Journal of Statistical Modelling: Theory and Applications
Vol. 5, No. 2, 2024, pp. 97-119
Yazd University Press 2024

JSMTA

Research Paper

Linear mixed model based on mean mixture of multivariate normal distributions: A flexible estimate based on missing value

Farzane Hashemi*, Faranak Goodarzi
Department of Statistics, Faculty of Mathematics, University of Kashan, Kashan, Iran

Received: October 01, 2024/ Revised: February 24, 2025/ Accepted: March 24, 2025

Abstract: The purpose of this paper is to extend the linear mixed model for handling missing and heavy-tailed data. In this model, the random effects have multivariate mean mixture of normal distribution and errors arise from a multivariate normal distribution. An expectation conditional maximization algorithm is developed for parameter estimation based on missing information. The mechanism of missing data is missing-at-random. Simulation studies and real data sets represent the efficiency and performance of the proposed model.

Keywords: ECM-algorithm; Heavy-tail distribution; Linear mixed models; Mean mixture normal distribution; Skewness.
Mathematics Subject Classification (2010): 62J05, 62H30.

1 Introduction

Linear mixed models (LMM) (Laird and Ware, 1982) have played an important role in analyzing longitudinal data that is measured repeatedly. The introduction of LMM paved a broad way for the development of applied statistics for analyzing data in a large variety of fields, especially medicine, and biostatistics. The key feature of LMM is to use a multivariate normal distribution for each sample that has been repeatedly tested.

*Corresponding author: `farzane.hashemi1367@kashanu.ac.ir`

In the case that the sample has participated in all periodic tests, for a set of random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, the generic form of an LMM model can theoretically be defined by a linear combination of random effects as

$$\begin{aligned} &\boldsymbol{Y}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{B}_j + \boldsymbol{\epsilon}_j, \\ &\boldsymbol{B}_j \overset{iid}{\sim} N_q(\boldsymbol{0}, \boldsymbol{D}), \quad \boldsymbol{\epsilon}_j \overset{iid}{\sim} N_p(\boldsymbol{0}, \boldsymbol{\Psi}), \qquad \boldsymbol{B}_j \perp \boldsymbol{\varepsilon}_j, \end{aligned} \tag{1}$$

where $\perp$ is the indicator independence of variables, $\boldsymbol{Y}_j \in \mathbb{R}^p$, $\boldsymbol{\beta}$ is a g-dimensional constant effects vector and $p \times g$ known matrix $\boldsymbol{X}_j$. Based on model (1), $p \times q$ known matrix $\boldsymbol{Z}_j$ describes the relation between $\boldsymbol{Y}_j$ and $\boldsymbol{B}_j$ for $j = 1, \ldots, n$. It is clear that $\boldsymbol{Y}_j$s are independent and $\boldsymbol{Y}_j | \boldsymbol{B}_j = \boldsymbol{b}_j \sim N_p(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b}_j, \boldsymbol{\Psi})$. LMMs are applicable in many medical branches, like the study of Laird and Ware (1982).

When faced with frequent measures, we might observe missing values in the measurements caused by either refusal, attrition, or lack of a record. In this regard, the problem of missing data should be addressed appropriately before engaging learning algorithms. One of the approaches to deal with the missing values is to delete that case completely that may lead to inefficient and biased inferences. Therefore, ignoring individuals with missing information may tend to introduce bias into inference and the remaining information may not be representative of the population. Another approach is that imputation with plausible values can replace missing observations. However, many researchers have proposed the maximum likelihood approach to deal with missing information.

The Gaussian Linear mixed models lack robustness against skewness and kurtosis and may seriously distort the estimates and results. Pinheiro et al. (2001) used a multivariate t distribution in the distribution of random effect and errors as a flexible model of heavy-tail data. For handling skewed data, Lin and Lee (2008) proposed an extension of the original LMM in which the distribution of random effect was the skew-normal and called (SN-LMM). To overcome the lack of heavy-tail in SN-LMM, Ho and Lin (2010) introduced a flexible skew LMM model by using the multivariate skew-t as a distribution of random effects; called skew-t Linear mixed models with missing data.

In asymmetric LMM, Lachos et al. (2010) studied LMM by using skew-normal independent distribution including skew-t, skew-contaminated normal and skew slash distribution for random effect variable. Recently, Negarestani et al. (2019) introduced a new class of skew distribution by assuming the mean mixture of multivariate normal (MMN) distribution for covering skewed and atypical observations. Based on Negarestani et al. (2019), a random vector $\boldsymbol{Y}$ is said to follow a p-variate MMN distribution if it takes the linear stochastic representation

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\lambda} W + \boldsymbol{U}, \quad W \perp \boldsymbol{U}, \tag{2}$$

where $\boldsymbol{U}$ is the p-dimension random vector of $\sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and W is the univariate random variable with cumulative distribution function (cdf) $H(\cdot; \boldsymbol{\nu})$. Very recently, Sepahdar et al. (2022) proposed an asymmetric extension of the mixture-of-experts model by using the MMN family for clustering multivariate observations with skewed and heavy-tailed dealing. Recently, Schumacher et al. (2021) considered a robust skew LMM in which the error terms have a dependence structure.

In this paper, we develop a robust model against asymmetric and heavy-tailed data via a flexible skew distribution on the LMM in the presence of missing values.

We called this model the mean mixture of normal linear mixed model (MMN-LMM). The outliers can create seriously biased estimates and subsequently lead to distorted inference. By using the representation of the MMN-LMM, we develop an expectation conditional maximization (Meng and Rubin, 1993, ECM) as an extension of expectation maximization (Dempster et al., 1977, EM) algorithm to carry out maximum likelihood (ML) estimation. In this paper, we consider the missing-at-random with an ignorable mechanism in which the cause of omission is unrelated to the missing information. To address the presence of missing values, two auxiliary indicator matrices are incorporated into the study to make methodological and theoretical developments more efficient. We show the efficiency of the proposed model via two simulations studies and by analyzing a real data set. We see that the performance of our model is better than other models.

The outline of this paper is therefore structured as follows. In Section 2, we describe the MMN distribution model and present some important properties for two special cases. In Section 3, we describe MMNE-LMM within the missing information and develop an efficient ECM algorithm for calculating ML estimates of parameters in Section 4. Extensive simulation studies are contained in Section 5. Section 6 includes the analysis of real data set.

2 Class of the MMN distribution

2.1 Preliminaries

A random vector $\boldsymbol{Y}$ is said to follow a p-variate restricted skew-normal (rSN) distribution if it obtains the probability density function (pdf)

$$f_{rSN}(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda}) = 2\phi_p(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Omega})\Phi\left(\frac{\boldsymbol{\lambda}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}{\sqrt{1-\boldsymbol{\lambda}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\lambda}}}\right), \quad \boldsymbol{y}\in\mathbb{R}^p, \tag{3}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\lambda}\boldsymbol{\lambda}^\top$, $\phi_p(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the pdf of $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, and $\Phi(\cdot)$ is the cdf of standard Gaussian distribution with a zero mean and a variance equal to unity, $N(0,1)$ (Pyne et al., 2014). It is clear from pdf rSN that the stochastic representation of rSN distribution is (2) when we replace pdf of W with standard truncated normal, $W \sim TN(0,1;(0,\infty))$. The notation $\boldsymbol{Y} \sim rSN(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda})$ will be used if $\boldsymbol{Y}$ has pdf (3). Moreover, since the truncated normal distribution belongs to the elliptical class of distributions, the rSN distribution is a skew elliptically contoured model. In this paper, we will refer to the p-variate random vector $\boldsymbol{Y} \in \mathbb{R}$ with MMN distribution is refer to as $\boldsymbol{Y} \sim MMN(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\boldsymbol{\nu})$ and the hierarchical stochastic representation can be gives

$$\boldsymbol{Y} \mid W = w \sim \mathcal{N}_p(\boldsymbol{\mu}+\boldsymbol{\lambda}w,\boldsymbol{\Sigma}), \;\; W \sim h(w;\boldsymbol{\nu}).$$

The p-dimension random vector $\boldsymbol{Y}$ takes the pdf

$$f_{\mathrm{MMN}}(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\Sigma},\boldsymbol{\nu}) = \int_{-\infty}^{\infty}\phi_p(\boldsymbol{y};\boldsymbol{\mu}+\boldsymbol{\lambda}w,\boldsymbol{\Sigma})h(w;\boldsymbol{\nu})\,dw, \quad \boldsymbol{y}\in\mathbb{R}^p, \tag{4}$$

where $\phi_p(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ indicates the pdf of $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$. It is clear that

$$E(\boldsymbol{Y}) = \boldsymbol{\mu} + E(W)\boldsymbol{\lambda} \quad \text{and} \quad \mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma} + \mathrm{Var}(W)\boldsymbol{\lambda}\boldsymbol{\lambda}^\top. \tag{5}$$

By using (5), the mean of MMN distribution changes for any members of this family upon W being changed. Moreover, notice how the skewness is caused by the mixing random variable W. If W is distributed symmetrically and/or elliptically the resulting distribution of $\boldsymbol{Y}$ is also symmetric and/or elliptical. Moreover, the pdf (4) may result in non-elliptically contoured distribution. When H degenerates with $w = 1$, or when $\boldsymbol{\lambda} \to 0$, the p-dimension random vector normal distribution is computed. We observe that the MMN model subclass, designated W, conforms to any positive or skewed distribution.. This setting results in nice skew distributions. In the rest of this section, some special cases of the MMN models are presented that will be considered later.

2.2 Special cases

• Convolution with exponential distribution. Let $\boldsymbol{Y} \sim MMNE(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ denotes the mean mixture normal exponential (MMNE) distribution such that in (2), we replace $W \sim E(\nu)$ where $E(\nu)$ exponential distribution with the mean $1/\nu$. Its pdf function $MMNE(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ can be expressed as

$$f_{\text{MMNE}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu) = \frac{\nu\sqrt{2\pi}}{\tau} \exp\left\{\frac{A_E^2}{2}\right\} \phi_p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi(A_E), \quad \boldsymbol{y} \in \mathbb{R}^p,$$

where $\tau^2 = \boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\lambda}$, and $A_E = \tau^{-1}\left[\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) - \nu\right]$.

• Convolution with exponential and half-normal distribution. Let $\boldsymbol{Y} \sim MMNEH$ $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ denotes the mean mixture normal exponential half normal (MMNEH) distribution such that in (2), we replace W with the following pdf

$$f(w; \nu) = \nu_1 \nu_2 \exp\{-\nu_2 w\} + 2(1 - \nu_1)\phi(w), \quad w, \nu_2 > 0, \quad 0 < \nu_1 < 1. \tag{6}$$

Its pdf function $MMNEH(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ can be expressed as

$$\begin{aligned} f_{\text{MMNEH}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\nu}) &= \nu_1 f_{\text{MMNE}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu_2) \\ &\quad + (1 - \nu_1) f_{\text{rSN}}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}), \quad \boldsymbol{y} \in \mathbb{R}^p, \end{aligned}$$

where $\boldsymbol{\nu} = (\nu_1, \nu_2)$. It is clear that the pdf of MMNEH is a mixture of two pdfs and $E(W) = \nu_1 \nu_2^{-1} + \sqrt{2/\pi}(1 - \nu_1)$.

• Convolution with Weibull distribution. The mixed-Weibull MMN (MMNW) distribution, denoted by $\boldsymbol{Y} \sim MMNW(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$, arises from (2) when W follows Weibull distribution with scale parameter ν and shape parameter 2. The pdf of $\boldsymbol{Y}$ is given by

$$\begin{aligned} f_{\text{MMNW}_p}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) &= \frac{2\nu^2\sqrt{2\pi}}{\tau_{WE}^2} \exp\left\{\frac{A_{WE}^2}{2}\right\} \phi_p(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\quad \times \left(A_{WE}\Phi(A_{WE}) + \phi(A_{WE})\right), \quad \boldsymbol{y} \in \mathbb{R}^p, \end{aligned}$$

where $\tau_{WE}^2 = \tau^2 + 2\nu^2$ and $A_{WE} = \tau_{WE}^{-1}\left[\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right]$. The MMNW model is a non-elliptically contoured distribution with a log-concave pdf (since the Weibull distribution with shape parameter grater than 1 has a log-concave pdf).

