

Estimating linear mixed effects models with truncated normally distributed random effects

Hao Chen, Lanshan Han & Alvin Lim

To cite this article: Hao Chen, Lanshan Han & Alvin Lim (2024) Estimating linear mixed effects models with truncated normally distributed random effects, Communications in Statistics - Simulation and Computation, 53:4, 2050-2070, DOI: [10.1080/03610918.2022.2066696](https://doi.org/10.1080/03610918.2022.2066696)

To link to this article: <https://doi.org/10.1080/03610918.2022.2066696>



View supplementary material [↗](#)



Published online: 28 Apr 2022.



Submit your article to this journal [↗](#)



Article views: 197



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)



Estimating linear mixed effects models with truncated normally distributed random effects

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

^aRetailer Products Research & Development, NielsenIQ, Chicago, IL; ^bGoizueta Business School, Emory University, Atlanta, GA

ABSTRACT

Linear Mixed Effects (LME) models have been widely applied in clustered data analysis in many areas including marketing research, clinical trials, and biomedical studies. Inference can be conducted using maximum likelihood approach if assuming Normal distributions on the random effects. However, in many applications of economy, business and medicine, it is often essential to impose constraints on the regression parameters after taking their real-world interpretations into account. Therefore, in this paper we extend the classical (unconstrained) LME models to allow for sign constraints on its overall coefficients. We propose to assume a symmetric doubly truncated Normal (SDTN) distribution on the random effects instead of the unconstrained Normal distribution which is often found in classical literature. With the aforementioned change, difficulty has dramatically increased as the exact distribution of the dependent variable becomes analytically intractable. We then develop likelihood-based approaches to estimate the unknown model parameters utilizing the approximation of its exact distribution. Simulation studies have shown that the proposed constrained model not only improves real-world interpretations of results, but also achieves satisfactory performance on model fits as compared to the existing model.

ARTICLE HISTORY

Received 16 August 2021
Accepted 6 April 2022

KEYWORDS


Constrained regression analysis; Mixed effects model; Penalized least squares; Penalized restricted least squares

1. Introduction

In practice, it is often necessary for modelers to quantify the heterogeneity among subgroups in a statistical study. For example, in marketing research, modelers often need to consider the geographical difference in response to marketing activities. In clinical research, a modeler may aim to capture the demographic difference in responding to a certain treatment regime. To do so, modelers often rely on mixed effect models with random effects specified to capture the heterogeneity. It is also often assumed that the random effects follow Normal distributions, so that a closed-form likelihood function is available and thus efficient numerical algorithms are available to maximize it. However, in many applications the normality assumption can lead to results that are inconsistent with our common sense or not interpretable. We provide an example to elaborate this issue.

We consider a study in marketing research with the goal of inferring the effectiveness of various marketing activities from sales and marketing data. Typically, for this kind of study, modelers collect weekly sales volume and marketing activity readings across different geographical regions for various products. A statistical model, typically a mixed effect model, is specified to find how the marketing activities affect the sales volume of the individual products in the different regions.

CONTACT Hao Chen  hao.chen@stat.ubc.ca  Retailer Products Research & Development, NielsenIQ, Chicago, IL 60606.

 Supplemental data for this article can be accessed online at <http://dx.doi.org/10.1080/03610918.2022.2066696>

© 2022 Taylor & Francis Group, LLC

Table 1. Snapshot of the dataset with discounted sales information on a product.

Store cluster	Discount rate	Logit quantity
A	0.000	0.236
...
A	0.000	0.358
B	0.000	0.272
...
B	0.440	-0.333
...
F	0.010	0.082
...
F	0.000	-1.113

We consider a simplified real-world dataset recording weekly sales and discount information on a consumer packaged good (CPG) (Bronnenberg, Dhar, and Dubé 2007) from a retailer. The dataset contains weekly sales volume from different stores in three consecutive weeks of a non-holiday period in 2017. It includes three columns: (1) Store Cluster, (2) Discount Rate and (3) Logit Quantity. Each row represents a weekly record of a particular store. A snapshot of the dataset is given in Table 1. Some explanations about the variables are given below.

- Store Cluster: the stores are pre-clustered into 6 different store clusters (A, B, C, D, E, F) based on their proximity such that stores within a cluster are more homogeneous than those across cluster.
- Discount Rate: the discount rate ranging between 0 and 1. The larger the discount rate is, the cheaper the product is sold to customers. A 0 rate means the product was not sold at a promotional discount for that week.
- Logit Quantity: the weekly sales quantity of a product after a logistic transformation, defined as $\log\left(\frac{q}{\mu-q}\right)$, where q is the observed sales quantity and μ is the theoretical maximum sales quantity from domain knowledge and treated as known.

Given this dataset, we aim to quantify the relationship between Logit Quantity and Discount Rate while capturing heterogeneity across store clusters. Note that for simplicity of illustration, the dataset has been simplified with some other confounding factors removed to focus on the Discount Rate. The temporal effect is also ignored since neither seasonality nor holiday effects is expected to affect modeling results in a non-trivial way for the three-week non-holiday period, and is not the focus of this research. We specify a linear mixed effects (LME) model (McCulloch and Neuhaus 2014) with both random intercept and random slope in (1):

$$y_{\ell,i} = (\beta_0 + \beta_{0,\ell}) + (\beta_1 + \beta_{1,\ell})x_{\ell,i} + \varepsilon_{\ell,i}, \quad (1)$$

where ℓ is the index for Store Cluster, which is also the grouping factor, $x_{\ell,i}$ is the Discount Rate for observation i and Store Cluster ℓ , $y_{\ell,i}$ is the Logit Quantity for observation i and Store Cluster ℓ and $\varepsilon_{\ell,i}$ is the random error term. Working under the independent covariance structure (Wu 2009), the classical LME model assumes $\varepsilon_{\ell,i} \sim N(0, \sigma^2)$, $\beta_{0,\ell} \sim N(0, \varsigma_0^2)$, $\beta_{1,\ell} \sim N(0, \varsigma_1^2)$ and the random effects and the error term are independent to each other.

In addition to the above assumptions, in practice, it is also required that $\beta_1 \geq 0$ and the overall coefficient $\beta_1 + \beta_{1,\ell} \geq 0$, representing a common belief that promotional discounts will not reduce sales quantity. In other words, a non-negative sign constraint is needed on the fixed effect and the overall coefficient of Discount Rate. A classical LME model was fitted on the dataset that produces $\hat{\beta}_1 = -0.319$, and none of the absolute values of the estimated random effects is above 0.1 making all of the 6 overall coefficients negative as well. In practice, with the aforementioned modeling results, some heuristics often needs to be applied to “correct” the sign of the estimated coefficients for Discount Rate before the model is considered as conceivable. This example

demonstrates the necessity to deviate from the Normality assumption when equipping the traditional LME models with sign constraints while ensuring the technical rigor of the statistical assumptions and estimation approaches.

The above motivating example is one of the many real-world applications, where one often has restrictions and/or prior knowledge on the signs of the parameters to be estimated when the business or physical interpretations are taken into account. In addition to practical benefits, a sign constraint also bears some theoretical merits. For example, it can help with model identifiability by truncating parts of the possible parameter space. It also mitigates the issue of multicollinearity by restricting the feasible regions. This is especially useful when data quality is a concern.

The major contribution of this research is to estimate the model parameters with sign constraints assuming that the random effects follow a symmetric doubly truncated Normal (SDTN) distribution, instead of the unconstrained Normal distribution often found in the literature. The lower bounds and upper bounds of a SDTN distribution are carefully chosen based on its fixed effects so that the overall coefficients will not violate the sign constraints. The “minor” change in the distribution on the random effects brings some profound differences in the estimation process, and has dramatically increased the difficulty as a SDTN distribution does not have as elegant a set of mathematical properties as the unconstrained Normal distribution does. As a simple illustrating example, the sum of two independent SDTNs is not necessarily a SDTN with known analytical expressions of its resulting parameters, for which people take it for granted in terms of an unconstrained Normal distribution. Therefore, it is practically infeasible to derive a closed form probability density function for the sum of several independent SDTNs. This is a major hurdle if a likelihood based estimation approach is to be applied. In this paper, we use an approximated probability density function inspired by the Bayesian Central Limit Theorem (BCLT) (Theorem 3.1 of Carlin and Louis (2009)) to derive an approximated likelihood function. Parameters estimation is then conducted by maximizing the approximated likelihood function.

The linear mixed effects model enjoys its popularity during the past several years, and it is not unexpected that there are quite many monographs and academic articles about it covering a full spectrum of areas from studies on its mathematical properties and computational aspects of its implementation to its applications in economics, marketing, medicine and pharmaceutical research. Jiang (2007) provided a comprehensive overview of the mixed effects model focusing on its mathematical properties. Lindstrom and Bates (1988) discussed the computational details and proposed the use of Newton-Raphson and EM algorithms. Demidenko (2013) reviewed its many implementations in R, while Bates et al. (2014) specifically introduced the popular R package, *lme4*. Wu (2009) covered how the model behaves when missing data and measurement errors are present. On the application side, Brabec, Konár, Pelikán, and Malý (2008) used it to study the natural gas consumption by individual customers. It was also applied by Mitsumata et al. (2012) to research the effects of parental hypertension on longitudinal trends in offspring blood pressure. Moreover, linear mixed effects models have been extended to generalized linear mixed effects models, see McCulloch and Neuhaus (2014), to handle situations where the response variable is non-Normal, such as dummy (binomial) and count (Poisson). In addition, Wolfinger and O’Connell (1993) discussed its estimation using a pseudo-likelihood approach. In this paper, we restrict our research to the linear mixed effects model as our current business applications do not require the generalized version.

