


Estimating linear mixed effect models with non-normal random effects through saddlepoint approximation and its application in retail pricing analytics

Hao Chen, Lanshan Han & Alvin Lim


To cite this article: Hao Chen, Lanshan Han & Alvin Lim (2024) Estimating linear mixed effect models with non-normal random effects through saddlepoint approximation and its application in retail pricing analytics, Journal of Applied Statistics, 51:11, 2116-2138, DOI: [10.1080/02664763.2023.2260576](https://doi.org/10.1080/02664763.2023.2260576)

To link to this article: <https://doi.org/10.1080/02664763.2023.2260576>

 View supplementary material 

 Published online: 24 Sep 2023.

 Submit your article to this journal 

 Article views: 123

 View related articles 

 View Crossmark data 



Estimating linear mixed effect models with non-normal random effects through saddlepoint approximation and its application in retail pricing analytics

Hao Chen^a, Lanshan Han^a and Alvin Lim^{b,c}

^aRetail Solutions Research & Development, NielsenIQ, Chicago, IL, USA; ^bMeasured, Inc., Austin, TX, USA;

^cEmory University Goizueta Business School, Atlanta, GA, USA

ABSTRACT

Linear Mixed Effects (LME) models are powerful statistical tools that have been employed in many different real-world applications such as retail data analytics, marketing measurement, and medical research. Statistical inference is often conducted via maximum likelihood estimation with Normality assumptions on the random effects. Nevertheless, for many applications in the retail industry, it is often necessary to consider non-Normal distributions on the random effects when considering the unknown parameters' business interpretations. Motivated by this need, a linear mixed effects model with possibly non-Normal distribution is studied in this research. We propose a general estimating framework based on a saddlepoint approximation (SA) of the probability density function of the dependent variable, which leads to constrained nonlinear optimization problems. The classical LME model with Normality assumption can then be viewed as a special case under the proposed general SA framework. Compared with the existing approach, the proposed method enhances the real-world interpretability of the estimates with satisfactory model fits.

ARTICLE HISTORY

Received 5 September 2022
Accepted 4 September 2023

KEYWORDS

Mixed effects model; linear regression; constrained optimization; statistical inference; saddlepoint approximation


MATHEMATICAL SUBJECT CLASSIFICATION

62J05

1. Introduction

In retail analytics, linear mixed effects models are prevalent statistical models to quantify the association between the dependent variable and the independent variables, especially when it is necessary to account for the possible heterogeneity among different subgroups. Consider an application in retail analytics where the goal is to establish the relationship between a product's sales volume and price using regression. The established relationship is known as the price elasticity of demand (PED), which characterizes how sales volume of a product changes with varying price [23]. It is usually assumed that PED is negative for consumer packaged goods (CPG) [6] that are sold in a grocery store, representing a common belief in the retail industry that sales volume is negatively associated with the price of a product. In this application, since the same product is sold in different stores, *store*

CONTACT Hao Chen  chenhaonelson2018@gmail.com  Retail Solutions Research & Development, NielsenIQ, Chicago, IL 60606, USA

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/02664763.2023.2260576>.

© 2023 Informa UK Limited, trading as Taylor & Francis Group

becomes a natural grouping factor, for which the possible variation needs to be considered. In what follows, we present a simplified real-world case to illustrate the motivation of this research.

A sales dataset on 1 liter regular whole milk from a retail chain was available to us. The dataset consists of weekly sales price and corresponding weekly sales volume of the product from 6 different stores, and each store has data for the whole year of 2021 (53 weeks in total). Given the available data, we are interested in finding the association between volume and price, while taking the variation across the 6 different stores into account. To best illustrate the motivation, all other potential factors affecting sales such as holidays, events, seasonality and the possible temporal effect are ignored for now to focus on a simplified case. A linear mixed effects model with implementation from the R package *lme4* [3] was fitted on the data with the following model specification.

$$y_{i,t} = (\alpha + \alpha_i) + (\beta + \beta_i)x_{i,t} + \epsilon_{i,t}, \quad (1)$$

where i is the store index, i.e. $i = 1, \dots, 6$, and is the grouping factor for both the random intercepts and the random slopes, t is the individual week index, i.e. $t = 1, \dots, 53$, y is the natural logarithm of the sales volume, x is the natural logarithm of the sales price. Both the random intercepts (α_i) and random slopes (β_i) are assumed to follow independent Normal distributions, and the error ϵ follows a Normal distribution as well under the canonical model specification of a LME model. After the model was fitted, the two fixed effects are as follows: $\hat{\alpha} = 2.671$ and $\hat{\beta} = -0.055$. The 6 random slopes are $\hat{\beta}_1 = 0.080$, $\hat{\beta}_2 = -0.034$, $\hat{\beta}_3 = -0.059$, $\hat{\beta}_4 = 0.103$, $\hat{\beta}_5 = -0.053$, $\hat{\beta}_6 = -0.036$. If we further define the overall slope of store i as $\beta + \beta_i$, we observe that for store 1 and store 4, the overall slopes are 0.025 and 0.048, respectively, i.e. both are above 0. However, as the real world interpretation of the overall slope of store i is the product's PED for that store, obviously the values for store 1 and store 4 are not meaningful for direct use in practice, and some ad hoc procedures are entailed to 'correct' the sign of the overall slopes before the modeling results are treated as reasonable.

In retail analytics, fitting hundreds or even thousands of independent models within a tight delivery timeline is common. While it's possible to investigate the causes of wrong signs manually when working with a few models, this approach is not feasible at scale. However, relying solely on classical linear mixed-effects (LME) models with a Normal assumption on the random effects can lead to unbounded overall coefficients [8]. This violates reasonable business interpretations that require coefficients to be bounded. One solution is to assume a bounded non-Normal distribution on the random effects, but this can pose significant technical challenges, as the likelihood function of an LME model cannot be derived analytically without the Normality assumption. This represents a major hurdle for maximum likelihood (ML) approaches, which typically require an analytical form of the likelihood function. To overcome this challenge, an ML approach for LME models with non-Normally distributed random effects must be developed.

To overcome this hurdle, we propose to approximate the likelihood function using the saddlepoint approximation (SA) [10], which provides an accurate point-wise approximation of a probability density function (PDF) using its moment generating function (MGF). The SA approach has been successfully used in many circumstances where an analytical form of the PDF of the random variables of interest is not readily available. However, to the

best of our knowledge, it has not been used under the ML paradigm for statistical inference. In fact, the SA approach does not directly provide an analytical approximation of the likelihood function. Instead, the SA approach defines an implicit function as an approximation of the PDF of interest. To use this implicit function in the maximum likelihood paradigm, we propose a novel approach to include the defining nonlinear equations as constraints, leading to constrained nonlinear optimization problems. With the recent advances in optimization theory and algorithms, we demonstrate in this paper that the resulting optimization problems can be solved efficiently and therefore produce high quality estimations on both simulated data sets and real-world data sets under varying assumptions on the distribution of the random effects. In particular, we examine the proposed approach under assumptions of four different non-Normal distributions: (1) Uniform, (2) Laplace, (3) Gamma, and (4) Triangular.

The classical LME models have been successfully applied in many areas. There is a vast body of statistical literature covering its theoretical properties and computational implementations. The mathematical properties were carefully studied and presented by Jiang [13]. Details about the computational aspects were reported by Lindstrom and Bates [16]. In a more recent paper, Bates et al. [2] implemented *lme4* – a widely used R package for fitting a linear mixed effects model. On the other hand, there have been studies on the LME model with non-Normal random effects in the literature as well. Instead of the Normal distribution, the multivariate t-distribution was studied by Pinheiro et al. [22]. Working with synthetic data, Yucel and Demirtas [29] researched the influence of non-Normal random effects on the model parameter estimation. However, the Normality assumption on random effects was still employed when estimating the unknown parameters. Moreover, Lin and Lee [15] extended the classical LME model by employing a multivariate skew-Normal assumption on the random effects. Matos et al. [17] considered multivariate t distribution with censored response, and likelihood-based approach was proposed to conduct statistical inference. Among all the pertinent literature, the following three more relevant points are worthy of further discussions.

First, Verbeke and Lesaffre [26] showed that the maximum likelihood estimators are consistent for estimating the fixed effects and variance component even when the distribution of random effects is misspecified. McCulloch and Neuhaus [19] considered the robustness to the assumed distribution on a random intercept, and argued that the misspecification of the distribution has only minor impact, where the inferences for within-group covariates are robust to the misspecification. However, it is our argument that the business interpretation on the variables of interest must be taken into account as well. In the motivating example, if, for example, a Uniform distribution was assumed on the random effects instead of the (unconstrained) Normal distribution, we then would ensure the needed negativity of the overall coefficients of the PED for store 1 and store 4, which is of great practical importance in retail analytics.

Second, a LME model can also be viewed as a special case of the latent variables model proposed by Skrondal and Rabe-Hesketh [25], where random effects are treated as latent variables. Statistical inference is then conducted by maximizing the marginal likelihood function after integrating out the latent variables. In most of the non-toy examples, methods for numerical integration such as Monte Carlo integration [12] and Gaussian-Hermite quadrature [11] are entailed to approximate the marginal likelihood. The EM algorithm is another option to iteratively maximize the likelihood. For example, it is utilized by Mattos

et al. [18] to estimate nonparametric functions using smoothing splines in the context of linear mixed models for longitudinal censored data. Considering a LME model as a special case of the latent variable model conceptually can handle non-Normal random effects, since random effects are viewed as latent variables that are integrated out.

