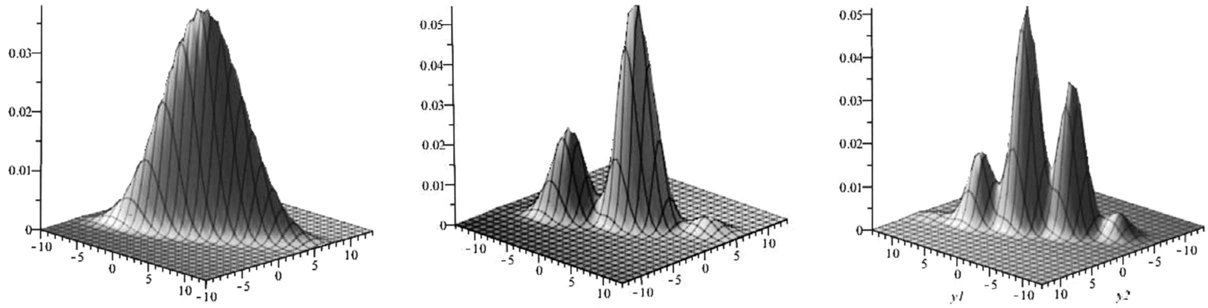

Fig. 1. The density plot of $\mathcal{MMN}_2\left(\boldsymbol{\delta}, \mathbf{I}, \mathbf{0}, 1, \lambda, 1\right)$ defined in Proposition 2 for $\boldsymbol{\delta} = (4, 2)^\top$ and $\lambda = 0.5$ (left) $\lambda = 3$ (center) $\lambda = 5$ (right).

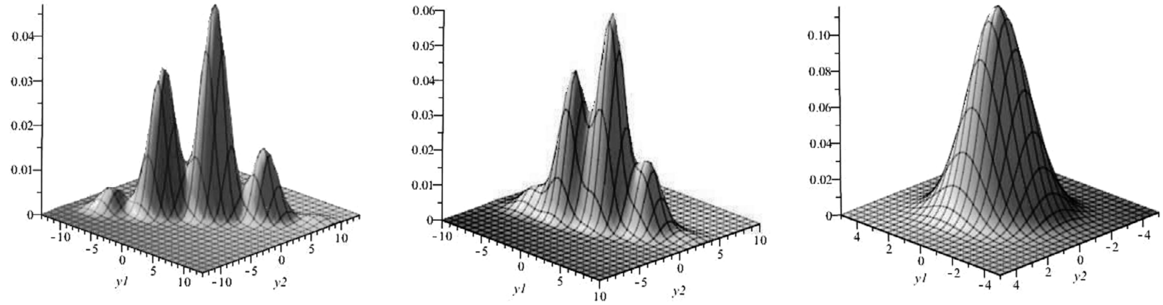

Fig. 2. The density plot of $\mathcal{MMN}_2\left(\boldsymbol{\delta}, \mathbf{I}, \mathbf{0}, 1, 5, 1\right)$ defined in Proposition 2 for $\boldsymbol{\delta} = (4, 3)^\top$ (left) $\boldsymbol{\delta} = (3, 1)^\top$ (center) $\boldsymbol{\delta} = (0.8, 0.5)^\top$ (right).

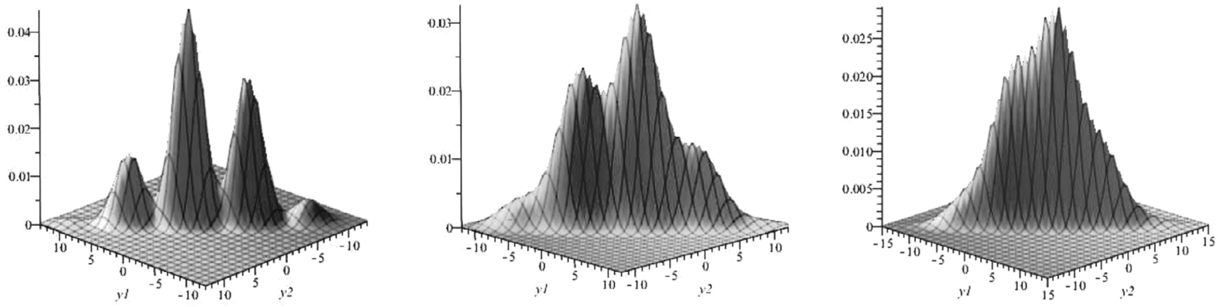

Fig. 3. The density plot of $\mathcal{MMN}_2\left(\boldsymbol{\delta}, \mathbf{I}, \mathbf{0}, \alpha, 5, 1\right)$ defined in Proposition 2 for $\boldsymbol{\delta} = (4, 4)^\top$ and $\alpha = 1$ (left) $\alpha = 3$ (center) $\alpha = 7$ (right).

where $\mu_w = \sigma_w^2 \boldsymbol{\delta}^\top \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu})$ and $\sigma_w^2 = \left(\boldsymbol{\delta}^\top \mathbf{V}^{-1} \boldsymbol{\delta} + 1/\sigma^2\right)^{-1}$. The integral can be computed using (2) and Proposition 6 in Appendix A. □

For $\boldsymbol{\delta} = \mathbf{0}$, the multivariate normal is retrieved and setting $p = 1$ coincides with the univariate case (2). However, if one let the components of the random vector $\mathbf{Y} = \left(Y_1, \ldots, Y_p\right)^\top$ be independent and each follows the univariate multimodal normal distribution then the distribution of $\mathbf{Y}$ is not $\mathcal{MMN}$. In the bivariate case, Figs. 1–3 show density (3) by setting $\boldsymbol{\mu}$ a null vector and $\mathbf{V}$ the identity matrix $\mathbf{I}_2$. Figures clearly indicate the multimodal feature of density functions for various values of α, λ and $\boldsymbol{\delta} = (\delta_1, \delta_2)^\top$. Fig. 1 shows that as λ increases, the number of peaks increases. While, the number of peaks increases as λ increases when one or both values of δ_1 and δ_2 will not be close to one. Also, as δ_1 or δ_2 increases the peaks become more visible and the range of densities increases (Fig. 2). Fig. 3 shows that the peaks are prominent for α close to 1, but as α increases they tend to be smoother. The change of σ depicts that the peaks tend to smooth out as σ tends to 0 and become more visible for σ close to 1. Also, the range of densities increases and peaks tend to smooth out as σ increases. Some particular examples are:

1. As λ decreases or α increases $\alpha_y \to \infty$ and thus $f(\mathbf{y}) \to \phi_p(\mathbf{y}|\boldsymbol{\mu}, \mathbf{V})$ (Fig. 1 (left) and Fig. 3 (right)).
2. As $\lambda \to 0$, or scale parameter σ decreases, $\boldsymbol{\Lambda}_{\mathbf{y}} \to \mathbf{0}$ and thus $f(\mathbf{y}) \to \phi_p(\mathbf{y}|\boldsymbol{\mu}, \mathbf{V})$ (Fig. 1 (left)).

3. As $\boldsymbol{\delta} \to \mathbf{0}$, or at least one elements of $\boldsymbol{\delta}$ and the off-diagonal elements of $\mathbf{V}$ tend to zero, then $\boldsymbol{\Lambda}_{\mathbf{y}} \to 0$ and thus $f(\mathbf{y}) \to \phi_p(\mathbf{y}|\boldsymbol{\mu}, \mathbf{V})$ (Fig. 2 (right)).

By using statistical techniques, the following properties of the multivariate multimodal normal distribution hold.

Proposition 3. *Let* $\mathbf{Y} \sim \mathcal{MMN}_p(\boldsymbol{\delta}, \mathbf{V}, \boldsymbol{\mu}, \alpha, \lambda, \sigma)$.

(i) *The hierarchical representation of* $\mathbf{Y}$ *is given by*

$$\mathbf{Y} = \boldsymbol{\mu} + W\boldsymbol{\delta} + \mathbf{V}^{1/2}\mathbf{Z},$$

where $W \sim \mathcal{MN}(0, \alpha, \lambda, \sigma)$ *and* $\mathbf{Z} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ *are independent.*

(ii) *The expectation and covariance matrix of* $\mathbf{Y}$ *are*

$$\begin{aligned} \mathrm{E}(\mathbf{Y}) &= \boldsymbol{\mu} + \frac{\sigma\lambda}{\alpha}\exp\left(\frac{-\lambda^2}{2}\right)\boldsymbol{\delta}, \\ \mathrm{Cov}(\mathbf{Y}) &= \mathbf{V} + \sigma^2\left\{1 - \frac{\lambda^2}{\alpha^2}\exp\left(-\lambda^2\right)\right\}\boldsymbol{\delta}\boldsymbol{\delta}^\top. \end{aligned}$$

(iii) *The characteristic function of* $\mathbf{Y}$ *equals*

$$\psi_{\mathbf{Y}}(t) = \exp\left\{i\mathbf{t}^\top\boldsymbol{\mu} - \frac{\mathbf{t}^\top\left(\mathbf{V} + \sigma^2\boldsymbol{\delta}\boldsymbol{\delta}^\top\right)\mathbf{t}}{2}\right\}\left\{1 + \frac{1}{\alpha}\exp\left(\frac{-\lambda^2}{2}\right)\sin\left(i\lambda\sigma\mathbf{t}^\top\boldsymbol{\delta}\right)\right\}.$$

(iv) *Let* $\mathbf{Z} = \mathbf{A}\mathbf{Y} + \mathbf{b}$, *where* $\mathbf{A}$ *is an* $q \times p$ *arbitrary matrix and vector* $\mathbf{b} \in \mathbb{R}^q$. *Then*

$$\mathbf{Z} \sim \mathcal{MMN}_q\left(\mathbf{A}\boldsymbol{\delta}, \mathbf{A}\mathbf{V}\mathbf{A}^\top, \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \alpha, \lambda, \sigma\right).$$

Proof. The basic properties (i) and (ii) hold simply using usual statistical methods.
(iii) we have

$$\begin{aligned} \psi_{\mathbf{Y}}(\mathbf{t}) &= \mathrm{E}\left[\mathrm{E}\left\{\exp\left(i\mathbf{t}^\top\mathbf{Y}\right)|W\right\}\right] \\ &= \exp\left(i\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top\mathbf{V}\mathbf{t}\right)\mathrm{E}\left\{\exp\left(iW\mathbf{t}^\top\boldsymbol{\delta}\right)\right\} = \exp\left(i\mathbf{t}^\top\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top\mathbf{V}\mathbf{t}\right)\psi_W\left(\boldsymbol{\delta}^\top\mathbf{t}\right). \end{aligned}$$

By replacing the characteristic function of W given by Proposition 1, the result is obtained.
(iv) The characteristic function of $\mathbf{Z}$ is $\psi_{\mathbf{Z}}(\mathbf{t}) = \exp\left(i\mathbf{t}^\top\mathbf{b}\right)\psi_{\mathbf{Y}}\left(\mathbf{A}^\top\mathbf{t}\right)$. After some algebra

$$\psi_{\mathbf{Z}}(\mathbf{t}) = \exp\left\{i\mathbf{t}^\top\boldsymbol{\mu}_z - \frac{1}{2}\mathbf{t}^\top\left(\mathbf{V}_z + \sigma^2\boldsymbol{\delta}_z\boldsymbol{\delta}_z^\top\right)\mathbf{t}\right\}\left\{1 + \frac{1}{\alpha}\exp\left(-\frac{\lambda^2}{2}\right)\sin\left(i\lambda\sigma\mathbf{t}^\top\boldsymbol{\delta}_z\right)\right\},$$

where $\boldsymbol{\delta}_z = \mathbf{A}\boldsymbol{\delta}$, $\boldsymbol{\mu}_z = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ and $\mathbf{V}_z = \mathbf{A}\mathbf{V}\mathbf{A}^\top$. Thus the proof is completed using part (iii). □

The following Proposition states that any marginal and conditional distribution of $\mathcal{MMN}$ belongs to this family.

Proposition 4. *Let* $\mathbf{Y} \sim \mathcal{MMN}_p(\boldsymbol{\delta}, \mathbf{V}, \boldsymbol{\mu}, \alpha, \lambda, \sigma)$. *Consider the partition*

$$\mathbf{Y} = \binom{\mathbf{Y}_1}{\mathbf{Y}_2}, \quad \boldsymbol{\mu} = \binom{\boldsymbol{\mu}_1}{\boldsymbol{\mu}_2}, \quad \boldsymbol{\delta} = \binom{\boldsymbol{\delta}_1}{\boldsymbol{\delta}_2}, \quad \text{and} \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix},$$

where $\mathbf{Y}_1$, $\boldsymbol{\mu}_1$, *and* $\boldsymbol{\delta}_1$ *are* k *dimensional while* $\mathbf{Y}_2$, $\boldsymbol{\mu}_2$, *and* $\boldsymbol{\delta}_2$ *are* $(p-k)$ *dimensional vectors,* $\mathbf{V}_{11}$ *and* $\mathbf{V}_{22}$ *are* $k \times k$ *and* $(p-k) \times (p-k)$ *positive definite matrices, respectively.*

(i) $\mathbf{Y}_1 \sim \mathcal{MMN}_p(\boldsymbol{\delta}_1, \mathbf{V}_{11}, \boldsymbol{\mu}_1, \alpha, \lambda, \sigma)$.