There are a few other studies on non-Normal random effect models. Pinheiro, Liu, and Wu (2001) specifically discussed the multivariate t-distribution random effects. More recently, Nelson et al. (2006) applied the probability integral transformation (PIT) to estimate non-linear mixed effects models with non-Normal random effects. Liu and Yu (2008) then proposed a computationally more practical method to obtain the maximum likelihood estimations for mixed effects models by reformulating the conditional likelihood function on non-Normal random effects. Moreover, Yucel and Demirtas (2010) employed simulations to study the impact of random

effects generated from non-Normal distributions on statistical inference under missing data conditions, while the Normality assumption is still assumed in the parameter estimation process. A direct comparison between the proposed methods and PIT is discussed further in [Sec. 5.3](#).

In our opinion, the sign constraint proposed on the overall coefficient in this research has at least the following two merits. First, it explicitly takes prior knowledge and physical interpretations into account guaranteeing that the final estimates will automatically make business sense in real-world applications. Admittedly, one can at times fit unconstrained models, and then manually investigate the possible causes whenever a “wrong” sign is produced and adjust the models accordingly. However, after ensuring that the input data are accurate, it is extremely challenging if not impossible in our real-world business practice to manually investigate the causes of “wrong” signs and adjust the models. This is the case, for example, with the over 10,000 independent models we utilize in practice for retail grocery products. Therefore, it is necessary to construct a production system with a constrained model specification to automate the whole process so that parallelly training a non-trivial number of models is possible. Second, similar to the idea of ridge regression (Hoerl and Kennard (1970)) and LASSO (Tibshirani (1996)) where regularization is employed to restrict the parameter space when optimizing its objective function, the proposed sign constraint also shrinks the parameter space by restricting to possible orthants only from a geometric perspective, which reduces the complexity of the model. We will further illustrate the merits of imposing sign constraints in [Sec. 5.2](#).

Admittedly, it is often possible to impose sign constraints in the estimation process under the existing framework of the linear mixed effects model, namely, imposing the constraints without modifying the likelihood function in the unconstrained case. However, we argue that this is not a theoretical sound approach. In particular, under the classical framework, random effects are assumed to follow a Normal distribution. Imposing constraints on them requires us to deviate from the Normal distribution assumption and hence invalidate the widely used likelihood functions derived based on Normal distributions. Hence, we propose to use the SDTN distribution instead of the unconstrained Normal one on the random effects in the model specification to comply with the use of sign constraints in the estimation.

The rest of the paper is organized as follows. The model specification is given in [Sec. 2](#). We provide some theoretical results on the truncated Normal distribution in [Sec. 3](#), building a solid foundation. The proposed estimation methods are detailed in [Sec. 4](#). Some simulation results and application examples are presented in [Sec. 5](#) and [Sec. 6](#), respectively. We make several concluding remarks in [Sec. 7](#).

2. Linear mixed effects model with sign constraints

Let $\mathbf{0}_n$ be the all zero vector of length n , $\mathbf{1}_n$ be the all one vector of length n , \mathbf{I}_n be the identity matrix of size $n \times n$. Throughout this paper, we use lower case letters to represent scalars, bold lower case letters to represent vectors, and upper case letters to represent matrices. For a vector $\mathbf{x} \in \mathbb{R}^n$, let $\text{diag}[\mathbf{x}]$ be the $n \times n$ diagonal matrix with the main diagonal being \mathbf{x} . We use subscripts to index scalars and superscripts to index vectors and matrices. For a vector \mathbf{x} , x_i represents its i -th component. For a matrix $X \in \mathbb{R}^{n \times p}$, $x_{i,j}$ represents its (i, j) element. Let $\alpha \subseteq \{1, \dots, p\}$ be an index set and $\bar{\alpha}$ be its complement. We write as \mathbf{x}_α the vector formed by taking the elements with indices in α from $\mathbf{x} \in \mathbb{R}^p$, and $X_{\bullet, \alpha}$ the submatrix of X composed of the columns with indexes in α from $X \in \mathbb{R}^{n \times p}$.

Consider a classical linear mixed effect model that captures heterogeneity among different clusters. Let the clusters, which could be geographical regions, product classes, individual subjects, etc, be indexed by $\ell = 1, \dots, g$. For each cluster ℓ , the dependent variable \mathbf{y}^ℓ is linearly dependent on the independent variables with an error term following a Normal distribution with 0 mean and unknown variance (to be estimated). Mathematically, this model is given by:

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

where

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

We often call $\boldsymbol{\beta}$ the fixed effects, $\gamma_{\ell,i}, \ell = 1, \dots, g; i = 1, \dots, k$ the random effects, and $\boldsymbol{\varepsilon}^\ell \sim \mathcal{N}(\mathbf{0}_{n_\ell}, \sigma^2 \mathbf{I}_{n_\ell})$ the error vector with σ^2 unknown. Note that n_ℓ is the sample size for group ℓ , and the total size is $n = \sum_{\ell=1}^g n_\ell$. p is the dimension. The number of variables for which random effects will be considered is denoted by $k, 0 \leq k \leq p$. If $k=0$, the linear mixed effects model reduces to a linear regression model with fixed effects only. One could also include an intercept to measure the baseline. The classical linear mixed effects model also assumes that

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

where $\boldsymbol{\Sigma}$ is structured, and is parameterized by some unknown parameters. Stacking up the data from different clusters, we 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 these definitions, we can rewrite the model into a more concise expression given below.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad (4)$$

where $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}_{kg}, \mathbf{G}), \mathbf{G} = \begin{bmatrix} \boldsymbol{\Sigma} & & \\ & \ddots & \\ & & \boldsymbol{\Sigma} \end{bmatrix}$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{R})$ with $\mathbf{R} = \sigma^2 \mathbf{I}_n$. In addition, $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are independent, and hence

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

2.1. Maximum likelihood approach for linear mixed effects models: a brief review

Working under the model specification in (4), a classical approach to estimate the parameters is to maximize certain likelihood functions. According to the assumptions, \mathbf{y} also follows a multivariate Normal distribution, in particular,

$$\mathbf{y} \sim \mathcal{N}(X\boldsymbol{\beta}, ZGZ^T + R),$$

where G is parameterized by the unknown $\varsigma_1^2, \dots, \varsigma_k^2$ and R is parameterized by σ^2 . Let $\boldsymbol{\theta}$ collectively denote $\varsigma_1^2, \dots, \varsigma_k^2$ and σ^2 . The covariance matrix of \mathbf{y} can hence be written as

$$V(\boldsymbol{\theta}) \triangleq ZG(\boldsymbol{\theta})Z^T + R(\boldsymbol{\theta}).$$

For any squared matrix A , let $|A|$ be its determinant. The **profile** log-likelihood function is defined by:

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{Z}) \triangleq -\frac{1}{2} \left(\ln |V(\boldsymbol{\theta})| + (\mathbf{y} - X\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}))^T V(\boldsymbol{\theta})^{-1} (\mathbf{y} - X\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})) \right), \quad (5)$$

where

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (X^T V(\boldsymbol{\theta})^{-1} X)^{-1} X^T V(\boldsymbol{\theta})^{-1} \mathbf{y}. \quad (6)$$

According to Zhang (2015), the **restricted** log-likelihood function is defined

$$\begin{aligned} \mathcal{L}_{\text{REML}}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{Z}) &\triangleq \\ &-\frac{1}{2} \left(\ln |V(\boldsymbol{\theta})| + (\mathbf{y} - X\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}))^T V(\boldsymbol{\theta})^{-1} (\mathbf{y} - X\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})) + \ln |X^T V(\boldsymbol{\theta})^{-1} X| \right). \end{aligned} \quad (7)$$

By maximizing $\mathcal{L}(\boldsymbol{\theta})$ or $\mathcal{L}_{\text{REML}}(\boldsymbol{\theta})$ using either Newton's method (Lindstrom and Bates 1988; Wolfinger, Tobias, and Sall 1994) or EM algorithm (Dempster, Laird, and Rubin 1977), we can obtain the estimator of $\boldsymbol{\theta}$, i.e.

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \text{argmax } \mathcal{L}(\boldsymbol{\theta}), \quad \text{or} \quad \hat{\boldsymbol{\theta}}_{\text{REML}} = \text{argmax } \mathcal{L}_{\text{REML}}(\boldsymbol{\theta}).$$

Based on McCulloch and Neuhaus (2014), it is known that $\hat{\boldsymbol{\theta}}_{\text{ML}}$ is **biased**, while $\hat{\boldsymbol{\theta}}_{\text{REML}}$ is **unbiased** which is more favorable in theory. After estimates of $\boldsymbol{\theta}$ (denoted $\hat{\boldsymbol{\theta}}$) are obtained, we can let $\hat{G} = G(\hat{\boldsymbol{\theta}})$ and $\hat{R} = R(\hat{\boldsymbol{\theta}})$ and estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ together by maximizing the joint likelihood function with \hat{G} and \hat{R} considered as known, i.e.

$$\begin{aligned} f(\mathbf{y}, \boldsymbol{\gamma}) &= f_{\mathbf{y}}(\mathbf{y}|\boldsymbol{\gamma}) f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) \\ &\propto |\hat{G}|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} \boldsymbol{\gamma}^T \hat{G}^{-1} \boldsymbol{\gamma} \right) \left| \hat{R} \right|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (\mathbf{y} - X\boldsymbol{\beta} - Z\boldsymbol{\gamma})^T \hat{R}^{-1} (\mathbf{y} - X\boldsymbol{\beta} - Z\boldsymbol{\gamma}) \right). \end{aligned}$$

Ignoring constant terms, the joint log likelihood function of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is hence,

$$\ln f(\mathbf{y}, \boldsymbol{\gamma}) \propto -\frac{1}{2} \boldsymbol{\gamma}^T \hat{G}^{-1} \boldsymbol{\gamma} - \frac{1}{2} (\mathbf{y} - X\boldsymbol{\beta} - Z\boldsymbol{\gamma})^T \hat{R}^{-1} (\mathbf{y} - X\boldsymbol{\beta} - Z\boldsymbol{\gamma}).$$