Moving forward with the idea of latent variables, Nelson et al. [20] proposed the probability integral transformation (PIT) method, which utilizes the fact that a non-Normal realization can be converted from a Normal realization using the inverse of its cumulative distribution function (CDF). Therefore, one can presumably assume any continuous distribution on the random effects as long as the inverse of its CDF exists, and inference is then conducted against the marginal likelihood by numerically integrating out the random effects with Gaussian-Hermite quadrature. Compared to the PIT approach, the proposed SA method is expected to be numerically more stable especially when sample size becomes large. A direct comparison between the proposed SA and PIT is discussed further in Section 4.4 and in Section 5, in which the motivating example was revisited and comments were made.

Third, in a more recent paper [8], the authors proposed to estimate the model parameters when random effects follow a truncated Normal distribution. Although the motivation bears similarity at first glance, the main approach and the estimating framework are very different from the SA approach. Since the exact probability density function of the responses is intractable, the theory proposed by Chen et al. [8] was established primarily based on the idea that as the number of truncated Normally distributed random effects becomes large, the distribution of the responses can be approximated by a Normal distribution. However, the distribution of the responses is approximated by the SA approach in this paper. Moreover, the approach based on the SA approximation is a general framework as mentioned before, i.e. it is applicable as long as the MGF of the distribution assumed on the random effects exists. The objective here is not to substitute for either the PIT approach [20] or the approach of Chen et al. [8], but rather to provide practitioners with more flexibility and choices when facing the need for dealing with non-Normal random effects.

The rest of the paper is organized as follows. In Section 2, we provide preliminaries regarding LME models and SA approaches. The proposed estimation method is detailed in Section 3. Some simulation results are presented in Section 4. We then revisit the real world example in Section 5. Some concluding remarks are made in Section 6.

2. Preliminaries

We first introduce the notations used in this paper. We use lower and upper case letters to represent scalars or (one dimensional) random variables, bold lower case letters to represent vectors or (multiple dimensional) random variables, and bold upper case letters to represent matrices. All vectors are assumed to be column vectors. We index vectors and matrices by superscripts and scalars by subscripts. We denote the n -dimensional real vector space as \mathbb{R}^n , with its nonnegative orthant denoted by \mathbb{R}_+^n . For a vector $\mathbf{v} \in \mathbb{R}^n$, we denote its i th element by v_i and its transpose by \mathbf{v}^T . For a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, we denote its (i, j) element by a_{ij} , its transpose by \mathbf{A}^T , and its determinant by $|\mathbf{A}|$. The all zero vector and all one vector of length n are denoted by $\mathbf{0}_n$ and $\mathbf{1}_n$, respectively. The $n \times n$ identity matrix is denoted by \mathbf{I}_n . We use $\|\mathbf{v}\|_2$ to denote the two-norm of a vector $\mathbf{v} \in \mathbb{R}^n$, i.e. $\|\mathbf{v}\|_2 = \sqrt{\mathbf{v}^T \mathbf{v}}$. Given a function $g: \mathbb{R}^n \mapsto \mathbb{R}$, we let $\nabla g(\mathbf{x})$ be its gradient evaluated at \mathbf{x} . We use $\mathcal{N}(\mu, \sigma^2)$

to represent a Normal distribution with mean μ and variance σ^2 and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to represent a multivariate Normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. We use the relation \sim to indicate that a random variable follows a certain distribution.

2.1. Linear mixed effects models

Working under the classical LME model, we use $\ell = 1, \dots, g$ to denote the grouping factor. For each group ℓ , the dependent variable y^ℓ is assumed to linearly dependent on the independent variables and the error term follows a Normal distribution with mean 0 and variance σ^2 that is unknown. Let n denote the total sample size and n_ℓ is the size for group ℓ such that $\sum_{\ell=1}^g n_\ell = n$. Let p denote the dimension of the design matrix, and k is the number of variables that random effects are considered. When $k = 0$, the LME model then reduces to the ordinary linear regression model. Mathematically, the LME model is presented as follows:

$$\mathbf{y}^\ell = \mathbf{X}^\ell \boldsymbol{\beta} + \mathbf{Z}^\ell \boldsymbol{\gamma}^\ell + \boldsymbol{\varepsilon}^\ell, \quad (2)$$

where

$$\begin{aligned} \mathbf{y}^\ell &\triangleq \begin{bmatrix} y_{\ell,1} \\ \vdots \\ y_{\ell,n_\ell} \end{bmatrix} \in \mathbb{R}^{n_\ell}, \quad \boldsymbol{\beta} \triangleq \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \in \mathbb{R}^p, \quad \boldsymbol{\gamma}^\ell \triangleq \begin{bmatrix} \gamma_{\ell,1} \\ \vdots \\ \gamma_{\ell,k} \end{bmatrix} \in \mathbb{R}^k, \\ \mathbf{X}^\ell &\triangleq \begin{bmatrix} x_{\ell,1,1} & x_{\ell,1,2} & \cdots & x_{\ell,1,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{\ell,n_\ell,1} & x_{\ell,n_\ell,2} & \cdots & x_{\ell,n_\ell,p} \end{bmatrix} \in \mathbb{R}^{n_\ell \times p}, \\ \mathbf{Z}^\ell &\triangleq \begin{bmatrix} z_{\ell,1,1} & z_{\ell,1,2} & \cdots & z_{\ell,1,k} \\ \vdots & \vdots & \vdots & \vdots \\ z_{\ell,n_\ell,1} & z_{\ell,n_\ell,2} & \cdots & z_{\ell,n_\ell,k} \end{bmatrix} \in \mathbb{R}^{n_\ell \times k}, \\ \text{and } \boldsymbol{\varepsilon}^\ell &\triangleq \begin{bmatrix} \varepsilon_{\ell,1} \\ \vdots \\ \varepsilon_{\ell,n_\ell} \end{bmatrix} \in \mathbb{R}^{n_\ell}, \end{aligned}$$

where the symbol \triangleq means *defined to be equal to*. We often refer to $\beta_j, j = 1, \dots, p$, as the fixed effect coefficients, $\gamma_{\ell,j}, \ell = 1, \dots, g; j = 1, \dots, k$, the random effect coefficients, and $\boldsymbol{\varepsilon}^\ell \sim \mathcal{N}(\mathbf{0}_{n_\ell}, \sigma^2 \mathbf{I}_{n_\ell})$, the error vector with σ^2 unknown. Under the classical model specification, a multivariate Normal distribution is assumed on the random effects as follows.

$$\boldsymbol{\gamma}^\ell \stackrel{iid}{\sim} \mathcal{N}(\mathbf{0}_k, \boldsymbol{\Sigma}), \quad \forall \ell = 1, \dots, g, \quad (3)$$

where Σ is typically a structured $k \times k$ covariance matrix that parameterized by some unknown parameters. Many different structures of Σ have been considered in the literature [28]. The independent structure,

$$\Sigma = \begin{bmatrix} \varsigma_1^2 & & \\ & \ddots & \\ & & \varsigma_k^2 \end{bmatrix}$$

is assumed in this paper, where $\varsigma_1, \dots, \varsigma_k$ require estimation in practice. In other words, for each $\ell = 1, \dots, g$:

$$\gamma_{\ell,j} \stackrel{iid}{\sim} \mathcal{N}(0, \varsigma_j^2), \quad \forall j = 1, \dots, k. \quad (4)$$

Stacking up the data from the g groups, we then have

$$\begin{aligned} \mathbf{y} &\triangleq \begin{bmatrix} \mathbf{y}^1 \\ \vdots \\ \mathbf{y}^g \end{bmatrix} \in \mathbb{R}^n, \quad \boldsymbol{\gamma} \triangleq \begin{bmatrix} \boldsymbol{\gamma}^1 \\ \vdots \\ \boldsymbol{\gamma}^g \end{bmatrix} \in \mathbb{R}^{kg}, \quad \mathbf{X} \triangleq \begin{bmatrix} \mathbf{X}^1 \\ \vdots \\ \mathbf{X}^g \end{bmatrix} \in \mathbb{R}^{n \times p}, \\ \text{and } \mathbf{Z} &\triangleq \begin{bmatrix} \mathbf{Z}^1 & & \\ & \ddots & \\ & & \mathbf{Z}^g \end{bmatrix} \in \mathbb{R}^{n \times kg}, \quad \boldsymbol{\varepsilon} \triangleq \begin{bmatrix} \boldsymbol{\varepsilon}^1 \\ \vdots \\ \boldsymbol{\varepsilon}^g \end{bmatrix} \in \mathbb{R}^n, \end{aligned}$$

With the above definitions, we rewrite the model into a succinct expression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad (5)$$

where

$$\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}_{nk}, \mathbf{G}), \quad \mathbf{G} = \begin{bmatrix} \Sigma & & \\ & \ddots & \\ & & \Sigma \end{bmatrix}$$

and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{R})$ with $\mathbf{R} = \sigma^2 \mathbf{I}_n$. In addition, as $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are independent, we have

$$\begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}\right).$$

Thanks to the Normality assumption on the random effects, \mathbf{y} is multivariate Normally distributed, that is, the exact distribution of \mathbf{y} is tractable. Therefore, either the maximum likelihood estimation or restricted maximum likelihood estimation can be utilized to conduct statistical inference, see e.g. [8,30] for a discussion on these technical details. The above classical approach has been implemented in popular statistical packages such as the *statsmodels* module [24] in Python and the *lme4* library [3] in R.

