Computational Statistics and Data Analysis 213 (2026) 108254

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Random effects misspecification and its consequences for prediction in generalized linear mixed models

Quan Vu [a,*], Francis K.C. Hui [a], Samuel Muller [b], A.H. Welsh [a]

[a] Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Australia
[b] Faculty of Science and Engineering, Macquarie University, Australia

ARTICLE INFO

Keywords:
Clustered data
Empirical best predictor
Longitudinal data
Mean squared error of prediction
Prediction inference

ABSTRACT

When fitting generalized linear mixed models, choosing the random effects distribution is an important decision. As random effects are unobserved, misspecification of their distribution is a real possibility. Thus, the consequences of random effects misspecification for point prediction and prediction inference of random effects in generalized linear mixed models need to be investigated. A combination of theory, simulation, and a real application is used to explore the effect of using the common normality assumption for the random effects distribution when the correct specification is a mixture of normal distributions, focusing on the impacts on point prediction, mean squared prediction errors, and prediction intervals. Results show that the level of shrinkage for the predicted random effects can differ greatly under the two random effect distributions, and so is susceptible to misspecification. Also, the unconditional mean squared prediction errors for the random effects are almost always larger under the misspecified normal random effects distribution, while results for the mean squared prediction errors conditional on the random effects are more complicated but remain generally larger under the misspecified distribution (especially when the true random effect is close to the mean of one of the component distributions in the true mixture distribution). Results for prediction intervals indicate that the overall coverage probability is, in contrast, not greatly impacted by misspecification. It is concluded that misspecifying the random effects distribution can affect prediction of random effects, and greater caution is recommended when adopting the normality assumption in generalized linear mixed models.

1. Introduction

Generalized linear mixed models (GLMMs, Jiang and Nguyen, 2007; McCulloch et al., 2011) are widely used in many disciplines to model correlations between observations. These correlations are induced through the introduction of unobserved random effects. As such, one important decision to make when fitting GLMMs is what distribution to assume for the random effects, and in practice the vast majority of applications and statistical software packages for fitting such models e.g., `lme4` (Bates et al., 2015) and `glmmTMB` (Brooks et al., 2017), assume a normal distribution. As a motivating example, we consider a study to examine the effect of time spent in the workforce on hourly wages, where we found evidence that the random effects distribution exhibits deviations from the usually assumed normal distribution; see Section 5 for more details. Such examples raise the question of what the consequences are for estimation, inference, and prediction when we misspecify the random effects distribution in GLMMs?

* Corresponding author.
E-mail address: quan.vu@anu.edu.au (Q. Vu).

https://doi.org/10.1016/j.csda.2025.108254
Received 26 February 2025; Received in revised form 22 July 2025; Accepted 22 July 2025

Available online 29 July 2025

In this article, we investigate the impact of random effects misspecification on point prediction—specifically, the widely used best predictors that minimize mean squared prediction error—as well as on the inference for random effects in independent-cluster generalized linear mixed models (GLMMs). Such investigation is essential given that prediction underlies the main objective in many applications of GLMMs, e.g., social studies (e.g., Marino et al., 2019) and agriculture (e.g., Smith et al., 2005). Despite its importance, however, this particular area of research remains understudied, as noted by Hui et al. (2021). Moreover, the research literature presents conflicting results. Verbeke and Lesaffre (1996) and McCulloch and Neuhaus (2011a) showed empirically that for independent-cluster GLMMs, the normality assumption results in too "normal-like" predicted random effects distribution when the true distribution is not normal, and concluded that determining the true shape of the random effects distribution from predictors of the random effects is a challenging task. This contrasts with results of Hui et al. (2021), who found for linear mixed models that the shape of the predicted random effects distribution can exhibit clear non-normality, and resembles the true shape when the cluster size is sufficiently large; see also the recent related asymptotic results by Lyu and Welsh (2022) and Ning et al. (2024) which support this idea. Agresti et al. (2004) found substantial bias in the best predictors when the true random effects distribution was a mixture with a large variance. This contrasts with McCulloch and Neuhaus (2011b), who found for both best prediction and prediction using the mode that the unconditional mean square error of prediction (UMSEP) was fairly robust to the random effects distribution, provided the true random effects variance was not too large and cluster sizes were moderate. The UMSEP is widely used for uncertainty quantification and constructing prediction intervals, particularly in small-area estimation (e.g., Rao and Molina, 2015). However, other researchers have proposed using the conditional mean squared error of prediction for cluster-level prediction, either conditioning on the random effects (e.g., Zheng and Cadigan, 2023) or conditioning on the responses from that cluster (Booth and Hobert, 1998; Lee et al., 2011; Korhonen et al., 2023). How random effects misspecification impacts such conditional MSEPs is largely unknown.

In this article, we study a particular form of random effects misspecification in independent-cluster GLMMs, namely assuming a normal random effects distribution when the true random effects follow a finite mixture of normals distribution. The novel contribution of our article is the comparison of closed-forms of the MSEPs conditional on the random effects (henceforth abbreviated as CeMSEPs) under both the true (a finite mixture of normals) and misspecified (a normal) random effects distributions. We further compare closed-forms of the UMSEPs under the two distributions. These results offer analytical insights into the impacts of misspecification. Note that while other non-normal random effects distributions could be considered (e.g., see the works of Litiere et al., 2008; McCulloch and Neuhaus, 2011a; Hui et al., 2021), we purposely use a finite mixture of normals since it can represent a wide range of continuous distributions including those with multimodality, skewness, and heavy-tailed behavior (e.g., Nguyen and McLachlan, 2019); see also Komárek and Lesaffre (2008); Liu and Yu (2008); Pan et al. (2020) where finite mixture random effects distributions have been used in GLMMs. We further investigate the impact on prediction intervals constructed from the point predictors and their corresponding UMSEPs and CeMSEPs. Our theoretical results are supported by numerical studies with random effects simulated from mixtures of distributions ranging from mild to extreme deviation from the normal distribution, and a case study of hourly wages data.

We summarize our results into three main findings: First, the UMSEP obtained under the misspecified normal distribution is consistently larger, with the differences being more pronounced when cluster sizes are small and the number of clusters is large. Second, the CeMSEP is also larger under the misspecified normal random effects distribution, although the extent depends on the true finite mixture distribution and the response distribution. For instance, if the true random effects distribution is asymmetric, then CeMSEPs under misspecification are larger when the true random effect is near the mean of one of the mixture components. Third, coverage probabilities for prediction intervals using UMSEPs vary widely across clusters (which is not too surprising given that UMSEP is based on averaging over all clusters) and regardless of the true random effects distribution, while prediction intervals using CeMSEPs tend to achieve nominal coverage even under misspecification.

Overall, our results suggest that random effects misspecification can have negative consequences for point prediction and prediction inference in GLMMs, and we encourage greater caution when it comes to adopting the standard normality assumption in GLMMs. If fitting a mixed model with this assumption leads to clear non-normal random effect predictions, then the assumption should be reconsidered and modified appropriately. If the dataset has mostly small clusters (with "small" depending on the response distribution), the impact of misspecification will be more severe. For example, in the case study with wages data in Section 5, the normality assumption results in larger CeMSEPs at the tail of the random effects distribution, which denotes that prediction for high-earning individuals is negatively impacted under misspecification. Instead, fitting a model with a mixture random effects distribution reduces the CeMSEPs at the tail.

The remainder of this article is organized as follows: Section 2 reviews independent-cluster GLMMs and the (empirical) best predictors of the random effects, before presenting forms for the UMSEP and CeMSEP. Section 3 focuses on the special case of LMMs, and derives closed-form expressions for the best predictors, UMSEPs, and CeMSEPS under the normal and mixture of normal distributions for the random effects. These derivations are supported by simulation studies with binary and count responses in Section 4, and an application to the motivating longitudinal hourly wages data in Section 5. Section 6 offers some concluding remarks.