(ii) $\mathbf{Y}_2|\mathbf{Y}_1 = \mathbf{y}_1 \sim \mathcal{MMN}_{p-k}\left(\boldsymbol{\delta}_{2.1}, \mathbf{V}_{2.1}, \boldsymbol{\mu}_{2.1}, \alpha, \lambda, \sigma\right)$, *where* $\boldsymbol{\delta}_{2.1} = \boldsymbol{\delta}_2 - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\boldsymbol{\delta}_1$, $\mathbf{V}_{2.1} = \mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}$ *and* $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)$.

Proof. (i) The vector $\mathbf{Y}_1$ can be expressed as $\mathbf{A}\mathbf{Y}$, where $\mathbf{A} = \left(\mathbf{I}_k \quad \mathbf{0}_{k\times(p-k)}\right)$ is a $k \times p$ matrix, $\mathbf{I}_k$ is an identity matrix and $\mathbf{0}_{k\times(p-k)}$ is a null matrix. Following Proposition 3(iv), $\mathbf{Y}_1$ follows a $\mathcal{MMN}$ distribution with parameters $\boldsymbol{\delta}_1 = \mathbf{A}\boldsymbol{\delta}$, $\mathbf{V}_{11} = \mathbf{A}\mathbf{V}\mathbf{A}^\top$, $\boldsymbol{\mu}_1 = \mathbf{A}\boldsymbol{\mu}$, α, λ and σ.
(ii) The proof is done by using the conditional distribution $\mathbf{Y}|W = w \sim \mathcal{N}_p(\boldsymbol{\mu} + w\boldsymbol{\delta}, \mathbf{V})$, where $W \sim \mathcal{MN}(\alpha, \lambda, \sigma)$, so that

$$\mathbf{Y}_2|\mathbf{Y}_1 = \mathbf{y}_1, W = w \sim \mathcal{N}_{p-k}\left(\boldsymbol{\mu}^*, \mathbf{V}_{2.1}\right)$$

with $\mathbf{V}_{2.1} = \mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}$ and $\boldsymbol{\mu}^* = w\boldsymbol{\delta}_2 + \boldsymbol{\mu}_2 + \mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1 - w\boldsymbol{\delta}_1)$ which can be simplified to $\boldsymbol{\mu}^* = \boldsymbol{\mu}_{2.1} + W\boldsymbol{\delta}_{2.1}$. Marginalizing over W gives the conditional distribution of $\mathbf{Y}_2$ given $\mathbf{Y}_1 = \mathbf{y}_1$ as stated in (ii). □

3. Specification of multivariate multimodal linear mixed-effects models

The multivariate linear mixed-effects (MLME) models take advantage of describing variation in multiple responses that is longitudinally measured on each subject in terms of a set of fixed covariates. These models are commonly constructed by fitting separate mixed regression models for each response and being joined by specifying a joint distribution for their underlying random effects. This familiar strategy generates efficiently an association structure between the measurements of multivariate responses and takes into account the correlation within-subjects which often presents in repeated measures [10].

Let us consider r responses and N subjects. Define the response vector $\mathbf{y}_i^k = \left(y_{i1}^k, \ldots, y_{in_i}^k\right)^\top$ for subject $i \in \{1, \ldots, n\}$ corresponding to the kth ($k \in \{1, \ldots, r\}$) response measurements at n_i different time periods. For each response, consider the linear mixed-effects (LME) model

$$\mathbf{y}_i^k = \mathbf{x}_i^k \boldsymbol{\beta}^k + \mathbf{z}_i^k \mathbf{b}_i^k + \mathbf{e}_i^k,$$

where $\mathbf{x}_i^k$ $(n_i \times p)$ and $\mathbf{z}_i^k$ $(n_i \times q)$ are known covariate matrices related to the p-dimensional vector of unknown fixed regression coefficients $\boldsymbol{\beta}^k$ and the q-dimensional vector of random effects $\mathbf{b}_i^k$, respectively, and $\mathbf{e}_i^k$ is the n_i-dimensional vector of within-subject error terms. In order to introduce the multivariate linear mixed model, we use the vec operator for the ith response and error vectors as $\mathbf{Y}_i = \text{vec}\left(\mathbf{y}_i^1, \ldots, \mathbf{y}_i^r\right)$ and $\mathbf{E}_i = \text{vec}\left(\mathbf{e}_i^1, \ldots, \mathbf{e}_i^r\right)$, the vector of all fixed regression coefficients as $\boldsymbol{\beta} = \text{vec}\left(\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^r\right)$ and the vector of all random effects as $\mathbf{b}_i = \text{vec}(\mathbf{b}_i^1, \ldots, \mathbf{b}_i^r)$. Denote the matrix $\mathbf{X}_i = \text{diag}\left(\mathbf{x}_i^1, \ldots, \mathbf{x}_i^r\right)$ for the fixed regression coefficients and $\mathbf{Z}_i = \text{diag}\left(\mathbf{z}_i^1, \ldots, \mathbf{z}_i^r\right)$ for the random effects. Thus, the MLME model for r responses is specified by

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \mathbf{E}_i.$$

where the repeated measures $y_{i1}^k, \ldots, y_{in_i}^k$ for each $k \in \{1, \ldots, r\}$ are conditionally independent given $\mathbf{b}_i^k$. In practical applications, it may be reasonable and convenient for the computational purposes to restrict the covariance structures of random effects and error terms as some specific forms. This may also lead to more efficient parameter estimation and statistical inference. In particular, two commonly used approaches are addressed by [11]. In the first approach a multivariate distribution is specified for the error vector $\mathbf{E}_i$ to take into account correlated responses. Thus, the covariance matrix of $\mathbf{E}_i$ is a blocked-diagonal structure of the form $\mathbf{V}_e = \mathbf{V}_\varepsilon \otimes \mathbf{I}_{n_i}$, where $\mathbf{V}_\varepsilon$ is a covariance matrix of size r, $\mathbf{I}_{n_i}$ is a $n_i \times n_i$ identity matrix and the symbol $\otimes$ denotes the Kronecker product. Here, the random effects $\mathbf{b}_i^1, \ldots, \mathbf{b}_i^r$ are assumed to be independent with the covariance matrix $\mathbf{V}_b = \mathbf{I}_r \otimes \mathbf{G}$, where $\mathbf{G}$ is a covariance matrix of size q, where its off-diagonal elements are constructed related to each random effects vector $\mathbf{b}_i^k$. In the second approach, all responses are conditionally independent, given random effects and are joined by assuming that the vectors $\mathbf{b}_i^1, \ldots, \mathbf{b}_i^r$ being correlated and the covariance matrix of $\mathbf{b}_i$ takes the form $\mathbf{V}_b = \boldsymbol{\Omega} \otimes \mathbf{G}$, where $\boldsymbol{\Omega}$ is a covariance matrix of size r with non-zero off-diagonal elements. In this case, the errors $\mathbf{e}_i^1, \ldots, \mathbf{e}_i^r$ are usually assumed to be independent having covariance matrices $\sigma_{e_1}^2 \mathbf{I}_{n_i}, \ldots, \sigma_{e_r}^2 \mathbf{I}_{n_i}$. Thus, the covariance matrix of error vector $\mathbf{E}_i$ is $\mathbf{V}_e = \mathbf{I}_{n_i} \otimes \mathbf{V}_\varepsilon$, where $\mathbf{V}_\varepsilon$ is a diagonal matrix with elements $\sigma_{e_1}^2, \ldots, \sigma_{e_r}^2$.

Now, we propose an extension of MLME models based on utilizing the $\mathcal{MMN}$ distribution for the random effects. Specifically, the new model is constructed by allowing the response vector $\mathbf{Y}_i$, conditioned on the effects $\mathbf{b}_i$, follows the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{V}_e$, while the random effects follow a rq-variate multimodal normal distribution. We introduce the hierarchical mixed model

$$\begin{aligned} \mathbf{Y}_i | \boldsymbol{b}_i &\overset{ind}{\sim} \mathcal{N}_{rn_i}\left(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i, \mathbf{V}_e\right), \\ \mathbf{b}_i | W_i = w_i &\overset{ind}{\sim} \mathcal{N}_{rq}\left(w_i \boldsymbol{\delta}, \mathbf{V}_b\right), \\ W_i &\overset{iid}{\sim} \mathcal{MN}\left(\mu, \alpha, \lambda, \sigma\right), \end{aligned} \tag{4}$$

where *ind* denotes independent and *iid* stands for the independent and identically distributed. For $\boldsymbol{\delta} = \mathbf{0}$, the model reduces to the commonly used linear mixed model, i.e., when $\mathbf{b}_i \sim \mathcal{N}_{rq}\left(\mathbf{0}, \mathbf{V}_b\right)$. In (4), it is seen that $\mathrm{E}\left(\mathbf{b}_i\right) = \mu_w \boldsymbol{\delta}$ and $\text{cov}\left(\mathbf{b}_i\right) = \mathbf{V}_b + \sigma_w^2 \boldsymbol{\delta}\boldsymbol{\delta}^\top$, where $\mu_w = \mu - \sigma\lambda\alpha^{-1}\exp\left(-\lambda^2/2\right)$ and $\sigma_w^2 = \sigma^2 - \mu_w^2$. If all $\mathbf{E}_i$'s are assumed to be independent and normally distributed with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{V}_e$ then the marginal expectation of $\mathbf{Y}_i$ is given by $\mathrm{E}\left(\mathbf{Y}_i\right) = \mathbf{X}_i \boldsymbol{\beta} + \mu_w \mathbf{Z}_i \boldsymbol{\delta}$.

In the regression modeling methodology, the marginal expectation of the response vector is commonly assumed to depend only on the covariates, i.e., $\mathrm{E}\left(\mathbf{Y}_i\right) = \mathbf{X}_i \boldsymbol{\beta}$. It is desirable to keep this property even for our proposed model by centralizing the latent variable W_i, i.e., by setting $\mu = \sigma\lambda\alpha^{-1}\exp\left(-\lambda^2/2\right)$, we have $\mu_w = 0$ which implies $\mathrm{E}\left(\mathbf{b}_i\right) = \mathbf{0}$ and $\sigma_w^2 = \sigma^2$. The covariance matrix of the response is given by $\text{cov}\left(\mathbf{Y}_i\right) = \mathbf{V}_e + \mathbf{Z}_i\left(\mathbf{V}_b + \sigma^2 \boldsymbol{\delta}\boldsymbol{\delta}^\top\right)\mathbf{Z}_i^\top$. The marginal distribution of the response vector $\mathbf{Y}_i$ is obtained in the following proposition.

Proposition 5. *Let $\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \mathbf{E}_i$, where $\mathbf{E}_i \overset{ind}{\sim} \mathcal{N}_{rn_i}\left(\mathbf{0}, \mathbf{V}_e\right)$ and $\mathbf{b}_i \sim \mathcal{MMN}_{rq}\left(\boldsymbol{\delta}, \mathbf{V}_b, \mu, \alpha, \lambda, \sigma\right)$ are independent. The marginal distribution of $\mathbf{Y}_i$ is given by*

$$f\left(\mathbf{y}_i | \boldsymbol{\theta}\right) = \phi_{rn_i}\left(\mathbf{Y}_i | \boldsymbol{\mu}_{y_i}, \mathbf{V}_{y_i}\right) \int_{\mathbb{R}^{rq}} \phi_{rq}\left(\mathbf{b}_i | \boldsymbol{\mu}_{b_i}, \Lambda_{b_i}\right) \left[1 + \frac{1}{\alpha_y} \sin\left\{\boldsymbol{\delta}_b\left(\mathbf{b}_i - \mu\boldsymbol{\delta}\right)\right\}\right] d\mathbf{b}_i, \tag{5}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{V}_e, \mathbf{V}_b, \boldsymbol{\delta}, \alpha, \lambda, \sigma)$ *and*

$$\begin{array}{rclrcl}
\boldsymbol{\mu}_{b_i} & = & \mu\boldsymbol{\delta} + \boldsymbol{\Lambda}_{b_i}\mathbf{Z}_i^\top \mathbf{V}_e^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mu\mathbf{Z}_i\boldsymbol{\delta}\right), & \boldsymbol{\mu}_{y_i} & = & \mathbf{X}_i\boldsymbol{\beta} + \mu\mathbf{Z}_i\boldsymbol{\delta}, \\
\boldsymbol{\Lambda}_{b_i} & = & \left\{\left(\mathbf{V}_b + \sigma^2\boldsymbol{\delta}\boldsymbol{\delta}^\top\right)^{-1} + \mathbf{Z}_i^\top \mathbf{V}_e^{-1}\mathbf{Z}_i\right\}^{-1}, & \mathbf{V}_{y_i} & = & \mathbf{V}_e + \mathbf{Z}_i\left(\mathbf{V}_b + \sigma^2\boldsymbol{\delta}\boldsymbol{\delta}^\top\right)\mathbf{Z}_i^\top, \\
\boldsymbol{\delta}_b & = & \lambda\sigma\left(1 + \sigma^2\boldsymbol{\delta}^\top\mathbf{V}_b^{-1}\boldsymbol{\delta}\right)^{-1}\boldsymbol{\delta}^\top\mathbf{V}_b^{-1}, & \alpha_y & = & \alpha\exp\left\{-\frac{\lambda^2}{2}\left(1 + \sigma^2\boldsymbol{\delta}^\top\mathbf{V}_b^{-1}\boldsymbol{\delta}\right)^{-1}\right\}.
\end{array}$$

Proof. See Appendix B. □

Note that the proposed model introduces a unimodal distribution for the within-subject errors $\mathbf{E}_i$ and a multimodal distribution for the random effects $\mathbf{b}_i$. By using graphical techniques, the marginal distribution of $\mathbf{Y}_i$ is shown to be in the class of multimodal distributions. For illustration, we plot $f(\mathbf{y}_i|\boldsymbol{\theta})$ in a regression model with only random intercepts in the univariate and bivariate cases. Results reveal that $f(\mathbf{y}_i|\boldsymbol{\theta})$ is multimodal with similar properties to the related multimodal normal distribution. Also, the univariate marginal distribution of each response $\mathbf{Y}_i^k$, $k = \{1, 2\}$, is multimodal. Some graphs are provided in Appendix D. In fact, from (4) and using the properties of the multivariate normal distribution, the marginal distribution of each $\mathbf{Y}_i^k$, given $\mathbf{b}_i^k$, is multivariate normal $\mathcal{N}_{n_i}\left(\mathbf{X}_i^k\boldsymbol{\beta}^k + \mathbf{Z}_i^k\mathbf{b}_i^k, \sigma_e^{2k}\mathbf{I}_{n_i}\right)$. Thus, by Proposition 5 the marginal distribution of each response belongs to the class of multimodal distributions.