2.2. Saddlepoint approximation

The saddlepoint approximation method, proposed by Daniels [10] provides an accurate pointwise approximation formula for the probability density function (PDF) of a distribution based on its MGF. In general, given a random variable w with PDF $f_w(w)$, the MGF

$M_w(t)$ and cumulant generating function (CGF) $K_w(t)$ are defined as

$$M_w(t) \triangleq \mathbb{E}[e^{tw}], \quad \text{and} \quad K_w(t) \triangleq \ln(M_w(t)),$$

respectively. Then, the saddlepoint approximation of $f_w(w)$ at any given w is given by:

$$\hat{f}_w(w) = \frac{1}{\sqrt{2\pi K_w''(t^*)}} \exp(K_w(t^*) - t^*w) \quad (6)$$

and t^* satisfies:

$$K_w'(t^*) = w, \quad (7)$$

where $K_w'(t)$ and $K_w''(t)$ are first and second derivatives of $K_w(t)$, respectively. The saddlepoint approximation in Equation (6) provides a convenient analytical formula for pointwise approximation to the density function of the random variable w . This is particularly relevant when we examine the distribution of the (weighted) sum of a finite set of random variables. In fact, let $s = \sum_{i=1}^d o_i w_i$, where w_i are independent to each other, but not necessarily identically distributed, and o_i are known constants (weights). Essentially, s is the weighted sum of d independent random variable w_i . Assume the MFG of each w_i is $M_{w_i}(w)$. The MFG and CGF of s are then given by

$$M_s(t) = \prod_{i=1}^d M_{w_i}(o_i t) \quad \text{and} \quad K_s(t) = \sum_{i=1}^d K_{w_i}(o_i t).$$

To obtain the exact PDF of s , one needs to use convolution, which is usually intractable unless w_i conveniently follows a Normal distribution. On the other hand, the saddlepoint approximation for the density of s is given by

$$\hat{f}_s(s) = \frac{1}{\sqrt{2\pi K_s''(t^*)}} \exp(K_s(t^*) - t^*s), \quad (8)$$

with t^* being the solution to

$$K_s'(t^*) = \sum_{i=1}^d o_i K_{w_i}'(o_i t^*) = h. \quad (9)$$

Equations (8) and (9) together provide a valuable approximation formula for the density of $s = \sum_{i=1}^d o_i w_i$ at each point. Note that under proper regularity conditions, saddlepoint Equation (7) defines an implicit function $t^*(w)$, based on which we can define the *saddlepoint density*:

$$\hat{f}_w(w) = \frac{1}{\sqrt{2\pi K_w''(t^*(w))}} \exp(K_w(t^*(w)) - t^*(w)w). \quad (10)$$

Note that the saddlepoint density function may need to be normalized to become a proper density function. Due to the involvement of an implicit function, which often does not have a closed form, it is not straightforward to use (10) directly in a maximum likelihood

paradigm. We propose the inclusion of the saddlepoint Equation (7) as an equality constraint in the maximization of the likelihood function. This proposed approach is detailed in the next section.

In addition, Equation (10) will then reduce to the Normal density when it follows a Normal distribution. In other words, the density for a Normal distribution can be viewed as a special case of saddlepoint approximation, and it is no longer an approximation.

3. The proposed estimation method

We consider a mixed effects model involving k random effect coefficients, each of which follows a distribution $\mathcal{F}_j(\boldsymbol{\theta}^j)$, parameterized by unknown parameters $\boldsymbol{\theta}^j$, with its PDF, MGF, and CGF given by $\phi_j(\gamma; \boldsymbol{\theta}^j)$, $M_j(t; \boldsymbol{\theta}^j)$, and $K_j(t; \boldsymbol{\theta}^j)$, respectively with $\boldsymbol{\theta}^j$ unknown. Specifically, for the ℓ th cluster, with the cluster index ℓ being dropped for conciseness, we have

$$y_i = \sum_{j=1}^p x_{ij}\beta_j + \sum_{j=1}^k z_{ij}\gamma_j + \varepsilon_i, \quad i = 1, \dots, n_\ell, \quad (11)$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\gamma_j \sim \mathcal{F}_j(\boldsymbol{\theta}^j)$. Let $f(y_{\ell,i}; \boldsymbol{\beta}, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^k, \sigma^2)$ be the PDF of the i th observation in cluster ℓ . Assuming $\varepsilon_{\ell,i}$ are independent to each other, and $\gamma_{\ell,j}$ are independent to each other as well as to $\varepsilon_{\ell,i}$, the joint density function of $\mathbf{y} = (y_{\ell,i})_{\ell=1, \dots, g; i=1, \dots, n_\ell}$ is given by

$$\prod_{\ell=1}^g \prod_{i=1}^{n_\ell} f(y_{\ell,i}; \boldsymbol{\beta}, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^k, \sigma^2). \quad (12)$$

When $\mathcal{F}_j(\boldsymbol{\theta}^j)$ does not represent a Normal distribution, it is difficult to derive the analytical form for $f(y_{\ell,i}; \boldsymbol{\beta}, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^k, \sigma^2)$. We therefore resort to the SA method to find a good approximation of $f(y_{\ell,i}; \boldsymbol{\beta}, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^k, \sigma^2)$ at each observation of the dependent variable $y_{\ell,i}$, assuming the existence of MGF and CGF. According to (8), we have

$$f(y_{\ell,i}; \boldsymbol{\beta}, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^k, \sigma^2) \approx \frac{1}{\sqrt{2\pi K''_{\ell,i}(t_{\ell,i})}} \exp(K_{\ell,i}(t_{\ell,i}) - t_{\ell,i}y_{\ell,i}), \quad (13)$$

where $t_{\ell,i}$ is the solution of $K'_{\ell,i}(t_{\ell,i}) = y_{\ell,i}$ and

$$\begin{aligned} K_{\ell,i}(t_{\ell,i}) &= \ln \left(\exp \left(t_{\ell,i} \sum_{j=1}^p x_{\ell,i,j} \beta_j \right) \times \left(\prod_{j=1}^k M_{\ell,j}(z_{\ell,i,j} t_{\ell,i}) \right) \times (M_{\varepsilon_{\ell,i}}(t_{\ell,i}; \sigma^2)) \right) \\ &= t_{\ell,i} \left(\sum_{j=1}^p x_{\ell,i,j} \beta_j \right) + \sum_{j=1}^k K_j(z_{\ell,i,j} t_{\ell,i}; \boldsymbol{\theta}^j) + \frac{1}{2} \sigma^2 t_{\ell,i}^2. \end{aligned} \quad (14)$$

The last term in above equation is from the CGF of the Normal distribution $\mathcal{N}(0, \sigma^2)$. With Equation (13), the log-likelihood can be approximated by

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^k, \sigma^2 | \mathbf{y}_{\ell,i}) \approx \sum_{\ell=1}^g \sum_{i=1}^{n_\ell} \ln \left(\frac{1}{\sqrt{2\pi K''_{\ell,i}(t_{\ell,i})}} \exp(K_{\ell,i}(t_{\ell,i}) - t_{\ell,i}y_{\ell,i}) \right). \quad (15)$$

When random effects follow a Normal distribution, Equation (15) is no longer an approximation, i.e. it is same as the log-likelihood derived with Normally distributed random effects. In other words, the log-likelihood of a regular LME model with Normally distributed random effects can be viewed as a special case of Equation (15). With Equation (15), we propose to estimate the fixed effects, $\beta, \theta^1, \dots, \theta^k, \sigma^2$ by solving the following optimization problem.

$$\begin{aligned} \max_{\beta, \theta^1, \dots, \theta^k, \sigma^2} \quad & \sum_{\ell=1}^g \sum_{i=1}^{n_\ell} \left[-\frac{1}{2} \ln(K''_{\ell,i}(t_{\ell,i})) + K_{\ell,i}(t_{\ell,i}) - t_{\ell,i} y_{\ell,i} \right] \\ \text{s.t.} \quad & K'_{\ell,i}(t_{\ell,i}) = y_{\ell,i}, \quad \forall \ell = 1, \dots, g; i = 1, \dots, n_\ell. \end{aligned} \quad (16)$$

Equation (16) is typically nonlinear and nonconvex, which is, in general, challenging to solve to global optimality. However, there are efficient modern optimization algorithms, such as sequential quadratic programming (SQP) [4] and alternating direction method of multipliers (ADMM) [5] to find local optimal solutions or stationary points. For practitioners, if computational resources permit, it is highly recommended to consider both and see if they lead to similar estimated coefficients as well as similar log-likelihood values or not. In addition, we can have multiple runs of an algorithm with different initial solutions and then choose the one with the best objective value. For all the simulated examples that have been studied in Section 4, satisfactory results can be obtained by having 5 different initial values. Empirically, it is recommended to consider at least that amount of different starting points to ensure the optimization algorithm ends with an acceptable solution for both the simulated examples and the real-world examples.