2. Independent-cluster GLMMs and prediction

Consider a set of independent clusters $i = 1, \dots, m$, where in cluster i we observe the response vector $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^\top$ along with, at each observation $j = 1, \dots, n_i$, a vector of fixed effect covariates $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijq_f})^\top$, and a vector of random effect covariates $\mathbf{z}_{ij} = (z_{ij1}, \dots, z_{ijq_r})^\top$. We assume the first elements $x_{ij1} = z_{ij1} = 1$ to represent a fixed and a random intercept term, respectively. We fit an independent-cluster GLMM to this data, in which the conditional distribution of y_{ij} given a vector of random effects $\mathbf{u}_i = (u_{i1}, \dots, u_{iq_r})^\top$ comes from the exponential family of distributions, $p(y_{ij} \mid \boldsymbol{\beta}, \phi, \mathbf{u}_i) = \exp[\{y_{ij}\vartheta_{ij} - a(\vartheta_{ij})\}/\phi + b(y_{ij}, \phi)]$ for known

functions $a(\cdot)$ and $b(\cdot)$, and a dispersion parameter $\phi > 0$ that may also require estimation. The mean of this distribution, denoted here as $\mu_{ij} = a'(\vartheta_{ij})$, is modeled as a linear combination of the fixed and random effect covariates through a link function $g(\cdot)$ i.e., $g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \mathbf{u}_i$, where $\boldsymbol{\beta}$ denotes the q_f-vector of fixed effect coefficients. Finally, for $i = 1, \dots, m$, the random effects $\mathbf{u}_i$ are assumed to follow a distribution $p(\mathbf{u}_i \mid \boldsymbol{\sigma})$ with parameter vector $\boldsymbol{\sigma}$, satisfying $\mathrm{E}(\mathbf{u}_i) = \mathbf{0}_i$ with the zero expectation required for parameter identifiability, and $\mathrm{Cov}(\mathbf{u}_i) = \boldsymbol{\Sigma}$ which is parametrized by $\boldsymbol{\sigma}$. As reviewed in Section 1, a common choice implemented in a variety of software for fitting GLMMs assumes that $p(\mathbf{u}_i \mid \boldsymbol{\sigma})$ follows a q_r-dimensional normal distribution, $p(\mathbf{u}_i \mid \boldsymbol{\sigma}) = \mathcal{N}_{q_r}(\mathbf{0}_{q_r}, \boldsymbol{\Sigma})$, where $\boldsymbol{\sigma} = \mathrm{chol}(\boldsymbol{\Sigma})$ in the case of an unstructured covariance matrix. Throughout this article, we define random effects misspecification as occurring when the true form of $p(\mathbf{u}_i \mid \boldsymbol{\sigma})$ does not follow such a normal distribution (although it has the same first two moments), but we assume normality when fitting the GLMM.

Assuming the responses y_{ij} are conditionally independent given $\mathbf{u}_i$, the marginal log-likelihood of the independent-cluster GLMM is given by $\log L(\boldsymbol{\theta}) = \sum_{i=1}^m \log \left\{ \int \prod_{j=1}^{n_i} p(y_{ij} \mid \boldsymbol{\beta}, \phi, \mathbf{u}_i) p(\mathbf{u}_i \mid \boldsymbol{\sigma}) \mathrm{d}\mathbf{u}_i \right\}$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi, \boldsymbol{\sigma}^\top)^\top$ denotes the vector of model parameters. The integral in the marginal log-likelihood function generally does not possess a tractable form, and for likelihood-based estimation, there are a number of ways of maximizing $\log L(\boldsymbol{\theta})$, most of which involve some approximation to the integral (see for instance Breslow and Clayton, 1993; Wolfinger, 1993; Ormerod and Wand, 2012; Brooks et al., 2017, among many others).

2.1. Random effect predictions

For point prediction of the random effects $\mathbf{u}_i$, one of the most popular approaches which we investigate below is the best predictor (BP, e.g. Booth and Hobert, 1998; Jiang, 2003), defined as the conditional expectation of the random effects given the observed data,

$$\mathbf{w}_i(\boldsymbol{\theta}) = \mathrm{E}[\mathbf{u}_i \mid \mathbf{y}_i] = \int \mathbf{u}_i p(\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}) \mathrm{d}\mathbf{u}_i = \frac{\int \mathbf{u}_i \prod_{j=1}^{n_i} p(y_{ij} \mid \boldsymbol{\beta}, \phi, \mathbf{u}_i) p(\mathbf{u}_i \mid \boldsymbol{\sigma}) \mathrm{d}\mathbf{u}_i}{\int \prod_{j=1}^{n_i} p(y_{ij} \mid \boldsymbol{\beta}, \phi, \mathbf{u}_i) p(\mathbf{u}_i \mid \boldsymbol{\sigma}) \mathrm{d}\mathbf{u}_i}. \tag{1}$$

To understand how random effects misspecification impacts the BP, we consider two possible cases as follows: First, let $p_0(\mathbf{u}_i \mid \boldsymbol{\sigma})$ denote the true density of the random effects, which in the article we set to be a finite mixture of normal distributions. That is, $p_0(\mathbf{u}_i \mid \boldsymbol{\sigma}) = \sum_{k=1}^c \pi_k p_k(\mathbf{u}_i | \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) = \sum_{k=1}^c \pi_k \mathcal{N}_d(\mathbf{u}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote the mean vector and covariance matrix, respectively, for the kth mixture component, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_c)^\top$ denotes the vector of mixture probabilities, c denotes the number of mixture components, and $\boldsymbol{\sigma} = (\boldsymbol{\pi}^\top, \boldsymbol{\mu}_1^\top, \dots, \boldsymbol{\mu}_c^\top, \boldsymbol{\sigma}_1^\top, \dots, \boldsymbol{\sigma}_c^\top)^\top$ here. Note that we require the constraints $\sum_{k=1}^c \pi_k \boldsymbol{\mu}_k = \mathbf{0}_d$ and $\sum_{k=1}^c \pi_k (\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top) = \boldsymbol{\Sigma}$ to ensure the true random effects distribution continues to have zero expectation and covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^\top, \phi_0, \boldsymbol{\sigma}_0^\top)^\top$ denote the true parameters in the GLMM under this true mixture distribution. It follows that the conditional distribution of the random effects given the observed data is given by

$$\begin{aligned} p_0(\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}_0) &= \frac{\prod_{j=1}^{n_i} p(y_{ij} \mid \boldsymbol{\beta}_0, \phi_0, \mathbf{u}_i) \sum_{k=1}^c \pi_{0k} p_k(\mathbf{u}_i \mid \boldsymbol{\mu}_{0k}, \boldsymbol{\sigma}_{0k})}{\int \prod_{j=1}^{n_i} p(y_{ij} \mid \boldsymbol{\beta}_0, \phi_0, \mathbf{u}_i) \sum_{k=1}^c \pi_{0k} p_k(\mathbf{u}_i \mid \boldsymbol{\mu}_{0k}, \boldsymbol{\sigma}_{0k}) \mathrm{d}\mathbf{u}_i} \\ &= \frac{\sum_{k=1}^c \pi_{0k} p_k(\mathbf{y}_i \mid \boldsymbol{\theta}_0) p_k(\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}_0)}{\sum_{k=1}^c \pi_{0k} p_k(\mathbf{y}_i \mid \boldsymbol{\theta}_0)}, \end{aligned} \tag{2}$$

where $p_k(\mathbf{y}_i \mid \boldsymbol{\theta}_0) = \int \prod_{j=1}^{n_i} p(y_{ij} \mid \boldsymbol{\beta}_0, \phi_0, \mathbf{u}_i) p_k(\mathbf{u}_i \mid \boldsymbol{\mu}_{0k}, \boldsymbol{\sigma}_{0k}) \mathrm{d}\mathbf{u}_i$ can be regarded as the marginal likelihood for the ith cluster assuming it belongs to the mixture component $k = 1, \dots, c$, and $p_k(\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}_0) = \prod_{j=1}^{n_i} p(y_{ij} \mid \boldsymbol{\beta}_0, \phi_0, \mathbf{u}_i) p_k(\mathbf{u}_i \mid \boldsymbol{\mu}_{0k}, \boldsymbol{\sigma}_{0k}) / p_k(\mathbf{y}_i \mid \boldsymbol{\theta}_0)$ is the corresponding conditional distribution given the observed data. Based on (2), we deduce that the true BP $\mathbf{w}_{0i}(\boldsymbol{\theta}_0) = \int \mathbf{u}_i p_0(\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}_0) \mathrm{d}\mathbf{u}_i$ can be written as a weighted sum of best predictors under each mixture component distribution, where the weights are proportional to $\pi_{0k} p_k(\mathbf{y}_i \mid \boldsymbol{\theta}_0)$.