• Convolution with Gamma distribution. In this case, the mixing random variable W in (2) is distributed by $Gamma(2, \nu)$, where $Gamma(a, b)$ denotes the gamma distribution with mean and variance are $E(W) = 2/\nu$ and $Var(W) = 2/\nu^2$. Then, the pdf of $\boldsymbol{Y} \sim MMNG(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ takes the form

$$\begin{aligned} f_{\mathrm{MMNG}_p}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu) &= \frac{\nu^2\sqrt{2\pi}}{\tau^2} \exp\left\{\frac{A_E^2}{2}\right\} \phi_p\left(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\ &\quad \times \left(A_E \Phi(A_E) + \phi(A_E)\right), \quad \boldsymbol{y} \in \mathbb{R}^p. \end{aligned}$$

It is also clear that the pdf of MMNG distribution belongs to the class of skew non-elliptically contoured distributions and is a log-concave function.

• Convolution with Lindley distribution. Let the random variable W in (2) follows a Lindley distribution with the pdf

$$f_W(w; \nu) = \frac{\nu^2}{1+\nu}(1+w)\exp\{-\nu w\}, \quad w, \nu > 0.$$

The mean and variance of W are

$$E(W) = \frac{2+\nu}{\nu(\nu+1)} \text{ and } Var(W) = \frac{\nu^2+4\nu+2}{\nu^2(\nu+1)^2}.$$

We note that, the Lindley distribution is a mixture of $E(\nu)$ and $Gamma(2, \nu)$ distributions with mixing parameter $\nu/(\nu+1)$. Then, the pdf of $\boldsymbol{Y}$ following the mixed-Lindley MMN (MMNL) distribution is

$$\begin{aligned} f_{\mathrm{MMNL}_p}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu) &= \frac{1}{1+\nu}\Big(\nu f_{\mathrm{MMNE}_p}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu) \\ &\quad + f_{\mathrm{MMNG}_p}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)\Big), \quad \boldsymbol{y} \in \mathbb{R}^p. \end{aligned}$$

3 The MMN linear mixed model with incomplete data

To introduce a robust extension of the LMM (1), we consider that the joint vector of random effects and residual errors follow a $(p+q)$-variate of MMN distribution. Specifically,

$$\begin{aligned} &\boldsymbol{Y}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b}_j + \boldsymbol{\epsilon}_j, \\ &\boldsymbol{b}_j \overset{iid}{\sim} MMN_q(\boldsymbol{0}, \boldsymbol{D}, \boldsymbol{\lambda}, \nu), \quad \boldsymbol{\epsilon}_j \overset{iid}{\sim} N_p(\boldsymbol{0}, \boldsymbol{\Psi}), \qquad \boldsymbol{b}_j \perp \boldsymbol{\epsilon}_j, \end{aligned} \tag{7}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)$ is a $q \times 1$ vector of skewness parameters and covariance matrix of error is assumed $\boldsymbol{\Psi} = \sigma^2\boldsymbol{R}$, where $\boldsymbol{R}$ is known matrix, unless otherwise noted. Then, using the representation in (2), MMN-LMM in (7) has the following hierarchical stochastic representation Alternatively, by the linear representation (2), the proposed MMN-LMM model in (7) admits the following two-level representation

$$\begin{aligned} \boldsymbol{Y}_j \mid w_j, &\sim N_p(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{\lambda}w_j, \boldsymbol{Z}_j\boldsymbol{D}\boldsymbol{Z}_j^\top + \boldsymbol{\Psi}), \\ W_j &\sim h(w_j; \boldsymbol{\nu}). \end{aligned} \tag{8}$$

Let $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{\lambda}, \nu, \sigma^2\}$ denote all unknown parameters in the MMN-LMM. Based on the assumption (7) together with applying (8), it is straightforward to see

$$\boldsymbol{Y}_j \sim MMN_p(\boldsymbol{X}_j\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{Z}_j\boldsymbol{\lambda}, \boldsymbol{\nu}), \quad \text{with} \quad \boldsymbol{\Sigma} = \boldsymbol{Z}_j\boldsymbol{D}\boldsymbol{Z}_j^\top + \boldsymbol{\Psi}.$$

By using (5), $\boldsymbol{Y}_j$ has the following expectation and covariance

$$E(\boldsymbol{Y}_j) = \boldsymbol{X}_j\boldsymbol{\beta} + E(W_j)\boldsymbol{Z}_j\boldsymbol{\lambda} \quad \text{and} \quad \mathrm{Cov}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma} + \mathrm{Var}(W_j)\boldsymbol{Z}_j\boldsymbol{\lambda}\left(\boldsymbol{Z}_j\boldsymbol{\lambda}\right)^\top.$$

From (7), models based on MMN-LMM can be represented hierarchically at the following three levels

$$\boldsymbol{Y}_j \mid (\boldsymbol{b}_j, w_j) \sim \mathcal{N}_p(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b}_j, \boldsymbol{\Psi}),\ \boldsymbol{b}_j \mid w_j \sim \mathcal{N}_q\big(w_j\boldsymbol{\lambda}, \boldsymbol{D}\big),\ W_j \sim h(w_j; \boldsymbol{\nu}). \tag{9}$$

Consequently, the log-likelihood function for $\boldsymbol{\Theta}$ associated with the matrix of observations $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ is

$$\ell(\boldsymbol{\Theta}|\boldsymbol{Y}) = \sum_{j=1}^{n} \log f_{MMN}(\boldsymbol{y}_j; \boldsymbol{X}_j\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{Z}_j\boldsymbol{\lambda}, \boldsymbol{\nu}). \tag{10}$$

where $f_{MMN}(\cdot)$ is the pdf of MMN distribution. To carry out ML estimation of the model (10), some of the observed data $\boldsymbol{y}_i$ is incomplete because in some cases, the sample refuses to repeat the test or lacks the data. Under this assumption, we partition $\boldsymbol{Y}_j$ to the observed component $\boldsymbol{Y}_j^o \in \mathbb{R}^{p_j^o}$ and the missing component $\boldsymbol{Y}_j^m \in \mathbb{R}^{p_j^m}$, where $p_j^o + p_j^m = p$. For easy notation and computation, two ancillary permutation matrices are introduced as $\boldsymbol{O}_j$ $(p_j^o \times p_j)$ and $\boldsymbol{M}_j$ $((p - p_j^o) \times p_j)$ to identify $\boldsymbol{y}_j^o$ and $\boldsymbol{y}_j^m$ as

$$\boldsymbol{Y}_j^o = \boldsymbol{O}_j\boldsymbol{Y}_j \quad \text{and} \quad \boldsymbol{Y}_j^m = \boldsymbol{M}_j\boldsymbol{Y}_j, \quad \text{such that} \quad \boldsymbol{Y}_j = \boldsymbol{O}_j^\top\boldsymbol{Y}_j^o + \boldsymbol{M}_j^\top\boldsymbol{Y}_j^m \quad \text{and} \quad \boldsymbol{O}_j^\top\boldsymbol{O}_j + \boldsymbol{M}_j^\top\boldsymbol{M}_j = \boldsymbol{I}_p.$$

The following proposition presents some significant consequences, which help obtain the Q-function of the ECM algorithm.

Proposition 3.1. *From the MMN-LMM model* (7)*, we have*
a. With w_j, $\boldsymbol{Y}_j^o$ has the following conditional distribution:

$$\boldsymbol{Y}_j^o \mid w_j \sim N_{p_j^o}\left(\boldsymbol{O}_j\left(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{\lambda}w_j\right), \boldsymbol{\Sigma}_j^{oo}\right),$$

where $\boldsymbol{\Sigma}_j^{oo} = \boldsymbol{O}_j\boldsymbol{\Sigma}\boldsymbol{O}_j^\top$.
b. $\boldsymbol{Y}_j^o$ has a marginal distribution:

$$\boldsymbol{Y}_j^o \sim MMN(\boldsymbol{\xi}_j^o, \boldsymbol{\Sigma}_j^{oo}, \boldsymbol{\eta}_j^o, \boldsymbol{\nu}) \tag{11}$$

where $\boldsymbol{\xi}_j^o = \boldsymbol{O}_j\boldsymbol{X}_j\boldsymbol{\beta}$ and $\boldsymbol{\eta}_j^o = \boldsymbol{O}_j\boldsymbol{Z}_j\boldsymbol{\lambda}$.
c. With w_j and $\boldsymbol{b}_j$, $\boldsymbol{Y}_j^o$ has the following conditional distribution:

$$\boldsymbol{Y}_j^o \mid \boldsymbol{b}_j, w_j \sim N_{p_j^o}\left(\boldsymbol{O}_j\left(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b}_j\right), \boldsymbol{\Psi}_j^{oo}\right),$$

where $\boldsymbol{\Psi}_j^{oo} = \boldsymbol{O}_j\boldsymbol{\Psi}\boldsymbol{O}_j^\top$.
d. With $\boldsymbol{y}_j^o$, $\boldsymbol{b}_j$ *and,* w_j, $\boldsymbol{Y}_j^m$ *has the following conditional distribution:*

$$\boldsymbol{Y}_j^m \mid (\boldsymbol{y}_j^o, \boldsymbol{b}_j, w_j) \sim N_{p-p_j^o}(\boldsymbol{\varphi}_{ij}^{m.o}, \boldsymbol{\Psi}_j^{mm.o}),$$

where $\boldsymbol{\varphi}_j^{m.o} = \boldsymbol{M}_j\left[\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b}_j + \boldsymbol{\Psi}\boldsymbol{C}_j^{oo}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b}_j)\right]$, $\boldsymbol{C}_j^{oo} = \boldsymbol{O}_j^\top\left(\boldsymbol{O}_j\boldsymbol{\Psi}\boldsymbol{O}_j^\top\right)^{-1}$ $\boldsymbol{O}_j$, *and* $\boldsymbol{\Psi}_j^{mm.o} = \boldsymbol{M}_j\left(\boldsymbol{I}_p - \boldsymbol{\Psi}\boldsymbol{C}_j^{oo}\right)\boldsymbol{\Psi}\boldsymbol{M}_j^\top$.
e. we have

$$f(w_j \mid \boldsymbol{y}_j^o) = \frac{\phi_{p_j^o}\left(\boldsymbol{y}_j^o; \boldsymbol{O}_j\left(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{\lambda}w_j\right), \boldsymbol{\Sigma}_j^{oo}\right)f(w_i)}{f_{MMN_{p_j^o}}(\boldsymbol{y}_j^o; \boldsymbol{\xi}_j^o, \boldsymbol{\Sigma}_j^{oo}, \boldsymbol{\eta}_j^o, \boldsymbol{\nu})}.$$

f. we have

$$\boldsymbol{b}_j \mid (\boldsymbol{y}_j^o, w_j) \sim N_q(\boldsymbol{\lambda}w_j + \boldsymbol{F}_j^{o\top}(\boldsymbol{y}_j - \boldsymbol{\xi}_j - w_j\boldsymbol{\eta}_j), \left(\boldsymbol{I}_q - \boldsymbol{F}_j^{o\top}\boldsymbol{Z}_j\right)\boldsymbol{D}),$$

where $\boldsymbol{\xi}_j = \boldsymbol{X}_j\boldsymbol{\beta}$, $\boldsymbol{\eta}_j = \boldsymbol{Z}_j\boldsymbol{\lambda}$ *and* $\boldsymbol{F}_j^o = \boldsymbol{C}_j^{oo}\boldsymbol{Z}_j\boldsymbol{D}$.