Once a solution of (16), denoted by $\hat{\beta}, \hat{\theta}^1, \dots, \hat{\theta}^k, \hat{\sigma}^2$, collectively written as $\hat{\Phi}$ is obtained, we can further estimate the random effect coefficients γ^ℓ 's. The joint likelihood function of \mathbf{y} and $\boldsymbol{\gamma}$ conditional on $\hat{\Phi}$ is:

$$\begin{aligned} f(\mathbf{y}, \boldsymbol{\gamma} | \hat{\Phi}) &= f_{\mathbf{y}}(\mathbf{y} | \boldsymbol{\gamma}, \hat{\Phi}) f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma} | \hat{\Phi}) \\ &\propto \prod_{\ell=1}^g \exp \left(-\frac{1}{2\hat{\sigma}^2} \left(\mathbf{y}^\ell - \mathbf{X}^\ell \hat{\beta} - \mathbf{Z}^\ell \boldsymbol{\gamma}^\ell \right)^T \left(\mathbf{y}^\ell - \mathbf{X}^\ell \hat{\beta} - \mathbf{Z}^\ell \boldsymbol{\gamma}^\ell \right) \right) \\ &\quad \times \prod_{\ell=1}^g \prod_{i=1}^{n_\ell} \prod_{j=1}^k \phi_j(\gamma_{\ell,i,j}; \theta^j). \end{aligned} \quad (17)$$

Conditional on the estimates of fixed effects coefficients, we therefore propose to estimate the random effects coefficients $\gamma_{\ell,j}$, $\ell = 1, \dots, g, j = 1, \dots, k$, by solving the following optimization problem.

$$\begin{aligned} &(\hat{\boldsymbol{\gamma}}^1, \dots, \hat{\boldsymbol{\gamma}}^g) \\ &= \min_{\boldsymbol{\gamma}^1, \dots, \boldsymbol{\gamma}^g} \left[\sum_{\ell=1}^g \frac{1}{\hat{\sigma}^2} \left(\mathbf{y}^\ell - \mathbf{X}^\ell \hat{\beta} - \mathbf{Z}^\ell \boldsymbol{\gamma}^\ell \right)^T \left(\mathbf{y}^\ell - \mathbf{X}^\ell \hat{\beta} - \mathbf{Z}^\ell \boldsymbol{\gamma}^\ell \right) \right. \\ &\quad \left. - 2n_\ell \sum_{j=1}^k \ln \left(\phi_j \left(\gamma_{\ell,j}; \hat{\theta}^j \right) \right) \right]. \end{aligned} \quad (18)$$

Note that (18) is decomposable and can be solved for each \boldsymbol{y}^ℓ individually as follows.

$$\hat{\boldsymbol{y}}^\ell = \min_{\boldsymbol{y}^\ell} \left[\frac{1}{\widehat{\sigma}^2} \left(\boldsymbol{y}^\ell - \boldsymbol{X}^\ell \widehat{\boldsymbol{\beta}} - \boldsymbol{Z}^\ell \boldsymbol{y}^\ell \right)^T \left(\boldsymbol{y}^\ell - \boldsymbol{X}^\ell \widehat{\boldsymbol{\beta}} - \boldsymbol{Z}^\ell \boldsymbol{y}^\ell \right) - 2n_\ell \sum_{j=1}^k \ln \left(\phi_j \left(\gamma_{\ell,j}; \widehat{\boldsymbol{\theta}}^j \right) \right) \right]. \quad (19)$$

Now, we have completed the estimation process for both the fixed effects $(\boldsymbol{\beta}, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^k, \sigma^2)$ and the random effects $(\boldsymbol{y}^1, \dots, \boldsymbol{y}^g)$.

Remark 3.1: Since we apply the saddlepoint approximation of the joint density function (12), the Normality assumption on the error term ε_i in (11) is in fact not essential. The proposed framework can handle non-Normal errors as well, as long as the distribution of the error term possesses an MGF and a CGF. We only consider the traditional Normal error assumption in this paper for simplicity. Moreover, using the proposed approach, the random effects do not have to follow the same distribution, providing great flexibility in modeling different factors contributing to the dependent variable. For simplicity, we only demonstrate the cases where the random effects follow the same distribution in this paper.

To demonstrate the proposed approach, in the rest of this section, we study a special case where all the random effects follow the same distribution, but possibly with different parameters (unknown). More specifically, we assume that for each $j = 1, \dots, k$

$$\gamma_{\ell,j} \stackrel{iid}{\sim} \mathcal{F}(\boldsymbol{\theta}^j), \quad \forall \ell = 1, \dots, g.$$

We specifically consider four different distributions in this paper: (1) Uniform, (2) Laplace [14], (3) Gamma, and (4) Triangular. The equations representing the Uniform distribution are given in the next paragraph. The equations for the other three distributions are derived in detail in Section A of the *Supplemental Document*.

Assume for each $j = 1, \dots, k$

$$\gamma_{\ell,j} \stackrel{iid}{\sim} U(-|\beta_j|, |\beta_j|), \quad \forall \ell = 1, \dots, g,$$

where $U(\cdot)$ stands for a Uniform distribution, β_j is the corresponding fixed effect coefficient. In this way, the random effect coefficient is bounded by the absolute value of β_j . The overall coefficient $(\beta_j + \gamma_{\ell,j})$ is restricted to within $(0, 2\beta_j)$ if $\beta_j > 0$ or $(2\beta_j, 0)$ with $\beta_j < 0$, for all $\ell = 1, \dots, g$ and $j = 1, \dots, k$. We can derive the following:

$$\begin{aligned} K_{\ell,i}(t_{\ell,i}) &= t_{\ell,i} \left(\sum_{j=1}^p x_{\ell,i,j} \beta_j \right) + \left(\sum_{j=1}^k \ln \left(\frac{e^{\beta_j z_{\ell,i,j} t_{\ell,i}} - e^{-\beta_j z_{\ell,i,j} t_{\ell,i}}}{2\beta_j z_{\ell,i,j} t_{\ell,i}} \right) \right) + \left(\frac{1}{2} \sigma^2 t_{\ell,i}^2 \right), \\ K'_{\ell,i}(t_{\ell,i}) &= \left(\sum_{j=1}^p x_{\ell,i,j} \beta_j \right) + \left(\sum_{j=1}^k \frac{1 + \beta_j z_{\ell,i,j} t_{\ell,i} + e^{2\beta_j z_{\ell,i,j} t_{\ell,i}} (\beta_j z_{\ell,i,j} t_{\ell,i} - 1)}{(e^{2\beta_j z_{\ell,i,j} t_{\ell,i}} - 1) t_{\ell,i}} \right) + \sigma^2 t_{\ell,i}, \\ K''_{\ell,i}(t_{\ell,i}) &= \sum_{j=1}^k \frac{1 + e^{4\beta_j z_{\ell,i,j} t_{\ell,i}} - 2e^{2\beta_j z_{\ell,i,j} t_{\ell,i}} (2\beta_j^2 z_{\ell,i,j}^2 t_{\ell,i}^2 + 1)}{(e^{2\beta_j z_{\ell,i,j} t_{\ell,i}} - 1)^2 t_{\ell,i}^2} + \sigma^2 \end{aligned}$$

and $\beta_j z_{\ell,ij} t_{\ell,i} \neq 0$ for $\ell = 1, \dots, g; i = 1, \dots, n_\ell; j = 1, \dots, k$. To estimate the fixed effects, we solve the following optimization problem

$$\begin{aligned} \max_{\beta, t, \sigma^2} \quad & \sum_{\ell=1}^g \sum_{i=1}^{n_\ell} \left[-\frac{1}{2} \ln(K''_{\ell,i}(t_{\ell,i})) + K_{\ell,i}(t_{\ell,i}) - t_{\ell,i} y_{\ell,i} \right] \\ \text{subject to} \quad & K'_{\ell,i}(t_{\ell,i}) = y_{\ell,i}, \quad \forall \ell = 1, \dots, g; i = 1, \dots, n_\ell. \end{aligned}$$

Remark 3.2: Technically, we also need to include an inequality constraint $\beta_j z_{\ell,ij} t_{\ell,i} \neq 0$, for all $\ell = 1, \dots, g; i = 1, \dots, n_\ell; j = 1, \dots, k$ to make sure that the denominators are not 0. However, this kind of constraint is typically not easy to include in an optimization problem. Moreover, the constraint only removes a measure zero set in the $\beta - t$ space. Therefore, for practical reasons, we can ignore this constraint. Our numerical experiments also show that ignoring this constraint does not prevent the optimization algorithms from completing successfully.

For the random effects estimation, we have

$$\ln(\phi(\gamma_{\ell,j} | \hat{\theta}^j)) = \ln \left(\frac{1}{2\hat{\beta}_j} \right),$$

which is independent of the random effects coefficients γ^ℓ 's. Therefore the estimation of individual γ^ℓ can be obtained by solving the following minimization problem

$$\hat{\gamma}^\ell = \min_{\gamma^\ell} \left(\frac{1}{\hat{\sigma}^2} (\gamma^\ell - X^\ell \hat{\beta} - Z^\ell \gamma^\ell)^T (\gamma^\ell - X^\ell \hat{\beta} - Z^\ell \gamma^\ell) \right),$$

which is a linear regression and has a closed form solution.