Next, let $p_*(\mathbf{u}_i \mid \boldsymbol{\sigma})$ denote the misspecified normal random effects distribution, $p_*(\mathbf{u}_i \mid \boldsymbol{\sigma}) = \mathcal{N}(0, \boldsymbol{\Sigma})$, and let $\boldsymbol{\theta}_* = (\boldsymbol{\beta}_*^\top, \phi_*, \boldsymbol{\sigma}_*^\top)^\top$ denote the corresponding vector of pseudo-true model parameters for this misspecified GLMM; see Verbeke and Lesaffre (1997); Heagerty and Kurland (2001); Hui et al. (2022) for details on pseudo-true parameters in the context of (G)LMMs specifically. It follows that

$$p_*(\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}_*) = \frac{\prod_{j=1}^{n_i} p(y_{ij} \mid \boldsymbol{\beta}_*, \phi_*, \mathbf{u}_i) p_*(\mathbf{u}_i \mid \boldsymbol{\sigma}_*)}{\int \prod_{j=1}^{n_i} p(y_{ij} \mid \boldsymbol{\beta}_*, \phi_*, \mathbf{u}_i) p_*(\mathbf{u}_i \mid \boldsymbol{\sigma}_*) \mathrm{d}\mathbf{u}_i},$$

where the denominator is the marginal likelihood corresponding to the ith cluster for the misspecified GLMM. The resulting misspecified BP $\mathbf{w}_{*i}(\boldsymbol{\theta}_*) = \int \mathbf{u}_i p_*(\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}_*) \mathrm{d}\mathbf{u}_i$ follows from this.

Note that the true and misspecified best predictors are functions of the true and pseudo-true parameters, respectively. Upon replacing these with corresponding likelihood-based estimates based on fitting the true and misspecified GLMMs, which we denote here as $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}_*$ respectively, we obtain estimates often referred to as empirical best predictors (EBPs, Booth and Hobert, 1998; Jiang, 2003).

For the remainder of this article, we shall write the generic form of the EBP based on (1) as $\hat{\mathbf{w}}_i = \mathbf{w}_i(\hat{\boldsymbol{\theta}})$, the true EBP (i.e., the EBP for the GLMM with the true random effects distribution) as $\hat{\mathbf{w}}_{0i} = \mathbf{w}_{0i}(\hat{\boldsymbol{\theta}}_0)$, and the misspecified EBP (i.e., the EBP for the GLMM with the misspecified distribution) as $\hat{\mathbf{w}}_{*i} = \mathbf{w}_{*i}(\hat{\boldsymbol{\theta}}_*)$. Further, we write the generic form of the BP as $\mathbf{w}_i$, which is equal to $\mathbf{w}_{0i}(\boldsymbol{\theta}_0)$ under the true random effects distribution, and equal to $\mathbf{w}_{*i}(\boldsymbol{\theta}_*)$ under the misspecified random effects distribution.

2.2. Mean squared prediction errors

Measuring the uncertainty of the EBP is important for prediction inference, e.g., constructing prediction intervals of the random effects as demonstrated in Sections 4 and 5. In this article, we examine the common approach of using the mean squared error of prediction (MSEP; see Das et al., 2004; Cantoni et al., 2017; Flores-Agreda and Cantoni, 2019, for other examples of its usage for prediction inference in GLMMs). In particular, we introduce three different flavors of the MSEP for the random effects $\mathbf{u}_i$ as follows.

2.2.1. Unconditional mean squared error of prediction

The UMSEP of $\mathbf{u}_i$, which is evaluated over the marginal distribution of $\mathbf{y}$, is given by the diagonal entries of the matrix

$$\mathcal{U} = \mathrm{E}[(\hat{\mathbf{w}}_i - \mathbf{u}_i)(\hat{\mathbf{w}}_i - \mathbf{u}_i)^\top] = \mathbf{U}_1 + \mathbf{U}_2 + \mathbf{U}_3, \tag{3}$$

where $\mathbf{U}_1 = \mathrm{E}[(\mathbf{w}_i - \mathbf{u}_i)(\mathbf{w}_i - \mathbf{u}_i)^\top]$, $\mathbf{U}_2 = \mathrm{E}[(\hat{\mathbf{w}}_i - \mathbf{w}_i)(\hat{\mathbf{w}}_i - \mathbf{w}_i)^\top]$, and $\mathbf{U}_3 = \mathrm{E}[(\hat{\mathbf{w}}_i - \mathbf{w}_i)(\mathbf{w}_i - \mathbf{u}_i)^\top + (\mathbf{w}_i - \mathbf{u}_i)(\hat{\mathbf{w}}_i - \mathbf{w}_i)^\top]$. In (3), the term $\mathbf{U}_1$ represents the error arising from using the BP, and its diagonal entries are the MSEPs when we know the true/pseudo-true parameters. This term depends on the cluster size, and tends to zero when $n_i \to \infty$. The term $\mathbf{U}_2$ represents the error arising from estimating the parameters, and converges to zero as the number of clusters $m \to \infty$. Finally, the term $\mathbf{U}_3$ involves the interaction between the prediction and the estimation. We refer to Zheng and Cadigan (2021) and Ning et al. (2024) for more details on the asymptotic behavior of these terms in GLMMs.

In the setting of longitudinal studies when the number of clusters m is large relative to the cluster sizes n_i, it is reasonable to assume $\mathbf{U}_2$ tends to zero while $\mathbf{U}_1$ remains non-trivial. Importantly, the term $\mathbf{U}_1$ is most sensitive to misspecification of the random effects distribution, as $\mathbf{w}_i$ is dependent on the choice of $p(\mathbf{u}_i \mid \boldsymbol{\sigma})$; see Sections 3 and 4 for evidence of this. On the other hand, when m is small relative to n_i, then terms $\mathbf{U}_2$ and $\mathbf{U}_3$ can still be non-trivial in finite samples; this will be shown in the simulation studies in Section 4.

For the special case of LMMs, the UMSEP has been studied under the normal random effects assumption (e.g., Kackar and Harville, 1984; Prasad and Rao, 1990). In that case and when the random effects distribution is truly normally distributed, the term $\mathbf{U}_1$ has a closed-form solution. In the following section, we show that even when the true random effects distribution follows a mixture of normal distributions, we can derive a closed-form for $\mathbf{U}_1$.

2.2.2. Conditional mean squared error of prediction

The CeMSEP is defined here as the MSEP given the random effects for cluster i. This quantity is of interest when we are interested in prediction for a particular (subset of the) clusters (e.g., Rao and Molina, 2015; Ning et al., 2024, 2025). Critically, note that this definition of the CeMSEP is different from the definition from conditional MSEP of Booth and Hobert (1998) among others which conditions on $\mathbf{y}_i$; we choose to condition on the random effects as this can be used to construct prediction intervals with the correct conditional coverage (e.g., Zheng and Cadigan, 2023).

For cluster i, the CeMSEP of the random effects $\mathbf{u}_i$ is given by the diagonal entries of the following matrix

$$\mathcal{C} = \mathrm{E}[(\hat{\mathbf{w}}_i - \mathbf{u}_i)(\hat{\mathbf{w}}_i - \mathbf{u}_i)^\top \mid \mathbf{u}_i] = \mathbf{C}_1 + \mathbf{C}_2 + \mathbf{C}_3, \tag{4}$$

where $\mathbf{C}_1 = \mathrm{E}[(\mathbf{w}_i - \mathbf{u}_i)(\mathbf{w}_i - \mathbf{u}_i)^\top \mid \mathbf{u}_i]$, $\mathbf{C}_2 = \mathrm{E}[(\hat{\mathbf{w}}_i - \mathbf{w}_i)(\hat{\mathbf{w}}_i - \mathbf{w}_i)^\top \mid \mathbf{u}_i]$, and $\mathbf{C}_3 = \mathrm{E}[(\hat{\mathbf{w}}_i - \mathbf{w}_i)(\mathbf{w}_i - \mathbf{u}_i)^\top + (\mathbf{w}_i - \mathbf{u}_i)(\hat{\mathbf{w}}_i - \mathbf{w}_i)^\top \mid \mathbf{u}_i]$. Each of the three terms in (4) is analogous to the terms $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3$ in (3). Notably, the term $\mathbf{C}_1$ converges to zero as $n_i \to \infty$, while the term $\mathbf{C}_2$ converges to zero as $m \to \infty$. Similar to the term $\mathbf{U}_1$ in the UMSEP, the term $\mathbf{C}_1$ in the CeMSEP is most affected by misspecification, the impact being more evident when the cluster size is smaller. For LMMs, we can again derive a closed-form expression for $\mathbf{C}_1$ under both the misspecified normal and the mixture random effects distributions, thus enabling some analytical insight as we shall see below.