Therefore, it suffices to minimize the following function with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

$$Q(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \boldsymbol{\gamma}^T \hat{G}^{-1} \boldsymbol{\gamma} + (\mathbf{y} - X\boldsymbol{\beta} - Z\boldsymbol{\gamma})^T \hat{R}^{-1} (\mathbf{y} - X\boldsymbol{\beta} - Z\boldsymbol{\gamma}).$$

By taking the first-order derivatives of Q with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and set them equal to zero, we obtain the following linear system:

$$\begin{bmatrix} X^T \hat{R}^{-1} X & X^T \hat{R}^{-1} Z \\ Z^T \hat{R}^{-1} X & Z^T \hat{R}^{-1} Z + \hat{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} X^T \hat{R}^{-1} \mathbf{y} \\ Z^T \hat{R}^{-1} \mathbf{y} \end{bmatrix} \quad (8)$$

Let $\hat{V} = V(\hat{\boldsymbol{\theta}})$, we can derive from (8) that the estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as follows.

$$\hat{\boldsymbol{\beta}} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} \mathbf{y}, \quad (9)$$

$$\hat{\mathbf{y}} = \hat{G} Z^T \hat{V}^{-1} (\mathbf{y} - X \hat{\boldsymbol{\beta}}). \quad (10)$$

The aforementioned classical approach has been successfully used in many applications, and has been implemented in a statistical software packages such as R, Python and SAS. However, as discussed in Sec. 1, practitioners often face the situation where sign constraints are needed for some of the unknown coefficients to make practical sense. Hence, in what follows, we will augment the approach to handle sign constraints in a mathematical rigorous way.

2.2. Proposed model specification

In this paper, we consider a specific case where Z^ℓ is usually formed by taking a subset of the columns from X^ℓ , i.e. $Z^\ell = X_{\bullet, \alpha}^\ell$ for some index set $\alpha \subseteq \{1, \dots, p\}$. Mixed effects are considered for any column indexed with α . In this case, for each $\ell = 1, \dots, g$, model (2) can be written as

$$\mathbf{y}^\ell = X_{\bullet, \alpha}^\ell (\beta_\alpha + \gamma_\alpha^\ell) + X_{\bullet, \bar{\alpha}}^\ell \beta_{\bar{\alpha}} + \boldsymbol{\varepsilon}^\ell.$$

In many applications, it is necessary to make sure that the elements in $\beta_\alpha + \gamma_\alpha^\ell$ have correct signs. Without loss of generality, we consider the situation where $\alpha = \{1, \dots, p\}$ and $\beta_\alpha + \gamma_\alpha^\ell \geq 0$. In other words, a full model in which mixed effects are accounted for every column is assumed. The subscript α is dropped for the remainder of this section. We further assume

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

where ς_i^2 's are unknown. We continue to follow the classical model by assuming that

$$\mathbf{y}^\ell = X^\ell \boldsymbol{\beta} + Z^\ell \boldsymbol{\gamma}^\ell + \boldsymbol{\varepsilon}^\ell, \quad (11)$$

where

$$\boldsymbol{\varepsilon}^\ell \sim \mathcal{N}(\mathbf{0}_{n_\ell}, \sigma^2 \mathbf{I}_{n_\ell}). \quad (12)$$

However, to ensure the nonnegativity requirement is satisfied, we assume that the random effects $\gamma_{\ell, i}$ independently follow a symmetric doubly truncated normal (SDTN) distribution that is bounded by $-|\beta_i|$ and $|\beta_i|$. We will explicitly define SDTN in Sec. 3. As a result, each γ_i^ℓ is mathematically constrained within its corresponding $[-|\beta_i|, |\beta_i|]$. This way, we can guarantee that the overall coefficient will be non-negative. The proposed modification seems to be rather straightforward, however, it imposes significant challenges in the estimation of the parameters. Specifically, the distribution of the sum of finitely many random variables following SDTN distributions does not have a concise closed form. This fact represents a major hurdle when a maximum likelihood approach is applied. We will elaborate on more details about the parameter estimation process in Sec. 4. We will study and present results on some fundamental properties regarding truncated Normal distribution in Sec. 3 to lay a solid foundation, upon which our proposed estimating methods will build.

3. Truncated normal distributions

In this section, we present some properties of the truncated Normal distribution of interests. Throughout the paper, whenever we use the term Normal distribution, we refer to the usual unconstrained Normal distribution unless specified otherwise. Let the error function for a

Normal distribution be

$$\operatorname{erf}(z) \triangleq \frac{1}{\sqrt{\pi}} \int_0^z e^{-t^2} dt.$$

A truncated Normal (TN) distribution is parameterized by 4 parameters: location, μ ; scale, η ; lower bound a ; upper bound b . The Normal distribution is a special case of the TN distribution with $a = -\infty$ and $b = \infty$. The probability density function (PDF) of a $\mathcal{TN}(\mu, \eta^2, [a, b])$, with $\eta > 0$, is given by

$$f_{\mathcal{TN}}(x; \mu, \eta^2, a, b) = \begin{cases} \frac{1}{\eta} \frac{\phi(\xi)}{\Phi(b') - \Phi(a')}, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the PDF and cumulative distribution function (CDF) of the standard Normal distribution, i.e.

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\xi^2\right), \quad \text{and} \quad \Phi(\xi) = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{\xi}{\sqrt{2}}\right)\right],$$

respectively, and

$$\xi \triangleq \frac{x - \mu}{\eta}, \quad a' \triangleq \frac{a - \mu}{\eta}, \quad \text{and} \quad b' \triangleq \frac{b - \mu}{\eta}.$$

The mean and variance of $x \sim \mathcal{TN}(\mu, \eta^2, [a, b])$ are known and given by Olive (2008):

$$\begin{aligned} \mathbf{E}[x] &= \mu + \frac{\phi(a') - \phi(b')}{\Phi(b') - \Phi(a')} \eta, \\ \mathbf{Var}[x] &= \eta^2 \left[1 + \frac{a'\phi(a') - b'\phi(b')}{\Phi(b') - \Phi(a')} - \left(\frac{\phi(a') - \phi(b')}{\Phi(b') - \Phi(a')} \right)^2 \right]. \end{aligned}$$

In this paper, we are particularly interested in a special case, namely the symmetric doubly truncated normal (SDTN) distribution $\mathcal{SDTN}(\mu, \eta^2, [\mu - \rho\eta, \mu + \rho\eta])$ with $\rho > 0$, denoted by $\mathcal{SDTN}(\mu, \eta^2, \rho)$. It is a special case of a TN with $a = \mu - \rho\eta, b = \mu + \rho\eta$, i.e. the lower bound and upper bound are symmetric around mean μ . The properties of a SDTN distribution is given by Lemma 3.1.

Lemma 3.1. Suppose $x \sim \mathcal{SDTN}(\mu, \eta^2, \rho)$, the following results hold

i. The density function is

$$f_{\mathcal{SDTN}}(x; \mu, \eta^2, \rho) = \begin{cases} \frac{1}{\eta} \frac{\phi(\xi)}{2\Phi(\rho) - 1}, & x \in [\mu - \rho\eta, \mu + \rho\eta] \\ 0, & \text{otherwise} \end{cases}$$

ii. The expectation is

$$\mathbf{E}[x] = \mu,$$

iii. The variance is

$$\mathbf{Var}[x] = \eta^2 \left[1 - \frac{2\rho\phi(\rho)}{2\Phi(\rho) - 1} \right],$$

The proof is omitted as it is straightforward to verify the above results. Note that we define SDTN distributions with $\rho > 0$. In fact, when $\rho = 0$, it becomes a deterministic value, and hence the variance is 0. This is indeed consistent with the fact that

$$\lim_{\rho \rightarrow 0} \left[1 - \frac{2\rho\phi(\rho)}{2\Phi(\rho) - 1} \right] = 1 - \lim_{\rho \rightarrow 0} \frac{2\rho\phi(\rho)}{2\Phi(\rho) - 1} = 1 - \lim_{\rho \rightarrow 0} \frac{2\phi(\rho) + 2\rho\phi'(\rho)}{2\phi(\rho)} = 0$$

where the second equal sign is due to L'Hôpital's rule. We state some properties regarding the SDTN distribution in [Lemma 3.2](#).

Lemma 3.2. *Let $x \sim \mathcal{SDTN}(\mu, \eta^2, \rho)$ with $\rho > 0, \eta > 0$. Then, the following properties hold:*

- i. $x - \mu \sim \mathcal{SDTN}(0, \eta^2, \rho)$.
- ii. For any $x, y \in [\mu - \rho\eta, \mu + \rho\eta]$, if $x + y = 2\mu$ then $f_{\mathcal{SDTN}}(x; \mu, \eta^2, \rho) = f_{\mathcal{SDTN}}(y; \mu, \eta^2, \rho)$.
- iii. $\mathbf{Var}[x] \leq \eta^2$.
- iv. Suppose $x' \sim \mathcal{SDTN}(\mu, \eta^2, \rho')$, then $\mathbf{Var}[x] \leq \mathbf{Var}[x']$ if $\rho \leq \rho'$.
- v. If $x \sim \mathcal{SDTN}(0, \eta^2, \rho)$. Define $x' = m_0 + m_1x$ with $m_1 \neq 0$, where m_0, m_1 are finite real numbers, then $x' \sim \mathcal{SDTN}(m_0, m_1^2\eta^2, \rho)$.

The proof of [Lemma 3.2](#) is provided in Section A of the [Supplemental Document](#). It is worth pointing out that, while the sum of independent non-identically distributed Normal random variables is Normally distributed, it is not the case for SDTNs. The exact distribution of the sum of independent non-identically SDTNs is analytically intractable. However, the following Normality results hold.