We visualize the distributions of the four distributions in Figure 1. From (a) and (d) of Figure 1, we note that both the Uniform distribution and the Triangular distribution have a lower bound and an upper bound. Hence, it is feasible to assume either on the random

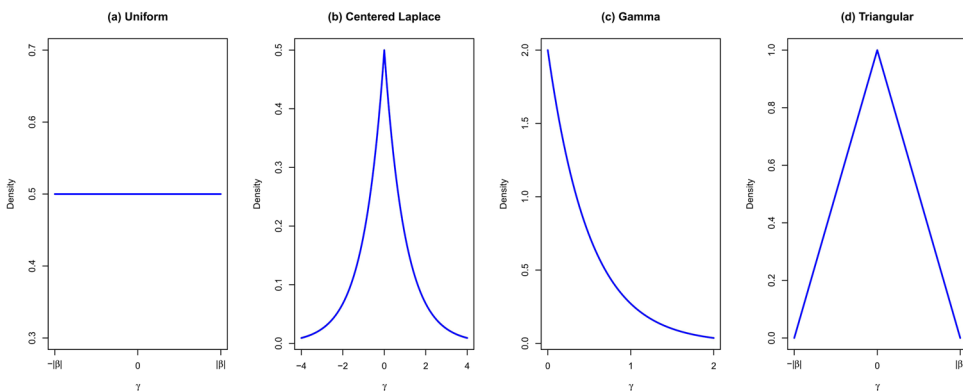


Figure 1. (a) is the PDF of a Uniform distribution that is bounded by $-\beta$ and β ; (b) the PDF of a Centered Laplace distribution with scale 1 for illustrative purpose; (c) is a Gamma PDF with $a = 1, b = 2$ for illustrative purpose; (d) is the PDF of a Triangular distribution that is bounded by $-\beta$ and β .

effects when sign constraints are needed for the overall regression coefficients. The difference lies in the underlying belief: The Uniform distribution corresponds to no preference on the magnitude of the random effect coefficients, while the Triangular distribution corresponds to a preference that the magnitude is closer to 0. In fact, the Triangular distribution is based on some knowledge of the minimum, maximum and a preference of the most likely value. From (c) of Figure 1, we see that samples from the Gamma distribution are strictly positive ranging from 0 to infinity. Hence, if domain knowledge suggests that all random effects should be greater than 0, then the Gamma distribution is a natural choice. Similar to the Normal distribution, the Centered Laplace distribution in (b) of Figure 1 is unconstrained and ranges from negative infinity to positive infinity, which can be considered as an alternative to the Normal distribution. The above four distributions cover all the scenarios for the sign of the overall effects.

4. Simulation with synthetic data

We conduct simulation studies in this section. Results for a LME model with random intercept only is presented in Section 4.1. We then detail results for the same model with both random intercept and one random slope in Section 4.2. All the optimization problems involved were solved by an interior point algorithm described in [7,21] and implemented within the SciPy package [27] optimization module. This algorithm is a rather sophisticated general purpose nonlinear optimization algorithm, featuring a logarithmic barrier function with adaptive barrier parameter, a linearly constrained convex quadratic approximation at each iteration, and a trust region with adaptive radius. To solve the optimization problems for fixed effects estimation, we started with 5 different initial solutions and the one with the largest log-likelihood value is retained. This is a typical strategy for solving nonconvex optimization problems. As we have shown earlier, the random effects estimation problems are often convex and hence were solved with only 1 initial solution.

4.1. Models with random intercept

Datasets are simulated according to the statistical assumption of an LME model with random intercept only. We detail the specific procedures about data generation in Section A of the *Supplemental Document*. The choice of dimension is $p = 5$ and $p = 8$, and the sample sizes are $n = 100, 200$, and 500 . Each sample size is then equally divided into $g = 20$ clusters. Therefore, we have $2 \times 3 = 6$ combinations of different dimensions and sample sizes. For each combination, we simulate 10 different datasets aiming to check the variation in the data-generating process. The true parameters used to simulate the datasets and more simulation details are reported in Section A of the *Supplemental Document*. The root mean square error (RMSE) of the estimated fixed effect coefficients $(\hat{\beta}_1, \dots, \hat{\beta}_p)$ is computed against the true regression parameters used to simulate data, and is used as the criterion for performance assessment. Note that the first dimension of data \mathbf{X} is intercept, 1. For instance, when $p = 5$, it means

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,2} & \dots & x_{1,4} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,2} & \dots & x_{n,4} \end{bmatrix}.$$

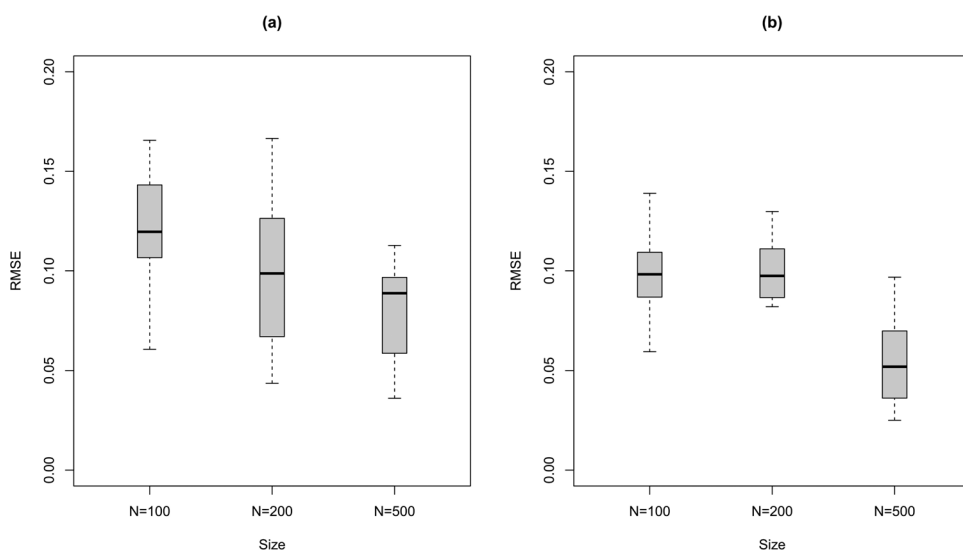


Figure 2. Uniform (a) $p = 5$; (b) $p = 8$. Each box is 10 RMSEs of 10 different datasets for that combination. LME models with random intercept only.

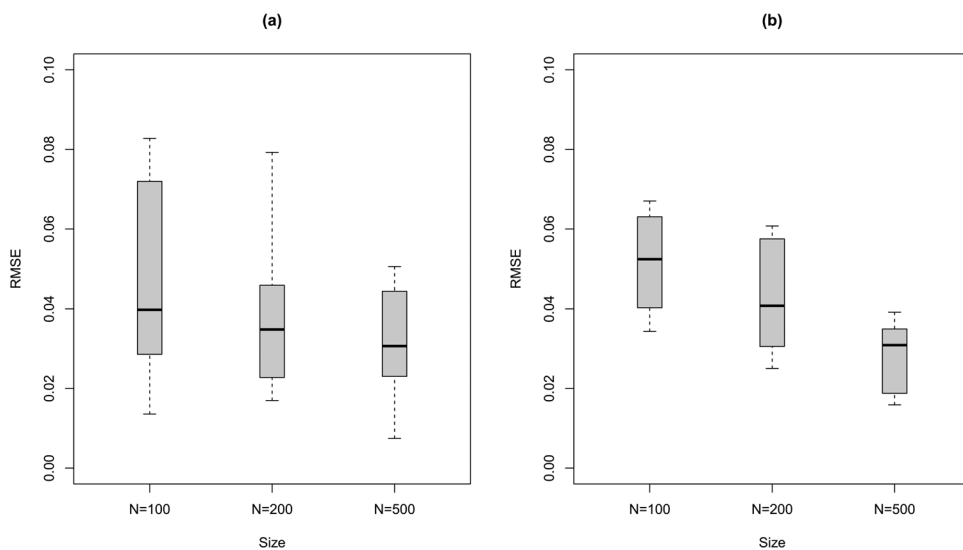


Figure 3. Centered Laplace (a) $p = 5$; (b) $p = 8$. Each box is 10 RMSEs of 10 different datasets for that combination. LME models with random intercept only.

The results for Uniform and Centered Laplace distributions are reported in Figures 2 and 3, respectively.

From Figure 2, it is obvious that as the sample size increases, the RMSEs decrease for both $p = 5$ and $p = 8$ in general. Also, the variation that comes from different datasets of the same model is slightly smaller with larger sample size, which is visualized in each box in Figure 2. The dimensionality also plays a role here as the performances are similar

Table 1. Each cell is median of 10 RMSEs of estimated fixed effect coefficients against true parameters of 10 different datasets for that combination

Dimension	Sample Size	Uniform		Centered Laplace	
		Random Intercept	Random Intercept and Slope	Random Intercept	Random Intercept and Slope
$p = 5$	$n = 100$	0.120	0.112	0.040	0.053
	$n = 200$	0.099	0.083	0.035	0.035
	$n = 500$	0.089	0.065	0.031	0.032
$p = 8$	$n = 100$	0.099	0.095	0.052	0.058
	$n = 200$	0.097	0.087	0.041	0.046
	$n = 500$	0.052	0.068	0.031	0.041
Dimension	Sample Size	Gamma with fixed shape parameter		Triangular	
		Random Intercept	Random Intercept and Slope	Random Intercept	Random Intercept and Slope
$p = 5$	$n = 100$	0.096	0.143	0.230	0.280
	$n = 200$	0.041	0.121	0.153	0.279
	$n = 500$	0.037	0.115	0.133	0.246
$p = 8$	$n = 100$	0.078	0.167	0.229	0.240
	$n = 200$	0.057	0.133	0.140	0.237
	$n = 500$	0.036	0.091	0.136	0.199

when $n = 100$ and $n = 200$, but the RMSEs of $p = 8$ become smaller than those for $p = 5$ when n increases to 500. For each combination, the median of its 10 RMSEs is reported in Table 1. Median is used instead of the mean since median is robust in the presence of extreme values. The results in Table 1 agrees with the observations from Figure 2. It is also observed that all of the RMSEs for Uniform are less than 0.2 indicating satisfactory performances.

Observations from Figure 3 are consistent with those made for Figure 2, i.e. as the sample size increases, the RMSEs decrease for the same dimension. In addition, the Centered Laplace distribution actually has a better performance than Uniform as none of the RMSEs for Centered Laplace distribution is above 0.1. Actually, the median of RMSEs for $n = 100, 200$, and 500 are 0.040, 0.035 and 0.031, respectively. It manifests that the estimation accuracy is satisfactory.