2.2.3. Bootstrap estimates of the MSEPs

For a general situation where the response distribution is non-normal, there is usually no closed-form solution for the expressions in (3) and (4). Indeed, even in LMMs the term $\mathbf{U}_2$ (and similarly $\mathbf{C}_2$) is not available in closed form and usually has to be approximated using a Taylor expansion (e.g., Kackar and Harville, 1984). With this in mind, later on in this article we also consider a more general, practical approach to estimate the UMSEP and CeMSEP via the parametric bootstrap (Chatterjee et al., 2008; Flores-Agreda and Cantoni, 2019). A full parametric bootstrap algorithm to estimate the UMSEP and CeMSEP is provided in the Supplementary Material. But briefly, it involves either resampling the random effects from the estimated random effects distribution (for UMSEP) or conditioning on the set of predicted random effects (for CeMSEP), and then sampling responses conditional on these random effects.

3. Random effects misspecification in LMMs

In this section, we examine the special case of independent-cluster linear mixed models, given that this is one of the primary cases for which we can arrive at closed-form expressions for the terms in, and hence analytical insights related to, (3) and (4). Recall from the preceding section that the terms $\mathbf{U}_1$ and $\mathbf{C}_1$ are the most sensitive, in their respective MSEPs, to misspecification of the random effects distribution.

Consider the independent-cluster LMM $y_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \mathbf{u}_i + \epsilon_{ij}$, where $\epsilon_{ij} \sim \mathcal{N}(0, \tau^2)$ is assumed to be correctly specified. Next, suppose the true distribution of the random effects follows a finite mixture of normal distributions such that, analogous to below (1),

we can write $p_0(\mathbf{u}_i|\boldsymbol{\sigma}_0) = \sum_{k=1}^{c} \pi_{0k}\mathcal{N}(\mathbf{u}_i|\mathbf{L}_0\boldsymbol{\mu}_{0k}, \mathbf{L}_0\boldsymbol{\Sigma}_{0k}\mathbf{L}_0^\top)$ with constraints $\sum_{k=1}^{c}\pi_{0k}\boldsymbol{\mu}_{0k} = \mathbf{0}$ and $\sum_{k=1}^{c}\pi_{0k}(\boldsymbol{\mu}_{0k}\boldsymbol{\mu}_{0k}^\top + \boldsymbol{\Sigma}_{0k}) = \mathbf{I}$, and $\mathbf{L}_0$ is the Cholesky factor of the covariance matrix $\boldsymbol{\Sigma}_0$ of $\mathbf{u}_i$. The corresponding misspecified random effects distribution is denoted by $p_*(\mathbf{u}_i|\boldsymbol{\sigma}_*) = \mathcal{N}(0, \boldsymbol{\Sigma}_*)$, and $\mathbf{L}_* = \text{chol}(\boldsymbol{\Sigma}_*)$. We begin by establishing the following result on the equality between the pseudo-true and true parameters for LMMs. The proof is given in Supplementary Material S1.

Theorem 1. *If the true random effects distribution follows a (standardized) finite mixture of normal distributions in an independent-cluster LMM, then the pseudo-true model parameter vector $\boldsymbol{\theta}_* = (\boldsymbol{\beta}_*^\top, \mathbf{L}_*^\top, \tau_*^2)^\top$ from a LMM assuming a misspecified normal random effects distribution is equal to the true parameter vector $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^\top, \mathbf{L}_0^\top, \tau_0^2)^\top$.*

We will use the above result to establish formulas for the BPs and their corresponding MSEPs. In particular, we will see that even though the pseudo-true and true parameters are equal in the LMM setting, the corresponding BPs and the MSEPs can be quite different, meaning point prediction and prediction inference can be greatly impacted by random effects misspecification even when estimation is not. In what follows, we use $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \mathbf{L}^\top, \tau^2)^\top$ to denote both $\boldsymbol{\theta}_*$ and $\boldsymbol{\theta}_0$, in light of Theorem 1.

3.1. Best predictors

Applying the results in Section 2.1 to the special case of $\mathbf{w}_{*i}$ under the misspecified LMM (note the BP in this setting is equivalent to the best unbiased linear predictor or BLUP), we can show that

$$\begin{aligned}\mathbf{w}_{*i}(\boldsymbol{\theta}) = \mathrm{E}_*(\mathbf{u}_i \mid \mathbf{y}_i) &= \{(\mathbf{L}\mathbf{L}^\top)^{-1} + \tau^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\}^{-1}\tau^{-2}\mathbf{Z}_i^\top(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \\ &= \{(\mathbf{L}\mathbf{L}^\top)^{-1} + \tau^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\}^{-1}\tau^{-2}\mathbf{Z}_i^\top(\mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i),\end{aligned} \tag{5}$$

where $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})^\top$, and $\mathbf{Z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{in_i})^\top$. On the other hand, the best predictor $\mathbf{w}_{0i}$ under the true LMM is given by

$$\mathbf{w}_{0i}(\boldsymbol{\theta}) = \frac{\sum_{k=1}^{c}\pi_{0k}p_k(\mathbf{y}_i)\mathbf{m}_k}{\sum_{k=1}^{c}\pi_{0k}p_k(\mathbf{y}_i)}, \tag{6}$$

where $\mathbf{m}_k = \{(\mathbf{L}\boldsymbol{\Sigma}_{0k}\mathbf{L}^\top)^{-1} + \tau^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\}^{-1}\{(\mathbf{L}\boldsymbol{\Sigma}_{0k}\mathbf{L}^\top)^{-1}\mathbf{L}\boldsymbol{\mu}_{0k} + \tau^{-2}\mathbf{Z}_i^\top(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\}$, and $p_k(\mathbf{y}_i) = \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{L}\boldsymbol{\mu}_{0k}, \tau^2\mathbf{I} + \mathbf{Z}_i(\mathbf{L}\boldsymbol{\Sigma}_{0k}\mathbf{L}^\top)\mathbf{Z}_i^\top)$. The EBPs for both are found by substituting the estimated parameters into these formulas.

Comparing the two forms of the BP, we observe the shrinkage effect i.e., the amount each predictor is shrunk toward the mean of the distribution, is impacted by random effects misspecification. For the true LMM, each of the terms $\mathbf{m}_k$ in (6) has two components, with one component being a scaled version of $(\mathbf{L}\boldsymbol{\Sigma}_{0k}\mathbf{L}^\top)^{-1}\mathbf{L}\boldsymbol{\mu}_{0k}$ which shrinks towards the *mean of the individual mixture component* in the true random effects distribution, i.e., $\mathbf{L}\boldsymbol{\mu}_{0k}$, and the other component is a scaled version of $\tau^{-2}\mathbf{Z}_i^\top(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$ which behaves more similarly to (5). It follows that this BP, being a weighted sum of the terms $\mathbf{m}_k$, $k = 1, \ldots, c$, is not necessarily shrunk toward zero. On the other hand, the BP for the misspecified LMM in (5) is always shrunk towards zero. This difference in shrinkage effect is less prominent if the term $\tau^{-2}\mathbf{Z}_i^\top(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$ is relatively large, as can occur when the cluster size is large.