To carry out inference for the vector parameter $\boldsymbol{\theta}$, a direct maximization of the log-likelihood function $\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_i \ln[f(\mathbf{y}_i|\boldsymbol{\theta})]$ may involve solving complex integrals using advanced numerical techniques. The hierarchical structure (4) allows better implementation of several estimation methods, such as the ECM algorithm of the frequentist approach or the Markov chain Monte Carlo (MCMC) approach in a Bayesian perspective. In the next section we describe in detail the MCMC approach to make inferences on model parameters. Some details of the ECM algorithm are given in Appendix C.

4. Bayesian computation

The proposed multivariate multimodal mixed-effects model is very convenient to implement in a Bayesian framework, since it can be reconstructed by the hierarchical form (4) which allows us to utilize the MCMC techniques straightforwardly. In this setting, to make inference from the posterior distributions and obtain a subsequent approximation of moments for the model parameters, the Gibbs sampler method is usually implemented. The posterior distribution combines the prior information regarding the parameter values with the information in data obtained directly from the complete likelihood function associated with $(\mathbf{y}, \mathbf{b}, w)$. It is given by

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{b}, w) = \prod_{i=1}^{N} \phi_{rn_i}\left(\mathbf{y}_i|\mathbf{x}_i\boldsymbol{\beta} + \mathbf{z}_i\mathbf{b}_i, \mathbf{V}_e\right)\phi_{rq}\left(\mathbf{b}_i|w_i\boldsymbol{\delta}, \mathbf{V}_b\right)\phi\left(w_i|\mu, \sigma^2\right)h(w_i),$$

where $h(w_i) = 1 + \alpha^{-1}\sin\{\lambda(w_i - \mu)/\sigma\}$ and $\mu = \sigma\lambda\alpha^{-1}\exp\left(-\lambda^2/2\right)$. To choose a joint prior distribution for the unknown parameters $\boldsymbol{\theta}$, we adopt independent prior distributions, i.e.,

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\,\pi(\mathbf{V}_e)\,\pi(\boldsymbol{\delta})\,\pi(\mathbf{V}_b)\,\pi(\alpha)\,\pi(\lambda)\,\pi(\sigma),$$

where $\pi(\boldsymbol{\theta})$ is the prior density function of $\boldsymbol{\theta}$. We assume that $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, respectively, follow the normal distributions $\mathcal{N}_{rp}\left(\boldsymbol{\beta}_0, \mathbf{S}_\beta\right)$ and $\mathcal{N}_{rq}(\boldsymbol{\delta}_0, \mathbf{S}_\delta)$, the matrix $\mathbf{V}_b$ follows the rq-variate inverse-Wishart distribution $\mathcal{IW}_{rq}(\tau_b, \mathbf{S}_b)$ and the same prior for $\mathbf{V}_e = \sigma_e^2\mathbf{I}_{rn_i}$, and the scale parameter σ_e^2 follows the inverse-Gamma distribution $\mathcal{IG}(\tau_e/2, \tau_e/2)$. For computational simplicity, we adopt uniform priors over their parameter spaces for α, λ and σ. All hyper-parameters in the priors are assumed to be known. In this setting Bayesian computation is simply done, since the full conditional posteriors involved in the Gibbs sampler are of known forms and hence easy to simulate. The joint posterior distribution for $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}, \mathbf{b}, \mathbf{w}|\mathbf{y}) = \pi(\boldsymbol{\theta})\,\mathrm{L}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{b}, w). \tag{6}$$

The Gibbs sampler proceeds by drawing samples iteratively from all full conditional posteriors deriving from (6). Under the full model as previously described, the full conditional posteriors of $\boldsymbol{\theta}$ and $\mathbf{b}_i$ for $i \in \{1, \ldots, n\}$ are given by

$$\boldsymbol{\beta}|\mathbf{V}_e, \mathbf{V}_b, \boldsymbol{\delta}, \alpha, \lambda, \sigma, \mathbf{b}, w, \mathbf{y} \sim \mathcal{N}_{rp}\left(\boldsymbol{\mu}_\beta, \mathbf{V}_\beta\right),$$

where $\boldsymbol{\mu}_\beta = \mathbf{V}_\beta\left\{\mathbf{S}_\beta^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^N \mathbf{x}_i^\top\mathbf{V}_e^{-1}(\mathbf{y}_i - \mathbf{z}_i\mathbf{b}_i)\right\}$ and $\mathbf{V}_\beta = \left(\mathbf{S}_\beta^{-1} + \sum_{i=1}^N \mathbf{x}_i^\top\mathbf{V}_e^{-1}\mathbf{x}_i\right)^{-1}$;

$$\sigma_e^2|\boldsymbol{\beta}, \mathbf{V}_b, \boldsymbol{\delta}, \alpha, \lambda, \sigma, \mathbf{b}, w, \mathbf{y} \sim \mathcal{IG}\left(\frac{1}{2}(\tau_e + N), \frac{1}{2}\left(\tau_e + \sum_{i=1}^N \mathbf{r}_i^\top\mathbf{r}_i\right)\right),$$

with $\mathbf{r}_i = \mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta} - \mathbf{z}_i\mathbf{b}_i$;

$$\boldsymbol{\delta}|\boldsymbol{\beta}, \mathbf{V}_e, \mathbf{V}_b, \alpha, \lambda, \sigma, \mathbf{b}, w, \mathbf{y} \sim \mathcal{N}_{rq}\left(\boldsymbol{\mu}_\delta, \mathbf{V}_\delta\right),$$

for $\mathbf{V}_\delta = \left(\mathbf{S}_\delta^{-1} + \mathbf{V}_b^{-1}\sum_{i=1}^N w_i\right)^{-1}$ and $\boldsymbol{\mu}_\delta = \mathbf{V}_\delta\left(\mathbf{S}_\delta^{-1}\boldsymbol{\delta}_0 + \mathbf{V}_b^{-1}\sum_{i=1}^N w_i\mathbf{b}_i\right)$;

$$\mathbf{V}_b|\boldsymbol{\beta},\mathbf{V}_e,\boldsymbol{\delta},\alpha,\lambda,\sigma,\mathbf{y},w,\mathbf{y} \sim \mathcal{IW}_{rq}\left(\tau_b+N,\mathbf{S}^{-1}\right),$$

with $\mathbf{S} = \mathbf{S}_b^{-1} + \sum_{i=1}^N \left(\mathbf{b}_i - w_i\boldsymbol{\delta}\right)\left(\mathbf{b}_i - w_i\boldsymbol{\delta}\right)^\top$, and

$$\mathbf{b}_i|\boldsymbol{\beta},\mathbf{V}_e,\mathbf{V}_b,\boldsymbol{\delta},\alpha,\lambda,\sigma,w_i,\mathbf{y}_i \overset{iid}{\sim} \mathcal{N}_{rq}\left(\boldsymbol{\mu}_{b_i},\mathbf{S}_{b_i}\right),$$

where $\mathbf{S}_{b_i} = \left(\mathbf{V}_b^{-1} + \mathbf{z}_i^\top\mathbf{V}_e^{-1}\mathbf{z}_i\right)^{-1}$ and $\boldsymbol{\mu}_{b_i} = \mathbf{S}_{b_i}\left\{w_i\mathbf{V}_b^{-1}\boldsymbol{\delta} + \mathbf{z}_i^\top\mathbf{V}_e^{-1}\left(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}\right)\right\}$.

For most quantities, the simulation is straightforward since their associated full conditional posteriors are in closed form following known distributions. For w_i, we obtain

$$\pi\left(w_i|\boldsymbol{\beta},\mathbf{V}_e,\mathbf{V}_b,\boldsymbol{\delta},\alpha,\lambda,\sigma,\mathbf{b}_i,\mathbf{y}_i\right) \propto \phi\left(w_i|\mu_{w_i},S_w\right)\left[1+\frac{1}{\alpha}\sin\left\{\frac{\lambda\left(w_i-\mu\right)}{\sigma}\right\}\right], \tag{7}$$

where $S_w = \left(\boldsymbol{\delta}^\top\mathbf{V}_b^{-1}\boldsymbol{\delta} + \sigma^{-2}\right)^{-1}$ and $\mu_{w_i} = \mu + S_w\boldsymbol{\delta}^\top\mathbf{V}_b^{-1}\left(\mathbf{b}_i - \mu\boldsymbol{\delta}\right)$. To sample w_i from the non-closed form (7), the Metropolis–Hastings (MH) algorithm can be implemented. To do it, a new value w_i^{new} is required to be generated from a proposal distribution. In the current version of OpenBUGS, the Metropolis-within-Gibbs algorithm [7,21] is based on a normal proposal distribution [22]. Moreover, for α, λ and σ, the full conditional posteriors were not in closed form. Thus, a similar process could be embedded for drawing samples from these posteriors.

5. Simulation studies

We conduct our simulation studies with two scenarios for the distribution of the underlying random effects. The first simulation is designed to assess the performance of our proposed modeling strategy in the case of multimodality. In the second study, the random effects are randomly drawn by a commonly used bivariate normal distribution. We consider the bivariate LME model

$$y_{ij}^k = \beta_0^k + \beta_1^k x_j^k + \beta_2^k x_i^k + b_i^k + e_{ij}^k, \tag{8}$$

for $k \in \{1,2\}$, $i \in \{1,\ldots,100\}$, $j \in \{1,\ldots,6\}$, where $e_{ij}^1 \overset{iid}{\sim} \mathcal{N}(0,2)$, $e_{ij}^2 \overset{iid}{\sim} \mathcal{N}(0,1)$, values of each subject-level covariate x_i^k have been drawn from $\mathcal{N}(2,3)$. Also, we let x_j^1 and x_j^2 be the same for all subjects while changing within subjects and independently follow $\mathcal{N}(5,2)$. True values of fixed parameters are given in the first column of Table 1. In the first scenario we let the random intercepts b_i^1 and b_i^2 to accommodate multimodality by generating their values from the bivariate mixture distribution $\sum_{j=1}^2 \pi_j\phi_2\left(\boldsymbol{\mu}_j,\mathbf{V}_b\right)$. We set $\pi_1 = \pi_2 = 0.5$, $\boldsymbol{\mu}_1 = (3,-5)^\top$, $\boldsymbol{\mu}_2 = (-3,5)^\top$, and $\mathbf{V}_b = \begin{pmatrix}3 & 1\\ 1 & 4\end{pmatrix}$. In the second scenario, we let b_i^1 and b_i^2 be bivariate normally distributed with mean $\mathbf{0}$ and the covariance matrix $\mathbf{V}_b$. Then, we generated 100 data sets and for each one, the model (8) was fitted by letting $e_{ij}^k \overset{iid}{\sim} \mathcal{N}\left(0,\sigma_{ek}^2\right)$ for $k \in \{1,2\}$, and two random intercepts b_i^1 and b_i^2 be distributed as:

M1: the bivariate normal $\mathcal{N}_2\left(\mathbf{0},\mathbf{V}_b\right)$;
M2: the bivariate multimodal normal $\mathcal{MMN}_2\left(\boldsymbol{\delta},\mathbf{V}_b,\sigma\lambda\alpha^{-1}\exp\left(-\lambda^2/2\right),\alpha,\lambda,\sigma\right)$;
M3: the mixture distribution $\sum_{j=1}^2 \pi_j\phi_2\left(\boldsymbol{\mu}_j,\mathbf{V}_b\right)$, where $\sum_{j=1}^2 \pi_j = 1$. Here, the additional constraint $\sum_{j=1}^2 \pi_j\boldsymbol{\mu}_j = 0$ is required to let the mean value of the random effects being zero. Also, it is necessary to assume a common covariance matrix for all components to avoid an unbounded likelihood [32].