Proof. See Appendix 7.1. □

4 Estimation process via ECM algorithm

4.1 Parameter estimation

The ECM algorithm is a well-known framework for computing ML parameter estimation when data are incomplete or treated as such. This algorithm iterates two steps wherein the missing and latent data are estimated by their conditional expectations in the E-step to obtain the expected value of the complete-data log-likelihood and the CM-step maximizes the conditional expectation of complete-data log-likelihood function to update parameter estimates. These E- and CM-steps are iterated until a convergence criterion is attained.

To implement our feasible ECM procedure for learning the MMN-LMM, let $\boldsymbol{Y}_c = (\boldsymbol{y}^o, \boldsymbol{y}^m, \boldsymbol{b}, \boldsymbol{w})$ be the complete data, where $\boldsymbol{y}^m = (\boldsymbol{y}_1^m, \ldots, \boldsymbol{y}_n^m)$ denotes the missing portion of the data, $\boldsymbol{y}^o = (\boldsymbol{y}_1^o, \ldots, \boldsymbol{y}_n^o)$ are part of the data that has been observed, $\boldsymbol{b} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\}$ is the set of random effect, and $\boldsymbol{w} = (w_1, \ldots, w_n)$ are the missing variables. The complete-data log-likelihood function for $\boldsymbol{\Theta}$ based on the complete data $\boldsymbol{Y}_c$, and omission of constant values, is

$$\begin{aligned}\ell_c(\boldsymbol{\Theta} \mid \boldsymbol{Y}_c) \quad = \quad & \sum_{j=1}^n \log h(W_i; \boldsymbol{\nu}) - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{j=1}^n \operatorname{tr}\left(\boldsymbol{\Psi}^{-1}\boldsymbol{\Upsilon}_{1j}\right) - \frac{n}{2}\log|\boldsymbol{D}| \\ & -\frac{1}{2}\sum_{j=1}^n \operatorname{tr}\left(\boldsymbol{D}^{-1}\boldsymbol{\Upsilon}_{2j}\right)\end{aligned}$$

where $\boldsymbol{\Upsilon}_{1j} = \left(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta} - \boldsymbol{Z}_j\boldsymbol{b}_j\right)\left(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta} - \boldsymbol{Z}_j\boldsymbol{b}_j\right)^\top$, $\boldsymbol{\Upsilon}_{2j} = \left(\boldsymbol{b}_j - W_j\boldsymbol{\lambda}\right)\left(\boldsymbol{b}_j - W_j\boldsymbol{\lambda}\right)^\top$ and $\operatorname{tr}(\boldsymbol{M})$ denotes the trace of matrix $\boldsymbol{M}$.

The following proposition can be used to evaluate the Q-function.

Proposition 4.1. *Given the hierarchical representation of the MMN-LMM* (9)*, we have*

$$\begin{aligned}
E(\boldsymbol{b}_i \mid \boldsymbol{y}_i^o) &= \boldsymbol{\lambda} E(W_j \mid \boldsymbol{y}_j^o) + \boldsymbol{F}_j^{o\top}(\boldsymbol{y}_j - \boldsymbol{\xi}_i - E(W_j \mid \boldsymbol{y}_j^o)\boldsymbol{\eta}_j^o),\\
E(W_i\boldsymbol{b}_i \mid \boldsymbol{y}_i^o) &= \boldsymbol{\lambda} E(W_j^2 \mid \boldsymbol{y}_j^o) + \boldsymbol{F}_j^{o\top}((\boldsymbol{y}_j - \boldsymbol{\xi}_i)E(W_j \mid \boldsymbol{y}_j^o) - E(W_j^2 \mid \boldsymbol{y}_j^o)\boldsymbol{\eta}_j^o),\\
E(\boldsymbol{b}_j\boldsymbol{b}_j^\top \mid \boldsymbol{y}_j^o) &= E(\boldsymbol{b}_j \mid \boldsymbol{y}_j^o)(\boldsymbol{F}_j^{o\top}(\boldsymbol{y}_j - \boldsymbol{\xi}_j))^\top + E(W_i\boldsymbol{b}_i \mid \boldsymbol{y}_j^o)(\boldsymbol{\lambda} - \boldsymbol{F}_j^{o^\top}\boldsymbol{\eta}_j^o)^\top\\
&\quad + \left(\boldsymbol{I}_q - \boldsymbol{F}_j^{o\top}\boldsymbol{Z}_j\right)\boldsymbol{D}.
\end{aligned}$$

Proof. By using Proposition 3.1 part (f), the proof is straightforward and omitted. □

Let $\hat{\boldsymbol{\Theta}}^{(k)} = \left(\hat{\boldsymbol{\beta}}^{(k)}, \hat{\boldsymbol{D}}^{(k)}, \hat{\boldsymbol{\lambda}}^{(k)}, \hat{\nu}^{(k)}, \hat{\sigma}^{(k)^2}\right)$ denote all estimations of model parameters, $\boldsymbol{\Theta}$, at the kth iteration. We then have the following conditional expectations.

$$\begin{aligned}
&\hat{w}_{1j}^{(k)} = E(W_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}^{(k)}), \quad \hat{t}_j^{(k)} = E(W_j^2 \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}^{(k)}),\\
&\hat{\boldsymbol{\zeta}}_{0j}^{(k)} = E(\boldsymbol{b}_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}^{(k)}), \quad \hat{\boldsymbol{\zeta}}_{1j}^{(k)} = E(W_i\boldsymbol{b}_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}^{(k)}), \quad \hat{\boldsymbol{\Phi}}_j^{(k)} = E(\boldsymbol{b}_j\boldsymbol{b}_j^\top \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}^{(k)}),
\end{aligned}$$

for $j = 1, \ldots, n$. Evaluations can be produced using Proposition 4.1. Notice that $\hat{w}_{1j}^{(k)}$ and $\hat{t}_j^{(k)}$ can be easily evaluated using the results stated in 7.2. Considering $\hat{\boldsymbol{\Theta}}^{(0)}$ in the beginning, our proposed ECM algorithm for ML estimates of the MMN-LMM, iterates the following steps:

E-step: Given the observed data $\boldsymbol{y}^o$ and the current estimate $\hat{\boldsymbol{\Theta}}^{(k)}$, the E-step calculates the conditional expectation of the complete-data log-likelihood function. This leads to the so-called Q-function:

$$\begin{aligned}
Q(\boldsymbol{\Theta}) &= \sum_{j=1}^{n}\Bigg[E\left(\log h(W_i;\boldsymbol{\nu}) \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}^{(k)}\right) - \frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Psi}^{-1}\widehat{\boldsymbol{\Upsilon}_{1j}}^{(k)}\right)\\
&\quad -\frac{1}{2}\log|\boldsymbol{D}| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{D}^{-1}\widehat{\boldsymbol{\Upsilon}_{2j}}^{(k)}\right)\Bigg], \qquad (12)
\end{aligned}$$

where $\widehat{\boldsymbol{\Upsilon}}_{1j}^{(k)} = E\left(\left(\boldsymbol{Y}_j - \boldsymbol{X}_j\boldsymbol{\beta} - \boldsymbol{Z}_j\boldsymbol{b}_j\right)\left(\boldsymbol{Y}_j - \boldsymbol{X}_j\boldsymbol{\beta} - \boldsymbol{Z}_j\boldsymbol{b}_j\right)^\top \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\theta}}^{(k)}\right)$ and $\widehat{\boldsymbol{\Upsilon}_{2j}}^{(k)} = \hat{\boldsymbol{\Phi}}_{ij}^{(k)} - \hat{\boldsymbol{\zeta}}_{1j}^{(k)}\boldsymbol{\lambda}^\top - \boldsymbol{\lambda}\hat{\boldsymbol{\zeta}}_{1j}^{(k)^\top} + \hat{t}_j^{(k)}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top$.

CM-step 1: Let $\boldsymbol{\Psi} = \sigma^2\boldsymbol{R}$. Maximizing Q-function (12) over $\hat{\boldsymbol{\Theta}}$ leads to the parameter updates

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \left(\sum_{j=1}^{n}\boldsymbol{X}_j^\top\boldsymbol{R}^{-1}\boldsymbol{X}_j\right)^{-1}\left(\sum_{j=1}^{n}\boldsymbol{X}_j^\top\boldsymbol{R}^{-1}\left[\hat{\boldsymbol{q}}_j^{o(k)} - \hat{\boldsymbol{\Psi}}^{(k)}\hat{\boldsymbol{C}}_j^{oo^{(k)}}\boldsymbol{Z}_j^{(k)}\hat{\boldsymbol{\zeta}}_{0j}^{(k)}\right]\right)$$

$$\hat{\sigma}^{2(k+1)} = \frac{1}{np}\mathrm{tr}\left[\mathrm{Diag}\left(\sum_{j=1}^{n}\widehat{\boldsymbol{\Upsilon}}_{1j}^{(k+1)}\right)\right],$$

$$\hat{\boldsymbol{\lambda}}^{(k+1)} = \frac{\sum_{j=1}^{n}\hat{\boldsymbol{\zeta}}_{1j}^{(k)}}{\sum_{j=1}^{n}\hat{t}_j^{(k)}}, \quad \hat{\boldsymbol{D}}^{(k+1)} = \frac{1}{n}\sum_{j=1}^{n}\widehat{\boldsymbol{\Upsilon}}_{2j}^{(k+1)}$$

$$\hat{\nu}^{(k+1)} = \begin{cases} \frac{n}{\sum_{j=1}^{n} \hat{w}_{1j}^{(k)}}, & \text{for MMNE-LMM,} \\ \frac{2n}{\sum_{j=1}^{n} \hat{w}_{1j}^{(k)}}, & \text{for MMNG-LMM,} \\ \left(\frac{n}{\sum_{j=1}^{n} \hat{w}_{2j}^{(k)}} \right)^{1/2}, & \text{for MMNW-LMM,} \\ \frac{\sqrt{8\frac{\sum_{j=1}^{n} \hat{w}_{1j}^{(k)}}{n} + (1-\frac{\sum_{j=1}^{n} \hat{w}_{1j}^{(k)}}{n})^2} + (1-\frac{\sum_{j=1}^{n} \hat{w}_{1j}^{(k)}}{n})}{2\frac{\sum_{j=1}^{n} \hat{w}_{1j}^{(k)}}{n}}, & \text{for MMNL-LMM} \end{cases}$$

where $\hat{\boldsymbol{q}}_j^{o(k)} = \boldsymbol{X}_j \hat{\boldsymbol{\beta}}^{(k)} + \hat{\boldsymbol{\Psi}}^{(k)} \hat{\boldsymbol{C}}_j^{oo^{(k)}} (\boldsymbol{y}_j - \boldsymbol{X}_j \hat{\boldsymbol{\beta}}^{(k)})$, $\hat{\boldsymbol{E}}_j^{oo^{(k)}} = \left(\boldsymbol{I}_p - \hat{\boldsymbol{\Psi}}^{(k)} \hat{\boldsymbol{C}}_j^{oo^{(k)}} \right) \boldsymbol{Z}_j$ and

$$\begin{aligned} \hat{\boldsymbol{\Upsilon}}_{1j}^{(k+1)} = & (\hat{\boldsymbol{q}}_j^{o(k)} - \boldsymbol{X}_j \hat{\boldsymbol{\beta}}^{(k+1)})(\hat{\boldsymbol{q}}_j^{o(k)} - \boldsymbol{X}_j \hat{\boldsymbol{\beta}}^{(k+1)})^\top + \left(\boldsymbol{I}_p - \hat{\boldsymbol{\Psi}}^{(k)} \hat{\boldsymbol{C}}_j^{oo^{(k)}} \right) \hat{\boldsymbol{\Psi}}^{(k)} \\ & + \left(\hat{\boldsymbol{E}}_j^{oo^{(k)}} - \boldsymbol{Z}_j \right) \hat{\boldsymbol{\Phi}}_j^{(k)} \left(\hat{\boldsymbol{E}}_j^{oo^{(k)}} - \boldsymbol{Z}_j \right) + \left(\hat{\boldsymbol{q}}_j^{o(k)} - \boldsymbol{X}_j \hat{\boldsymbol{\beta}}^{(k+1)} \right) \hat{\hat{\boldsymbol{\zeta}}}_{0j}^{(k)\top} \\ & \times \left(\hat{\boldsymbol{E}}_j^{oo^{(k)}} - \boldsymbol{Z}_j \right)^\top + \left(\hat{\boldsymbol{E}}_j^{oo^{(k)}} - \boldsymbol{Z}_j \right) \hat{\hat{\boldsymbol{\zeta}}}_{0j}^{(k)} \left(\hat{\boldsymbol{q}}_j^{o(k)} - \boldsymbol{X}_j \hat{\boldsymbol{\beta}}^{(k+1)} \right)^\top . \end{aligned}$$