3.2. Mean squared prediction errors

Under the misspecified LMM, the term $\mathbf{U}_1$ in the UMSEP (3) is straightforwardly shown to be $\{(\mathbf{L}\mathbf{L}^\top)^{-1} + \tau^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\}^{-1}$. This contrasts with the true LMM, for which $\mathbf{U}_1$ is given by $\mathrm{E}[(\mathbf{w}_i - \mathbf{u}_i)(\mathbf{w}_i - \mathbf{u}_i)^\top] = \mathrm{E}[\mathrm{E}[(\mathbf{w}_i - \mathbf{u}_i)(\mathbf{w}_i - \mathbf{u}_i)^\top \mid \mathbf{y}_i]] = \mathrm{E}[\mathbf{v}_{0i}]$, where $\mathbf{v}_{0i}(\boldsymbol{\theta}) = \mathrm{E}[(\mathbf{w}_i - \mathbf{u}_i)(\mathbf{w}_i - \mathbf{u}_i)^\top \mid \mathbf{y}_i] = \{\sum_{k=1}^{c}\pi_{0k}p_k(\mathbf{y}_i)\}^{-1}\sum_{k=1}^{c}\pi_{0k}p_k(\mathbf{y}_i)(\mathbf{v}_k + \mathbf{m}_k\mathbf{m}_k^\top) - \mathbf{w}_{0i}\mathbf{w}_{0i}^\top$, with $\mathbf{v}_k = \{(\mathbf{L}\boldsymbol{\Sigma}_{0k}\mathbf{L}^\top)^{-1} + \tau^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\}^{-1}$. As demonstrated in the numerical study in Supplementary Material S2, the UMSEPs, which are dominated by the term $\mathbf{U}_1$ for our simulation designs, are consistently larger under the misspecified LMM than under the true LMM, although the differences diminish as the cluster size increases.

Turning to the CeMSEP, from (5) the difference $\mathbf{w}_{*i} - \mathbf{u}_i$ under the misspecified LMM is given by

$$\begin{aligned}\mathbf{w}_{*i} - \mathbf{u}_i &= [\{(\mathbf{L}\mathbf{L}^\top)^{-1} + \tau^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\}^{-1}\tau^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i - \mathbf{I}]\mathbf{u}_i + \{(\mathbf{L}\mathbf{L}^\top)^{-1} + \tau^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\}^{-1}\tau^{-2}\mathbf{Z}_i^\top\boldsymbol{\epsilon}_i \\ &\triangleq \boldsymbol{\Gamma}_1\mathbf{u}_i + \boldsymbol{\Gamma}_2\boldsymbol{\epsilon}_i,\end{aligned} \tag{7}$$

from which it follows that $\mathbf{C}_1 = \boldsymbol{\Gamma}_1\mathbf{u}_i\mathbf{u}_i^\top\boldsymbol{\Gamma}_1^\top + \tau^2\boldsymbol{\Gamma}_2\boldsymbol{\Gamma}_2^\top$. We thus see that the CeMSEP, which is dominated by $\mathbf{C}_1$ under the setting when m is large relative to n_i, is a quadratic function of $\mathbf{u}_i$ centered about zero for the misspecified LMM. This can be problematic if the true random effects distribution is skewed or multimodal, as under such cases there may be potentially a non-negligible number of random effects far from zero; see Section 4 for an example of this in the case of GLMMs. By contrast, using (6) we can show the same difference under the true LMM is given by

$$\mathbf{w}_{0i} - \mathbf{u}_i = \sum_{k=1}^{c}\zeta_k\{(\tau^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i)\mathbf{u}_i + (\mathbf{L}\boldsymbol{\Sigma}_{0k}\mathbf{L}^\top)^{-1}\mathbf{L}\boldsymbol{\mu}_{0k} + \tau^{-2}\mathbf{Z}_i^\top\boldsymbol{\epsilon}_i\} - \mathbf{u}_i, \tag{8}$$

where $\zeta_k = [\pi_{0k}p_k(\mathbf{y}_i)\{(\mathbf{L}\boldsymbol{\Sigma}_{0k}\mathbf{L}^\top)^{-1} + \tau^{-2}\mathbf{Z}_i^\top\mathbf{Z}_i\}^{-1}]\{\sum_{k=1}^{c}\pi_{0k}p_k(\mathbf{y}_i)\}^{-1}$. It follows that the form of $\mathbf{C}_1$ and hence the CeMSEP resembles a weighted quadratic function of $\mathbf{u}_i$: when the random effect is closer to the mean of the mixture component k, the weight

Table 1
Simulated UMSEPs ($\text{UMSEP}_{\text{sim}}$) for the misspecified GLMM/true GLMM, in the case of binary responses and for two true mixture of normal random effects distributions.

	$n_i = 20$	$n_i = 40$	$n_i = 60$	$n_i = 80$
	Random effects distribution I			
$m = 100$	0.3620/0.4289	0.2237/0.1507	0.1591/0.0996	0.1251/0.0776
$m = 200$	0.3739/0.3641	0.2259/0.1160	0.1609/0.0848	0.1234/0.0704
	Random effects distribution II			
$m = 100$	0.5298/0.4595	0.2955/0.2470	0.2176/0.1876	0.1643/0.1429
$m = 200$	0.5057/0.4391	0.2680/0.2295	0.1958/0.1717	0.1456/0.1288

corresponding to that component will dominate and the overall CeMSEP will be approximately equal to the quadratic function centered about $\mathbf{L}\boldsymbol{\mu}_{0k}$. This guarantees the CeMSEP will be relatively small around the mean of each component, in contrast to the single zero-centered quadratic form of the CeMSEP under the misspecified LMM; see Supplementary Material S2 for further illustration of this behavior.

4. Simulation studies

We performed a numerical study to compare point prediction and prediction inference obtained under a misspecified normal random effects distribution, with those obtained under a correctly specified mixture of normal distributions, for the cases of Bernoulli and Poisson GLMMs. We present results under two true mixture of normals distributions which were purposefully chosen to represent skewness and multimodality in a true random effects distribution. Further simulations under a symmetric mixture distribution but with a heavier tail, and under a normal distribution are presented in Supplementary Material S2. Also, simulations involving LMMs were considered and are available in Supplementary Material S2, in which the results largely confirm some of the analytical conclusions obtained in the preceding section.

4.1. Bernoulli response

We consider a binary logistic GLMM, i.e., the conditional distribution of the response follows a Bernoulli distribution with $g(\cdot)$ set to the canonical logit function, with $q_f = 2$ fixed effects covariates and a single random intercept, i.e., $q_r = 1$. We generate the elements x_{ij2} from a uniform distribution $[-5, 5]$, and set the true fixed effects coefficients to $\boldsymbol{\beta} = (0, 1)^\top$. As the true distribution for the random intercept, we consider two choices: (I) $p_0(u_i \mid \boldsymbol{\sigma}_0) = 0.9\mathcal{N}(-0.28, 0.28^2) + 0.1\mathcal{N}(2.56, 1.42^2)$ which exhibits clear right skew, and (II) $p_0(u_i \mid \boldsymbol{\sigma}_0) = 0.5\mathcal{N}(-1.77, 0.59^2) + 0.5\mathcal{N}(1.77, 1.18^2)$ which exhibits clear multimodality. Note the true parameters were chosen to satisfy the mean zero constraints and to set the variance of the random intercept equal to one and four for the two distributions, respectively.

We begin by investigating the impact of cluster size n_i and number of clusters m on UMSEP, by simulating datasets with $n_i = 20, 40, 60, 80$, and $m = 100, 200$. For each combination of m and n_i, we generate 100 datasets by sampling the random effects $\{u_i, i = 1, \ldots, m\}$ from one of the two true mixture distributions, noting each dataset has a different set of random effects, and then simulating the responses y_{ij} based on the set up above. For each dataset, we then fit a binary logistic GLMM assuming either a misspecified normal distribution, or a correct two-component mixture of normal distributions, for the random effects. The fitted parameters are then substituted into equation (1) to obtain the EBPs of the random intercept for the misspecified GLMM, i.e., $\hat{u}_{*i}$, and the true GLMM i.e., $\hat{u}_{0i}$. We also simulate the UMSEP by averaging the squared differences between the EBPs and the true random effect for each dataset, and then averaging across the 100 datasets, i.e., $\text{UMSEP}_{\text{sim}} = 100^{-1} \sum_{s=1}^{100} \{m^{-1} \sum_{i=1}^{m} (\hat{w}_i^{(s)} - u_i^{(s)})^2\}$ where $u_i^{(s)}$ denotes the simulated random intercept for the sth dataset, and $\hat{w}_i^{(s)}$ is either $\hat{w}_{*i}$ or $\hat{w}_{0i}$ for the sth dataset.

Results for the $\text{UMSEP}_{\text{sim}}$ can be found in Table 1, while the boxplots for the 100 UMSEPs of each dataset in each setting can be found in the Supplementary Material. As expected, $\text{UMSEP}_{\text{sim}}$ under the misspecified GLMM is generally larger compared with the true GLMM. The only exception was for random effects distribution I when $n_i = 20$, and indeed we note that the difference in $\text{UMSEP}_{\text{sim}}$ between the two models is small when $m = 200$ at the smallest cluster size. We conjecture that this might be due to the fact that for a binary GLMM, estimation of a mixture of normal random effects distributions is relatively challenging and the uncertainty in the estimation contributes a non-negligible amount to the UMSEP.