We adopt independent and non-informative priors for all models. In particular, we assign $\mathcal{N}(0,1000)$ for each regression coefficient and each peak parameter in $\boldsymbol{\delta}$, the inverse-Gamma $\mathcal{IG}(0.01,0.01)$ for each σ_{e1}^2 and σ_{e2}^2, the uniform prior $\mathcal{U}(-10,10)$ for λ, $\mathcal{U}(1,10)$ for α, $\mathcal{U}(0,10)$ for σ, $\mathcal{U}(0,1)$ for π_1 and the inverse-Wishart $\mathcal{IW}_2\left(2,\mathbf{I}_2\right)$ for $\mathbf{V}_b$. By implementing the MCMC approach within the OpenBUGS software, we used 10 000 iterations after discarding 2000 burn-in samples and the lagged value of 5 to avoid autocorrelation that appeared in the generated chains (BUGS code is provided in Appendix E). Convergence was assessed by the standard tools, such as trace plots [5]. To perform a comparative study, the parameters estimate and their standard errors, averaged over the 100 data sets, are reported in Table 1.

The estimate of $\boldsymbol{\delta}$, α, λ and σ in the proposed model M2 are significant from zero since the associated 95% confidence intervals do not include 0. For the first simulation, we obtain $\hat{\sigma} = 0.917$, $\hat{\lambda} = 4.192$, $\hat{\alpha} = 2.512$, and $\hat{\boldsymbol{\delta}} = (2.988, 3.014)$ and for the second simulation $\hat{\sigma} = 0.232$, $\hat{\lambda} = 0.549$, $\hat{\alpha} = 10.472$, and $\hat{\boldsymbol{\delta}} = (0.231, 0.583)$.

It is seen that most parameter estimates are relatively close to each other for all fitted models. The efficiency of fixed-effects estimates are nearly equal showing that the distribution of random effects does not considerably influence the estimate of these parameters. Although, for the proposed model M2, the biases and standard errors are smaller than others and the efficiency of most estimates have been improved in comparison with bivariate normally distributed random-effects. This evidence reveals that our proposed model is useful in practical applications and the adoption of incorrect assumptions (e.g., normality) for random-effects distribution may reduce the efficiency of regression parameter estimates.

Table 1
Simulation results based on 100 generated data sets of model (8) when the random effects have been generated from (a) mixture distribution and (b) bivariate normal. The parameters estimate (Est) and their standard errors (SE), averaged over the 100 data sets, are reported. M1: Bivariate normal, M2: Bivariate multimodal normal, M3: Mixture distribution with 2 components.

Parameters	M1		M2		M3	
	Est	SE	Est	SE	Est	SE
(a) Multimodal random effects						
$\beta_0^1 = 10$	8.391	0.793	8.987	0.765	8.545	0.787
$\beta_0^2 = 20$	18.624	0.665	18.402	0.544	17.832	0.612
$\beta_1^1 = 2$	2.191	0.324	2.181	0.276	2.182	0.311
$\beta_1^2 = 4$	4.321	0.432	4.217	0.324	4.219	0.376
$\beta_2^1 = 3$	2.976	0.243	2.979	0.198	3.022	0.231
$\beta_2^2 = 5$	4.968	0.321	4.974	0.288	5.024	0.290
$\sigma_{e^1}^2 = 2$	2.074	0.143	2.031	0.125	2.032	0.132
$\sigma_{e^2}^2 = 1$	1.110	0.189	1.056	0.146	1.059	0.155
(b) Normal random effects						
$\beta_0^1 = 10$	10.24	0.399	8.071	0.422	9.691	0.423
$\beta_0^2 = 20$	19.512	0.302	18.213	0.312	19.776	0.317
$\beta_1^1 = 2$	1.963	0.169	2.076	0.181	2.233	0.199
$\beta_1^2 = 4$	4.027	0.289	4.031	0.310	4.132	0.315
$\beta_2^1 = 3$	3.002	0.390	3.005	0.419	3.009	0.434
$\beta_2^2 = 5$	5.040	0.101	5.042	0.111	5.046	0.119
$\sigma_{e^1}^2 = 2$	2.029	0.203	1.964	0.210	1.942	0.217
$\sigma_{e^2}^2 = 1$	1.018	0.122	1.019	0.132	1.031	0.149

Similar findings have been addressed in [24] and [32] in a specific linear mixed-effects model. The estimate of β_0^1, β_0^2, and the scale parameters are not the same for the fitted models but are not comparable because of different statistically meaning.

In the second simulation study, the estimate of fixed-effects parameters in the proposed model is extremely close to the normal model. Furthermore, there is no efficiency loss associated when using the bivariate multimodal normal distribution. This evidence illustrates that the proposed model still provides reasonable estimation results. It reveals that our proposed model deserves to be used in practical applications as a reliable alternative even if the classical model is correct.

6. Two illustrative empirical applications

In this section, two real-life data sets are analyzed to illustrate the usefulness of our proposed model. The first data are taken from a prospective cohort study on low back pain conducted by [26]. The main aim of this study was to investigate the effects of the treatments package composed of herbal medicine, acupuncture, bee venom acupuncture, and a Korean version of spinal manipulation (Chuna) on low back pain. This data set has been analyzed in [18] using a simple bivariate mixed-effects model. We show that a complex structure involving the multimodality of bivariate responses is more realistic. The second data set, collected originally by [13], comprises the height of school girls to explain the significantly difference of the height course of girls according to the category of height of their mother (small, medium and tall). This data set appeared in several published papers using a univariate mixed-effects model and a finite mixture of normal distributions for the random-effects [32]. We re-analyze the data set using our methodology to model the multimodal behavior of the response and compare our findings with previous cited results. To obtain the results, in the Bayesian setting, we assign conjugate but non-informative priors, e.g., a normal prior with a large variance, to ensure that they are flat enough over a realistic range of parameter values. Model selection is done by the deviance information criterion (DIC) [30] to select the best fitted model. Smaller values of the DIC indicate a better fit. Other model selection criteria, such as the deviance, $D\left(\widehat{\boldsymbol{\theta}}\right) = -2\ln\left\{\mathrm{L}\left(\widehat{\boldsymbol{\theta}}|\mathbf{y}\right)\right\}$ is also computed, where $\widehat{\boldsymbol{\theta}}$ is the vector of parameters estimate.

6.1. The low back pain study

The data set was collected from a research study conducted by an institutional review board at Jaseng hospital in Korea from November 2006 to October 2007. The University of North Carolina managed the study. Eligible cases were 127 patients who had not been previously treated for low back pain. The treatment was performed at baseline followed by

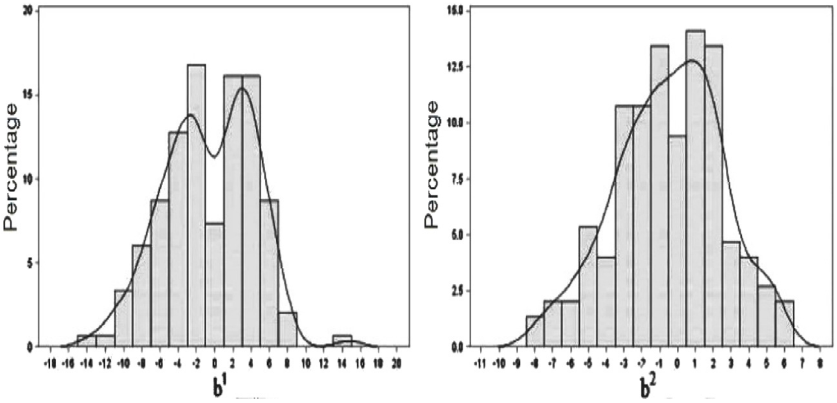

Fig. 4. The low back pain study. Histogram of predicted random intercepts b_i^1 (left) and b_i^2 (right) for the normal model.

weeks 4, 8, 12, 16, 20, and 24. Two responses are the visual analogue scale (VAS) $(0-10)$ of back pain [17] and Oswestry disability index (ODI) [4]. These are influenced by several medical and demographic factors. In our current study, we only consider the influence patients' age, sex, body mass index (BMI), surgery recommendation ($0 =$ recommended and $1 =$ not recommended), baseline values of responses and two covariates mental and physical health that indicate the quality of life.

A preliminary descriptive analysis shows that the number of patients who are in the normal BMI category $(18.5-23)$, overweight (> 23) and underweight (< 18.5) are, respectively, 63(42%), 37(24.7%) and 48(32%). Based on these categories and other unmeasured factors, a hidden classification may exist in the structure of collected data. Thus, without any prior put on the groups structure, we may utilize a multimodal distribution for the random intercepts to reflect the clustering scheme of responses.

The profile plot, not shown here, shows that both ODI and VAS levels increase over time for most patients and substantial inter-patient variation exists. Thus, we first fit two separate univariate normal random-intercepts for each response VAS and ODI with normality assumption of error terms. Then, we observed that the correlation between VAS and ODI at subsequent periods was close to one. Thus we fitted a bivariate mixed-effects model which may significantly be better than fitted separate models. Consider the following random-intercept model

$$
\begin{aligned}
ODI_{ij} &= \mathbf{X}_{ij}\boldsymbol{\beta}^1 + b_i^1 + e_{ij}^1, \\
VAS_{ij} &= \mathbf{X}_{ij}\boldsymbol{\beta}^2 + b_i^2 + e_{ij}^2,
\end{aligned}
\tag{9}
$$

for $i \in \{1, \dots, 127\}$ and $j \in \{1, \dots, 6\}$, where $e_{ij}^k \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_{ek}^2\right)$ for $k \in \{1, 2\}$ and $\mathbf{b}_i = \left(b_i^1, b_i^2\right)^\top \overset{iid}{\sim} \mathcal{N}_2\left(\mathbf{0}, \mathbf{V}_b\right)$. Results of the fitted model show that the correlation between the prediction of random intercepts of two separated models for the ODI and the VAS is close to one (0.83), which may suggest that a model with one shared random intercept should also fit well. Using the DIC value, we compared two fitted models with shared and separated random-intercepts. Findings, however, indicated that the sharing strategy did not substantively improve the data analysis.

Figs. 4 and 5 (left) show the marginal and joint distribution of predicted random intercepts related to VAS and ODI. Histograms demonstrate visually the deviation from normality and clearly multimodal shape of the distribution of each random intercept b_i^1 and b_i^2. Furthermore, the distribution surface and scatter plot of predicted random intercepts, in Fig. 5 (left), show that the bivariate normality of random intercepts is unrealistic and suggest a multimodal structure for the joint distribution of random intercepts.

The above evidence motivates us to examine the ability of our proposed model to the data analysis. Because the multimodal structure of the joint distribution of random intercepts may also suggest fitting a finite mixture model, we fit a bivariate LME model by assuming a finite mixture distribution with normal components for random intercepts. For model comparison, we fit the mixed-effects model (9) by letting the underlying random intercepts be distributed as those previously specified in cases M1–M3.

For the sake of comparison, the non-informative priors are adopted for all fitted models, similar to the simulation study. Results are given in Table 2. Using the values of the model selection criteria shown in the last row of Table 2, model M2 fits better. The Bayes estimate of $\boldsymbol{\delta}$, α, λ and σ in the preferred model are (5.81, 2.46), 1.67, 2.76 and 1.34, respectively, and parameters differ significantly from zero. For model M2 the variance and correlation between random effects can be estimated from $\mathbf{V}_b + \sigma^2 \left\{1 - \lambda^2 \exp\left(-\lambda^2\right)\alpha^2\right\} \boldsymbol{\delta}\boldsymbol{\delta}^\top$. It is seen that the variance components $\sigma_{e^1}^2$ and $\sigma_{e\,2}^2$ in multimodal models, and in particular for M2, are smaller than those in the normal model. Furthermore, Fig. 5 (right) displays the scatter plot of predicted random effects with the superimposed contour plots of the fitted bivariate multimodal normal density and shows that the additional flexibility afforded by the $\mathcal{MMN}$ is sufficient to capture more possible peaks of the random-effects distributions.

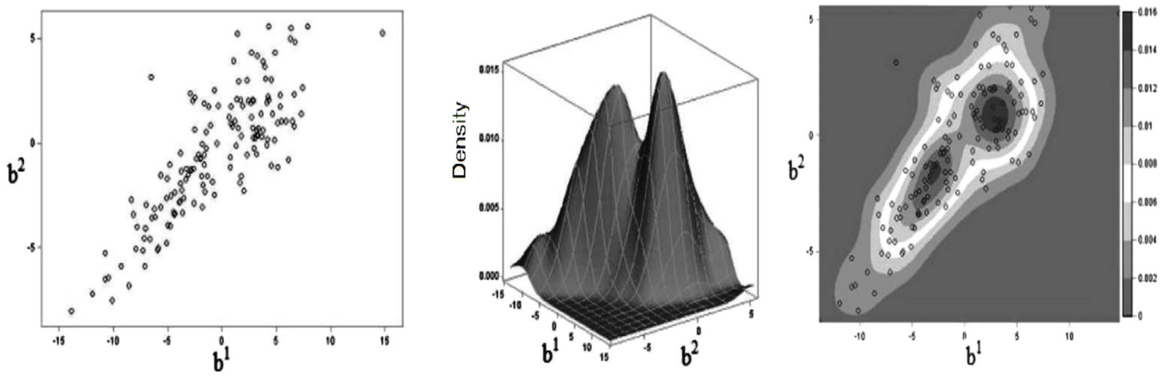

Fig. 5. The scatter plot and surface plot of predicted random intercepts b_i^1 and b_i^2 for the normal model (left and center) in the low back pain study. The super imposed contour plots of the fitted bivariate multimodal normal density (right).

Table 2
Posterior means and 95% confidence intervals of parameters under three models M1–M3 for the low back pain study. M1: Bivariate normal, M2: Bivariate multimodal normal, M3: Finite mixture distribution with 2 components.