Theorem 3.3. *For every $x_i \sim \mathcal{SDTN}(\mu_i, \eta_i^2, \rho_i)$, suppose that the random variables making up the collection $\mathbf{X}_k = \{x_i : 1 \leq i \leq k\}$ are independent with the following conditions:*

- μ_i are finite, i.e. $\max_{1 \leq i \leq k} \mu_i < +\infty$
- ρ_i is bounded from below by $\underline{\rho} > 0$
- η_i is bounded from below and above by $\underline{\eta} > 0$ and $\bar{\eta} < +\infty$, respectively.

Then

$$\frac{1}{t_k} \sum_{i=1}^k (x_i - \mu_i) \xrightarrow{d} \mathcal{N}(0, 1),$$

as $k \rightarrow \infty$, where

$$t_k^2 = \sum_{i=1}^k \mathbf{Var}[x_i].$$

The proof of [Theorem 3.3](#) is provided in Section B of the [Supplemental Document](#). Moreover, it is also straightforward to verify the following corollary to [Theorem 3.3](#).

Corollary 3.4. Let $x_i \sim \mathcal{SDTN}(\mu_i, \eta_i^2, \rho_i)$, $i = 1, 2, \dots$ be independent with μ_i 's, η_i 's, and ρ_i 's satisfying conditions in [Theorem 3.3](#). Let $\beta_i, i = 1, 2, \dots$, be real numbers and the absolute values are bounded from below and above, i.e. there exist $\underline{\beta}$ and $\bar{\beta}$ satisfying $0 < \underline{\beta} \leq |\beta_i| \leq \bar{\beta} < +\infty$ for all $i = 1, 2, \dots$. Then,

$$\frac{1}{t_k} \sum_{i=1}^k \beta_i (x_i - \mu_i) \xrightarrow{d} \mathcal{N}(0, 1), \quad (13)$$

as $k \rightarrow \infty$, where

$$t_k^2 = \sum_{i=1}^k \beta_i^2 \mathbf{Var}[x_i]. \quad (14)$$

Corollary 3.4 indicates that the (weighted) sum of finitely many independent but non-identically distributed SDTNs converges in distribution to a Normal distribution.

Theorem 3.3 and **Corollary 3.4** indicate that the (weighted) sum of finitely many random variables obeying SDTN distributions, while does not exactly follow a Normal distribution, approximately obeys a Normal distribution when the number of the random variables in the summation is large. For conciseness, let's only focus on one row of the data: in model (4), the response y is approximately Normal given β_i 's. It is easy to see that

$$\mathbf{E}[\gamma_i] = 0, \quad \text{and} \quad \mathbf{Var}[\gamma_i] = \eta_i^2 \left[1 - \frac{2\rho_i\phi(\rho_i)}{2\Phi(\rho_i) - 1} \right],$$

where $\rho_i = \frac{\beta_i}{\eta_i}$, $i = 1, \dots, k$. Therefore, the mean and variance of y given β_i 's and x_i 's are as follows

$$\begin{aligned} \mathbf{E}[y|\beta_i, x_i] &= \sum_{i=1}^p \beta_i x_i, \\ \mathbf{Var}[y|\beta_i, x_i] &= \sum_{i=1}^k x_i^2 \mathbf{Var}[\gamma_{\ell, i}] + \mathbf{Var}[\epsilon] = \sigma^2 + \sum_{i=1}^k x_i^2 \eta_i^2 \left[1 - \frac{2\rho_i\phi(\rho_i)}{2\Phi(\rho_i) - 1} \right], \end{aligned}$$

where k is the number of variables, for which random effect is considered. With the observed data, if we write the model in a matrix format, we have

$$\begin{aligned} \mathbf{E}[y|X, \boldsymbol{\beta}] &= X\boldsymbol{\beta}, \\ \mathbf{Var}[y|X, Z, \boldsymbol{\beta}] &= Z\Lambda Z^T + \sigma^2 \mathbf{I}_n, \end{aligned}$$

where

$$\Lambda = \begin{bmatrix} \Delta & & \\ & \ddots & \\ & & \Delta \end{bmatrix} \quad \text{and} \quad \Delta = \begin{bmatrix} \eta_1^2 \left[1 - \frac{2\rho_1\phi(\rho_1)}{2\Phi(\rho_1) - 1} \right] & & \\ & \ddots & \\ & & \eta_k^2 \left[1 - \frac{2\rho_k\phi(\rho_k)}{2\Phi(\rho_k) - 1} \right] \end{bmatrix}.$$

We approximate the distribution of y given data X, Z and all model parameters by a multivariate Normal distribution $\mathcal{N}(X\boldsymbol{\beta}, Z\Lambda Z^T + \sigma^2 \mathbf{I}_n)$. With this approximation, we will report the proposed estimation methods in the next Section.

4. Proposed estimation methods

4.1. Point estimation

We let $\boldsymbol{\eta} = (\eta_i)_{i=1}^k \in \mathbb{R}^k$. With an approximated distribution of y given the data (X, Z) and the parameters, $\mathcal{N}(X\boldsymbol{\beta}, Z\Lambda Z^T + \sigma^2 \mathbf{I}_n)$, we estimate the unknown parameters by maximizing the approximated log-likelihood function. In fact, the approximated log-likelihood function is given by:

$$\mathcal{L}_{\text{approx}}(\boldsymbol{\beta}, \boldsymbol{\eta}, \sigma) = -\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |V| - \frac{1}{2} (\mathbf{y} - X\boldsymbol{\beta})^T V^{-1} (\mathbf{y} - X\boldsymbol{\beta}), \quad (14)$$

with $V \triangleq Z\Lambda Z^T + \sigma^2 \mathbf{I}_n$. Due to the requirement that $\boldsymbol{\beta} \geq 0$, it is necessary to keep $\boldsymbol{\beta}$ in the likelihood function explicitly instead of profiling it out as in the unconstrained case, i.e. Eq. (5). Therefore, we propose to obtain estimates of $\boldsymbol{\beta}, \boldsymbol{\eta}$, and σ by solving the following constrained optimization problem:

$$\begin{aligned}
 (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}, \hat{\sigma}) &= \arg \min_{\substack{\boldsymbol{\beta}, \boldsymbol{\eta}, \sigma \\ \text{s.t.}}} (\mathbf{y} - X\boldsymbol{\beta})^T V^{-1} (\mathbf{y} - X\boldsymbol{\beta}) + \ln |V| \\
 &\quad \boldsymbol{\beta} \geq 0, \\
 &\quad \boldsymbol{\eta} \geq 0,
 \end{aligned} \tag{15}$$

with $V \triangleq Z\Lambda Z^T + \sigma^2 \mathbf{I}_n$. Note that there is no need to impose nonnegativity on σ in (15) because only σ^2 appears in the objective function. However, it is not the case for $\boldsymbol{\eta}$ due to the definition of ρ_i 's. We refer to this approach as approximated maximum likelihood (AML). In addition, inspired by the restricted log-likelihood function for the unconstrained case, We propose as a second approach to solve the following constrained optimization problem.

$$\begin{aligned}
 (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}, \hat{\sigma}) &= \arg \min_{\substack{\boldsymbol{\beta}, \boldsymbol{\eta}, \sigma \\ \text{s.t.}}} (\mathbf{y} - X\boldsymbol{\beta})^T V^{-1} (\mathbf{y} - X\boldsymbol{\beta}) + \ln |V| + \ln |X^T V^{-1} X| \\
 &\quad \boldsymbol{\beta} \geq 0, \\
 &\quad \boldsymbol{\eta} \geq 0.
 \end{aligned} \tag{16}$$

In a similar interpretation, we refer to the second approach as approximated restricted maximum likelihood (ARML). Let the optimal solution, most likely only local optimum due to the non-convexity of the objective functions, of either (15) or (16) be denoted as $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}, \hat{\sigma})$, we can then estimate the random effect coefficients $\boldsymbol{\gamma}$. In fact, the joint likelihood function in this case is

$$\prod_{\ell=1}^g \prod_{j=1}^{n_\ell} \left[\frac{1}{\hat{\sigma} \sqrt{2\pi}} \exp \left(-\frac{1}{2} \left(\frac{y_{\ell,j} - \sum_{i=1}^p (\hat{\beta}_i + \gamma_{\ell,i}) x_{\ell,j,i}}{\hat{\sigma}} \right)^2 \right) \prod_{i=1}^k \frac{1}{\hat{\eta}_i (2\Phi(\hat{\rho}_i) - 1)} \phi \left(\frac{\gamma_{\ell,i}}{\hat{\eta}_i} \right) \right],$$

where $\hat{\rho}_i \triangleq \frac{\hat{\beta}_i}{\hat{\eta}_i}$. Therefore the joint log-likelihood function after removing constants (with respect to $\boldsymbol{\gamma}$) is given by:

$$\mathcal{L}_\gamma(\boldsymbol{\gamma}) = \sum_{\ell=1}^g \sum_{j=1}^{n_\ell} -\frac{1}{2} \left[\left(\frac{\tilde{y}_{\ell,j} - \sum_{i=1}^p \gamma_{\ell,i} x_{\ell,j,i}}{\hat{\sigma}} \right)^2 + \sum_{i=1}^k \left(\frac{\gamma_{\ell,i}}{\hat{\eta}_i} \right)^2 \right], \tag{17}$$

where

$$\tilde{y}_{\ell,j} = y_{\ell,j} - \sum_{i=1}^p x_{\ell,j,i} \hat{\beta}_i, \quad \forall \ell = 1, \dots, g; j = 1, \dots, n_\ell.$$

Notice also that the $\mathcal{L}_\gamma(\boldsymbol{\gamma})$ is only defined in the following set:

$$\left\{ \boldsymbol{\gamma} \mid -\hat{\boldsymbol{\beta}} \leq \boldsymbol{\gamma}^\ell \leq \hat{\boldsymbol{\beta}}, \quad \forall \ell = 1, \dots, g \right\}.$$