In addition, we present the boxplots for Gamma and Triangular in Section C of the *Supplemental Document* to save space. For Gamma distribution, we actually consider a special case by fixing $a = 1$ to make its shape less flexible for model identifiability purposes. These observations agree with what we have discussed above.

4.2. Models with random intercept and one random slope

In this section, we keep the same set up as in Section 4.1. However, instead of using the LME models with random intercept only, we simulate the same model with both a random intercept (x_1) and one random slope (x_2). The true parameters used are reported in the Section B of the *Supplemental Document*. The root mean squared error (RMSE) of the estimated fixed effect coefficients ($\hat{\beta}_1, \dots, \hat{\beta}_p$) is computed against the true parameters used to simulate data, and is used as the criterion for performance assessment. Similarly, the boxplots

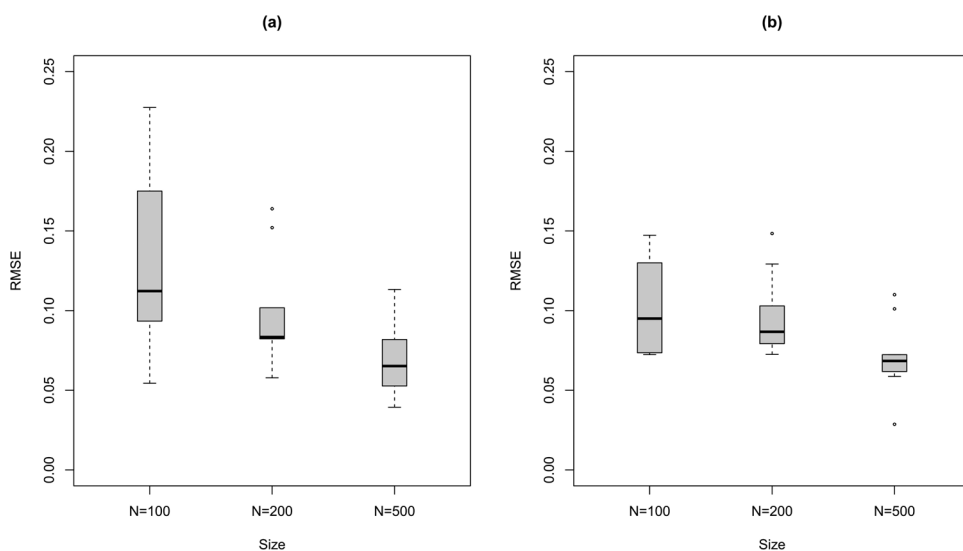


Figure 4. Uniform (a) $p = 5$; (b) $p = 8$. Each box is 10 RMSEs of 10 different datasets for that combination. LME models with random intercept and one random slope.

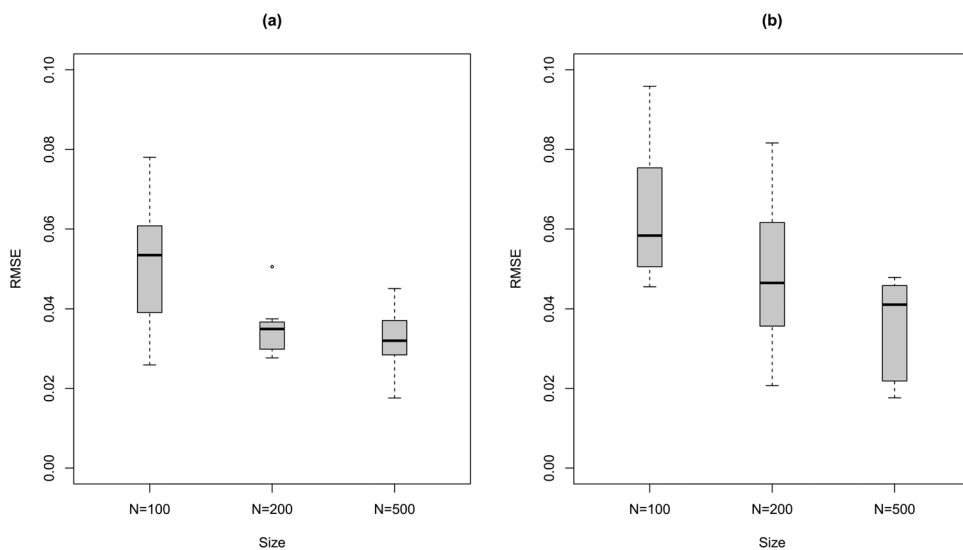


Figure 5. Centered Laplace (a) $p = 5$; (b) $p = 8$. Each box is 10 RMSEs of 10 different datasets for that combination. LME models with random intercept and one random slope.

for Uniform and Centered Laplace distributions are reported in Figures 4 and 5, respectively. The boxplots for the other two distributions are in Section C of the *Supplemental Document*.

From Figures 4 and 5, similar observations hold as those for models with random intercept only. Generally speaking, as the sample size increases, the RMSEs decrease. Taking the Centered Laplace distribution as an example, the median of the RMSEs drops from

0.053 for $n = 100$ to 0.032 for $n = 500$ with $p = 5$. We observe the same pattern for the other distributions. Moreover, if we compare the performances of models with both random intercept and random slope with models with random intercept only, it is actually observed that making the model more complicated is unnecessary especially when the sample size is not big enough. For example, for the same distribution, the median RMSE is 0.046 for $n = 200$, $p = 8$ compared to 0.041 of the same combination. The medians of the RMSEs are reported in Table 1.

4.3. Validation of random effects

Since the random effects are simulated as well, RMSEs can be calculated against the true random effects. Sticking to the LME model with random intercept only, we report the median of RMSEs for random effects in Table 2. The observations are similar to those for the fixed effects in the previous two subsections: as sample size increases, the median of the RMSEs decreases. In addition, none of the medians is above 0.3 suggesting that the estimation performance is satisfactory.

Moreover, in order to further justify the proposed method, focusing on one dataset for each distribution, the estimated $g = 20$ random effects are tested against the distribution that is assumed using a two-sided Kolmogorov–Smirnov test [9] and the p -values are reported in Table 3. It is observed that none of the p -values is smaller than 0.05 suggesting that there is no statistically significant evidence to reject the null hypothesis that the $g = 20$ random effects are not from the distribution assumed.

Furthermore, taking Gamma distribution as an example, we further validate the results by presenting Figure 6, where the left hand plot is the histogram of the $g = 20$ true random effects generated from $\Gamma(a = 1, b = 0.8)$ with the true density imposed on the same plot, while the right hand side is the histogram of the 20 estimated random effects

Table 2. Each cell is the median of 10 RMSEs of estimated random effects against true random effects of the 10 different datasets for that combination.

Dimension	Sample Size	Uniform	Centered Laplace	Gamma	Triangular
$p = 5$	$n = 100$	0.168	0.140	0.089	0.255
$p = 5$	$n = 200$	0.129	0.092	0.055	0.143
$p = 5$	$n = 500$	0.107	0.060	0.049	0.116
$p = 8$	$n = 100$	0.156	0.138	0.111	0.274
$p = 8$	$n = 200$	0.148	0.090	0.059	0.153
$p = 8$	$n = 500$	0.108	0.067	0.044	0.123

Table 3. The p -values of a two-sided Kolmogorov–Smirnov test on $g = 20$ random effect coefficients with each distribution assumed. One dataset simulated from the LME model with random intercept only.

Dimension	Sample Size	Uniform	Centered Laplace	Gamma	Triangular
$p = 5$	$n = 100$	0.586	0.591	0.328	0.548
$p = 5$	$n = 200$	0.430	0.569	0.256	0.310
$p = 5$	$n = 500$	0.667	0.649	0.209	0.778
$p = 8$	$n = 100$	0.376	0.604	0.365	0.398
$p = 8$	$n = 200$	0.452	0.623	0.312	0.771
$p = 8$	$n = 500$	0.541	0.596	0.243	0.779

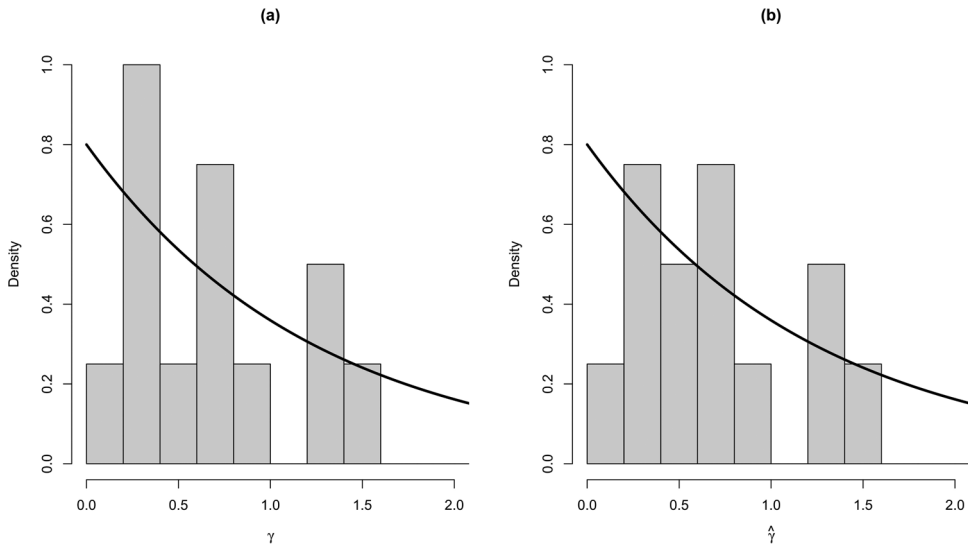


Figure 6. (a): Histogram of the $g = 20$ true random effects generated from $\Gamma(a = 1, b = 0.8)$ with the true density; (b): Histogram of the $g = 20$ estimated random effects with density of $\Gamma(a = 1, b = 0.8)$.