CM-step 2: We can update ν for the MMNEH-LMM using the following relation:

$$\hat{\nu}^{(k+1)} = \arg\max_{\nu} \sum_{j=1}^{n} \log f_{\mathrm{MMNEH}_{p_j^o}} \left(\boldsymbol{y}_j^o; \boldsymbol{\xi}_j^{o^{(k+1)}}, \boldsymbol{\Sigma}_j^{oo^{(k+1)}}, \boldsymbol{\eta}_j^{o^{(k+1)}}, \nu \right),$$

where $\hat{\boldsymbol{\xi}}^{o^{(k+1)}}$, $\hat{\boldsymbol{\eta}}_j^{o^{(k+1)}}$, and $\hat{\boldsymbol{\Sigma}}_j^{oo^{(k+1)}}$ are $\boldsymbol{\xi}_j^o$, $\boldsymbol{\eta}_j^o$ and $\boldsymbol{\Sigma}_j^{oo}$ in (11), respectively, obtained at the prevalent estimation at the beginning of the $(k+1)$-th repetition. By using functions `optim` and `nlminb` in R programming, we can update ν, but calculating $\hat{\nu}^{(k+1)}$ is still complicated.

Remark 4.2. *It can be seen that the update estimate of ν for the MMNEH-LMM is obtained by CM-step 2. To circumvent this entangled form in CM-step 2, we introduce an indicator variable V_j as follows the hierarchical representation*

$$\begin{aligned} \boldsymbol{Y}_j^o \mid (W_j = w_j, V_j = 1) & \sim \mathcal{N}_{p_j^o}(\boldsymbol{O}_j(\boldsymbol{X}_j \boldsymbol{\beta} + w_j \boldsymbol{Z}_j \boldsymbol{\lambda}), \boldsymbol{\Sigma}_j^{oo}), \\ W_j \mid V_j = 1 & \sim E(\nu_2) \\ V_j & \sim Ber(1, \nu_1), \end{aligned} \tag{13}$$

where $Ber(1,\nu)$ denotes the Bernoulli trail with probability ν. Here, $V_j = 1$ if $\boldsymbol{y}_j^o$ comes from the MMNE-type distribution, and $V_j = 0$ if $\boldsymbol{y}_j^o$ generated by the rSN model. By using (13)*, the parameter $\boldsymbol{\nu}$ can be updated as*

$$\hat{\nu}_1^{(k+1)} = \frac{\sum_{j=1}^{n} \pi(\boldsymbol{y}_j^o)}{n}, \quad and \quad \hat{\nu}_2^{(k+1)} = \frac{\sum_{j=1}^{n} \pi(\boldsymbol{y}_j^o)}{\sum_{j=1}^{n} \hat{w}_{1j}^{(k)} \pi(\boldsymbol{y}_j^o)},$$

where

$$\pi(\boldsymbol{y}_j^o) = \frac{\hat{\nu}_1^{(k)} f_{MMNE}(\boldsymbol{y}_i^o; \boldsymbol{\xi}_j^{o^{(k+1)}}, \boldsymbol{\Sigma}_j^{oo^{(k+1)}}, \boldsymbol{\eta}_j^{o^{(k+1)}}, \hat{\nu}_2^{(k)})}{f_{MMNEH}(\boldsymbol{y}_i^o; \boldsymbol{\xi}_j^{o^{(k+1)}}, \boldsymbol{\Sigma}_j^{oo^{(k+1)}}, \boldsymbol{\eta}_j^{o^{(k+1)}}, \hat{\boldsymbol{\nu}}^{(k)})}.$$

4.2 Predicting random effect and missing information

To calculate predicting random effect and missing information, we indicate the ML estimations by $\hat{\boldsymbol{\Theta}} = \left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{\lambda}}, \hat{\nu}, \hat{\sigma}^2\right)$. From Proposition 4.1, the estimator of random effect can be evaluated as follows

$$\hat{\boldsymbol{b}}_j = E(\boldsymbol{b}_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}}) \tag{14}$$

where $E(\boldsymbol{b}_j \mid \boldsymbol{y}_j^o, \hat{\boldsymbol{\Theta}})$ is obtained by replacing $\hat{\boldsymbol{\Theta}}$ with $\boldsymbol{\Theta}$ in Proposition 4.1 part (a). As a by-product of our ECM algorithm and Proposition 3.1 part (d), conditional imputation is used to estimate missing values as

$$\hat{\boldsymbol{y}}_j^m = \boldsymbol{M}_j \left[\boldsymbol{X}_j\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_j\hat{\boldsymbol{b}}_j + \hat{\boldsymbol{\Psi}}\hat{\boldsymbol{C}}_j^{oo}(\boldsymbol{y}_j - \boldsymbol{X}_j\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_j\hat{\boldsymbol{b}}_j)\right], \tag{15}$$

where $\hat{\boldsymbol{b}}_j$ is defined in (14). We use the mean squared deviation (MSD) as a measure of the difference between the true value of $\boldsymbol{y}_j^m$ and the imputed value of $\hat{\boldsymbol{y}}_j^m$. MSD can be calculated as follows

$$\text{MSD} = \frac{1}{n^*}\sum_{j=1}^{n}(\boldsymbol{y}_j^m - \hat{\boldsymbol{y}}_j^m)^\top(\boldsymbol{y}_j^m - \hat{\boldsymbol{y}}_j^m), \tag{16}$$

where n^* is the number of missing items.

4.3 Estimation of standard errors based on EM algorithm

The asymptotic covariance matrix of the ML estimates can be approximated by the inverse of the observed information matrix (Efron and Hinkley, 1978). Before showing the main result, we need to define some notation. Let $\ell_{cj}(\boldsymbol{\Theta} \mid \boldsymbol{y}_{cj})$ be the log-likelihood formed from the single complete observation $\boldsymbol{y}_{cj} = (\boldsymbol{y}_j^o, \boldsymbol{y}_j^m, w_j, \boldsymbol{b}_j)$ such that

$$\ell_{cj}(\boldsymbol{\Theta} \mid \boldsymbol{Y}_{cj}) = \log h(W_j; \boldsymbol{\nu}) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Upsilon}_{1j}\right) - \frac{1}{2}\log|\boldsymbol{D}| - \frac{1}{2}\text{tr}\left(\boldsymbol{D}^{-1}\boldsymbol{\Upsilon}_{2j}\right).$$

Also, Let $\boldsymbol{d}_i = \text{Diag}(\boldsymbol{D}_i)$ denote a $p \times 1$ vector containing entries on the main diagonal of $\boldsymbol{D}_i$.

According to Meilijson (1989) formula, the observed information matrix can be estimated by

$$\hat{I}_o = I_o(\hat{\boldsymbol{\Theta}} \mid \boldsymbol{y}^o) = \sum_{j=1}^{n}\boldsymbol{s}(\boldsymbol{y}_j^o \mid \hat{\boldsymbol{\Theta}})\hat{\boldsymbol{s}}^\top(\boldsymbol{y}_j^o \mid \hat{\boldsymbol{\Theta}}), \tag{17}$$

where

$$\boldsymbol{s}(\boldsymbol{y}_j^o \mid \hat{\boldsymbol{\Theta}}) = E\left(\frac{\ell_c(\boldsymbol{\theta} \mid \boldsymbol{y}_{cj})}{\partial\boldsymbol{\Theta}} \mid \boldsymbol{y}_j, \hat{\boldsymbol{\Theta}}\right),$$

is the individual score vector containing elements of $\left(\hat{s}_{j,\boldsymbol{\beta}}^\top, \hat{s}_{j,\boldsymbol{d}}^\top, \hat{s}_{j,\boldsymbol{\lambda}_g}^\top, \hat{s}_{j,\boldsymbol{\nu}}, \hat{s}_{j,\sigma^2}\right)$.

Explicit expressions for the above elements can be obtained by standard matrix differentiation. Technical derivations are given below:

$$\hat{\boldsymbol{s}}_{j,\boldsymbol{\beta}} = \boldsymbol{X}_j^\top\boldsymbol{R}^{-1}\left[\hat{\boldsymbol{q}}_j^o - \hat{\boldsymbol{\Psi}}\hat{\boldsymbol{C}}_j^{oo}\boldsymbol{Z}_j\hat{\boldsymbol{\zeta}}_{0j}\right] - \boldsymbol{\beta}\boldsymbol{X}_j^\top\boldsymbol{R}^{-1}\boldsymbol{X}_j,$$

$$
\begin{aligned}
\hat{\boldsymbol{s}}_{j,\boldsymbol{\sigma}^2} &= -p\log\hat{\sigma} - \frac{1}{2\hat{\sigma}^2}\operatorname{tr}\left(\widehat{\boldsymbol{\Upsilon}_{1j}}\right)\\
\hat{\boldsymbol{s}}_{j,\boldsymbol{d}} &= \operatorname{Diag}\left(-\frac{1}{2}\left\{\hat{\boldsymbol{D}}^{-1} - \hat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{\Upsilon}_{2j}}\hat{\boldsymbol{D}}^{-1}\right\}\right),\\
\hat{s}_{j,\boldsymbol{\lambda}} &= -2\hat{\boldsymbol{\zeta}}_{1j} + 2\hat{t}_j\hat{\boldsymbol{\lambda}},\\
\hat{s}_{j,\nu} &= \begin{cases} \frac{1}{\hat{\nu}} - \hat{w}_{1j}, & \text{for MMNE-LMM,}\\ \frac{2}{\hat{\nu}} - \hat{w}_{1j}, & \text{for MMNG-LMM,}\\ \frac{1}{\hat{\nu}^2} - \hat{t}_j, & \text{for MMNW-LMM,}\\ \frac{2+\hat{\nu}}{\hat{\nu}(\hat{\nu}+1)} - \hat{w}_{1j}, & \text{for MMNL-LMM.}\end{cases}
\end{aligned}
$$

It can be seen that the standards errors of $\boldsymbol{\nu}$ for the MMNEH-LMM is obtained by

$$\hat{s}_{j,\nu_1} = \hat{\nu_1} - \pi(\boldsymbol{y}_j^o) \text{ and } \hat{s}_{j,\nu_2} = \hat{\nu_2}\hat{w}_{1j}\pi(\boldsymbol{y}_j^o) - \pi(\boldsymbol{y}_j^o).$$

The standard errors can be approximated by calculating the square root of the diagonal elements of the inverse of (17). If the standard errors are obtainable, they are useful to assess the significance of parameter estimates as well as other inferential issues. The observed information-based estimator for the variances of ML estimators is asymptotically consistent if the model is correctly specified (White, 1996).

4.4 Notes on implementation

Like any other EM-type algorithm, if the ECM algorithm is given good parameter estimates, convergence may be sped up or made easier. When the skewness parameter in the MMN-LMM tends to zero, we have the original LMM. Therefore, we put $\hat{\boldsymbol{\lambda}}^{(0)} = \boldsymbol{0}$ corresponding to an initial supposition close to the original LMM. The initial values of $\hat{\boldsymbol{\beta}}^{(0)}$, $\hat{\sigma}^{2(0)}$ and $\hat{\boldsymbol{D}}^{(0)}$ are described in the R command "**lmm**". In addition, we started the algorithm with $\hat{\nu}^{(0)} = 0.5$ for MMNEH-LMM.