Therefore, maximizing (17) is equivalent to solving the following constrained optimization problem

$$\begin{aligned}
 (\hat{\boldsymbol{\gamma}}^1, \dots, \hat{\boldsymbol{\gamma}}^g) &= \arg \min_{\substack{\boldsymbol{\gamma}^1, \dots, \boldsymbol{\gamma}^g \\ \text{s.t.}}} \sum_{\ell=1}^g \left[\frac{1}{\hat{\sigma}^2} (\tilde{\mathbf{y}}^\ell - Z^\ell \boldsymbol{\gamma}^\ell)^T (\tilde{\mathbf{y}}^\ell - Z^\ell \boldsymbol{\gamma}^\ell) + (\boldsymbol{\gamma}^\ell)^T (\hat{\boldsymbol{\Sigma}})^{-1} \boldsymbol{\gamma}^\ell \right] \\
 &\quad -\hat{\boldsymbol{\beta}} \leq \boldsymbol{\gamma}^\ell \leq \hat{\boldsymbol{\beta}}, \quad \forall \ell = 1, \dots, g,
 \end{aligned} \tag{18}$$

where

$$\tilde{\mathbf{y}}^\ell \triangleq \begin{bmatrix} \tilde{y}_{\ell,1} \\ \vdots \\ \tilde{y}_{\ell,n_\ell} \end{bmatrix} \in \mathbb{R}^{n_\ell}, \forall \ell = 1, \dots, g, \text{ and } \hat{\boldsymbol{\Sigma}} \triangleq \begin{bmatrix} \hat{\eta}_1^2 & & \\ & \ddots & \\ & & \hat{\eta}_k^2 \end{bmatrix}.$$

Note that (18) is decomposable to solving for each $\boldsymbol{\gamma}^\ell$ independently. In fact, for each $\ell = 1, \dots, g$, we can solve

$$\begin{aligned} \hat{\gamma}^\ell &= \min_{\gamma^\ell} \frac{1}{\hat{\sigma}^2} (\tilde{\gamma}^\ell - Z^\ell \gamma^\ell)^T (\tilde{\gamma}^\ell - Z^\ell \gamma^\ell) + (\gamma^\ell)^T (\hat{\Sigma})^{-1} \gamma^\ell \\ \text{s.t.} \quad & -\hat{\beta} \leq \gamma^\ell \leq \hat{\beta}. \end{aligned} \quad (19)$$

4.2. Test of hypothesis

We discuss how to conduct a test of hypothesis on the fixed effects, $\beta = \{\beta_1, \dots, \beta_p\}$.

4.2.1. Test of all β

We first consider a hypothesis test on β . To be more specific, under the non-negativity sign constraints, we test the following hypothesis.

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \text{atleastone } \beta_i > 0, \quad (20)$$

It is assumed that under H_1 at least one inequality is strict, i.e. $\beta_i > 0$ for some $i = 1, \dots, p$. Note that we propose a one-sided alternative test only in this paper to comply with the sign constraint. Under such a hypothesis, we propose to employ a likelihood ratio test (LRT) (Casella and Berger (2001)). In our case, while an exact likelihood function is not analytically tractable, an approximated likelihood function is available, i.e. Eq. (14) for the proposed AML approach for instance. Let $(\hat{\beta}_0, \hat{\eta}_0, \hat{\sigma}_0)$ denote the estimation under the null hypothesis and $(\hat{\beta}, \hat{\eta}, \hat{\sigma})$ denote the estimation under the alternative hypothesis. The LRT statistic is given by

$$T_{LR} = -2 \left[\mathcal{L}_{\text{approx}}(\hat{\beta}_0, \hat{\eta}_0, \hat{\sigma}_0) - \mathcal{L}_{\text{approx}}(\hat{\beta}, \hat{\eta}, \hat{\sigma}) \right]. \quad (21)$$

It makes intuitive sense since if the ratio of likelihood under H_1 over that under H_0 is large enough, we then have strong evidence to prefer H_1 over H_0 . Denote the significance level as α and the unknown threshold as c_α . The exact distribution of T_{LR} is extremely difficult to compute if not impossible, especially since truncated Normal distributions are involved in our statistical model. In a more practical setting, we follow Silvapulle and Sen (2005) to approximate the probability of the test statistic T_{LR} larger than the threshold as follows:

$$\mathcal{P}(T_{LR} \geq c_\alpha) \approx \sum_{i=0}^p \omega_i \mathcal{P}(\chi_i^2 \geq c_\alpha),$$

where χ_i^2 is a Chi-squared random variable and $i > 0$ is the degrees of freedom. When $i=0$, we define χ_0^2 as the Chi-squared distribution with zero degrees of freedom, whose probability density function takes 0 with probability 1. Note that the above equation is still not quite applicable since the ω_i 's are still unknown and the calculation of the exact ω_i 's is not an easy task, see Davidov, Fokianos, and Iliopoulos (2010) for a discussion. Section 3 of Silvapulle and Sen (2005) provided the following upper bound.

$$\sum_{i=0}^p \omega_i \mathcal{P}(\chi_i^2 \geq c_\alpha) = \sum_{i=1}^p \omega_i \mathcal{P}(\chi_i^2 \geq c_\alpha) \leq \frac{1}{2} \left(\mathcal{P}(\chi_{p-1}^2 \geq c_\alpha) + \mathcal{P}(\chi_p^2 \geq c_\alpha) \right). \quad (22)$$

In practice, given α we adopt the above upper bound to compute the corresponding c_α making the test more conservative of committing a false positive error. This more applicable practice was also adopted in Sec. 4 of Davidov and Rosen (2011) and Example 4.3.1 of Silvapulle and Sen (2005). We will examine its finite sample performance in Sec. 5.4.

4.2.2. Test of individual β_i

We discuss test of individual β_i 's. For each $i = 1, \dots, p$, we consider the following null and alternative hypothesis:

$$H_0 : \beta_i = 0 \quad \text{v.s.} \quad H_1 : \beta_i > 0. \quad (23)$$

Under the proposed AML approach, let

$$\begin{aligned} (\hat{\beta}_{0i}, \hat{\eta}_{0i}, \hat{\sigma}_{0i}) &= \arg \min_{\beta, \eta, \sigma} (\mathbf{y} - X\beta)^T V^{-1} (\mathbf{y} - X\beta) + \ln |V| \\ \text{s.t.} \quad &\beta_{-i} > 0, \\ &\eta \geq 0, \\ &\beta_i = 0, \end{aligned} \quad (24)$$

where $\beta_{-i} = \{\beta_1, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_p\}$ denotes a $p-1$ length vector without the i -th element. The LR test statistic is given by

$$T_{LR} = -2 \left[\mathcal{L}_{\text{approx}}(\hat{\beta}_{0i}, \hat{\eta}_{0i}, \hat{\sigma}_{0i}) - \mathcal{L}_{\text{approx}}(\hat{\beta}, \hat{\eta}, \hat{\sigma}) \right].$$

Notice that the optimization problem in (24) is the same as (15) with an additional constrain $\beta_i = 0$. The threshold c_α is still computed using the upper bound of Eq. (22). Since only one individual parameter is tested, the upper bound of Eq. (22) changes to $\mathcal{P}(\kappa \geq c_\alpha)$, where κ denotes a Chi-squared distribution with 1 degree of freedom. It makes intuitive sense since the difference for the number of “free” parameters between H_0 and H_1 is 1 when testing of individual β_i .

5. Simulation

5.1. Finite sample performance

We first recapitulate the notation as follows: n is the total sample size of the data; p is the number of independent variables having fixed effects; k is the number of independent variables for which random effects are considered, and $k \leq p$; g is the total number of groups and each group has the same sample size $n_\ell = n/g$, where $\ell = 1, \dots, g$. As discussed in Sec. 3, the sum of k SDTNs converges in distribution to a Normal distribution as $k \rightarrow +\infty$. In this section, we use simulation to investigate its finite sample performance. We specify the following setting.

- $p = k = 5, 10, 20, 30$
- $g = 10, 20, 40$
- $n_\ell = 50, \ell = 1, 2, \dots, g$

Therefore, the total combination is $4 \times 3 \times 1 = 12$. Note that, in this section, we include an intercept in the model, i.e. for $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_p\}$, the column vector $\mathbf{x}_1 \in \mathbb{R}^n$ is always $\mathbf{1}$. For each combination, we simulated $M=100$ different datasets with random effects following SDTN with known true parameters. Then we fitted a model with the proposed AML and ARML approaches. The estimated parameters $(\hat{\beta}_1, \dots, \hat{\beta}_p, \hat{\eta}_1, \dots, \hat{\eta}_k, \hat{\sigma})$ were then compared against the true parameters $(\beta_1, \dots, \beta_p, \eta_1, \dots, \eta_k, \sigma)$ used to simulate the data, and two metrics, the root mean square error (RMSE) and mean absolute percentage error (MAPE), were utilized to quantify the performance. The true parameters and the data generation process are described in detail in Section C.1 of the [Supplemental Document](#). Since each combination has $M = 100$ replications, we reported the mean RMSE and mean MAPE and their associated standard errors (sample standard deviation of the M RMSE or MAPE divided by \sqrt{M}) in Table 2. The smaller RMSE and MAPE are, the better the performance is.

Table 2. For each combination, the mean RMSE, mean MAPE and their corresponding standard errors (in parentheses) of $M = 100$ replications are reported.