Parameters	M1		M2		M3	
	Mean	Probability interval	Mean	Probability interval	Mean	Probability interval
Fixed effects estimates						
Baseline^{ODI}	0.32	(0.23, 0.51)	0.31	(0.29, 0.39)	0.31	(0.19, 0.38)
Female^{ODI}	2.64	(−2.32, 7.65)	3.85	(−0.71, 5.85)	3.5	(0.11, 6.45)
Age^{ODI}	0.32	(0.05, 1.69)	0.26	(0.01, 0.54)	0.31	(−0.04, 0.67)
Body mass index^{ODI}	0.35	(−4.15, 5.02)	−2.68	(−5.96, 2.34)	−1.68	(−7.96, 3.14)
Surgery recommendation^{ODI}	1.18	(0.16, 3.92)	2.28	(1.79, 2, 65)	1.22	(0.09, 2.05)
Physical health^{ODI}	−0.03	(−0.24, 0.01)	−0.12	(−0.14, −0.11)	−0.11	(−0.19, −0.08)
Mental health^{ODI}	−0.15	(−0.29, −0.03)	−0.14	(−0.18, −0.13)	−0.14	(−0.19, −0.13)
Baseline^{VAS}	0.19	(0.01, 0.36)	0.15	(0.12, 0.19)	0.18	(0.11, 0.22)
Female^{VAS}	0.34	(−0.42, 0.88)	1.32	(−0.12, 0.45)	0.92	(−0.22, 0.47)
Age^{VAS}	0.02	(−0.01, 0.08)	0.07	(0.01, 0.09)	0.01	(−0.01, 0.06)
Body mass index^{VAS}	−0.23	(−0.93, 0.47)	−0.27	(−0.43, 0.21)	−0.47	(−0.62, 0.35)
Surgery recommendation^{VAS}	0.32	(0.11, 45)	0.28	(0.22, 0.33)	0.19	(0.14, 0.39)
Physical health^{VAS}	0.1	(−0.12, 0.18)	−0.02	(−0.04, −0.01)	−0.09	(−0.19, −0.01)
Mental health^{VAS}	−0.05	(−0.11, 0.19)	−0.24	(−0.18, −0.27)	−0.25	(−0.19, −0.27)
Estimate of variance and correlation components						
$\sigma^2_{b^1}$	34.23	(10.56, 47.32)	18.67	(12, 67, 24.52)	29.67	(13.35, 37.59)
$\sigma^2_{b^2}$	21.03	(14.54, 35.12)	6.56	(2.23, 11, 78)	16.16	(10.99, 27.12)
$\sigma_{b^{12}}$	16.33	(11.19, 19.49)	7.92	(5.39, 9.41)	13.43	(10.33, 15.57)
$\sigma^2_{e^1}$	7.52	(0.15, 0, 93)	7.54	(5.23, 9.12)	7.55	(3.99, 9.82)
$\sigma^2_{e^2}$	1.89	(0.87, 2.56)	1.87	(1.64, 2.06)	1.88	(1.62, 2.17)
Model selection criteria						
DIC		7879		7776		7854
$D\left(\widehat{\boldsymbol{\theta}}\right)$		5967.5		5888		5910

We report our findings only based on the best-fitted model to show the effect of different covariates on the ODI and VAS. We should mention that the ODI represents an index for indicating the disability of a patient and the VAS as an index for indicating pain severity. Results show that there is no development in the VAS or ODI according to the sex and the body mass index since the associated 95% confidence intervals include 0. A large amount of pain or disability at the beginning of the study without any intervention may be significantly associated with the degree of patient improvement according to the VAS and ODI changes in follow-up values. Furthermore, the relationship between the physical and the mental health is negative and significant due to the 95% confidence interval which includes 0. A similar result observed for the relationship between the ODI and VAS.

6.2. The height of school girls

The data set includes the growth curve of 20 pre-adolescent school girls with height measured on a yearly basis from age 6 to 10. The girls were classified according to the height of their mother into three categories as short, medium and

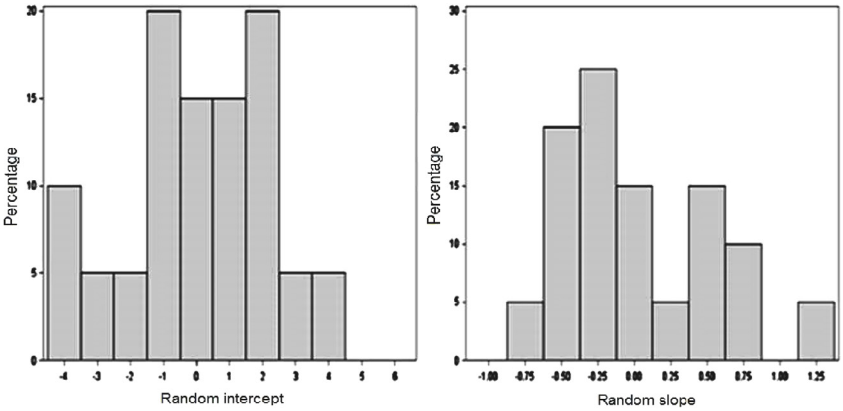

Fig. 6. Histogram of predicted random intercept (left) and random slope (right) for the normal model in the analysis of height of school girls.

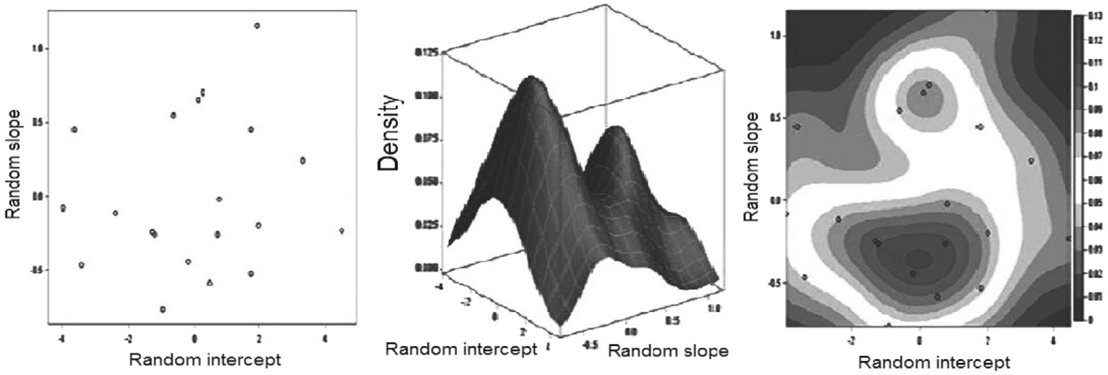

Fig. 7. The scatter plot and surface plot of predicted random intercept and random slope for the normal model (left and center) in the analysis of height of school girls. The super imposed contour plots of the fitted bivariate multimodal normal density (right).

tall mothers. The data have been previously analyzed by [32] who fitted a univariate linear mixed model by assuming that the random effects are sampled from a mixture of $g \in \{1, 2, 3\}$ normal components. Similar results have been also reported in some literature, e.g., [28], to illustrate that the distribution of growth curves is multimodal. As addressed in the literature of mixture models, the choice of number of clusters and mixture components is challenging. Thus, we utilize our proposed methodology as an alternative data analysis to deal with such issue.

Let height$_{ij}$ and age$_{ij}$ be the height and age of the ith ($i \in \{1, \ldots, 20\}$ girl at the time $j \in \{6, \ldots, 10\}$). Consider the linear mixed model

$$height_{ij} = \beta_0 + \beta_1 age_{ij} + b_{0i} + b_{1i} age_{ij} + e_{ij}, \tag{10}$$

where β_0 and β_1 are the overall average intercept and linear age effect, respectively, and b_{0i} and b_{i1} are the random intercept and slope, respectively. The usual linear mixed model assumes that the error terms e_{ij} are independent and normally distributed with mean 0 and variance σ_e^2 and $\mathbf{b}_i = (b_{0i}, b_{1i})^\top \overset{iid}{\sim} \mathcal{N}_2\left(\mathbf{0}, \mathbf{V}_b\right)$.

Prediction of b_{0i} and b_{i1} in fitting the simple mixed model, shown in Figs. 6 and 7 (left and center), illustrates the non-normality structure and multimodal pattern of the marginal and joint distributions of the random effects. Further evidence is illustrated in these Figures on the number of sub-populations for random effects. Thus, a multimodal bivariate distribution seems to be adequate for the random-effects distributions. This motivates us to use the $\mathcal{MMN}$ distribution for b_{0i} and b_{i1}. Using this idea, the girls can be classified into some latent clusters without any recognition of sub-populations.

To make a comparative study, we fit (10) by imposing the bivariate distributions specified in cases M1–M3 together with the following distribution

M4: $\sum_{j=1}^{3} \pi_j \phi_2\left(\boldsymbol{\mu}_j, \mathbf{V}_b\right)$, $\sum_{j=1}^{3} \pi_j = 1$, $\sum_{j=1}^{3} \pi_j \boldsymbol{\mu}_j = 0$.

A number of 10 000 samples was generated after discarding 3000 burn-in samples, and the lag value was fixed to 6 to avoid the correlation issue in the generated chains [5] . Results, after the convergence of the algorithm are given in Table 3.

Table 3
Posterior mean and 95% confidence interval of parameters under four models M1 to M4 for the height of school girls data. M1: Bivariate normal, M2: Bivariate multimodal normal, M3: Mixture distribution with 2 components, M4: Mixture distribution with 3 components.

Parameters	M1		M2		M3		M4	
	Mean	Probability interval	Mean	Probability interval	Mean	Probability interval	Mean	Probability interval
Fixed effects estimates								
β_0	79.78	(64.54, 85.45)	80.97	(77.79, 82.56)	75.18	(68.54, 79.15)	74.62	(63.54, 78.45)
β_1	3.63	(1.28, 8.72)	4.92	(2.41, 6.28)	4.27	(2.72, 8.28)	5.23	(1.32, 7.28)
Estimate of variance and correlation components								
σ_e^2	0.47	(0.06, 0.69)	0.46	(0.19, 0.56)	0.47	(0.24, 0.72)	0.47	(0.22, 0.77)
$\sigma_{b_0}^2$	7.21	(0.12, 10.63)	2.53	(1.19, 5.49)	6.74	(2.32, 6.54)	6.38	(2.89, 6.65)
$\sigma_{b_1}^2$	0.59	(0.06, 6.01)	0.18	(0.09, 2.99)	0.37	(0.18, 1.98)	0.32	(0.24, 1.77)
$\sigma_{b_{01}}$	0.29	(0.09, 0.79)	0.16	(0.10, 0.58)	0.12	(0.01, 0.82)	0.34	(0.09, 0.89)
Model selection criteria								
DIC		178.6		175.8		176.1		176.9
$D\left(\widehat{\boldsymbol{\theta}}\right)$		135		121		128		127

It is seen that the age is statistically significant in all models since the associated 95% confidence intervals do not include 0. The Bayes estimate of standard deviation of age, for the normal model are larger than those model assuming multimodality, and are smaller in our proposed model. The same is true for the estimate of random-effects variances. Also, $\widehat{\boldsymbol{\delta}} = (1.45, 2, 45)$ and $\left(\widehat{\alpha}, \widehat{\lambda}, \widehat{\sigma}\right) = (1.54, 3.61, 1.44)$ are significant. Results, shown in Table 3, illustrate that the proposed model performs satisfactory to this data analysis. Based on the model selection criteria $\mathcal{MMN}$ is the best fitted model, while the mixture model with three components is the second-best model.

Another advantage of estimating the random effects density by $\mathcal{MMN}$ distribution is appreciated by Fig. 7 (right) that displays the scatter plot of predicted effects with the super imposed contour plots of the fitted bivariate multimodal normal density. This figure demonstrates that the additional flexibility afforded by the $\mathcal{MMN}$ is sufficient to capture accurately the multimodal underlying feature of the random effects.

7. Concluding remarks and discussion

Multivariate linear mixed-effects models provide flexible tools for the analysis of multiple correlated response variables. In this paper, we proposed a new modeling strategy based on utilizing the $\mathcal{MMN}$ distribution to handle the joint analysis of repeatedly measured responses over time-periods with multimodal structures. Our strategy was initiated by combining separate LME models for each response together in a single model through imposing a $\mathcal{MMN}$ distribution on the random effects associated with the responses.

The modeling methodology can produce greater flexibility in designing mixed-effects models and analyzing real empirical data. It is useful when several peaks exist on both joint and marginal distributions of responses but are not directly identifiable. The proposed model is convenient for analyzing heterogeneous data when several sub-populations exist. It is an attractive alternative to multivariate mixture modeling settings. The reason is that our model, analogous to the finite mixture model, can cover most hidden peaks in the distribution of responses but with a few numbers of parameters. Moreover, it is unnecessary to employ any selection method of detecting the number of mixture components.

We argue that although experimental results show that our model performs better than other competitive models, a further comparative study is required to highlight its strengths and weaknesses in comparison with finite mixture models. This is the aim of our future study that will specifically focus on real-data analysis situations wherein common normality assumption of mixture components can possibly be violated. Moreover, an extension of the modeling strategy is required to research studies with aims concentrated on classification, clustering, and discrimination of population.