The Akaike information criterion (AIC) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) are measures to select the number of classes and factors. It calculates as

$$\text{AIC} = -2\ell_{\max} + 2m, \quad \text{BIC} = -2\ell_{\max} + m\log n,$$

where m is the number of free parameters, and $\ell_{\max}$ is the maximized log-likelihood value. Models with fewer AIC and BIC values are generally better fitted.

5 Simulation Study

In this section, we investigate the asymptotic properties of the ML estimates as well as the performance of our models in transactions with skewed and heavily tailed data. In all simulations, we considered the different percentages of missing values.

5.1 Asymptotic properties

In this simulation, we regain true parameters for two special cases of the MMN-LMM based on the ECM algorithm and compare these estimates with their true parameters. For this simulation, we generate 100 samples from MMNE-LMM with $q = 2$, $p = 5$, and $n = 100,\ 200,\ 400,\ 800,$ and 1600. The true parameters values are specified as $\boldsymbol{\beta} = (3,2,1)^\top$, $\boldsymbol{\lambda} = (1,2)^\top$, $\boldsymbol{\Psi} = \sigma^2 \boldsymbol{I}$ where $\sigma^2 = 0.25$ and $\boldsymbol{I}$ is identity matrix, $\boldsymbol{D} = \text{diag}(1,1)$, $\nu = 2$

$$\boldsymbol{Z}_j = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}^\top, \quad \boldsymbol{X}_j = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \\ t & t & t & t & t \end{bmatrix}^\top.$$

To empirically study, the flexibility of MAR model when dealing with missing value, data deletion method is used. In the MAR experience, missing items are obtained by random deletion under three levels, without missing value (r = 0%), moderate (r = 10%), and upper (r = 20%) rates of omission.

To evaluate the estimation accuracy, we obtain the relative mean absolute bias (RBias) and the root mean squared error (RMSE):

$$\text{RBias} = \frac{1}{100}\sum_{r=1}^{100} |\hat{\theta}^{(r)} - \theta_{true}| \ \text{ and } \ \text{RMSE} = \sqrt{\frac{\sum_{r=1}^{100}(\hat{\theta}^{(r)} - \bar{\hat{\theta}})^2}{100}},$$

where $\hat{\theta}^{(r)}$ explains the ML estimate obtained using Section 4.1 as a specific parameter at the rth replication and θ_{true} is its true value.

In addition, investigating the standard error's estimation consistency is of our interests. So, by using the approximate standard errors (ASE), we measured the sample standard deviation of parameters (STD) and the average standard errors:

$$\text{STD}(\hat{\theta}) = \sqrt{\frac{1}{99}\sum_{r=1}^{100}\left[\hat{\theta}^{(r)} - \frac{1}{99}\sum_{r=1}^{100}\hat{\theta}^{(r)}\right]^2} \ \text{ and } \ \text{ASE}(\hat{\theta}) = \frac{1}{100}\sum_{r=1}^{100}\text{SE}(\hat{\theta}^{(r)}),$$

where $\text{SE}(\hat{\theta}^{(r)})$ denotes the asymptotic standard errors of $\hat{\theta}$ at the rth replication. We examine the accuracies of STD estimators with ASE as well as n increase for the above model using discrepancy measures: sum of absolute deviation of STD with ASE computed by

$$\text{SAD}(\hat{\theta}) = |\text{STD}(\hat{\theta}) - \text{ASE}(\hat{\theta})|,$$

where $\text{STD}(\hat{\theta})$ is standard deviation of $\hat{\theta}$ and $\text{ASE}(\hat{\theta})$ is average standard errors using the observed information matrix for parameter $\hat{\theta}$ with sample size n.

The numerical results displayed in Figures 1 disclose that both values of RBias and RMSE tend to zero as the sample size increases, confirming the empirical consistency of the ML estimators. According to Figures 1, the estimate parameters are close to the true value that shows the better and more effective performance of the ECM algorithm proposed in Section 4.1. Also, Table 1 show that the values of SAD are tending to zero as n increases. This suggests that as the value of n increases, the difference between STD and ASE diminishes, indicating a convergence between the two.

5.2 Behavior of proposed models vs heavy-tailed data

In this subsection, the performance of robust extension of LMM based on MMN distribution in dealing with heavy-tailed data is tested in terms of estimating missing values.

Table 1: Simulation results for assessing the sum of absolute deviation of STD with ASE of parameters estimates across various sample sizes.

n	measure	β_0	β_1	β_2	d_{11}	d_{22}	σ^2	λ_1	λ_2	ν
100	SAD	0.2702	0.3642	0.4626	0.3538	0.4384	0.0520	0.2957	0.2267	0.2239
200	SAD	0.1170	0.1521	0.1533	0.1117	0.1689	0.0233	0.1940	0.1590	0.0757
400	SAD	0.0648	0.0848	0.1003	0.0664	0.0990	0.0176	0.0895	0.0663	0.0405
800	SAD	0.0390	0.0491	0.0577	0.0347	0.0559	0.0093	0.0310	0.0229	0.0238
1600	SAD	0.0316	0.0402	0.0383	0.0307	0.0490	0.0052	0.0178	0.0108	0.0167

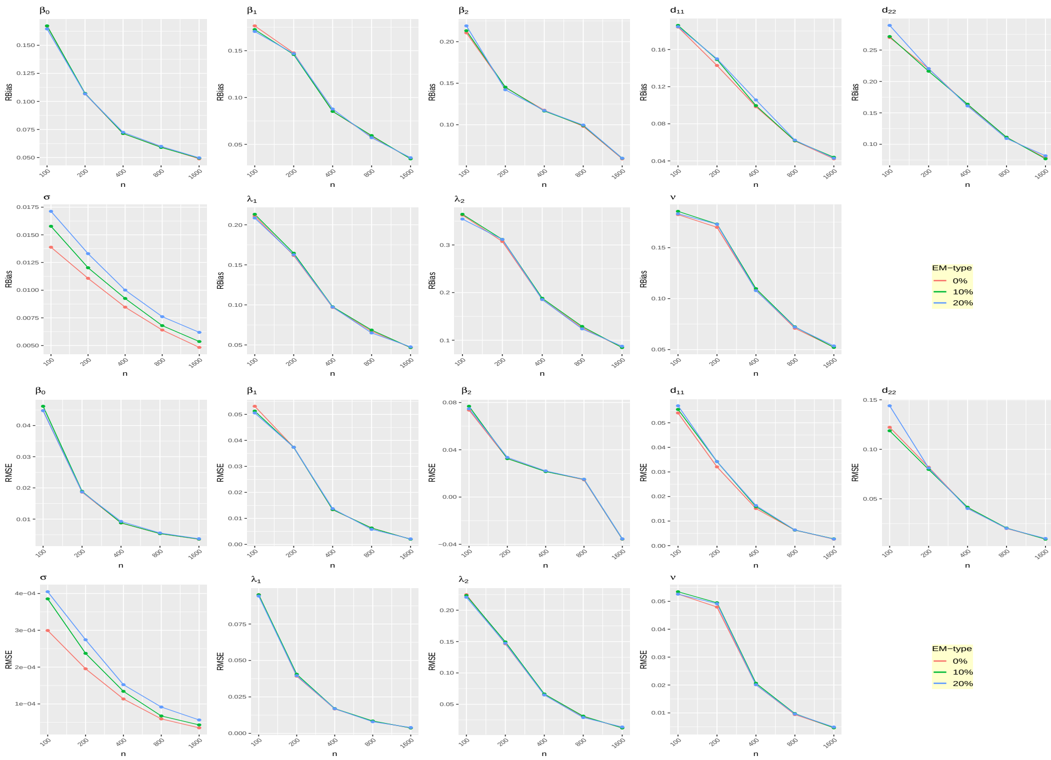

Figure 1: Mean of Rbias and RMSE for parameters based on ECM algorithm with increase sample size, MMNE-LMM.

For this aim, we considered three levels of missing data based on MAR and two scenarios to generate heavy-tailed data. We simulate $n = 200$ from MMN-LMM with $q = 2$ and $p = 5$. The mixing variable W in (2) is considered Birnbaum-Saunders distribution (Birnbaum and Saunders, 1969) with α and 1 for scale and shape parameter for the first scenario (S1) and generalized inverse Gaussian (GIG) (Good, 1953) with parameter $\chi = 1$, $\psi = 2$ and $\kappa = 0.5$ for the second scenario (S2). These models, referred to as MMNBS-LMM and MMNGIG-LMM, are not discussed in Section 4 since the pdf and conditional expectations are not available. The other true parameters values for proposed models are specified as $\boldsymbol{\beta} = (4, 2, 1)^\top$, $\boldsymbol{\lambda} = (2, 3)^\top$, $\boldsymbol{\Psi} = \sigma^2 \boldsymbol{I}$ where $\sigma^2 = 0.5$

and $\boldsymbol{I}$ is the identity matrix,

$$\boldsymbol{D} = \begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix}^{\top}, \quad \boldsymbol{Z}_j = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}^{\top}, \quad \boldsymbol{X}_j = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \\ t & t & t & t & t \end{bmatrix}^{\top},$$

where $t = 1$ for $j \leq n/2$ and $t = 0$ for $j > n/2$. For each run of 100 simulations, four models normal linear mixed model (N-LMM), SN-LMM, MMNE-LMM, and MMNEH-LMM as well as three levels of missing data 0%, 10%, and 20% are applied to the generated data. For each model, we employed two information criteria the Akaike information criterion (AIC) (Aitken, 1926) and the Bayesian information criterion (BIC) (Schwarz, 1978) and obtained MSD in (16) to measure the difference between the true value of missing data and the imputed value of missing data. For the sake of different models comparison, Table 2 report the average value of AIC, BIC and MSD for three levels missing rate. As seen in Table 2, the MMN-LMMs carry out better than the standard LMM and its extended based on SN, which confirms the positive effect of the presence of family of MMN distributions in random effects. It can be observed that the proposed models are superior more consistently compared to LMM and SN-LMM. In addition, LMMs based on heavy-tailed distributions, such as MMMNG and MMMNEH distributions, consistently presented more suitable values for three criterion values. It can clearly be seen that the fitting performance of the MMNEH-LMM and MMNG-LMM are improved in S1 and S2, respectively. For S1 and S2, the MMNEH-LMM and MMNG-LMM performed similarly in terms of estimating missing data.

Figures 2 and 3 give visuals comparison of different LMMs for BIC and MSD based on two scenarios. It is obviously observed that the MMNE-LMM provides a better fit in terms of BIC for two scenarios. Besides, the MMNEH-LMM provides the best estimated missing values in two scenarios.