		$g = 10$		$g = 20$		$g = 40$	
		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
AML	$k = 5$	1.028(0.056)	1.723(0.096)	0.878(0.060)	1.388(0.085)	0.790(0.040)	1.283(0.061)
	$k = 10$	0.375(0.038)	0.656(0.051)	0.181(0.016)	0.481(0.018)	0.150(0.011)	0.498(0.013)
	$k = 20$	0.135(0.004)	0.457(0.011)	0.139(0.002)	0.414(0.010)	0.142(0.002)	0.451(0.010)
	$k = 30$	0.132(0.001)	0.396(0.007)	0.121(0.001)	0.347(0.005)	0.139(0.001)	0.385(0.003)
ARML	$k = 5$	1.039(0.056)	1.747(0.091)	0.864(0.055)	1.391(0.082)	0.830(0.049)	1.308(0.074)
	$k = 10$	0.384(0.041)	0.668(0.056)	0.211(0.023)	0.487(0.028)	0.141(0.006)	0.452(0.010)
	$k = 20$	0.138(0.007)	0.462(0.014)	0.140(0.002)	0.455(0.010)	0.139(0.002)	0.448(0.009)
	$k = 30$	0.136(0.001)	0.402(0.007)	0.119(0.001)	0.344(0.004)	0.137(0.001)	0.375(0.003)

Both metrics are calculated by comparing the estimated parameters against the true parameters used to simulate the data, and $n_\ell = 50$ is same for all g , i.e. the total sample size is $n = g \times n_\ell$. AML and ARML are the proposed approaches in Sec. 4. The smaller RMSE and MAPE are, the better the performance is.

From Table 2, it is clear that given the same g , as k increases from 5 to 30, both RMSE and MAPE decrease dramatically. For instance, with $k = 5, g = 20$, the RMSE is 1.028 and MAPE is 1.723, and they then drop to 0.132 and 0.396, respectively when k increases to 30. The decreasing trend of RMSE and MAPE is consistent for all the combinations, which shows a satisfactory performance when k is finite, i.e. $k \leq 30$. Both the performance of AML and ARML are satisfactory, and we do not observe an obvious performance difference between the two proposed methods.

5.2. Merits of sign constraints

As discussed in previous sections, sign constraints on the regression parameters not only lead to practical benefits such as resulting estimates complying with business knowledge automatically, but also bear theoretical merits that yield more accurate estimates by shrinking the feasible parameter space. We aim to show the latter point in this subsection. As manifested in Sec. 5.1 that the larger k is, the better the performance of the proposed methods is. We then deliberately make all p, k, n_ℓ and g very small by simulating one dataset ($M = 1$) with $p = 3, k = 1, g = 2$ to investigate the performance under “adverse” circumstance. The objective function in Eq. (15) is deemed as a function of β_2, β_3 to facilitate producing contour plots. The true values for β_2 and β_3 are chosen as 0.001, which are very close to 0. All other parameters are fixed at their true values. The true parameters used to generate the data are provided in Section C.2 in the Supplemental Document. $n_\ell = 15$ is chosen, i.e. $n = g \times n_\ell = 30$, which is also very small. The estimation results are reported in Table 3. The contour plots are presented in Figures 1 and 2.

It is clear from Table 3 that both estimations find better results as measured by the objective value of Eq. (15), for which the lower, the better. Due to the small sample size, $n_\ell = 15$ used to simulate data, variation is relatively large, and it is not unexpected that although both $\beta_2 = 0.001$ and $\beta_3 = 0.001$ are positive, the estimates without any constraint end up with $\hat{\beta}_2 = -0.026$ that yields the lowest objective value of 13.340. However, with non-negative sign constraints on both parameters, it actually shrinks the feasible parameter space, leading to a similar objective value, 13.366. More importantly, it “corrects” the unconstrained estimate of β_2 to 0.000, while leaving β_3 largely unaffected. The contour plots in Figure 1 confirm the observation.

In addition, we present Figure 2, in which two contour lines, 13.340 and 13.366 are specifically drawn. We also add a vertical line $\beta_2 = 0$ in red to represent the boundary for β_2 in the same figure. For contour line 13.340, it is apparent that there does exist regions of (β_2, β_3) that comply with non-negative sign constraints, but the unconstrained algorithm was not able to arrive at any of those solutions. Instead it reported an alternative solution with wrong sign for β_2 of $(-0.026, 0.109)$ as reported in Table 3. Now, with the help of sign constraints, the constrained optimization picked $(0.000, 0.093)$ with a slightly worse objective value. We are confident that,

Table 3. Estimation results of unconstrained case and constrained case. Lower objective values are better.

	β_2	β_3	Objective value of (15)
True parameters plugged-in	0.001	0.001	14.152
Estimation without sign constraints	−0.026	0.109	13.340
Estimation with non-negative sign constraints	0.000	0.093	13.366

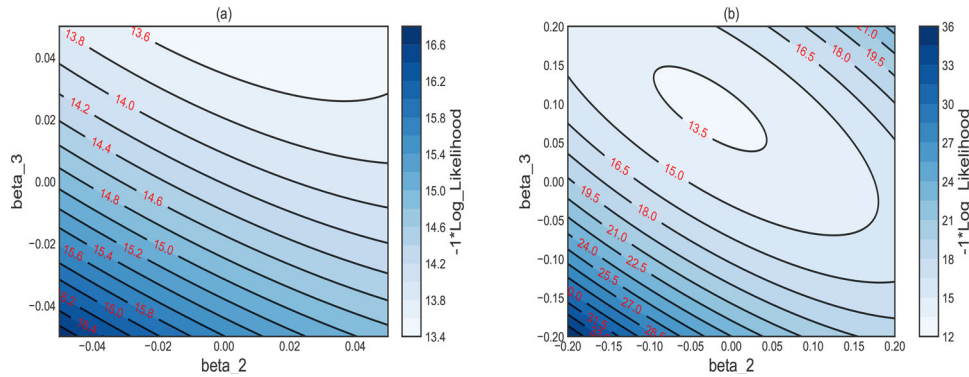


Figure 1. Contour plots of the log-likelihood function in Eq. (15) viewed as a function of β_2 and β_3 . The data range for (a) is -0.05 to 0.05 , while the range for (b) is -0.02 to 0.02 .

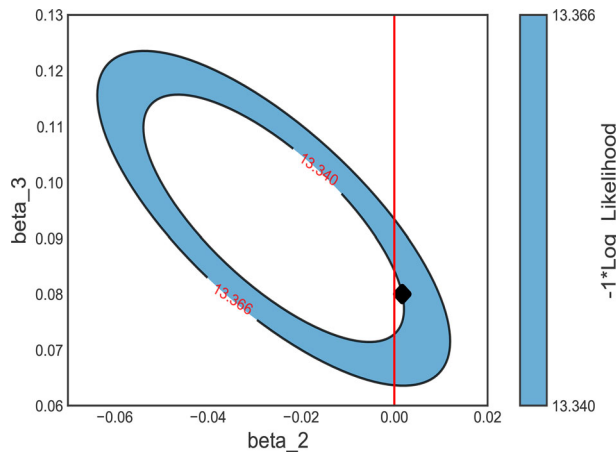


Figure 2. Contour lines of the log-likelihood function in Eq. (15) viewed as a function of β_2 and β_3 . Two contour lines are specifically plotted: 13.340 and 13.366. A correct sign alternative solution is plotted as a diamond marker on the 13.340 contour line.

with proper tuning of the optimization routine, we will arrive at a correct-sign alternative solution on the 13.340 contour line. Actually, with $[\beta_2 = 0.0017, \beta_3 = 0.0800]$ we have already found a correct sign alternative solution (the diamond marker) on the 13.340 contour line in Figure 2, which is an even more convincing merit of sign constraints. In other words, with close-to-boundary true parameters, the merits of imposing sign constraints on the regression parameters allowed the selection of alternative solutions that conforms with interpretability of the parameters.

5.3. Comparisons with the method of PIT

It is interesting to directly compare the performance of the proposed methods to the method of PIT (Nelson et al. 2006), which has been briefly reviewed in Sec. 1. Both the proposed methods and the PIT method are able to handle LME models with non-Normal random effects. First of

Table 4. For each combination, the mean RMSE, mean MAPE and their corresponding standard errors (in parentheses) of $M = 100$ repeats are reported.

	PIT $Q = 2$		PIT $Q = 4$	
	RMSE	MAPE	RMSE	MAPE
$g = 10$	0.473(0.125)	1.257(0.210)	0.274(0.035)	0.916(0.061)
$g = 20$	0.238(0.011)	0.871(0.021)	0.244(0.021)	0.876(0.037)
$g = 40$	0.231(0.001)	0.862(0.004)	0.228(0.002)	0.851(0.006)
	AML		ARML	
	RMSE	MAPE	RMSE	MAPE
$g = 10$	0.182(0.003)	0.695(0.016)	0.184(0.003)	0.704(0.015)
$g = 20$	0.175(0.002)	0.633(0.015)	0.178(0.002)	0.648(0.014)
$g = 40$	0.174(0.002)	0.632(0.015)	0.173(0.002)	0.622(0.014)

Both metrics are calculated by comparing the estimated parameters against the true parameters used to simulate the data. $k = 1$, $p = 2$ are same for all combinations, and $n_\ell = 50$ is same for all g , i.e. the total sample size is $n = g \times n_\ell$.

all, the proposed approaches are discriminative in nature, approximating the marginal distributions of the response directly in estimating fixed effects, while the PIT approach is a generative method that depends on the joint distribution of random effects and response variable. Secondly, both methods work best with independent covariance structure on random effects, and both require some non-trivial future work in order to deal with random effects with dependent covariance structure. Thirdly, it is easier using the proposed methods to incorporate more than one random effect as demonstrated in previous sections, while the PIT method will require the approximation of multiple integrals, in which the so-called “curse of dimensionality” could factor in. Last but not the least, following the notations of Nelson et al. (2006), the original formulation of PIT method involves the sum of Q products of n_i probability densities, which are inside the logarithmic function so that those multiplications cannot further be converted into summations of logarithms. See the last equation in Sec. 3 of Nelson et al. (2006). Therefore, compared to the PIT method, the proposed methods are less sensitive to numerical issues, and is numerically more stable especially when n_i is large.