As illustrated with the paper, the associated likelihood function was complex to carry out statistical inference. The analyst may use the Gauss–Hermite quadrature approach or the Laplace approximation to approximate integral equations involved in the marginal likelihood. To our best experiences, the computation is somehow time-consuming and estimation results are disappointing for slightly large dimensional integrals. Thus, for the computational convenience, we provided a hierarchical representation to facilitate the implementing of iterative estimation approaches, such as the ECM algorithm and the MCMC method in the Bayesian perspective. As extensively discussed in the literature, the MCMC scheme is more effective and direct to use in practice. The estimation procedure can be easily proceeding by accessible software packages, such as OpenBugs, or JAGS in R.

Another interesting topic for our future research is to utilize other families of multivariate distributions to allow for jointly accommodating multimodality, asymmetry, and heavy tails in longitudinal studies.

Acknowledgments

Authors are grateful to two anonymous reviewers and Editor-in-Chief for their constructive comments. The first author is also grateful to the Graduate Office of the University of Isfahan for the support.

Appendix A. A useful lemma and propositions

Lemma 1. *Let* $\mathbf{Y}|\mathbf{X} = \mathbf{x} \sim \mathcal{N}_p\left(\boldsymbol{\mu} + \mathbf{A}\mathbf{x}, \mathbf{V}\right)$ *and* $\mathbf{X} \sim \mathcal{N}_q\left(\boldsymbol{\eta}, \boldsymbol{\Omega}\right)$. *Then*

$$\phi_p\left(\mathbf{y}|\boldsymbol{\mu} + \mathbf{A}\mathbf{x}, \mathbf{V}\right)\phi_q\left(\mathbf{x}|\boldsymbol{\eta}, \boldsymbol{\Omega}\right) = \phi_p\left(\mathbf{y}|\boldsymbol{\mu} + \boldsymbol{A\eta}, \boldsymbol{V} + \boldsymbol{A\Omega A}^\top\right)\phi_q\left(\mathbf{x}|\boldsymbol{\eta} + \boldsymbol{\Lambda A}^\top\mathbf{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu} - \boldsymbol{A\eta}\right), \boldsymbol{\Lambda}\right)$$

where $\boldsymbol{\Lambda} = \left(\boldsymbol{\Omega}^{-1} + \mathbf{A}^\top\mathbf{V}^{-1}\mathbf{A}\right)^{-1}$.

Proof. See [2]. □

Proposition 6. *The function* $f(x) = \phi\left(x|\mu_1, \sigma_1^2\right)\sin\left\{\lambda\left(x - \mu_2\right)/\sigma_2\right\}$ *for* $\sigma_1, \sigma_2 > 0$ *is integrable in* $\mathbb{R}$ *and*

$$\int_{\mathbb{R}} f(x)\,dx = \exp\left(\frac{\lambda^2\sigma_1^2}{2\sigma_2^2}\right)\sin\left\{\frac{\lambda\left(\mu_1 - \mu_2\right)}{\sigma_2}\right\}.$$

Proof. The integral can be computed using mathematical relations, or, in a computational software, such as Maple or Mathematica. □

Proposition 7. *Consider* $\mathbf{Y}_i$ *and other quantities as in Expression (4). The conditional density of* $W_i|\boldsymbol{b}_i, \mathbf{Y}_i$ *is given by*

$$f\left(w_i|\boldsymbol{b}_i, \mathbf{y}_i\right) = \phi\left(w_i|\mu_{w_i}, S_{w_i}\right)\left[1 + \frac{1}{\alpha}\sin\left\{\lambda\left(w_i - \mu\right)/\sigma\right\}\right]$$

where $S_{w_i} = \left(\boldsymbol{\delta}^\top\mathbf{V}_b^{-1}\,\boldsymbol{\delta} + \sigma^{-2}\right)^{-1}$ *and* $\mu_{w_i} = \mu + S_{w_i}\boldsymbol{\delta}^\top\mathbf{V}_b^{-1}\left(\mathbf{b}_i - \mu\boldsymbol{\delta}\right)$.

Proof. The proof is conducted after some simple algebraic manipulations and using Lemma 1 for deriving the joint distribution

$$f\left(\mathbf{y}_i, \mathbf{b}_i, w_i|\boldsymbol{\theta}\right) = \phi_{rn_i}\left(\mathbf{Y}_i|\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{V}_e\right)\phi_{rq}\left(\mathbf{b}_i|w_i\boldsymbol{\delta}, \mathbf{V}_b\right)\phi\left(w_i|\mu, \sigma^2\right)\left[1 + \frac{1}{\alpha}\sin\left\{\lambda\left(w_i - \mu\right)/\sigma\right\}\right]. \quad \square$$

Appendix B. Proof of Proposition 5

From Expression (4) and Definition 1 together with using Lemma 1 in Appendix A, the marginal density of $\mathbf{Y}_i$ is obtained by computing the following integral

$$\begin{aligned} f\left(\mathbf{y}_i|\boldsymbol{\theta}\right) &= \int_{\mathbb{R}^{rq}}\int_{\mathbb{R}} f\left(\mathbf{y}_i|\mathbf{b}_i, \theta\right) f\left(\mathbf{b}_i|w_i, \boldsymbol{\theta}\right) f\left(w_i|\boldsymbol{\theta}\right) dw_i d\mathbf{b}_i \\ &= \int_{\mathbb{R}^{rq}} \phi_{rn_i}\left(\mathbf{Y}_i|\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{V}_e\right)\int_{\mathbb{R}} \phi_{rq}\left(\mathbf{b}_i|w_i\boldsymbol{\delta}, \mathbf{V}_b\right)\phi\left(w_i|\mu, \sigma^2\right)\left[1 + \frac{1}{\alpha}\sin\left\{\lambda\left(w_i - \mu\right)/\sigma\right\}\right] dw_i d\mathbf{b}_i \\ &= \int_{\mathbb{R}^{rq}} \phi_{rn_i}\left(\mathbf{Y}_i|\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{V}_e\right)\phi_{rq}\left(\mathbf{b}_i|\mu\boldsymbol{\delta}, \mathbf{V}_b + \sigma^2\boldsymbol{\delta}\boldsymbol{\delta}^\top\right)\int_{\mathbb{R}} \phi\left(w_i|\mu_w, \sigma_w^2\right)\left[1 + \frac{1}{\alpha}\sin\left\{\lambda\left(w_i - \mu\right)/\sigma\right\}\right] dw_i d\mathbf{b}_i, \end{aligned}$$

where $\mu_w = \mu + \sigma_w^2\boldsymbol{\delta}^\top\mathbf{V}_b^{-1}\left(\mathbf{b}_i - \mu\boldsymbol{\delta}\right)$ and $\sigma_w^2 = \left(\sigma^{-2} + \boldsymbol{\delta}^\top\mathbf{V}_b^{-1}\boldsymbol{\delta}\right)^{-1}$. Using Proposition 6 in Appendix A and replacing μ_w and σ_w^2 we have

$$\int_{\mathbb{R}} \phi\left(w_i|\mu_w, \sigma_w^2\right)\sin\left\{\lambda\left(w_i - \mu\right)/\sigma\right\} dw_i = \exp\left(\frac{\lambda^2\sigma_w^2}{2\sigma^2}\right)\sin\left\{\boldsymbol{\Delta}_b\left(\mathbf{b}_i - \mu\boldsymbol{\delta}\right)/\sigma\right\},$$

where $\boldsymbol{\delta}_b = \lambda\sigma_w^2\sigma^{-1}\boldsymbol{\delta}^\top\mathbf{V}_b^{-1}$. Then, by using Lemma 1 in Appendix A

$$f\left(\mathbf{y}_i|\boldsymbol{\theta}\right) = \phi_{rn_i}\left(\mathbf{Y}_i|\boldsymbol{\mu}_{yi}, \mathbf{V}_{yi}\right)\int_{\mathbb{R}^{rq}} \phi_{rq}\left(\mathbf{b}_i|\boldsymbol{\mu}_{bi}, \boldsymbol{\Lambda}_{bi}\right)\left[1 + \frac{1}{\alpha_y}\sin\left\{\boldsymbol{\delta}_b\left(\mathbf{b}_i - \mu\boldsymbol{\delta}\right)\right\}\right] d\mathbf{b}_i,$$

with $\alpha_y = \alpha\exp\left\{-\lambda^2/2\left(1 + \sigma^2\boldsymbol{\delta}^\top\mathbf{V}_b^{-1}\boldsymbol{\delta}\right)\right\}$ and

$$\begin{aligned} \boldsymbol{\mu}_{yi} &= \mathbf{X}_i\boldsymbol{\beta} + \mu\mathbf{Z}_i\boldsymbol{\delta}, & \boldsymbol{\mu}_{bi} &= \mu\boldsymbol{\delta} + \boldsymbol{\Lambda}_{bi}\mathbf{Z}_i^\top\mathbf{V}_e^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\,\boldsymbol{\beta} - \mu\mathbf{Z}_\mathbf{i}\boldsymbol{\delta}\right), \\ \mathbf{V}_{yi} &= \mathbf{V}_e + \mathbf{Z}_\mathbf{i}\left(\mathbf{V}_b + \sigma^2\boldsymbol{\delta}\boldsymbol{\delta}^\top\right)\mathbf{Z}_\mathbf{i}^\top, & \boldsymbol{\Lambda}_{bi} &= \left\{\left(\mathbf{V}_b + \sigma^2\boldsymbol{\delta}\boldsymbol{\delta}^\top\right)^{-1} + \mathbf{Z}_i^\top\mathbf{V}_e^{-1}\mathbf{Z}_i\right\}^{-1}. \end{aligned}$$

Appendix C. The ECM algorithm

To implement the EM algorithm, we set $\mathbf{y}_i$ as the observed data, $\mathbf{y}_{miss,i} = (\mathbf{b}_i, w_i)$ as the missing data, and $\mathbf{y}_{C,i} = (\mathbf{y}_{miss,i}, \mathbf{y}_i)$ as the complete data on the ith subject. It follows from (4) that the log-likelihood function of complete data, ignoring the terms that are free of $\boldsymbol{\theta}$, is of the form $\ell_c(\boldsymbol{\theta}|\mathbf{y}_C) = -1/2 \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta})$, where

$$\ell_i(\boldsymbol{\theta}) = \ln|\mathbf{V}_e| + (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta} - \mathbf{z}_i\mathbf{b}_i)^\top \mathbf{V}_e^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta} - \mathbf{z}_i\mathbf{b}_i) + \ln|\mathbf{V}_b| + (\mathbf{b}_i - \boldsymbol{\delta} w_i)^\top \mathbf{V}_b^{-1}(\mathbf{b}_i - \boldsymbol{\delta} w_i) + h(w_i), \tag{C.1}$$

with $h(w_i) = \ln(\sigma^2) + (w_i - \mu)^2/\sigma^2 - 2\ln\left[1 + \alpha^{-1}\sin\{\lambda(w_i - \mu)/\sigma\}\right]$ and $\mu = \sigma\lambda\alpha^{-1}\exp(-\lambda^2/2)$.

The expected complete-data log-likelihood, in the $(h+1)$th iteration, given the current estimate $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(h)}$ in the E-step can be written as

$$Q\left(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(h)}\right) = \mathrm{E}\left\{\ell_c(\boldsymbol{\theta}|\mathbf{y}_C)|\mathbf{y}, \widehat{\boldsymbol{\theta}}^{(h)}\right\} = -\frac{1}{2}\sum_{i=1}^{N}\left\{Q_{1i}\left(\boldsymbol{\beta}, \mathbf{V}_e|\widehat{\boldsymbol{\theta}}^{(h)}\right) + Q_{2i}\left(\boldsymbol{\delta}, \mathbf{V}_b|\widehat{\boldsymbol{\theta}}^{(h)}\right) + Q_{3i}\left(\alpha, \lambda, \sigma|\widehat{\boldsymbol{\theta}}^{(h)}\right)\right\},$$

with

$$\begin{aligned}
Q_{1i}\left(\boldsymbol{\beta}, \mathbf{V}_e|\widehat{\boldsymbol{\theta}}^{(h)}\right) &= \ln\left|\widehat{\mathbf{V}}_e^{(h)}\right| + \left(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^{(h)}\right)^\top \left(\widehat{\mathbf{V}}_e^{(h)}\right)^{-1}\left(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^{(h)}\right) \\
&\quad - \mathbf{2}\left(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^{(h)}\right)^\top \left(\widehat{\mathbf{V}}_e^{(h)}\right)^{-1}\mathbf{z}_i\widehat{\mathbf{b}}_i^{(h)} + \mathrm{trace}\left\{\left(\widehat{\mathbf{V}}_e^{(h)}\right)^{-1}\mathbf{z}_i\widehat{\mathbf{b}_i\mathbf{b}_i^\top}^{(h)}\mathbf{z}_i^\top\right\}, \\
Q_{2i}\left(\boldsymbol{\delta}, \mathbf{V}_b|\widehat{\boldsymbol{\theta}}^{(h)}\right) &= \ln\left|\widehat{\mathbf{V}}_b^{(h)}\right| + \mathrm{trace}\left\{\left(\widehat{\mathbf{V}}_b^{(h)}\right)^{-1}\left(\widehat{\mathbf{b}_i\mathbf{b}_i^\top}^{(h)} - 2\widehat{w_i\mathbf{b}_i}^{(h)}\widehat{\boldsymbol{\delta}}^{\top(h)} + \widehat{w_i^2}^{(h)}\widehat{\boldsymbol{\delta}}^{(h)}\widehat{\boldsymbol{\delta}}^{\top(h)}\right)\right\}, \\
Q_{3i}\left(\alpha, \lambda, \sigma_w|\widehat{\boldsymbol{\theta}}^{(h)}\right) &= \ln\left\{\widehat{\sigma}^{2(h)}\right\} + \frac{1}{\widehat{\sigma}^{2(h)}}\left(\widehat{w_i^2}^{(h)} - 2\widehat{\mu}^{(h)}\widehat{w}_i^{(h)} + \widehat{\mu}^{2(h)}\right)^2 - 2\widehat{lw}_i^{(h)},
\end{aligned}$$

where trace{.} indicates the trace of a matrix and $\widehat{\mu}^{(h)}$ must be replaced by $\widehat{\sigma}^{(h)}\widehat{\lambda}^{(h)}/\widehat{\alpha}^{(h)}\exp\left(-\widehat{\lambda}^{(h)2}/2\right)$. The calculation of these functions require expressions for $\widehat{\mathbf{b}_i\mathbf{b}_i^\top}^{(h)} = \mathrm{E}\left(\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(h)}\right)$, $\widehat{\mathbf{b}}_i^{(h)} = \mathrm{E}\left(\mathbf{b}_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(h)}\right)$, $\widehat{w_i\mathbf{b}_i}^{(h)} = \mathrm{E}\left(w_i\mathbf{b}_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(h)}\right)$, $\widehat{w_i^2}^{(h)} = \mathrm{E}\left(w_i^2|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(h)}\right)$ and $\widehat{lw}_i^{(h)} = \mathrm{E}\left[\ln\left\{1 + \alpha^{-1}\sin(\lambda(w_i - \mu)/\sigma)\right)|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(h\}}\right]$ which can be evaluated from (C.1) and (4), after some algebraic manipulation as following.