($p = 5, n = 500$) with the true density of $\Gamma(a = 1, b = 0.8)$. The p-value of two-sided Kolmogorov–Smirnov test is 0.204 and 0.209, respectively. Both plots have similar patterns, and both p-values are above 0.05 meaning it fails to reject the null hypothesis that the random effects are from $\Gamma(a = 1, b = 0.8)$. The RMSE of the 20 estimated random effects against the true random effects is 0.023, which indicates the estimation is very much close to the true parameters.

Last but not the least, we also intend to validate the approximated density against the true density for simple cases where we can actually obtain an analytical expression of the exact density of y_i . Using the Uniform distribution as an example, working under the LME model with random intercept, the PDF of $y_i = \sum_{j=1}^p x_{ij}\beta_j + z_{i,1}\gamma_1 + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$ and $\gamma_1 \sim U(-1, 1)$, can be explicitly written as follows by the convolution theorem.

$$f(y_i) = \frac{1}{2|z_{i,1}|} \left(\Phi \left(\frac{y_i - \sum_{j=1}^p x_{ij}\beta_j + |z_{i,1}|}{\sigma} \right) - \Phi \left(\frac{y_i - \sum_{j=1}^p x_{ij}\beta_j - |z_{i,1}|}{\sigma} \right) \right), \quad (20)$$

where $\Phi(\cdot)$ is the CDF of a standard Normal distribution. Densities from saddlepoint approximation in (13) can be evaluated against Equation (20) in a pointwise manner. The RMSEs are reported in Table 4 and all of the 6 RMSEs are less than 0.1 suggesting that the saddlepoint densities are indeed close to the true densities.

4.4. Comparisons with the PIT method

In this section, we compare the proposed SA method with the PIT method [20] that was reviewed in Section 1. Although both methods are able to handle non-Normal random effects, the proposed SA method is numerically more stable since the formulation of the

Table 4. RMSE of saddlepoint approximation density against the true density; Uniform distribution on random effects; One dataset simulated from the LME model with random intercept only.

	$p = 5$	$p = 8$
$n = 500$	0.058	0.091
$n = 1000$	0.051	0.073
$n = 2000$	0.035	0.037

Table 5. Each cell is median of 10 RMSEs of estimated fixed effect coefficients against true coefficients of 10 different datasets for that combination for the PIT method.

	$Q = 2$		$Q = 4$	
	$p = 5$	$p = 8$	$p = 5$	$p = 8$
$n = 100$	0.204	0.174	0.207	0.219
$n = 200$	0.112	0.115	0.126	0.133
$n = 500$	0.090	0.106	0.077	0.084

PIT method sums up Q products of n_i probability densities inside the logarithm function following the notation of [20], which is impossible to convert from multiplication to summation. We implemented the PIT method according to its original formulation [20]. Similar to Section 5.3 of [8], when implementing the PIT method, the only modification we did was to take the natural logarithm of the last equation in Section 3 of [20]. Numerical issues will arise during the optimization procedure without such modification.

Following the specification in Section 5.3 of [8], the number of points used to approximate integrals is 2 and 4, i.e. $Q = 2, 4$ to balance between approximation accuracy and computational resource. The values of other parameters such as $z_q, \eta_q, q = 1, \dots, Q$ were extracted from Table 25.10 of [1]. Taking Uniform and the LME model with random intercept only as an example. We follow the same procedures as in Section 4.1 and report the median of 10 RMSEs of the estimated fixed effect coefficients of PIT method against the true parameters in Table 5. The results are comparable to the same combination reported in Table 1, and the true parameters used to simulate data are reported in Section B of *Supplemental Document*.

From Table 5, the PIT method produces good results, although the RMSEs are mostly larger than the proposed SA method shown in Table 1. For example, the median of RMSEs for $n = 100, p = 5, Q = 2$ is 0.204 as compared to 0.120 of the proposed SA method for the same combination. The only exception is when $n = 500, p = 5$ where the median of RMSEs is 0.077, while that of the SA method is 0.089. Despite the slight difference in the performance of the PIT method and the proposed SA approach, both yield acceptable results in general.

Finally, the proposed SA method is admittedly not a perfect method as one of its limitation is the computational time. Consider, for instance, a simulated dataset with $p = 5$ and $n = 100$. It takes the SA approach 251 seconds to finish the whole estimation process on an Amazon Linux cloud computing machine with Intel Xeon Platinum 8259CL processor. For comparison, the PIT method can complete the same process using 10% of the computing time in the same environment. We believe part of the performance difference is due to

the constrained optimization in Equation (16), which is the most time consuming part in the whole process.

4.5. More considerations

We discuss in this section some additional important topics regarding the proposed SA approach and empirically evaluate them.

4.5.1. Model misspecification

We investigate how the proposed SA method performs when the distribution of the random effects is misspecified. In this regard, we work with a random intercept only model with $p = 5$ and $n = 100, 200$. Identical to the settings in Section 4.1, each sample size is then equally divided into $g = 20$ clusters. For each combination, we simulate 10 different datasets aiming to check the variation in the data-generating process. The random effects is generated from a truncated Normal distribution, but is fitted separately with the four distributions considered in this paper: (1) Uniform (2) Centered Laplace (3) Gamma and (4) Triangular. The RMSE of the estimated fixed effect coefficients is computed against the true regression parameters used to simulate data. The medians are reported in Table 6.

From Table 6, generally the median of the RMSE is slightly higher than that reported in Table 1, indicating the estimation accuracy is slightly affected by the model misspecification. However, we observe that for the Uniform distribution, the medians are 0.120 and 0.099 for $n = 100, 200$, respectively, when the model is correctly specified in Table 1, and the numbers remain at a similar level, i.e. 0.128 and 0.096 in Table 6 when model is misspecified, which may be due to the non-informative nature of the Uniform distribution. Therefore, when it is needed to model random effects with a non-Normal distribution, Uniform is empirically recommended to be the default distribution if no prior information is obviously against such choice. In addition, given the numbers in Table 6, we believe the estimation accuracy is still acceptable and the proposed SA approach remains practically viable even if the model is misspecified.

4.5.2. Test of hypothesis

Conducting a test of hypothesis is not an easy task when the exact likelihood function is intractable. Following the same approximation method detailed in Section 4.2 of [8], a likelihood ratio test (LRT) is applied to serve the purpose. In fact, we consider testing both all β and individual β . Since the exact expression of the likelihood function is not available, the SA approximation is then utilized in lieu of the exact analytical expression. Taking the

Table 6. Each cell represents median of 10 RMSEs of estimated fixed effect coefficients against true parameters of 10 different datasets for that combination for model misspecification.

Sample Size	Uniform	Centered Laplace
$n = 100$	0.128	0.049
$n = 200$	0.096	0.048
Sample Size	Gamma with fixed shape parameter	Triangular
$n = 100$	0.104	0.341
$n = 200$	0.063	0.272

test of all β as an illustrative example, the hypotheses are as follows, assuming the fixed effects are positively constrained.

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \text{at least one } \beta_i > 0, \quad (21)$$

where $i = 1, \dots, p$. The proposed test statistic is

$$T_{LR} = -2 [\mathcal{L}_{\text{approx}}(\hat{\beta}_0, \hat{\theta}_0, \hat{\sigma}_0) - \mathcal{L}_{\text{approx}}(\hat{\beta}, \hat{\theta}, \hat{\sigma})], \quad (22)$$

where $\hat{\beta}_0, \hat{\theta}_0, \hat{\sigma}_0$ denote the estimation under the null hypothesis, and $\hat{\beta}, \hat{\theta}, \hat{\sigma}$ denotes the estimation under the alternative hypothesis. The exact distribution of T_{LR} is not available and we adopt the upper bound given in the equation (22) of [8], i.e. $\frac{1}{2}(\mathcal{P}(\chi_{p-1}^2 \geq c_\alpha) + \mathcal{P}(\chi_p^2 \geq c_\alpha))$ to compute the corresponding critical value, c_α , where α is the significance level. Therefore, the LRT used here is more conservative of committing a false positive error.

Their finite sample performance is investigated against a random intercept only model with $p = 5, n = 100$ or $p = 5, n = 200$. Same as in Section 4.1, g is set as 20. The empirical true negative (TN) rate is utilized as the criterion, which is defined as 1 minus the calculated false positive (type 1 error) rate based on 500 replications. With $\alpha = 0.05$ the theoretical TN is 0.95. The hypotheses for testing all β are given in Equation (21). For testing of an individual β , we specifically test β_2 (the first non-intercept parameter for fixed effects). $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 > 0$. The performance is presented in Table 7 for testing all β and Table 8 for testing individual β .

From Tables 7 and 8, it is clear that except for Uniform and centered Laplace distributions that slightly under-cover when the sample size is small, all remaining combinations have an empirical TN that is higher than 0.95. The results are empirically acceptable indicating satisfactory performance of the proposed SA approach.

Table 7. Test of all β . The empirical true negative rate is computed based on $M = 500$ replications.