To the best of our knowledge, we are not aware of an implementation of the PIT method in either Python or R. Hence, we implemented it according to the original formulation. The only enhancement was to take natural logarithm of the last equation in Sec. 3 of Nelson et al. (2006), otherwise it will quickly lead to numerical issues when optimizing it. With a slight abuse of notations (only in this section), following the notations of Nelson et al. (2006), we solve the following optimization problem

$$\min_{\beta, \theta} \sum_{i=1}^n \left(-\ln \left(\sum_{q=1}^Q \prod_{k=1}^{n_i} f(y_{ik} | x_{ik}, F_{\theta}^{-1}(\Phi(d_q)), \beta) \phi(d_q) w_q \right) \right). \quad (25)$$

The number of points used to approximate integrals is either 2 or 4, i.e. $Q = 2, 4$ to keep a balance between accuracy and computing time in this section, and the values of $z_q, \eta_q, q = 1, \dots, Q$ are found in Table 25.10 of Abramowitz, Stegun, and Romer (1988). Similar to previous section, we have $g = 10, 20, 40$ different combinations, and $p = 2, k = 1, n_\ell = 50$ are fixed for all combinations. For each combination, $M = 100$ different datasets are simulated, and the estimation results are reported in Table 4. The true parameters are same as those reported in Section C.1 in the [Supplemental Document](#).

From Table 4, for each combination, the mean RMSE and mean MAPE of the proposed methods are lower than those of PIT with $Q = 2$ and $Q = 4$ showing the superior performance in terms of the estimation accuracy. Similar to the observation made in Sec. 5.1, we do not observe an obvious performance difference between the two proposed methods.

Table 5. For each combination, the empirical true negative rate, which is defined as $1 - \text{false positive (type I error) rate}$, is calculated based on $M = 2000$ replications and is reported in the table. $g = 20$ is fixed for all combinations, n_ℓ is the same for all g and the total sample size is $n = g \times n_\ell$. The significance level is set as 0.05.

		Test of individual β_1		Test of all β	
		$p = 5$	$p = 10$	$p = 5$	$p = 10$
AML	$n_\ell = 25$	0.973	0.975	0.937	0.956
	$n_\ell = 50$	0.967	0.971	0.964	0.960
ARML	$n_\ell = 25$	0.974	0.974	0.943	0.956
	$n_\ell = 50$	0.968	0.970	0.963	0.959

5.4. Assessment of the proposed test of hypothesis

We next evaluate the test of hypothesis proposed in Sec. 4.2 when k is relatively small. In this section, we use the following setting.

- $p = 5, 10$
- $n_\ell = 25, 50$, i.e. $n = g \times n_\ell$
- $k = 1$, $g = 20$, $M = 2000$, $\alpha = 0.05$ (significance level)

In other words, for each combination, we compute the empirical true negative (TN) rate, which is defined as $1 - \text{false positive (type I error) rate}$ based on the M repeats. The theoretical TN rate is $1 - 0.05 = 0.95$. We test both an individual regression parameter, and all regression parameters. The hypotheses are as follows.

- Test of an individual regression parameter. With $p = 5$ as an illustrating example, we test β_2 (the first non-intercept regression parameter). $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 > 0$
- Test of all regression parameter. Let $\beta = \{\beta_1, \dots, \beta_p\}$, we test $H_0 : \beta = 0$ versus $H_1 : \text{at least one } \beta_j > 0, \text{ where } j = 1, \dots, p$.

The results are reported in Table 5, from which we observe that except for one combination ($p = 5, n_\ell = 25$) where the empirical TN is slightly under-covered, the rest are all above the theoretical TN rate, which exhibits the satisfactory performance when k is merely 1.

6. Real-world applications

6.1. The discounted sales data

Let us revisit the motivating example introduced in Sec. 1. Since the performance of the PIT method is not very satisfying in Sec. 5.3, the traditional LME model is included for comparison purposes, and it was fitted using the *lme4* package in R. For the real world example, we do not know the true parameters, so we will use marginal R^2 and conditional R^2 to compare between competing methods. For a linear mixed effects model, we follow Nakagawa and Schielzeth (2013) in using the marginal R^2 and the conditional R^2 , where the marginal R^2 measures the proportion of the variance that the fixed effects can explain: the numerator is the variance of fixed effects, while the denominator is the total variance of the model, i.e. the sum of the variance of the fixed effects, variance of all random effects and variance of the error. In a similar fashion, conditional R^2 depicts the proportion of the variance that the whole regression model, i.e. both the fixed effects and random effects can explain. These two metrics are natural extensions of the usual R^2 to the mixed effects models, and they are between 0 to 1 inclusively. The model results of AML, ARML and *lme4* are reported in Table 6 for the estimated fixed coefficients.

Table 6. Model results for the discounted sales example.

	<i>lme4</i>	AML	ARML
β_0	0.106	0.164	0.164
β_1	-0.319	0.259	0.270
S.D. of random intercept	0.671	0.095	0.094
S.D. of random slope	0.361	0.075	0.156
S.D. of residuals	1.388	1.445	1.447
Marginal R^2	0.000	0.000	0.000
Conditional R^2	0.190	0.488	0.606

Note that in Table 6, the variance of the fixed effect is 1.452×10^{-4} for *lme4*, which is very small compared to the variance of the random effects and the residuals in the same table rendering the marginal R^2 very much close to 0 (it is essentially 0.000 with 3 decimal points). Even the conditional R^2 (variance of the fixed effect and random effect) is able to explain merely less than 20% of the total variance for *lme4*. Hence the major uncertainty comes from the unexplained residuals for *lme4*. The proposed methods perform better with conditional R^2 of 0.488 and 0.606, respectively, while the marginal R^2 is also very close to 0 indicating the immediate need of using a mixed effects model. In addition, from Table 6, it is very apparent that both the proposed models produce non-negative $\hat{\beta}_1$ as compared to -0.319 of the traditional LME. With the proposed methods, heuristics for correcting the “wrong” signs are no longer needed, and the modeling results can be applied directly in practice. The proposed methods not only preserve model interpretability and sign correctness, but also have better model fitting as measured by the conditional R^2 .

6.2. The sleep deprivation study

Consider a dataset on the reaction time per day for subjects in a sleep deprivation study (Bolker et al. 2013). This dataset is accessible in the *lme4* package in R (Bates et al. 2014). There are 18 subjects in the dataset. On day 0, subjects had the normal amount of sleep, followed by the next 10 days when they were restricted to 3 hours of sleep per night only. The response variable is the reaction times in milliseconds.

Before we conduct any formal analysis, we plotted the data grouped by subjects as shown in Figure 3. While most of the graphs match our intuition that the response time increases as the number of sleep-deprived days, we observe that the response curves across subjects are rather different. Thus, we use the full model that includes both the random intercept and random slope. We consider *lme4*, AML and ARML and report the results in Table 7. From Table 7, the estimated regression coefficients are close for all the three methods considered: for example, the estimated coefficient for Days are 10.467, 10.789, 10.795, respectively, for *lme4*, AML and ARML. In terms of model fits, the proposed methods clearly outperform *lme4* since the conditional R^2 for PLS and PRLS are 0.803 and 0.817 that are apparently higher than 0.702 of *lme4*. The same pattern holds for marginal R^2 as well.

We also report the estimated random effects. We show the overall effect (fixed effect + random effect) of all 18 subjects in Table 8. It is observed that the estimates are similar across the three methods for all subjects. In addition, the estimates are more consistent between AML and ARML than between AML/ARML and *lme4*. This is not unexpected as the two proposed methods share more similarity than the *lme4* method. Among all of the 18 subjects, the most interesting individual is subject 335, where the estimated slope is negative from *lme4*, while the estimates from the proposed methods are 0.000. This is a direct result of the use of SDTN in the model specification, and it is proposed that the modeling system should issue an anomaly warning in such a case, indicating a further investigation might be warranted for issues such as data collection or input error.

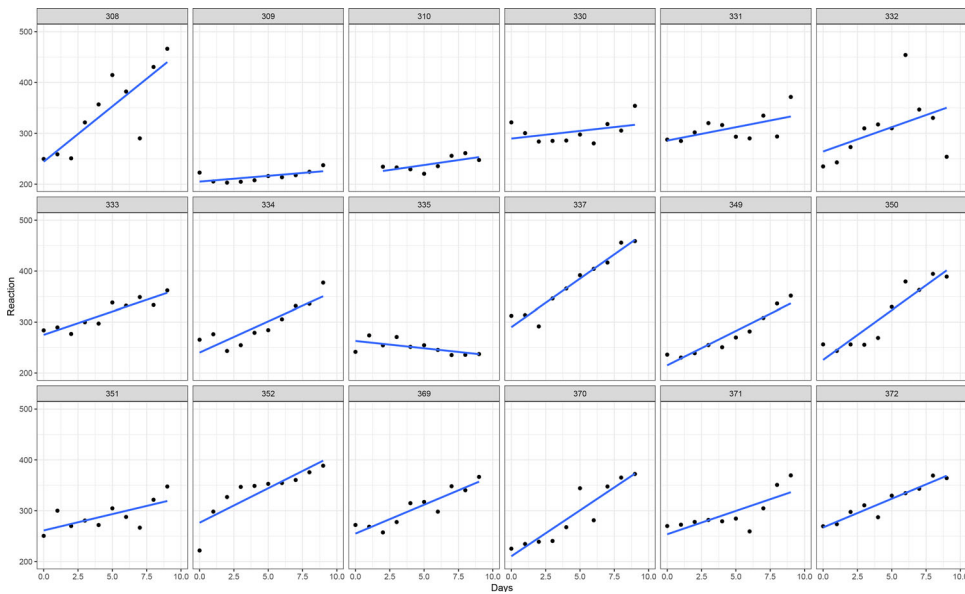


Figure 3. Scatter plots by the 18 subject.

Table 7. Model results for the sleep deprivation study.