From Expression (4), we have

$$f(\mathbf{y}_i, \mathbf{b}_i, w_i|\boldsymbol{\theta}) = f(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\theta})f(\mathbf{b}_i|w_i, \boldsymbol{\theta})f(w_i|\boldsymbol{\theta}).$$

Using successively Lemma 1 in Appendix A it is noted that

$$\mathbf{b}_i|W_i = w_i, \mathbf{Y}_i = \mathbf{y}_i \overset{ind}{\sim} \mathcal{N}_{rq}\left(\boldsymbol{\Omega}_i^{-1}\left\{\mathbf{z}_i^\top\mathbf{V}_e^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}) + w_i\mathbf{V}_b^{-1}\boldsymbol{\delta}\right\}, \boldsymbol{\Omega}_i\right),$$

where $\boldsymbol{\Omega}_i = \left(\mathbf{z}_i^\top\mathbf{V}_e^{-1}\mathbf{z}_i + \mathbf{V}_b^{-1}\right)^{-1}$. Now, we use the law of iterated expectations and the necessary conditional expectations to show that

$$\begin{aligned}
\mathrm{E}(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta}) &= \mathrm{E}\{\mathrm{E}(\mathbf{b}_i|\mathbf{y}_i, w_i, \boldsymbol{\theta})\} = \boldsymbol{\Omega}_i^{-1}\mathbf{z}_i^\top\mathbf{V}_e^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}) + \widehat{w}_i\mathbf{V}_b^{-1}\boldsymbol{\delta}, \\
\mathrm{E}\left(\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{y}_i, \boldsymbol{\theta}\right) &= \mathrm{E}\left\{\mathrm{E}\left(\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{y}_i, w_i, \boldsymbol{\theta}\right)\right\}, \\
\mathrm{E}\left(\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{y}_i, w_i, \boldsymbol{\theta}\right) &= \mathrm{cov}(\mathbf{b}_i|\mathbf{y}_i, w_i, \boldsymbol{\theta}) + \mathrm{E}(\mathbf{b}_i|\mathbf{y}_i, w_i, \boldsymbol{\theta})\,\mathrm{E}\left(\mathbf{b}_i^\top|\mathbf{y}_i, w_i, \boldsymbol{\theta}\right), \\
\mathrm{E}(w_i\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta}) &= \mathrm{E}\{w_i\mathrm{E}(\mathbf{b}_i|\mathbf{y}_i, w_i, \boldsymbol{\theta})\} \\
&= \widehat{w}_i\boldsymbol{\Omega}^{-1}\mathbf{z}_i^\top\mathbf{V}_e^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}) + \widehat{w_i^2}\boldsymbol{\Omega}_i^{-1}\mathbf{V}_b^{-1}\boldsymbol{\delta}.
\end{aligned}$$

Moreover

$$\widehat{\mathbf{b}}_i^{(h)} = \widehat{\boldsymbol{\Omega}}_i^{(h)}\left\{\widehat{w}_i^{(h)}\left(\widehat{\mathbf{V}}_b^{(h)}\right)^{-1}\widehat{\boldsymbol{\delta}}^{(h)} + \mathbf{z}_i^\top\left(\widehat{\mathbf{V}}_e^{(h)}\right)^{-1}\left(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^{(h)}\right)\right\},$$

with $\widehat{\boldsymbol{\Omega}}_i^{(h)} = \mathrm{cov}\left(\mathbf{b}_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(h)}\right) = \left\{\left(\widehat{\mathbf{V}}_b^{(h)}\right)^{-1} + \mathbf{z}_i^\top\left(\widehat{\mathbf{V}}_e^{(h)}\right)^{-1}\mathbf{z}_i\right\}^{-1}$ and $\widehat{w}_i^{(h)} = \mathrm{E}\left(w_i|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}^{(h)}\right)$,

$$\widehat{\mathbf{b}_i\mathbf{b}_i^\top}^{(h)} = \widehat{\boldsymbol{\Omega}}_i^{(h)} + \left(\widehat{\boldsymbol{\Omega}}_i^{(h)}\right)^{-1}\widehat{\boldsymbol{\Lambda}}_i^{(h)}\left(\widehat{\boldsymbol{\Omega}}_i^{(h)}\right)^{-1},$$

where

$$\begin{aligned}
\widehat{\boldsymbol{\Lambda}}_i^{(h)} &= \mathbf{z}_i^\top\left(\widehat{\mathbf{V}}_e^{(h)}\right)^{-1}\left(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^{(h)}\right)\left(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^{(h)}\right)^\top\left(\widehat{\mathbf{V}}_e^{(h)}\right)^{-1}\mathbf{z}_i \\
&\quad + \widehat{w_i^2}^{(h)}\left(\widehat{\mathbf{V}}_b^{(h)}\right)^{-1}\widehat{\boldsymbol{\delta}}^{(h)}\widehat{\boldsymbol{\delta}}^{\top(h)}\left(\widehat{\mathbf{V}}_b^{(h)}\right)^{-1} - 2\widehat{w}_i^{(h)}\mathbf{z}_i^\top\left(\widehat{\mathbf{V}}_b^{(h)}\right)^{-1}\left(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^{(h)}\right)\widehat{\boldsymbol{\delta}}^{\top(h)}\left(\widehat{\mathbf{V}}_b^{(h)}\right)^{-1},
\end{aligned}$$

and

$$\widehat{w_i\mathbf{b}_i}^{(h)} = \widehat{w}_i^{(h)} \left(\widehat{\boldsymbol{\Omega}}_i^{(h)}\right)^{-1} \mathbf{z}_i^\top \left(\widehat{\mathbf{V}}_e^{(h)}\right)^{-1} \left(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^{(h)}\right) + \widehat{w_i^2}^{(h)} \left(\widehat{\boldsymbol{\Omega}}_i^{(h)}\right)^{-1} \left(\widehat{\mathbf{V}}_b^{(h)}\right)^{-1} \widehat{\boldsymbol{\delta}}^{(h)}.$$

To compute $\widehat{lw_i}^{(h)}$, $\widehat{w}_i^{(h)}$ and $\widehat{w_i^2}^{(h)}$, first we need to compute the expectation of $\ln\left[1+\alpha^{-1}\sin\left\{\lambda\left(w_i-\mu\right)/\sigma\right\}\right]$, w_i and w_i^2 conditioned on $(\mathbf{y}_i, \mathbf{b}_i)$, using the density $f\left(w_i|\mathbf{y}_i, \mathbf{b}_i\right)$ (Proposition 7 in Appendix A) and then average it over the conditional distribution of $\mathbf{b}_i|\mathbf{y}_i$. The one-dimensional integral involved in this averaging required to be computed numerically. All expectations are evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(h)}$.

Each iteration of the EM algorithm increases the likelihood function $l_c\left(\boldsymbol{\theta}|\mathbf{y}_C\right)$ and the algorithm typically converges to a local or global maximum of the likelihood function. When the M-step is difficult to implement, it is useful to replace the maximization of $Q\left(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(h)}\right)$ by several simpler conditional maximization (CM) steps. The sequence of the CM-steps is such that the overall maximization is over the full parameter space. This leads to a simple extension of the EM algorithm, called the ECM algorithm [25]. In this work, we implemented the ECM algorithm to obtain the ML estimate of $\boldsymbol{\theta}$ as follows:

E-step: Given $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(h)}$, compute $\widehat{\mathbf{b}}_i^{(h)}$, $\widehat{\mathbf{b}_i\mathbf{b}_i^\top}^{(h)}$, $\widehat{w_i\mathbf{b}_i}^{(h)}$, $\widehat{w}_i^{(h)}$ and $\widehat{w_i^2}^{(h)}$ for $i \in \{1, \ldots, N\}$.
CM-steps: Conditionally maximize $Q\left(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(h)}\right)$ over $\boldsymbol{\theta}$, which leads to the following new estimates $\widehat{\boldsymbol{\theta}}^{(h)}$.

CM-step 1: Fix $\mathbf{V}_e = \widehat{\mathbf{V}}_e^{(h)}$ and update $\widehat{\boldsymbol{\beta}}^{(h)}$ by maximizing $\sum_{i=1}^N Q_{1i}\left(\boldsymbol{\beta}, \mathbf{V}_e|\widehat{\boldsymbol{\theta}}^{(h)}\right)$ over $\boldsymbol{\beta}$, yielding

$$\widehat{\boldsymbol{\beta}}^{(h+1)} = \left\{\sum_{i=1}^N \mathbf{x}_i^\top \left(\widehat{\mathbf{V}}_e^{(h)}\right)^{-1} \mathbf{x}_i\right\}^{-1} \sum_{i=1}^N \mathbf{x}_i^\top \left(\widehat{\mathbf{V}}_e^{(h)}\right)^{-1} \left(\mathbf{y}_i - \mathbf{z}_i\widehat{\mathbf{b}}_i^{(h)}\right).$$

CM-step 2: Fix $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(h+1)}$ and update $\widehat{\mathbf{V}}_e^{(h)}$ by numerically maximizing $\sum_{i=1}^N Q_{1i}\left(\widehat{\boldsymbol{\beta}}^{(h+1)}, \mathbf{V}_e|\widehat{\boldsymbol{\theta}}\right)$ over $\mathbf{V}_e$. Under the common situation $\mathbf{V}_e = \sigma_e^2\mathbf{I}_{rn}$, this step is reduced to get the following closed form for the estimation of σ_e^2,

$$\widehat{\sigma}_e^{2(h+1)} = \frac{1}{rn}\sum_{i=1}^N \left(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^{(h+1)}\right)^\top \left(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^{(h+1)}\right) - 2\left(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^{(h+1)}\right)^\top \mathbf{z}_i\widehat{\mathbf{b}}_i^{(h)} + \mathrm{trace}\left(\mathbf{z}_i\widehat{\mathbf{b}_i\mathbf{b}_i^\top}^{(h)}\mathbf{z}_i^\top\right).$$

CM-step 3: Update $\widehat{\boldsymbol{\delta}}^{(h)}$ by maximizing $\sum_{i=1}^N Q_{2i}\left(\boldsymbol{\delta}, \mathbf{V}_b|\widehat{\boldsymbol{\theta}}^{(h)}\right)$ over $\boldsymbol{\delta}$, yielding $\widehat{\boldsymbol{\delta}}^{(h+1)} = \sum_{i=1}^N \widehat{w_i\mathbf{b}_i}^{(h)} / \sum_{i=1}^N \widehat{w_i^2}^{(h)}$.

CM-step 4: Fix $\boldsymbol{\delta} = \widehat{\boldsymbol{\delta}}^{(h+1)}$ and update $\widehat{\mathbf{V}}_b$ by numerically maximizing $\sum_{i=1}^N Q_{2i}\left(\widehat{\boldsymbol{\delta}}^{(h+1)}, \mathbf{V}_b|\widehat{\boldsymbol{\theta}}^{(h)}\right)$ over $\mathbf{V}_b$. Under the unstructured situation for $\mathbf{V}_b$, the distinct elements $\widehat{\mathbf{V}}_b^{(h+1)}$get the following closed form,

$$\widehat{\mathbf{V}}_b^{(h+1)} = \frac{1}{N}\sum_{i=1}^N \left(\widehat{\mathbf{b}_i\mathbf{b}_i^\top}^{(h)} - 2\widehat{w_i\mathbf{b}_i}^{(h)}\widehat{\boldsymbol{\delta}}^{\top(h+1)} + \widehat{w_i^2}^{(h)}\widehat{\boldsymbol{\delta}}^{(h+1)}\widehat{\boldsymbol{\delta}}^{\top(h+1)}\right).$$

CM-step 5: Update $\widehat{\alpha}$, $\widehat{\lambda}$, $\widehat{\sigma}$ by numerically maximizing $\sum_{i=1}^N Q_{3i}\left(\alpha, \lambda, \sigma\,|\widehat{\boldsymbol{\theta}}^{(h+1)}\right)$ over α, λ, σ to get $\widehat{\alpha}^{(h+1)}$, $\widehat{\lambda}^{(h+1)}$ and $\widehat{\sigma}^{(h+1)}$.