Next, we compare the CeMSEP under the misspecified and true random effects distributions by fixing the cluster size to $n_i = 40$ and the number of clusters to $m = 400$; we choose $n_i \ll m$, so $\mathbf{C}_1$ dominates the CeMSEP. The remainder of the simulation setup is the same as above, with one crucial modification being that instead of simulating a new set of random effects for each dataset, we only simulate one set of random intercepts, and then generate 100 datasets conditioned on this set. We then simulate the CeMSEP by averaging the squared differences between the EBP and the true random effects across the 100 datasets, conditional on the single set of true random effects i.e., $\text{CeMSEP}_{\text{sim}} = 100^{-1} \sum_{s=1}^{100} (\hat{w}_i^{(s)} - u_i)^2$. Under the misspecified GLMM, the simulated CeMSEPs are lowest around zero i.e., the assumed mean of the random effects distribution, and increase in value in the tail of the distribution (Fig. 1 top row). By contrast, for the GLMM with the two true random effects distributions, we observe two local minima for $\text{CeMSEP}_{\text{sim}}$, each one at the mean of a component of the corresponding true mixture distribution. This is consistent with the analytical results for LMMs in Section 3.2, and shows that as we get closer to the mean of each of the mixture components, the value of $\text{CeMSEP}_{\text{sim}}$

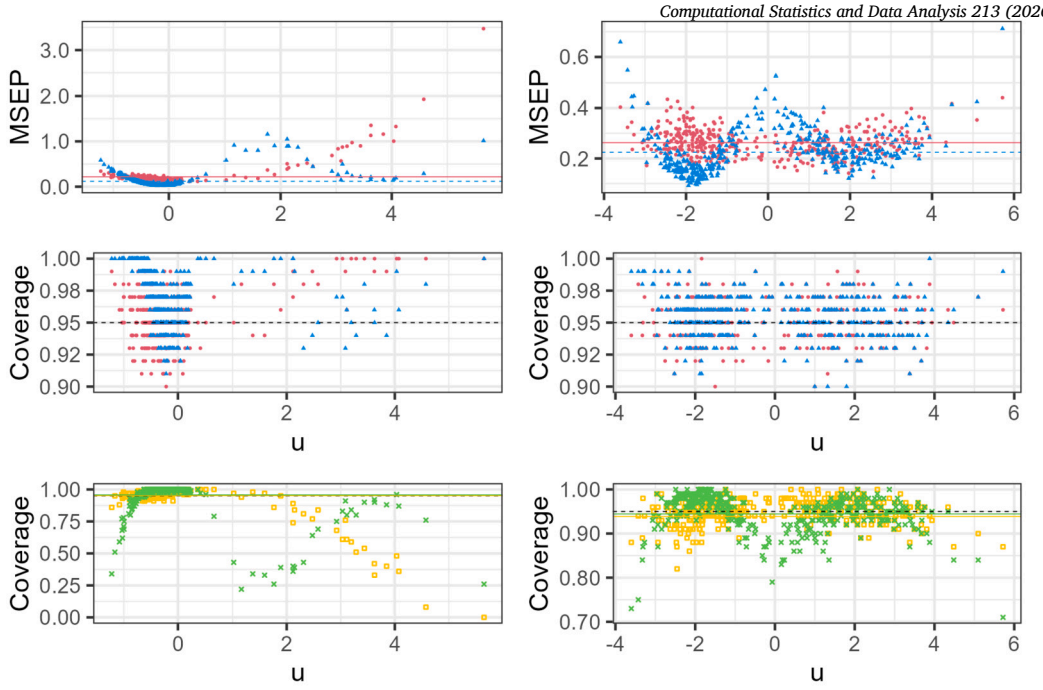

Fig. 1. Simulated CeMSEPs ($\text{CeMSEP}_{\text{sim}}$; first row) and empirical coverage probability of 95% prediction intervals constructed by CeMSEPs (second row) and UMSEPs (third row) for the misspecified versus true GLMMs in the case of binary responses, and for two true mixture of normal random effects distributions. Left column is distribution I; right column is distribution II. In the top two rows, the red dots and blue triangles correspond to results obtained using the misspecified and true GLMMs respectively, while in the third row the gold squares and green crosses correspond to intervals constructed under the misspecified and true GLMMs respectively. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

produced by the misspecified GLMM is larger than under the true GLMM. In Supplementary Material S2, we display results where we decompose the simulated CeMSEP into the simulated expected squared bias and the simulated variance term. Results of these show the variance terms are larger under the misspecified GLMM especially in regions close to the mean of each of the mixture components. On the other hand, the bias under the misspecified model is close to a linear function whose absolute value increases as the random effect value moves away from zero, while the bias under the correctly specified mixture distribution appears to be a weighted linear function.

Finally, we examine the performance of prediction intervals for the random intercept constructed using MSEP. Following Ha et al. (2011); Cantoni et al. (2017); Korhonen et al. (2023) among others, if we assume the difference $\hat{w}_i - u_i$ follows a normal distribution then a $100(1-\alpha)\%$ prediction interval for u_i is given by $\hat{w}_i \pm \Phi^{-1}(1-\alpha/2) \times (\widehat{\text{MSEP}})^{1/2}$, where $\Phi^{-1}(\cdot)$ denotes the inverse cumulative distribution function of the normal distribution and $\widehat{\text{MSEP}}$ is either the simulated UMSEP or simulated CeMSEP. Note in practice we do not observe the true random effect, and so we need to estimate the MSEP e.g., using the parametric bootstrap approach discussed at the end of Section 2.2. The coverage probability conditional on specific values of u_i, generally 95% prediction intervals constructed using simulated CeMSEPs (Fig. 1 middle panel) achieve reasonable coverage regardless of the true value of u_i and for both true mixture of normal random effects distributions. This means the coverage probability of intervals constructed from CeMSEPs is not impacted by misspecification. By contrast, the conditional coverage probability for prediction intervals constructed using the simulated UMSEPs (Fig. 1 bottom panel) is less than the nominal coverage for certain regions of the random effects distribution e.g., at the right tail of distribution I. This is a direct consequence of the UMSEP being lower than the CeMSEP at a certain value of the random effect. While such results regarding the prediction intervals constructed through UMSEPs and CeMSEPs may not be overly surprising (given UMSEP is in principle designed for averaging/marginalizing across all the random effects values), they nevertheless highlight the importance of consequences of random effects specification, and the importance of using CeMSEPs when examining coverage conditional on a specific value of u_i.

4.2. Poisson response

In the second setting, we consider a Poisson GLMM with the canonical log link. We again used $q_f = 2$ fixed effects covariates with the first set to an intercept term and x_{ij2} drawn independently from the standard uniform distribution, $\boldsymbol{\beta} = (0,1)^\top$, along with a single random intercept. As the true random effects distributions, we investigate the same two mixtures of normal distributions as in the binary GLMM case, while the remainder of the simulation setup and performance assessment was also analogous to Section 4.1.

Given the availability of more information in the response compared to the binary response in Section 4.1, then we consider smaller values of $n_i = 5, 10, 20, 40$ and $m = 50, 100$ and investigate their impact on UMSEP. Results for the $\text{UMSEP}_{\text{sim}}$ can be found

Table 2
Simulated UMSEPs ($UMSEP_{sim}$) for the misspecified GLMM/true GLMM, in the case of Poisson responses and for two true mixture of normal random effects distributions.