Table 2: Comparison of the estimation performance of four LMMs based on AIC, BIC, and MSD for three levels of missing rate simulated in two scenarios.

scenario	measure	missing rate	N-LMM	SN-LMM	MMNE-LMM	MMNEH-LMM	MMNG-LMM	MMNW-LMM	MMNL-LMM
S1	AIC	0%	4268.178	4186.393	4167.104	4163.246	4165.374	4167.846	4167.566
		10%	3958.915	3877.380	3858.105	3854.215	3855.573	3857.035	3862.760
		20%	3643.714	3562.691	3543.473	3537.625	3541.799	3540.292	3542.914
	BIC	0%	4297.863	4216.078	4196.789	4190.229	4193.058	4195.829	4199.251
		10%	3988.600	3907.065	3887.790	3880.198	3883.258	3882.018	3889.445
		20%	3673.398	3592.376	3573.158	3570.608	3572.484	3575.275	3574.599
	MSD	10%	0.499	0.497	0.447	0.398	0.445	0.396	0.455
		20%	1.132	1.124	1.073	1.024	1.075	1.026	1.030
S2	AIC	0%	4253.281	4201.226	4196.041	4197.843	4192.895	4194.842	4198.053
		10%	3940.631	3888.915	3883.764	3885.580	3881.466	3883.424	3886.577
		20%	3622.899	3571.728	3566.796	3568.597	3562.810	3564.757	3567.792
	BIC	0%	4282.966	4230.911	4225.726	4230.826	4222.580	4227.825	4227.738
		10%	3970.315	3918.599	3913.449	3918.563	3911.151	3916.407	3916.262
		20%	3652.584	3601.413	3596.481	3601.580	3592.494	3597.740	3597.476
	MSD	10%	0.520	0.517	0.467	0.417	0.406	0.416	0.446
		20%	1.156	1.146	1.035	1.045	1.011	1.031	1.052

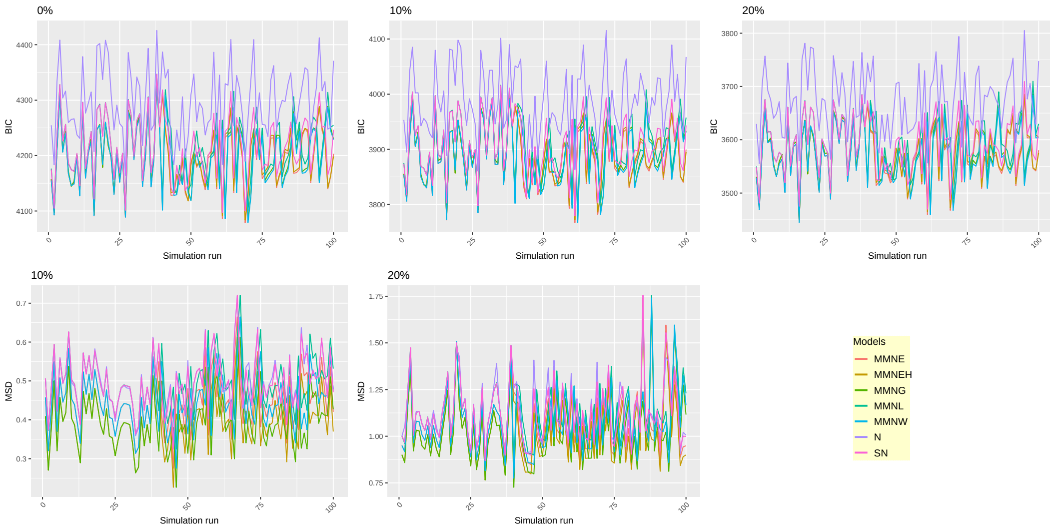

Figure 2: Comparison of BIC and MSD of seven LMMs for three levels missing rate with 100 simulations from the MMNBS-LMM model.

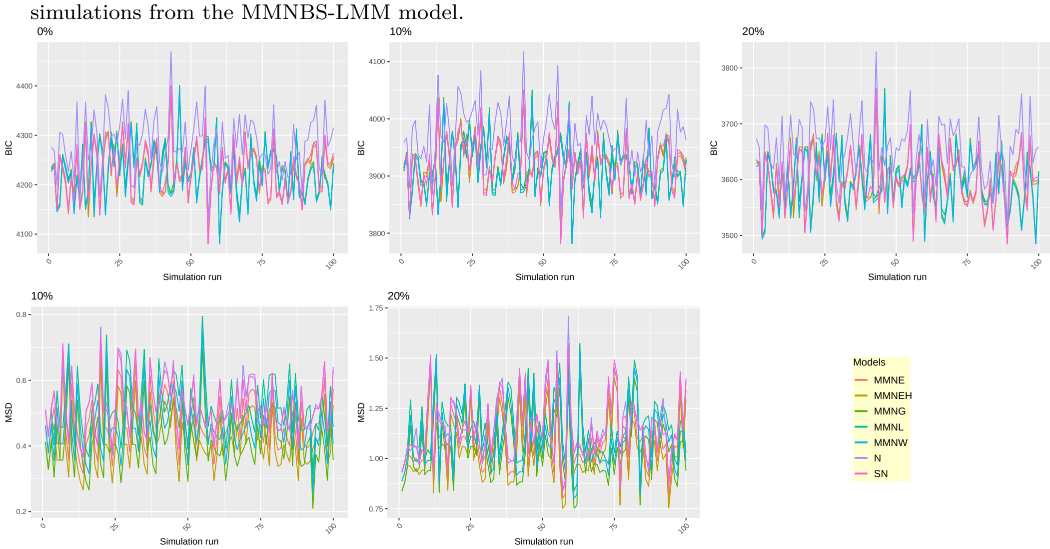

Figure 3: Comparison of BIC and MSD of seven LMMs for three levels missing rate with 100 simulations from the MMNGIG-LMM model.

6 Analysis of milk data sets

The Milk data was analyzed first by Diggle et al. (1994) and is available in the *nlme* package. The milk data sets describe the protein of cow's milk in the weeks following calving. This study has been done in 19 weeks ($p = 19$) with 79 cows ($n = 79$), and in some weeks, the protein content of cow's milk has not been recorded. We consider these cases as missing data. Cows are fed with three levels diets including barley, barley+lupins, and lupins. In the experiment, we assume a linear mixed model with random effects $\boldsymbol{B}_j = (B_{1j}, B_{2j})^\top$ and fixed effects $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$. Moreover, let $\boldsymbol{t} = (t_1, \ldots, t_{19})$ with $t_i = (week_i - 10)/10$ $i = 1, \ldots, 19$ and $diet_j$ be the type indicator with 0=barley+lupins, 1=barley and 2=lupins. Thus $\boldsymbol{X}_j = (1, \boldsymbol{t}, diet_j)^\top$ is matrix 19×3 and $\boldsymbol{Z} = (1, \boldsymbol{t})^\top$ is matrix 19×2. By using model (7), we fitted four models of MMN-LMM consisting of N-LMM, SN-LMM, MMNE-LMM, and MMNEH-LMM to the observations. At first, we fit N-LMM to the value of the observations as well as the missing value and obtain an estimation of random effects. Figure 4 clearly shows a skew distribution for random effects and thus MMN-LMMs are better to fit this data set. However, the estimation intercept random effects are positively skewed with 1.501 value and kurtosis 2.981, and therefore the offered original LMM does not fit well.

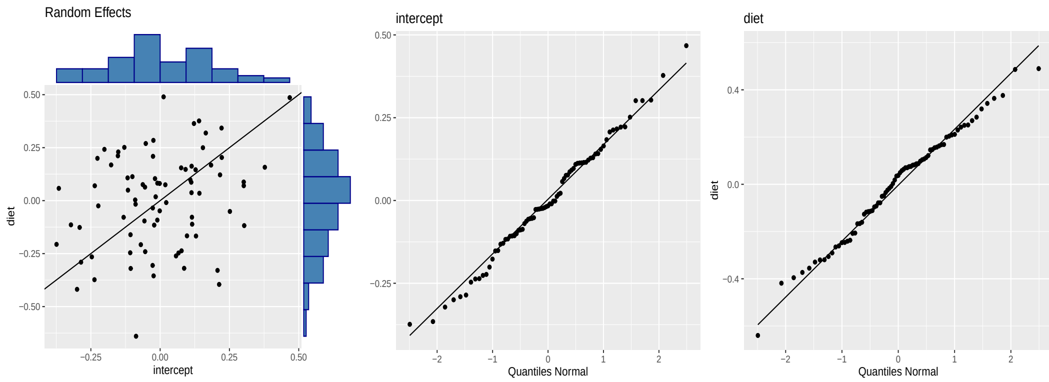

Figure 4: Scatter plot and normal Q-Q plots of for estimation random effects of milk data set based on N-LMM.

According to the above description, we assume an MMN distribution with two specific models for $\boldsymbol{b}_j$ and multivariate normal distribution for $\boldsymbol{\epsilon}_j$ with high-dimension as well as asymmetric data. In our fit for the analysis aims, we consider MMNE, MMNEH, and SN distributions from the MMN class.

The estimation of parameters, standard errors, maximum likelihood ($\ell(\boldsymbol{\Theta})$), AIC, and BIC values of entire LMMs used data are reported in Table 3. The AIC and BIC criterion show that the MMNEH-LMM and MMNG-LMM with heavy tails give a suitable fit in comparison to the SN-LMM and N-LMM models. Furthermore, the $\boldsymbol{\beta}$ estimations for three models with longer tails are almost similar, but the estimates for the σ^2s are not similar.

To assess the estimates of missing value with SN-LMM, MMNE-LMM, and MMNEH-LMM, we drop out the last five measurements of $\boldsymbol{y}_1^m = (y_{15}, \ldots, y_{19})$ from cow 1, then obtain the ML estimates by the new data. The estimate of $\boldsymbol{y}_1^m$ is created by formula (15). We use the MSD in (16) as a measure of precision. The comparison of the estimates based on the four models is presented in Table 4. The result showed that the MMNE and MMNEH distributions give better estimate than the SN and the Normal. Thus, the MMN-LMM not only provides a better fitting model, but it also yields almost accurately estimates of missing value for the milk data.

Table 3: Results from fitting the four models to the milk data set.

parameter	N-LMM		SN-LMM		MMNE-LMM		MMNEH-LMM		MMNG-LMM		MMNW-LMM		MMNL-LMM	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
β_0	3.449	1.158	3.462	0.998	3.460	1.452	3.462	1.052	3.425	1.253	3.470	1.489	3.458	1.326
β_1	-0.043	0.008	-0.043	0.027	-0.044	0.056	-0.043	0.072	-0.045	0.039	-0.044	0.061	-0.043	0.041
β_2	-0.125	0.015	0.179	0.098	0.075	0.041	0.179	0.099	0.152	0.074	0.201	0.125	-0.231	0.168
d_{11}	0.033	0.015	0.033	0.019	0.032	0.009	0.033	0.040	0.039	0.020	0.042	0.017	0.040	0.013
d_{12}	0.012	0.007	0.010	0.009	0.010	0.008	0.010	0.005	0.015	0.009	0.014	0.019	0.08	0.020
d_{22}	0.063	0.029	0.013	0.033	0.028	0.019	0.013	0.031	0.059	0.042	0.060	0.037	0.057	0.030
σ^2	0.062	0.022	0.079	0.046	0.073	0.044	0.081	0.022	0.078	0.042	0.080	0.023	0.039	0.010
λ_1	–	–	-2.016	1.230	-2.010	1.054	-3.016	1.523	-2.536	1.203	-2.790	1.429	-2.410	1.019
λ_2	–	–	-1.382	0.993	-1.201	0.985	-1.382	1.009	-1.503	0.963	-1.423	1.052	-1.296	1.132
ν_1	–	–	–		1.223	0.425	0.425	0.271	1.429	0.975	2.014	1.553	1.986	1.289
ν_2	–	–	–		–	–	1.426	0.764						
$\ell(\boldsymbol{\Theta})$	-175.225		-170.342		-165.220		-162.725		-163.792		-164.086		-164.736	
Num. Par.	7		9		10		11		10		10		10	
AIC	364.450		358.684		350.440		347.450		346.384		348.172		349.472	
BIC	381.965		380.009		374.1345		373.5139		370.038		371.866		373.166	

Table 4: Comparison of estimates of missing values accuracy in terms of MSD dropout of the last five measurements of the protein of milk for cow 1.

points being estimates	N-LMM	SN-LMM	MMNE-LMM	MMNEH-LMM	MMNG-LMM	MMNW-LMM	MMNL-LMM
$y_{15} = 4.13$	4.551	4.428	4.029	4.033	4.051	4.037	4.042
$y_{16} = 4.08$	4.367	4.241	3.942	3.961	3.976	3.966	4.140
$y_{17} = 4.22$	4.883	4.854	4.155	4.184	4.260	4.170	4.177
$y_{18} = 4.44$	4.991	4.867	4.268	4.367	4.396	4.383	4.385
$y_{19} = 4.30$	4.915	4.880	4.181	4.180	4.442	4.462	4.470
MSD	0.507	0.420	0.119	0.089	0.041	0.053	0.047

7 Concluding remarks and future work

In this study, we have presented a new class of asymmetric LMMs by using the class of MMN distribution. By using two auxiliary indicator matrices, an ECM algorithm is developed for obtaining the ML estimates of model parameters in the presence of missing data. The performance of the proposed LMMs has been studied using two simulation experiments and a world data set. Outcomes show that the efficiency of the MMN-LMM is better than SN-LMM and N-LMM. We believe that the approaches proposed here can also be used to study other asymmetric multivariate models. All calculations were carried out using R 4.2.2.