Note that, in the last CM-step a numerical maximization is needed. This can be easily implemented using statistical software packages, such as `optim` routine in R software. Also, the starting values are often corresponding to the estimates under the normal assumption. Moreover, as mentioned in the related literature, one needs to run the ECM-algorithm with several starting values to ensure the convergence of algorithm to a nearly global optimum. In practice, a standard strategy is to compute the associated log-likelihood functions and select the one with the largest value.

Appendix D. The multimodal feature of the marginal distribution

Consider the following hierarchical random-intercept model

$$\begin{aligned} \mathbf{y}|\mathbf{b} &\sim \mathcal{N}_2\left(\boldsymbol{\beta}_0 + \mathbf{b}, \mathbf{V}_e\right), \\ \mathbf{b}|w &\sim \mathcal{N}_2\left(w\boldsymbol{\delta}, \mathbf{V}_b\right), \\ w &\sim \mathcal{MN}\left(\mu, \alpha, \lambda, \sigma\right), \end{aligned}$$

where $\mathbf{y} = \left(y^1, y^2\right)^\top$, $\boldsymbol{\beta}_0 = \left(\beta_0^1, \beta_0^2\right)^\top$, $\mathbf{b} = \left(b^1, b^2\right)^\top$, $\boldsymbol{\delta} = \left(\delta^1, \delta^2\right)^\top$, $\mathbf{V}_e = \mathbf{I}_2$, $\mu = \sigma\lambda\alpha^{-1}\exp\left(-\lambda^2/2\right)$ and $\mathbf{V}_b$ with components $\{1, \rho, 1\}$. Then using Maple software, we plot

$$f\left(\mathbf{y}\right) = \phi_2\left(\mathbf{y}|\boldsymbol{\mu}_y, \mathbf{V}_y\right)\int_{\mathbb{R}^2}\phi_2\left(\mathbf{b}|\boldsymbol{\mu}_b, \boldsymbol{\Lambda}_b\right)\left[1 + \frac{1}{\alpha_y}\sin\left\{\boldsymbol{\delta}_b\left(\mathbf{b} - \mu\boldsymbol{\delta}\right)\right\}\right] db^1\, db^2, \tag{D.1}$$

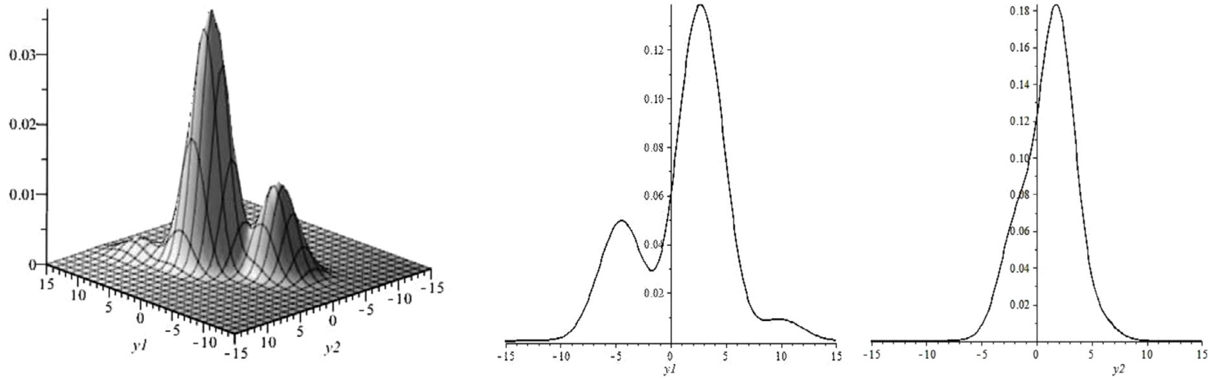

Fig. D.8. The density plot of (5) (left), the marginal density plot of y^1 (center), and the marginal density plot of y^2(right).

where $\boldsymbol{\mu}_y = \boldsymbol{\beta}_0 + \mu\boldsymbol{\delta}$, $\mathbf{V}_y = \mathbf{I}_2 + \left(\mathbf{V}_b + \sigma^2\boldsymbol{\delta}\boldsymbol{\delta}^\top\right)$, $\boldsymbol{\mu}_b = \mu\boldsymbol{\delta} + \boldsymbol{\Lambda}_b\mathbf{V}_e^{-1}\left(\mathbf{y} - \boldsymbol{\beta}_0 - \mu\boldsymbol{\delta}\right)$, $\boldsymbol{\Lambda}_b = \left\{\left(\mathbf{V}_b + \sigma^2\boldsymbol{\delta}\boldsymbol{\delta}^\top\right)^{-1} + \mathbf{I}_2\right\}^{-1}$, $\boldsymbol{\delta}_b = \lambda\sigma\left(1 + \sigma^2\boldsymbol{\delta}^\top\mathbf{V}_b^{-1}\boldsymbol{\delta}\right)^{-1}\boldsymbol{\delta}^\top\mathbf{V}_b^{-1}$ and $\alpha_y = \alpha\exp\left\{-\lambda^2/2\left(1 + \sigma^2\boldsymbol{\delta}^\top\mathbf{V}_b^{-1}\boldsymbol{\delta}\right)\right\}$.

We also plot two marginal density $f\left(y^1\right) = \int_{\mathbb{R}} f\left(\mathbf{y}\right) dy^2$ and $f\left(y^2\right) = \int_{\mathbb{R}} f\left(\mathbf{y}\right) dy^1$. Fig. D.8 indicates the multimodal feature of density functions for $\boldsymbol{\beta}_0 = (1, 1)^\top$, $\boldsymbol{\delta} = (4, 2)^\top$, $\lambda = 3$, $\alpha = 1.5$, $\sigma = 1$ and $\mathbf{V}_b = \mathbf{I}_2$.

Appendix E. BUGS code for model M2 in the simulation study

```
model{for(i in 1:Ntot) {y1[i] ~ dnorm(mu.e[i,1] , tau.e[1])
mu.e[i,1] <- beta[1]+ beta[2]* x1[i]+ beta[3]* x2[i]+ b[sub[i],1]
y2[i] ~ dnorm(mu.e[i,2] , tau.e[2])
mu.e[i,2] <- beta[4]+ beta[5]* x1[i]+ beta[6]* x2[i]+ b[sub[i],2] }
# MMN random effects
for(i in 1:N) {b[i,1:2]~dmnorm(mu.b[i,1:2],tau.b[1:2,1:2])
for(k in 1:2) {mu.b[i,k]<-delta[k]*w[i]}
w[i] ~ dloglik(logLike[i])
logLike[i] <- -0.5*log(2*pi*sigma*sigma)-0.5*(w[i]-muw)*(w[i]-muw)/(sigma*sigma)
        +log(1+sin(lambda*(w[i]-muw)/sigma)/alpha)}
pi <- arccos(-1)
muw <- -lambda*sigma*exp(-lambda*lambda/2)/alpha
# Priors
for(k in 1:6) {beta[k] ~ dnorm(0,0.001)}
for(k in 1:2) {delta[k] ~ dnorm(0,0.001)}
for(k in 1:2) {tau.e[k] ~ dgamma(0.01,0.01)}
tau.b[1:2,1:2] ~ dwish(omega[1:2,1:2],2)
lambda ~ dunif(-10,10); alpha ~ dunif(1,10); sigma ~ dunif(0,10) }
```

References

[1] D.F. Andrews, C.L. Mallows, Scale mixtures of normal distributions, J. R. Stat. Soc. Ser. B (Methodol.) 36 (1974) 99–102.

[2] R.B. Arellano-Valle, H. Bolfarine, V.H. Lachos, Skew-normal linear mixed models, Data Sci. 3 (2005) 415–438.

[3] A. Azzalini, A class of distributions which includes the normal ones, Scand. J. Stat. 12 (1985) 171–178.

[4] A.J.H.M. Beurskens, H.C.W. De Vet, A.J.A. Köke, Responsiveness of functional status in low back pain: a comparison of different instruments, Pain 65 (1996) 71–76.

[5] S. Brooks, G. Roberts, Convergence assessment techniques for Markov chain Monte Carlo, Stat. Comput. 8 (1998) 319–325.

[6] S. Chakraborty, P.J. Hazarika, M.M. Ali, A multimodal skewed extension of normal distribution: its properties and applications, Statistics 49 (2015) 859–877.

[7] S. Chib, E. Greenberg, Understanding the Metropolis Hastings algorithm, Amer. Statist. 49 (1995) 327–335.

[8] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. Ser. B (Methodol.) 39 (1977) 1–38.

[9] M.J. Denwood, Runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS, J. Stat. Softw. 71 (2016) 1–25.

[10] S. Fieuws, G. Verbeke, Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles, Biometrics 62 (2006) 424–431.

[11] G. Fitzmaurice, M. Davidian, G. Verbeke, G. Molenberghs, Longitudinal Data Analysis, Chapman & Hall/CRC Press, Boca Raton, 2009.

[12] M.G. Genton, N.M.R. Loperfido, Generalized skew-elliptical distributions and their quadratic forms, Anal. Inst. Stat. Math. 57 (2005) 389–401.

[13] H. Goldstein, The Design and Analysis of Longitudinal Studies, Academic Press, London, 1979.

[14] C. Hennig, Identifiablity of models for clusterwise linear regression, J. Classif. 17 (2000) 273–296.

[15] W.J. Huang, Y.H. Chen, Generalized skew-cauchy distribution, Statist. Probab. Lett. 77 (2007) 1137–1147.

[16] W.J. Huang, N.C. Su, A.K. Gupta, A study of generalized skew-normal distribution, Statistics 47 (2013) 942–953.

[17] M.P. Jensen, P. Karoly, S. Braver, The measurement of clinical pain intensity: A comparison of six methods, Pain 27 (1986) 117–126.

[18] I. Kazemi, Z. Mahdiyeh, M. Mansourian, J.J. Park, Bayesian analysis of multivariate mixed models for a prospective cohort study using skew-elliptical distributions, Biometrical 55 (2013) 495–508.

[19] A.T. Komárek, G. Verbeke, On a fitting of a linear mixed model with a finite normal mixture as random-effects distribution, Robust 1 (2002) 186–193.

[20] N.M. Laird, J.H. Ware, Random-effects models for longitudinal data, Biometrics 38 (1982) 963–974.

[21] D. Lunn, C. Jackson, N. Best, D. Spiegelhalter, A. Thomas, The BUGS Book: A Practical Introduction to Bayesian Analysis, Chapman & Hall/CRC, 2012.

[22] D. Lunn, A. Thomas, N. Best, D. Spiegelhalter, Winbugs a bayesian modelling framework: concepts, structure and extensibility, Stat. Comput. 10 (2000) 325–337.

[23] N. Lysenko, P. Roy, R. Waeber, Multivariate extremes of generalized skew-normal distributions, Statist. Probab. Lett. 79 (2009) 525–533.

[24] C.E. McCulloch, J.M. Neuhaus, Misspecifying the shape of a random effects distribution: why getting it wrong may not matter, Statist. Sci. 26 (2011) 388–402.

[25] X.L. Meng, D.B. Rubin, Maximum likelihood estimation via the ECM algorithm: A general framework, Biometrika 80 (1993) 267–278.

[26] J.J. Park, J. Shin, Y. Choi, Y. Youn, S. Lee, S.R. Kwon, H. Lee, M.H. Kang, I.H. Ha, I. Shin, Integrative package for low back pain with leg pain in Korea: A prospective cohort study, in: Complementary Therapies in Medicine, Vol. 18, 2010, pp. 78–86.

[27] M. Plummer, JAGS: A program for analysis of Bayesian graphical models using gibbs sampling, in: Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vol. 124, 2003, pp. 125–135.

[28] C. Proust, H. Jacqmin-Gadda, Estimation of linear mixed models with a mixture of distribution for the random effects, Comput. Methods Programs Biomed. 78 (2005) 165–173.

[29] A. Soberóna, W. Stute, Assessing skewness kurtosis and normality in linear mixed models, J. Multivariate Anal. 161 (2017) 123–140.

[30] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, A. Van Der Linde, Bayesian measures of model complexity and fit, J. R. Stat. Soc. Ser. B (Methodol.) 64 (2002) 583–639.

[31] D.M. Titterington, A.F.M. Smith, U.E. Makov, Statistical Analysis of Finite Mixture Distributions, Wiley, New York, 1985.

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

[33] G. Verbeke, G. Molenberghs, Linear Mixed Models for Longitudinal Data, Springer-Verlag, New York, 2000.

[34] M. Ye, Z.H. Lu, Y. Li, X. Song, Finite mixture of varying coefficient model: Estimation and component selection, J. Multivariate Anal. 171 (2019) 452–474.

[35] Y. Zhang, G. Qin, Z. Zhu, J. Zhang, Robust estimation in linear regression models for longitudinal data with covariate measurement errors and outliers, J. Multivariate Anal. 168 (2018) 261–275.