	$n_i = 5$	$n_i = 10$	$n_i = 20$	$n_i = 40$
	Random effects distribution I			
$m = 50$	0.1458/0.0836	0.0838/0.0607	0.0551/0.0458	0.0400/0.0371
$m = 100$	0.1434/0.0682	0.0782/0.0496	0.0436/0.0330	0.0264/0.0235
	Random effects distribution II			
$m = 50$	0.4062/0.2665	0.2790/0.1900	0.1725/0.1244	0.1426/0.1271
$m = 100$	0.3543/0.2100	0.2284/0.1364	0.1569/0.1122	0.0997/0.0841

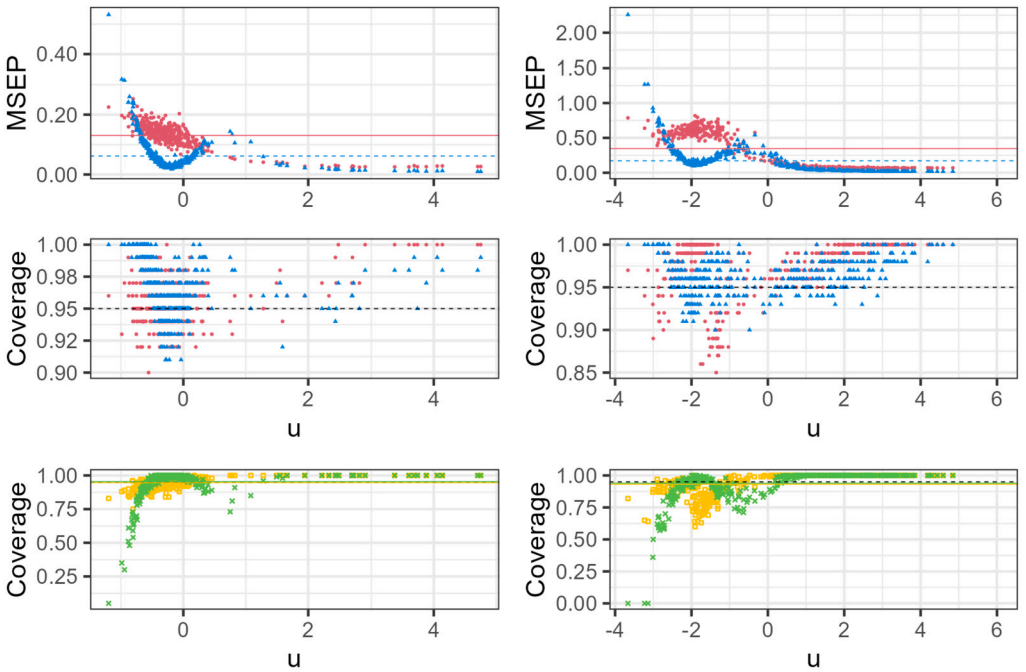

Fig. 2. Simulated CeMSEPs ($CeMSEP_{sim}$; first row) and empirical coverage probability of 95% prediction intervals constructed by CeMSEPs (second row) and UMSEPs (third row) for the misspecified versus true GLMMs in the case of Poisson responses, and for two true mixture of normal random effects distributions. Left column is distribution I; right column is distribution II. In the top two rows, the red dots and blue triangles correspond to results obtained using the misspecified and true GLMMs respectively, while in the third row the gold squares and green crosses correspond to intervals constructed under the misspecified and true GLMMs respectively.

in Table 2, while the boxplots for the 100 UMSEPs of each dataset in each setting can be found in the Supplementary Material. Table 2 shows that larger values of $UMSEP_{sim}$ generally arise under the misspecified GLMM compared with the true GLMM, with the differences being more evident when the cluster size is small.

Next, turning to comparisons of CeMSEP, we fix the cluster size at $n_i = 5$ and the number of clusters at $m = 400$. For both true random effects distributions I and II, the true GLMM produces considerably smaller values of the simulated CeMSEP around the mean of the first component in the mixture distribution (Fig. 2 top row). Around the mean of the second mixture component, the true GLMM also returns smaller simulated CeMSEPs compared to its misspecified counterpart, although the differences between the two models are substantially diminished: the reflects the impact of the link function on the behavior of $CeMSEP_{sim}$ i.e., with the log link the CeMSEPs for larger random effect values tend to zero. In Supplementary Material S2, we show results decomposing the simulated CeMSEP into a bias and variance term. Similar to Section 4.1, these results demonstrate the variances are larger under the misspecified GLMM, especially around the mean of the first mixture component, while the absolute bias can be larger for the correctly specified GLMM in some regions of the random effects distribution.

Finally, the results for conditional coverage probability using simulated CeMSEPs and UMSEPs are largely similar to those for the binary GLMM case (Fig. 2 bottom two rows). In particular, the 95% prediction intervals constructed using simulated CeMSEPs achieve reasonable conditional coverage probabilities even under misspecification, while analogous intervals constructed using simulated UMSEPs can perform quite poorly in certain regions of the true random effects distribution.

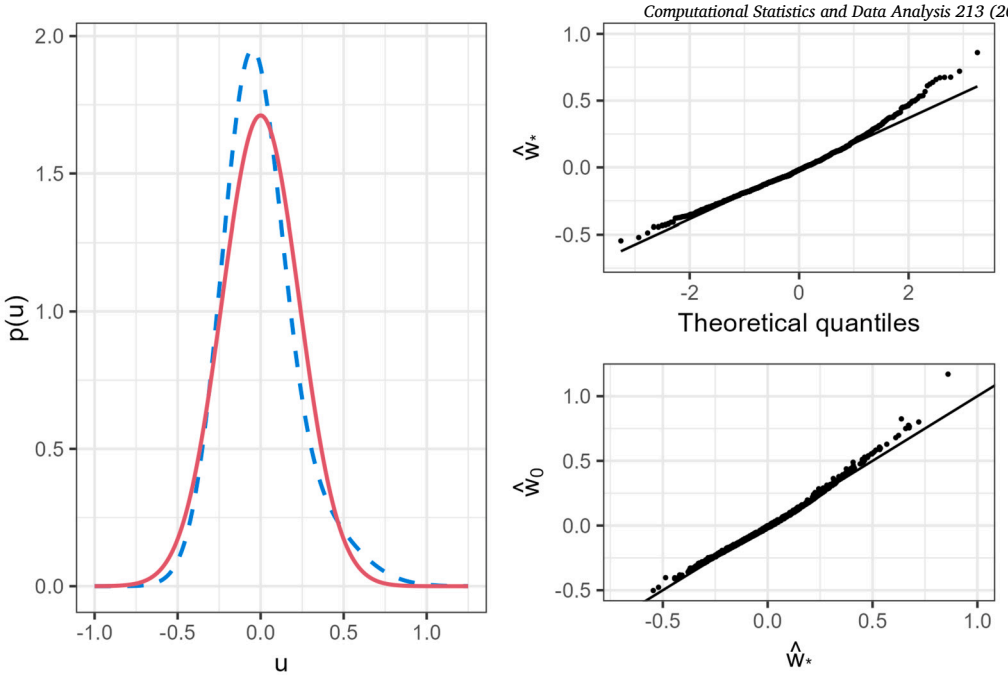

Fig. 3. Comparison of the random effects under the normal and mixture of normals LMMs fitted for the wages data. Left panel: the fitted random effects distribution (red solid line for normal LMM and blue dashed line for mixture of normals LMM). Right panel: the EBPs under the normal distribution against theoretical quantiles of the normal distribution (top), and against the EBPs under the mixture distribution (bottom).

5. Case study: longitudinal survey of hourly wages

In this section, we illustrate the analytical and empirical findings above using data from the National Longitudinal Survey of Youth (Singer and Willett, 2003). Another case study using multivariate abundance data (Gee et al., 1985) is provided in the Supplementary Material.

In this application, the response is the log of hourly wages for $m = 888$ individuals, each measured at n_i occasions with cluster sizes ranging from 1 to 13. Prediction of income has been considered in the context of LMMs by a number of researchers (e.g., Hui and Bondell, 2021; Erciulescu and Opsomer, 2022), and in this case the prediction of random effects serves as a primary objective as we aim to identify high earning individuals and quantify their difference in wages from the bulk of the individuals in the survey.

We fitted independent-cluster LMMs where time spent in the workforce, race, and education along with an intercept were included as fixed effect covariates (so $q_f = 4$), and a single $q_r = 1$ random intercept is included to account for between-individual heterogeneity. We begin by fitting a LMM assuming the random effects are normally distributed, as is the standard assumption in the literature. An examination of the resulting EBPs using both a normal quantile-quantile plot and a Shapiro-Wilk test at the 5% significance level offered statistical evidence of a deviation from normality, indicating that the assumed normality does not hold. In particular, there is evidence the true random distribution may be right-skewed (see top right panel of Fig. 3). Note we use both the quantile-quantile plot and the Shapiro-Wilk test here in an exploratory manner, and both should be interpreted with caution given they are constructed using the predicted random effects.