There are a few issues and possible modifications to the proposed methodology that deserves further attention. As has been indicated in the models (7), $\boldsymbol{\epsilon}_j$ is considered Normal distribution, its skew distribution can be challenged in LMMs. Thus, a change of $\boldsymbol{Z}$ for any sample is one of the future directions of our work. Bai et al. (2016) introduced a finite mixture linear mixed model in which the multivariate t distribution is used for random effects and error distribution. Using the MMN class it will then be of interest to extend the finite mixture linear mixed model for handling multi-modal, skewed, and heavy-tailed distributed data. We are currently working on these subjects and expect to present the findings in our future papers.

Acknowledgement

The authors are grateful to the Editors and two anonymous referees for their valuable comments and constructive suggestions which had improved the content of this paper greatly.

References

Aitken, A.C. (1926). On Bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh,* **46**:289–305.

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

Bai, X., Chen, K. and Yao, W. (2016). Mixture of linear mixed models using multivariate t distribution. *Journal of Statistical Computation and Simulation,* **86**(4):771–787.

Birnbaum, Z. and Saunders, S. (1969). A new family of life distributions. *Journal of Applied Probability,* **6**(02):319–327.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, **39**(1):1–38.

Diggle, P., Liang, K.-Y. and Zeger, S.L. (1994). *Longitudinal Data Analysis.* New York: Oxford University Press.

Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika,* **65**(3):457–483.

Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika,* **40**(3-4):237–264.

Ho, H.J. and Lin, T.-I. (2010). Robust linear mixed models using the skew t distribution with application to schizophrenia data. *Biometrical Journal,* **52**(4):449–469.

Lachos, V.H., Ghosh, P. and Arellano-Valle, R.B. (2010). Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica*, 303–322.

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

Lin, T.I. and Lee, J.C. (2008). Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data. *Statistics in Medicine,* **27**(9):1490–1507.

Meilijson, I. (1989). A fast improvement to the em algorithm on its own terms. *Journal of the Royal Statistical Society. Series B (Methodological),* **51**(1):127–138.

Meng, X.-L. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika,* **80**(2):267–278.

Negarestani, H., Jamalizadeh, A., Shafiei, S. and Balakrishnan, N. (2019). Mean mixtures of normal distributions: properties, inference and application. *Metrika,* **82**:501–528.

Pinheiro, J.C., Liu, C. and Wu, Y.N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, **10**(2): 249–276.

Pyne, S., Lee, S.X., Wang, K., Irish, J., Tamayo, P., Nazaire, M.-D., Duong, T., Ng, S.-K., Hafler, D., Levy, R. and Nolan, G.P. (2014). Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. *PloS one*, **9**(7):e100334.

Schumacher, F.L., Lachos, V.H. and Matos, L.A. (2021). Scale mixture of skew-normal linear mixed models with within-subject serial dependence. *Statistics in Medicine,* **40**(7):1790–1810.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics,* **6**(2):461–464.

Sepahdar, A., Madadi, M., Balakrishnan, N. and Jamalizadeh, A. (2022). Parsimonious mixture-of-experts based on mean mixture of multivariate normal distributions. *Stat*, **11**(1):e421.

White, H. (1996). *Estimation, Inference and Specification Analysis.* Cambridge University Press.

Appendix

7.1 Proof Proposition 3.1

(a) We have $bmY_j \mid w_j \sim N_p(\boldsymbol{\xi}_j + \boldsymbol{\eta} w_j, \boldsymbol{\Sigma})$, thus

$$\begin{bmatrix} \boldsymbol{Y}_j^o \\ \boldsymbol{Y}_j^m \end{bmatrix} \mid w_j \sim N_p\left(\begin{bmatrix} \boldsymbol{O}_j(\boldsymbol{\xi}_j + \boldsymbol{\eta} w_j) \\ \boldsymbol{M}_j(\boldsymbol{\xi}_j + \boldsymbol{\eta} w_j) \end{bmatrix}, \begin{bmatrix} \boldsymbol{O}_j \boldsymbol{\Sigma} \boldsymbol{O}_j^\top & \boldsymbol{O}_j \boldsymbol{\Sigma} \boldsymbol{M}_j^\top \\ \boldsymbol{M}_j \boldsymbol{\Sigma} \boldsymbol{O}_j^\top & \boldsymbol{M}_j \boldsymbol{\Sigma} \boldsymbol{M}_j^\top \end{bmatrix} \right).$$

It follows

$$\boldsymbol{Y}_j^o \mid w_j \sim N_{p_j^o}(\boldsymbol{O}_j(\boldsymbol{\xi}_j + \boldsymbol{\eta} w_j), \boldsymbol{O}_j \boldsymbol{\Sigma} \boldsymbol{O}_j^\top).$$

(b) Based on part (a), the proof is straightforward.
(c) We have

$$\begin{bmatrix} \boldsymbol{Y}_j^o \\ \boldsymbol{Y}_j^m \end{bmatrix} \mid \boldsymbol{b}_j, w_j \sim N_p\left(\begin{bmatrix} \boldsymbol{O}_j\left(\boldsymbol{X}_j \boldsymbol{\beta} + \boldsymbol{Z}_j \boldsymbol{b}_j\right) \\ \boldsymbol{M}_j\left(\boldsymbol{X}_j \boldsymbol{\beta} + \boldsymbol{Z}_j \boldsymbol{b}_j\right) \end{bmatrix}, \begin{bmatrix} \boldsymbol{O}_j \boldsymbol{\Psi} \boldsymbol{O}_j^\top & \boldsymbol{O}_j \boldsymbol{\Psi} \boldsymbol{M}_j^\top \\ \boldsymbol{M}_j \boldsymbol{\Psi} \boldsymbol{O}_j^\top & \boldsymbol{M}_j \boldsymbol{\Psi} \boldsymbol{M}_j^\top \end{bmatrix} \right).$$

Thus, by the marginal distribution of normal, we can see that

$$\boldsymbol{Y}_j^o \mid (\boldsymbol{b}_j, w_j) \sim N_{p_j^o}(\boldsymbol{O}_j\left(\boldsymbol{X}_j \boldsymbol{\beta} + \boldsymbol{Z}_j \boldsymbol{b}_j\right), \boldsymbol{O}_j \boldsymbol{\Psi} \boldsymbol{O}_j^\top).$$

(d) We have $\boldsymbol{Y}_j \mid (\boldsymbol{b}_j, w_j) \sim N_p(\boldsymbol{X}_j \boldsymbol{\beta} + \boldsymbol{Z}_j \boldsymbol{b}_j, \boldsymbol{\Psi})$, thus

$$\begin{bmatrix} \boldsymbol{Y}_j^o \\ \boldsymbol{Y}_j^m \end{bmatrix} \mid \boldsymbol{b}_j, w_j \sim N_p\left(\begin{bmatrix} \boldsymbol{O}_j\left(\boldsymbol{X}_j \boldsymbol{\beta} + \boldsymbol{Z} \boldsymbol{b}_j\right) \\ \boldsymbol{M}_j\left(\boldsymbol{X}_j \boldsymbol{\beta} + \boldsymbol{Z}_j \boldsymbol{b}_j\right) \end{bmatrix}, \begin{bmatrix} \boldsymbol{O}_j \boldsymbol{\Psi} \boldsymbol{O}_j^\top & \boldsymbol{O}_j \boldsymbol{\Psi} \boldsymbol{M}_j^\top \\ \boldsymbol{M}_j \boldsymbol{\Psi} \boldsymbol{O}_j^\top & \boldsymbol{M}_j \boldsymbol{\Psi} \boldsymbol{M}_j^\top \end{bmatrix} \right).$$

By Theorem 2.5.1 of Anderson (2003), we can see that

$$\begin{aligned} E(\boldsymbol{Y}_j^m \mid \boldsymbol{y}_j^o, \boldsymbol{b}_j, w_j) &= \boldsymbol{M}_j\left(\boldsymbol{X}_j \boldsymbol{\beta} + \boldsymbol{Z}_j \boldsymbol{b}_j\right) \\ &\quad + \boldsymbol{M}_j \boldsymbol{\Psi} \boldsymbol{O}_j^\top (\boldsymbol{O}_j \boldsymbol{\Psi} \boldsymbol{O}_j^\top)^{-1} (\boldsymbol{y}_j^o - \boldsymbol{O}_j\left(\boldsymbol{X}_j \boldsymbol{\beta} + \boldsymbol{Z}_j \boldsymbol{b}_j\right)) \\ &= \boldsymbol{M}_j \Bigg[\left(\boldsymbol{X}_j \boldsymbol{\beta} + \boldsymbol{Z} \boldsymbol{b}_j\right) \\ &\quad + \boldsymbol{\Psi} \boldsymbol{O}_j^\top (\boldsymbol{O}_j \boldsymbol{\Psi} \boldsymbol{O}_j^\top)^{-1} \boldsymbol{O}_j (\boldsymbol{y}_j - (\boldsymbol{X}_j \boldsymbol{\beta} + \boldsymbol{Z}_j \boldsymbol{b}_j)) \Bigg], \\ cov(\boldsymbol{Y}_j^m \mid \boldsymbol{y}_j^o, \boldsymbol{b}_j, w_j) &= \boldsymbol{M}_j \boldsymbol{\Psi} \boldsymbol{M}_j^\top - \boldsymbol{M}_j \boldsymbol{\Psi} \boldsymbol{O}_j^\top (\boldsymbol{O}_j \boldsymbol{\Psi} \boldsymbol{O}_j^\top)^{-1} \boldsymbol{O}_j \boldsymbol{\Psi} \boldsymbol{M}_j^\top \\ &= \boldsymbol{M}_j \left(\boldsymbol{I}_p - \boldsymbol{\Psi} \boldsymbol{O}_j^\top (\boldsymbol{O}_j \boldsymbol{\Psi} \boldsymbol{O}_j^\top)^{-1} \boldsymbol{O}_j \right) \boldsymbol{\Psi} \boldsymbol{M}_j^\top. \end{aligned}$$