Sample Size	Uniform	Centered Laplace
$n = 100$	0.944	0.947
$n = 200$	0.958	0.969
Sample Size	Gamma with fixed shape parameter	Triangular
$n = 100$	0.972	0.955
$n = 200$	0.981	0.970

Notes: The significance level is set as 0.05, $p = 5, g = 20$.

Table 8. Test of individual β . The empirical true negative rate is computed based on $M = 500$ replications.

Sample Size	Uniform	Centered Laplace
$n = 100$	0.922	0.939
$n = 200$	0.939	0.940
Sample Size	Gamma with fixed shape parameter	Triangular
$n = 100$	0.968	0.974
$n = 200$	0.971	0.990

Notes: The significance level is set as 0.05, $p = 5, g = 20$.

Table 9. Estimated overall slopes of sales data of the 1 liter regular whole milk.

Store No.	lme4	SA	PIT
1	0.025	−0.018	−0.000
2	−0.089	−0.018	−0.000
3	−0.114	−0.185	−0.185
4	0.048	−0.218	−0.231
5	−0.108	−0.171	−0.172
6	−0.091	−0.098	−0.100

Notes: The proposed SA is fitted with a Uniform distribution on the random effects.

Table 10. Estimated fixed effect coefficients of sales data of the 1 liter regular whole milk.

	Intercept	slope	Pred RMSE
lme4	2.671	−0.055	0.080
PIT	1.524	−0.277	0.081
SA	1.557	−0.118	0.079

Notes: The proposed SA is fitted with a Uniform distribution on the random effects.

5. An application in retail data analytics

We now apply the proposed SA method to analyze the sales data of the 1 liter regular whole milk that was introduced in Section 1. The Uniform distribution is assumed on the random effects as there is no preference on the magnitude of the random effects from a business's point of view. The classical LME model and the PIT method are included as comparisons. The estimated fixed effects of all models are reported in Table 10, from which we can clearly observed that the magnitude of the slope from the SA approach is about 4 times larger than that of the *lme4* and about 2 times larger than that of the PIT method. The 6 overall slopes for the 6 stores are reported in Table 9.

Since the random effects are bounded by a Uniform distribution and none of the 6 overall slopes is positive unlike the model with the Normality assumption, for which store 1 and store 4 have positive overall slopes as reported in Section 1 and Table 9. Compared with the traditional LME method, the proposed SA and the PIT method are observed to preserve model interpretability and sign correctness.

In order to measure the predictive performance, we define the predictive RMSE as the RMSE of the estimated response variable value \hat{y} against the observed response variable value y , i.e.

$$\text{predictive RMSE} \triangleq \sqrt{\frac{\sum_{\ell=1}^g \sum_{i=1}^{n_{\ell}} (\hat{y}_{\ell,i} - y_{\ell,i})^2}{\sum_{\ell=1}^g n_{\ell}}}.$$

Clearly, the predictive RMSE is a metric of how well the estimated model fit the observed data. As we can see from Table 10, another advantage of the proposed method is that the predictive RMSE of the proposed model is slightly lower than that of the traditional LME and the PIT method in Table 10, suggesting superior performance of the proposed SA method in not only maintaining model interpretability and sign correctness, but also having better predictability.

6. Discussion and concluding comments

In this paper, we study linear mixed effects models. Instead of assuming a Normal distribution on the random effects, we explore the possibility of non-Normal distributions to provide flexibility for modeling purposes. We propose to use a saddlepoint approximation method for parameter estimation, overcoming a major challenge in the lack of a closed-form likelihood function. To demonstrate the proposed approach, we specifically study four special cases of non-Normal distributions: (1) Uniform, (2) Centered Laplace, (3) Gamma and (4) Triangular. Both the simulation studies and the real-world application example in retail data analytics demonstrate the satisfactory performance of the proposed approach, which works best when the use of an appropriate distribution is supported by practical or domain knowledge of the range of the overall regression parameters.

There are two interesting areas of future research directly motivated by this paper. First, if a non-Normal distribution is assumed on the random effects, how much worse will the SA approximation quality be? How will the shape of the distribution affect the approximation quality? We are currently not aware of any theoretical results for these in the literature. If rigorous theoretical results are not available for this analysis, a paper reporting some empirical findings regarding these questions should be quite attainable and will be very informative. Second, the paper considers a linear mixed effects model, there is a need to extend it to a generalized linear mixed effects model (GLMM) if the data type of the response variable is non-continuous, for example, binary or countable quantities. We expect the estimation process to be more challenging for a GLMM as a link function will be involved. For example, if the response variable is a countable quantity and the logarithm of its expected value is modeled by a linear combination of unknown parameters, then the existence of a non-linear link function will greatly complicate the mathematical derivations. Therefore, the extension to GLMM is an interesting topic for future research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

- [1] M. Abramowitz, I.A. Stegun, and R.H. Romer, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards, Washington DC, 1964.
- [2] D. Bates, M. Mächler, B. Bolker, and S. Walker, *Fitting linear mixed-effects models using lme4*, preprint (2014). Available at arXiv, arXiv:1406.5823.
- [3] D. Bates, M. Mächler, B. Bolker, and S. Walker, *Fitting linear mixed-effects models using lme4*, J. Stat. Softw. 67 (2015), pp. 1–48. <https://doi.org/10.18637/jss.v067.i01>
- [4] P.T. Boggs and J.W. Tolle, *Sequential quadratic programming*, Acta Numerica 4 (1995), pp. 1–51.
- [5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends. Mach. Learn. 3 (2011), pp. 1–122.
- [6] B.J. Bronnenberg, S.K. Dhar, and J.-P. Dubé, *Consumer packaged goods in the united states: National brands, local branding*, J. Mark. Res. 44 (2007), pp. 4–13.
- [7] R.H. Byrd, M.E. Hribar, and J. Nocedal, *An interior point algorithm for large-scale nonlinear programming*, SIAM. J. Optim. 9 (1999), pp. 877–900.

- [8] H. Chen, L. Han, and A. Lim, *Estimating linear mixed effects models with truncated normally distributed random effects*, Commun. Stat. Simul. Comput. (2022). <https://doi.org/10.1080/03610918.2022.2066696>.
- [9] W.W. Daniel, *Applied Nonparametric Statistics*, PWS-Kent Pub., Boston, 1990.
- [10] H.E. Daniels, *Saddlepoint approximations in statistics*, Ann. Math. Stat. 25 (1954), pp. 631–650.
- [11] C.-E. Fröberg, *Introduction to Numerical Analysis*, Addison-Wesley, Reading, MA, 1969.
- [12] J.M. Hammersley and D.C. Handscomb, *Monte Carlo Methods*, Springer, Methuen, 1964.
- [13] J. Jiang, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Science & Business Media, New York, NY, 2007.
- [14] S. Kotz, T. Kozubowski, and K. Podgorski, *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*, Springer Science & Business Media, Boston, MA, 2012.
- [15] T.I. Lin and J.C. Lee, *Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data*, Stat. Med. 27 (2008), pp. 1490–1507.
- [16] M.J. Lindstrom and D.M. Bates, *Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data*, J. Am. Stat. Assoc. 83 (1988), pp. 1014–1022.
- [17] L.A. Matos, M.O. Prates, M.-H. Chen, and V.H. Lachos, *Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution*, Stat. Sin. 23 (2013), pp. 1323–1345.
- [18] T.B. Mattos, V.H. Lachos, L.M. Castro, and L.A. Matos, *Extending multivariate student's-t t semiparametric mixed models for longitudinal data with censored responses and heavy tails*, Stat. Med. 41 (2022), pp. 3696–3719.
- [19] C.E. McCulloch and J.M. Neuhaus, *Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter*, Stat. Sci. 26 (2011), pp. 388–402.
- [20] K.P. Nelson, S.R. Lipsitz, G.M. Fitzmaurice, J. Ibrahim, M. Parzen, and R. Strawderman, *Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects*, J. Comput. Graph. Stat. 15 (2006), pp. 39–57.
- [21] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Science & Business Media, New York, NY, 2006.
- [22] J.C. Pinheiro, C. Liu, and Y. Nian Wu, *Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution*, J. Comput. Graph. Stat. 10 (2001), pp. 249–276.
- [23] V.R. Rao, *Pricing research in marketing: The state of the art*, J. Bus. 57 (1984), pp. S39–S60.
- [24] S. Seabold and J. Perktold, *statsmodels: Econometric and statistical modeling with Python*, in *9th Python in Science Conference*, Austin, TX, 2010.
- [25] A. Skrondal and S. Rabe-Hesketh, *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Crc Press, Boca Raton, FL, 2004.
- [26] G. Verbeke and E. Lesaffre, *The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data*, Comput. Stat. Data. Anal. 23 (1997), pp. 541–556.
- [27] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, Í. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, and P. van Mulbregt, *SciPy 1.0 Contributors, SciPy 1.0: Fundamental algorithms for scientific computing in Python*, Nat. Methods. 17 (2020), pp. 261–272.
- [28] L. Wu, *Mixed Effects Models for Complex Data*, Chapman and Hall/CRC, Boca Raton, FL, 2009.
- [29] R.M. Yucel and H. Demirtas, *Impact of non-normal random effects on inference by multiple imputation: A simulation assessment*, Comput. Stat. Data. Anal. 54 (2010), pp. 790–801.
- [30] X. Zhang, *A tutorial on restricted maximum likelihood estimation in linear regression and linear mixed-effects model*, 2015. <http://statdb1.uos.ac.kr/teaching/multi-grad/ReML.pdf>.