	<i>lme4</i>	AML	ARML
Intercept	251.405	250.389	250.356
Days	10.467	10.789	10.795
S.D. of random intercept	25.051	25.555	26.407
S.D. of random slope	5.988	6.228	6.232
S.D. of residuals	25.565	20.140	19.509
Marginal R^2	0.415	0.467	0.463
Conditional R^2	0.702	0.803	0.817

Table 8. The overall effects (fixed effect + random effect) of the 18 subjects.

		Subject 308	Subject 309	Subject 310	Subject 330	Subject 331	Subject 332
<i>lme4</i>	Overall intercept	252.918	211.031	212.224	275.924	274.320	260.627
	Overall slope	19.791	1.868	5.079	5.499	7.273	10.159
PLS	Overall intercept	244.621	208.148	202.232	297.813	288.122	246.798
	Overall slope	21.580	1.330	6.512	5.652	4.700	14.301
PRLS	Overall intercept	244.592	208.110	202.237	297.743	288.107	246.910
	Overall slope	21.581	1.340	6.503	5.672	4.688	14.256
		Subject 333	Subject 334	Subject 335	Subject 337	Subject 349	Subject 350
<i>lme4</i>	Overall intercept	268.561	243.953	251.984	286.173	225.651	237.540
	Overall slope	10.180	11.583	−0.439	19.095	11.748	17.224
PLS	Overall intercept	275.692	247.760	253.771	292.131	220.663	229.012
	Overall slope	8.881	10.031	0.000	18.323	11.898	18.401
PRLS	Overall intercept	274.682	247.612	253.777	292.071	220.612	228.478
	Overall slope	8.912	10.051	0.000	18.350	11.910	18.451
		Subject 351	Subject 352	Subject 369	Subject 370	Subject 371	Subject 372
<i>lme4</i>	Overall intercept	256.321	272.334	254.664	224.929	252.311	263.827
	Overall slope	7.392	13.979	11.340	15.451	9.462	11.726
PLS	Overall intercept	265.060	268.932	257.112	212.870	261.842	267.758
	Overall slope	5.447	15.921	10.666	17.289	6.947	11.101
PRLS	Overall intercept	265.030	269.541	257.084	212.810	261.749	267.778
	Overall slope	5.513	15.848	10.689	17.292	6.970	11.117

7. Concluding Remarks

In this paper, we work under the framework of linear mixed effects model. We assume the SDTN distribution on the random effects instead of the Normal distribution to impose sign constraints on the overall effects in a theoretically sound way. This change has profound impact on the estimation methods because the exact distribution of y becomes analytically intractable. We lay a solid foundation by establishing properties of a SDTN distribution and then proposed two methods: AML and ARML for estimating the unknown model parameters. Both the simulation studies and the application examples show their satisfactory performance. When there is practical justification or domain knowledge to impose sign constraints on the overall regression parameters, the proposed methods work best in finding alternative solutions that allow intuitive interpretation of the results.

We discuss three future extensions motivated by this research. First, a natural extension of this research is to consider the generalized linear mixed effects model (GLMM) so that it can be applied to broader types of data such as binary or discrete outcomes. The framework of the linear mixed effects model is sufficient for our current practical needs, but GLMM has broader applications in social and economic research, medical studies, and the pharmaceutical industry. Second, although we discussed hypothesis testing in Sec. 4.2 and empirically assessed its performance in Sec. 5.4, it is difficult if not impossible at all to analytically derive the standard error of the estimators due to the lack of an explicit expression for the log likelihood function with SDTN assumed on the random effects. This is one of the non-trivial future research directions we wish to pursue. In addition, we generally agree with the opinion expressed in Sec. 3.3 of Silvapulle and Sen (2005) that the likelihood ratio test for constrained inference problems is difficult, especially in our case where the distribution of the response variable can only be approximated under a linear mixed effects model. Relying on the upper bound for calculation of a conservative p-value is a practical solution. However, much more research is needed to complete a follow up paper specifically discussing constrained inference under a mixed effects model with truncated Normally distributed random effects. Finally, the random coefficients of a mixed effects model are typically correlated, so an extension of the SDTN approach to dependent random coefficients is another valuable future research topic.

Acknowledgment

We thank the editor and the anonymous reviewer for suggestions that substantially improved the manuscript.

ORCID

Alvin Lim  <http://orcid.org/0000-0001-7886-5643>

References

- Abramowitz, M., I. A. Stegun, and R. H. Romer. 1988. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. New York, NY, USA: American Association of Physics Teachers.
- Bates, D., M. Mächler, B. Bolker, and S. Walker. 2014. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67 (1):1–48. doi:10.18637/jss.v067.i01.
- Bolker, B. M., B. Gardner, M. Maunder, C. W. Berg, M. Brooks, L. Comita, E. Crone, S. Cubaynes, T. Davies, P. de Valpine, et al. 2013. Strategies for fitting nonlinear ecological models in R, AD Model Builder, and bugs. *Methods in Ecology and Evolution* 4 (6):501–12. doi:10.1111/2041-210X.12044.
- Brabec, M., O. Konár, E. Pelikán, and M. Malý. 2008. A nonlinear mixed effects model for the prediction of natural gas consumption by individual customers. *International Journal of Forecasting* 24 (4):659–78. doi:10.1016/j.ijforecast.2008.08.005.

- Bronnenberg, B. J., S. K. Dhar, and J.-P. Dubé. 2007. Consumer packaged goods in the united states: National brands, local branding. *Journal of Marketing Research* 44 (1):4–13. doi:[10.1509/jmkr.44.1.4](https://doi.org/10.1509/jmkr.44.1.4).
- Carlin, B. P., and T. A. Louis. 2009. *Bayesian methods for data analysis*. 3rd ed. Boca Raton, FL: Chapman and Hall/CRC.
- Casella, G., and R. L. Berger. 2001. *Statistical inference*. 2nd ed. Pacific Grove, CA, USA: Cengage Learning.
- Davidov, O., K. Fokianos, and G. Iliopoulos. 2010. Order-restricted semiparametric inference for the power bias model. *Biometrics* 66 (2):549–57.
- Davidov, O., and S. Rosen. 2011. Constrained inference in mixed-effects models for longitudinal data with application to hearing loss. *Biostatistics (Oxford, England)* 12 (2):327–40.
- Demidenko, E. 2013. *Mixed models: Theory and applications with R*. Hoboken, NJ, USA: John Wiley & Sons.
- Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1):1–22.
- Hoerl, A. E., and R. W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12 (1):55–67. [Database] doi:[10.1080/00401706.1970.10488634](https://doi.org/10.1080/00401706.1970.10488634).
- Jiang, J. 2007. *Linear and generalized linear mixed models and their applications*. Berlin, Germany: Springer Science & Business Media.
- Lindstrom, M. J., and D. M. Bates. 1988. Newton–Raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association* 83 (404):1014–22.
- Liu, L., and Z. Yu. 2008. A likelihood reformulation method in non-normal random effects models. *Statistics in Medicine* 27 (16):3105–24.
- McCulloch, C. E., and J. M. Neuhaus. 2014. *Generalized linear mixed models*. Hoboken, NJ, USA: Wiley StatsRef.
- Mitsumata, K., S. Saitoh, H. Ohnishi, H. Akasaka, and T. Miura. 2012. Effects of parental hypertension on longitudinal trends in blood pressure and plasma metabolic profile: Mixed-effects model analysis. *Hypertension (Dallas, Tex.: 1979)* 60 (5):1124–30. doi:[10.1161/HYPERTENSIONAHA.112.201129](https://doi.org/10.1161/HYPERTENSIONAHA.112.201129).
- Nakagawa, S., and H. Schielzeth. 2013. A general and simple method for obtaining r^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4 (2):133–42. doi:[10.1111/j.2041-210x.2012.00261.x](https://doi.org/10.1111/j.2041-210x.2012.00261.x).
- Nelson, K. P., S. R. Lipsitz, G. M. Fitzmaurice, J. Ibrahim, M. Parzen, and R. Strawderman. 2006. Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects. *Journal of Computational and Graphical Statistics* 15 (1):39–57. doi:[10.1198/106186006X96854](https://doi.org/10.1198/106186006X96854).
- Olive, D. J. 2008. Applied robust statistics. Preprint M-02-006
- Pinheiro, J. C., C. Liu, and Y. N. Wu. 2001. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* 10 (2):249–76. doi:[10.1198/10618600152628059](https://doi.org/10.1198/10618600152628059).
- Silvapulle, M. J., and P. K. Sen. 2005. *Constrained statistical inference: Inequality, order and shape restrictions*. Hoboken, NJ, USA: John Wiley & Sons.
- Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1):267–88.
- Wolfinger, R., and M. O’Connell. 1993. Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* 48 (3–4):233–43. doi:[10.1080/00949659308811554](https://doi.org/10.1080/00949659308811554).
- Wolfinger, R., R. Tobias, and J. Sall. 1994. Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal on Scientific Computing* 15 (6):1294–310. doi:[10.1137/0915079](https://doi.org/10.1137/0915079).
- Wu, L. 2009. *Mixed effects models for complex data*. Boca Raton, FL, USA: CRC Press.
- Yucel, R. M., and H. Demirtas. 2010. Impact of non-normal random effects on inference by multiple imputation: A simulation assessment. *Computational Statistics & Data Analysis* 54 (3):790–801. doi:[10.1016/j.csda.2009.01.016](https://doi.org/10.1016/j.csda.2009.01.016).
- Zhang, X. 2015. A tutorial on restricted maximum likelihood estimation in linear regression and linear mixed-effects model. URL <http://statdb1.uos.ac.kr/teaching/multi-grad/ReML.pdf>.
- Zolotarev, V. M. 1967. A generalization of the lindeberg-feller theorem. *Theory of Probability & Its Applications* 12 (4):608–18. doi:[10.1137/1112076](https://doi.org/10.1137/1112076).