(e) By using part (a), the proof is straightforward.
(f) It follows by part (c) that

$$\begin{aligned} E(\boldsymbol{Y}_j^o \boldsymbol{B}_j^\top \mid w_j) &= E\left[E(\boldsymbol{Y}_j^o \mid \boldsymbol{B}_j, w_j) \boldsymbol{B}_j^\top \mid w_j\right] = E\left[\boldsymbol{O}_j\left(\boldsymbol{X}_j \boldsymbol{\beta} + \boldsymbol{Z} \boldsymbol{B}_j\right) \boldsymbol{B}_j^\top \mid w_j\right] \\ &= \boldsymbol{O}_j\left[\boldsymbol{X}_j \boldsymbol{\beta} E(\boldsymbol{B}_j^\top \mid w_j) + \boldsymbol{Z}_j E(\boldsymbol{B}_j \boldsymbol{B}_j^\top \mid w_j)\right] \\ &= \boldsymbol{O}_j\left[\boldsymbol{X}_j \boldsymbol{\beta} \boldsymbol{\lambda}^\top w_j + \boldsymbol{Z}_j (\boldsymbol{D} + \boldsymbol{\lambda} \boldsymbol{\lambda}^\top w_j^2)\right], \\ cov(\boldsymbol{Y}_j^o \boldsymbol{B}_j^\top \mid w_j) &= E(\boldsymbol{Y}_j^o \boldsymbol{B}_j^\top \mid w_j) - E(\boldsymbol{Y}_j^o \mid w_j) E(\boldsymbol{B}_j^\top \mid w_j) \\ &= \boldsymbol{O}_j\left[\boldsymbol{X}_j \boldsymbol{\beta} \boldsymbol{\lambda}^\top w_j + \boldsymbol{Z}_j (\boldsymbol{D} + \boldsymbol{\lambda} \boldsymbol{\lambda}^\top w_j^2)\right] - \boldsymbol{O}_j (\boldsymbol{\xi}_j + \boldsymbol{\eta} w_j) \boldsymbol{\lambda}^\top w_j \end{aligned}$$

$$= \quad \boldsymbol{O}_j \boldsymbol{Z}_j \boldsymbol{D}.$$

Thus, we have

$$\begin{bmatrix} \boldsymbol{Y}_j^o \\ \boldsymbol{U}_{ij} \end{bmatrix} \mid w_j \sim N_p \left(\begin{bmatrix} \boldsymbol{O}_j(\boldsymbol{\xi}_j + \boldsymbol{\eta} w_j) \\ w_j \boldsymbol{\lambda} \end{bmatrix}, \begin{bmatrix} \boldsymbol{O}_j \boldsymbol{\Sigma}_i \boldsymbol{O}_j^\top & \boldsymbol{O}_j \boldsymbol{Z} \boldsymbol{D} \\ (\boldsymbol{O}_j \boldsymbol{Z} \boldsymbol{D})^\top & \boldsymbol{D} \end{bmatrix} \right).$$

We then have the following results:

$$\begin{aligned} E(\boldsymbol{B}_j \mid \boldsymbol{y}_j^o, w_j) &= \boldsymbol{\lambda} w_j + \boldsymbol{F}_j^{o\top}(\boldsymbol{y}_j - \boldsymbol{\xi}_j - w_j \boldsymbol{\eta}_j), \\ cov(\boldsymbol{B}_j \mid \boldsymbol{y}_j^o, w_j) &= \left(\boldsymbol{I}_q - \boldsymbol{F}_j^{o\top} \boldsymbol{Z}_j \right) \boldsymbol{D}. \end{aligned}$$

7.2 Some properties of particular members of the MMN family

Lemma 7.1. *If $W \sim TN(\xi, \omega; (a_1, a_2))$, then*

$$\begin{aligned} E(W) &= \xi - \omega \frac{\phi(\alpha_2) - \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)}, \\ E(W^2) &= \xi^2 + \omega^2 - \omega^2 \frac{\alpha_2 \phi(\alpha_2) - \alpha_1 \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)} - 2\xi\omega \frac{\phi(\alpha_2) - \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)}, \end{aligned}$$

where $\alpha_i = (a_i - \xi)/\omega$, for $i = 1, 2$.

Proposition 7.2. *a. Let $\boldsymbol{Y}^o \sim MMNE_{p^o}(\boldsymbol{\xi}^o, \boldsymbol{\Sigma}^{oo}, \boldsymbol{\eta}^o, \nu)$ and $W \sim E(\nu)$. Then, conditional distribution of W given $\boldsymbol{Y}^o = \boldsymbol{y}^o$ is $W_{\boldsymbol{y}^o} \sim TN\left(A_E^o \tau^{o-1}, \tau^{o-2}; (0, \infty)\right)$ where ${\tau^o}_E^2 = \boldsymbol{\eta}^{o\top} \boldsymbol{\Sigma}^{oo-1} \boldsymbol{\eta}^o$ and $A_E^o = {\tau_E^o}^{-1}\left[\boldsymbol{\eta}^{o\top} \boldsymbol{\Sigma}^{oo-1}(\boldsymbol{y}^o - \boldsymbol{\xi}^o) - \nu\right]$.*
b. Let $\boldsymbol{Y}^o \sim MMNEH_{p^o}(\boldsymbol{\xi}^o, \boldsymbol{\Sigma}^{oo}, \boldsymbol{\eta}^o, \boldsymbol{\nu})$ and W have a density in (6). Then, $W_{\boldsymbol{y}^o}$ has pdf

$$f_{W_{\boldsymbol{y}^o}}(w) = \pi(\boldsymbol{y}^o) \frac{\phi\left(w; A_{EH}^o {\tau_E^o}^{-1}, \tau^{o-2}\right)}{\Phi(A_{EH}^o)} + (1 - \pi(\boldsymbol{y}^o)) \frac{\phi\left(w; \xi, \omega^2\right)}{\Phi(\xi/\omega)},$$

where

$$\pi(\boldsymbol{y}^o) = \frac{\nu_1 \nu_2 \sqrt{2\pi}}{\tau_E^o f_{MMNEH_{p^o}}(\boldsymbol{y}^o; \boldsymbol{\xi}^o, \boldsymbol{\Sigma}^{oo}, \boldsymbol{\eta}^o, \nu)} \phi_p\left(\boldsymbol{y}^o; \boldsymbol{\xi}^o, \boldsymbol{\Sigma}^{oo}\right) \exp\left(\frac{A_{EH}^o}{2}\right) \Phi(A_{EH}^o).$$

$A_{EH}^o = {\tau_E^o}^{-1}\left[\boldsymbol{\eta}^{o\top} \boldsymbol{\Sigma}^{oo-1}(\boldsymbol{y}^o - \boldsymbol{\mu}^o) - \nu_2\right]$, $\xi = \boldsymbol{\eta}^{o\top} \boldsymbol{\Omega}^{oo-1}(\boldsymbol{y}^o - \boldsymbol{\xi}^o)$ and $\omega^2 = 1 - \boldsymbol{\eta}^{o\top} \boldsymbol{\Omega}^{oo-1} \boldsymbol{\eta}^o$ for $\boldsymbol{\Omega}^{oo} = \boldsymbol{\Sigma}^{oo} + \boldsymbol{\eta}^o \boldsymbol{\eta}^{o\top}$. Furthermore, for any $\boldsymbol{y}_j^o \in \mathbb{R}^{p_j^o}$, and $k = 1, 2, \ldots$,

$$E(W_{\boldsymbol{y}}^k) = \pi(\boldsymbol{y}) E\left(V_1^k\right) + (1 - \pi(\boldsymbol{y})) E(V_2^k),$$

such that $V_1 \sim TN\left(A_{EH}^o {\tau_E^o}^{-1}, {\tau_E^o}^{-2}; (0, \infty)\right)$, $V_2 \sim TN\left(\xi, \omega^2; (0, \infty)\right)$.

Proposition 7.3. *Let $\boldsymbol{Y}^o \sim MMNW_{p^o}(\boldsymbol{\xi}^o, \boldsymbol{\Sigma}^{oo}, \boldsymbol{\eta}^o, \nu)$ and W follows Weibull distribution with scale and shape parameters ν and 2, respectively. Then, $W_{\boldsymbol{y}^o}$ has pdf*

$$f_{W|\boldsymbol{Y}^o = \boldsymbol{y}^o}(w) = \frac{w \tau_{WE}^o}{A_{WE}^o \Phi(A_{WE}^o) + \phi(A_{WE}^o)} \phi\left(w; A_{WE}^o {\tau_{WE}^o}^{-1}, {\tau_{WE}^o}^{-2}\right), \quad w > 0.$$

where $\tau^{o^2}_{WE} = \tau^{o2}_{E} + 2\nu^2$ *and* $A^o_{WE} = \tau^{o^{-1}}_{WE}\left[\boldsymbol{\eta}^{o\top}\boldsymbol{\Sigma}^{oo-1}(\boldsymbol{y}^o - \boldsymbol{\xi}^o)\right]$. *Moreover, for* $k = 1, 2, ...,$

$$E(W^k_{\boldsymbol{y}^o}) = \frac{\tau^o_{WE}\Phi(A^o_{WE})}{A^o_{WE}\Phi(A^o_{WE}) + \phi(A^o_{WE})}E(V^{k+1}),$$

where $V \sim \mathcal{TN}\left(A^o_{WE}\tau^{o^{-1}}_{WE}, \tau^{o^{-2}}_{WE}; (0, \infty)\right)$.

Proposition 7.4. *If* $\boldsymbol{Y}^o \sim MMNG_{p^o}(\boldsymbol{\xi}^o, \boldsymbol{\Sigma}^{oo}, \boldsymbol{\eta}^o, \nu)$ *and* $W \sim Gamma(2, \nu)$. *Then, the pdf of* $W_{\boldsymbol{y}^o}$ *is*

$$f_{W_{\boldsymbol{y}^o}}(w) = \frac{\tau^{o\,2}_E w}{\sqrt{2\pi}(A^o_G\Phi(A^o_G) + \phi(A^o_G))}\exp\left(-\frac{(\tau^o_E w - A^o_G)^2}{2}\right), \quad w > 0,$$

where $A^o_G = \tau^{o\,-1}_E\left[\boldsymbol{\eta}^{o\top}\boldsymbol{\Sigma}^{oo-1}(\boldsymbol{y}^o - \boldsymbol{\xi}^o) - \nu\right]$. *Moreover,*

$$E(W^k_{\boldsymbol{y}^o}) = \frac{\tau^o\Phi(A^o_G)}{A^o_G\Phi(A^o_G) + \phi(A^o_G)}E(V^{k+1}),$$

where $V \sim TN\left(A^o_G\tau^{o\,-1}_E, \tau^{o\,-2}_E; (0, \infty)\right)$.

Proposition 7.5. *If* $\boldsymbol{Y}^o \sim MMNL_{p^o}(\boldsymbol{\xi}^o, \boldsymbol{\Sigma}^{oo}, \boldsymbol{\eta}^o, \nu)$ *and* $W \sim Lindley(\nu)$. *Then, the pdf of* $W_{\boldsymbol{y}^o}$ *is*

$$\begin{aligned} f_{W_{\boldsymbol{y}^o}}(w) &= \pi(\boldsymbol{y}^o)\frac{\phi\left(w; A^o_E\tau^{o\,-1}_E, \tau^{o\,-2}_E\right)}{\Phi(A^o_{_E})} \\ &\quad +(1 - \pi(\boldsymbol{y}^o))\frac{\tau^{o\,2}_E w\exp\left(-\frac{(\tau^o_E w - A^o_E)^2}{2}\right)}{\sqrt{2\pi}(A^o_E\Phi(A^o_E) + \phi(A^o_E))}, \quad w > 0, \end{aligned}$$

where

$$\pi(\boldsymbol{y}^o) = \frac{\nu f_{MMNE_{p^o}}(\boldsymbol{y}^o; \boldsymbol{\xi}^o, \boldsymbol{\Sigma}^{oo}_j, \boldsymbol{\eta}^o, \nu)}{(v+1)f_{MMNL_{p^o}}(\boldsymbol{y}^o; \boldsymbol{\xi}^o, \boldsymbol{\Sigma}^{oo}, \boldsymbol{\eta}^o, \nu)}.$$

Moreover,

$$E(W^r_{\boldsymbol{y}^o}) = \pi(\boldsymbol{y})E(V^r_1) + (1 - \pi(\boldsymbol{y}))E(V^r_2),$$

where $V_1 \sim TN\left(A^o_E\tau^{o\,-1}_E, \tau^{o\,-2}_E; (0, \infty)\right)$ *and*

$$E(V^r_2) = \frac{\tau^o_E\Phi(A^o_E)}{A^o_E\Phi(A^o_E) + \phi(A^o_E)}E(V^{k+1}),$$

in which

$$V \sim TN\left(A^o_E\tau^{o\,-1}_E, \tau^{o\,-2}_E; (0, \infty)\right).$$