With the above in mind, we then proceed to fit an LMM using a two-component mixture of normal distributions for the random intercept, as formulated above Theorem 1. Fig. 3 presents the estimated random effects distributions along with the EBPs for the two LMMs fitted, from which we see that the fitted mixture of normal distributions is indeed slightly right skewed (left panel). The EBPs from this fit are comparably less shrunk toward zero than those produced by the normal random effects LMM (bottom right panel); this result is consistent with the theoretical findings in Section 3.1.

Next, we examine the MSEP for the two fitted LMMs. Given the number of clusters m is relatively large compared to the cluster sizes in this longitudinal survey, then as discussed in Section 2.2 the $\mathbf{C}_1$ term in the CeMSEP is expected to dominate in equation (4). As a result, based on equations (7) and (8), which compute the difference between the best predictor and the true random effects value, we can approximate the CeMSEP under the normal and mixture of normals LMMs respectively, where the estimated parameters are substituted into the equations. For illustrative purposes, we subset to only consider 103 individuals with cluster size $n_i = 7$ (the mean cluster size). Similar to the results of the simulation study for LMMs (Supplementary Material S2) and for GLMMs broadly (Section 4), the CeMSEPs under the mixture of normals random effects distribution is lower than the CeMSEPs under the normal random effects distribution for many predicted random effect values (Fig. 4 solid and dotted curves). This is especially the case around the mean of the main component of the mixture, and in the heavy right tail.

Alternatively, we can use the bootstrap to estimate the MSEPs, as discussed towards the end of Section 2.2. The resulting bootstrap estimates of the CeMSEPs are shown as points in Fig. 4. Note that in comparison to using the analytical expressions for $\mathbf{C}_1$, the bootstrap

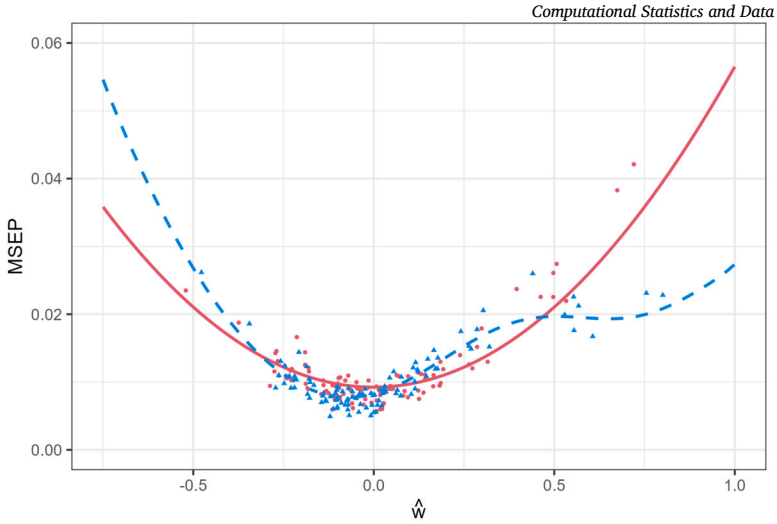

Fig. 4. Comparison of CeMSEP under LMMs fitted for the wages data, assuming either the normal or mixture of normals random effects distribution. The curves correspond to the analytical $\mathbf{C}_1$ term i.e., equations (7) (red solid line) and (8) (blue dashed line), while the red dots and blue triangles are parametric bootstrap estimates of the CeMSEP under the normal and mixture distributions.

estimates account for estimating the full MSEP e.g., incorporating terms $\mathbf{C}_2$ and $\mathbf{C}_3$ as well. On the other hand, because bootstrapping is performed given the EBPs of the random effects, then the bootstrap estimates must be interpreted carefully since they are naturally biased e.g., predictors under the normal assumption at the tail are shrunk toward zero more than it should be (see Carpenter et al., 2003, among others for solutions to this problem in practice). Overall, results from this parametric bootstrap approach are consistent with the analytical forms of $\mathbf{C}_1$ (comparing the curves with the points in Fig. 4). Moreover, the bootstrap estimate of the UMSEP under the normal distribution is 0.0112, which is slightly larger than the bootstrap estimate of the UMSEP, 0.0108 under the mixture distribution.

Finally, Supplementary Material S2 also presents results of 95% prediction intervals constructed using bootstrap estimates of the CeMSEP, from which we see that at the right tail of the random effects distribution, the intervals under the normal LMM are much wider. This may suggest that misspecifying the random effects distribution, assuming the mixture of normals LMM is a more accurate representation of the true data generation process, could lead to conclusions with higher prediction uncertainty.

6. Discussion

In this article, we studied the effect of random effects misspecification on point prediction and prediction inference of random effects in GLMMs, comparing a misspecified normal distribution to a correctly specified mixture of normal distributions. The analytical results for the case of LMMs, along with empirical findings for the case of GLMMs, reveal a number of key findings. First, the degree of shrinkage for random effects is affected: while the BPs and EBPs under a misspecified normal distribution are shrunk toward zero, the corresponding predictors under the true mixture distribution are shrunk differentially and not necessarily toward zero. Practically, this may lead to substantial biased point predictions, particularly in the tail of the random effects distribution, under misspecification. Second, MSEPs are negatively impacted under random effects misspecification, with UMSEPs consistently larger and the effects being most evident in settings with small cluster sizes. The case of CeMSEPs is more complicated, although they still tend to remain larger under the misspecified normal distribution, especially those in the regions of the mean of a component of the underlying mixture distribution. Finally, prediction intervals based on MSEPs are wider under the misspecified normal distribution. While prediction intervals using CeMSEPs achieve nominal coverage across clusters, conditional coverage probabilities of intervals based on UMSEPs vary across clusters and can suffer from severe undercoverage even under correct specification of the random effects distribution.

In this article, we focused on the case where the true random effects distribution is a mixture of normal distributions. Moreover, in our simulation studies and case studies, we assumed a two-component mixture for the random effects distribution while fitting the data. It is important to highlight that one can choose a different value for the number of components c in the mixture. For example, this can be done by examining the EBPs under the normal distribution, such as what we did in Section 5, and then potentially some form or order selection method (e.g., McLachlan and Rathnayake, 2014; Hui et al., 2015; Manole and Khalili, 2021). More broadly, one can also choose a random effects distribution that is not a finite mixture of normal distributions e.g., based on the shape of the empirical distribution of the EBPs under the normal distribution, choosing a form that can accommodate some degree of skewness and/or kurtosis (Ho and Lin, 2010; Meza et al., 2012; Pinheiro et al., 2001). However, we do recommend that although any distribution can be chosen for the random effects, if the aim is to obtain the EBPs and their MSEPs, it is generally preferred to use a random effects distribution with (at least) finite first and second moments.

While this article has focused on just the consequences of misspecification on prediction of random effects alone, predicting a function $t(\mathbf{u}_i)$ of the random effect $\mathbf{u}_i$ is a natural next quantity of interest e.g., in fisheries ecology one might need to predict the expected number of (non-zero) bycatch, which is a complicated non-linear function of the fixed and random effects (Cantoni et al.,

2017). Here, we can also use the (empirical) best predictor for prediction of this function (Vidoni, 2006; Skrondal and Rabe-Hesketh, 2009) i.e., $\mathrm{E}[t(\mathbf{u}_i) \mid \mathbf{y}_i] = \int t(\mathbf{u}_i)p(\mathbf{u}_i \mid \mathbf{y}_i)\mathrm{d}\mathbf{u}_i$, and we conjecture that the overall conclusions of this article will likely carry over to point prediction and prediction inference of functions of random effects. That is, misspecification is likely to have negative impacts on aspects such as shrinkage and conditional coverage probabilities.

Also, this article has focused on the (empirical) best predictors, and it would be interesting to investigate how other predictors, such as one based on quantile regression (e.g., Chambers and Tzavidis, 2006), or one based on the mode of the conditional distribution of the random effects (Bates et al., 2015; Brooks et al., 2017; Ning et al., 2025, commonly used when the Laplace approximation or penalized quasi-likelihood is applied to mixed models), perform under random effects misspecification. Finally, we focused on misspecification of the random effects distribution while assuming other elements of the GLMM, such as the conditional distribution of the responses, the link function, and so on, are correctly specified. It would be of interest to study the impact of (joint) misspecification of these elements on point prediction and prediction inference.

Acknowledgements

This work was supported by the Australian Research Council under Grants DP230101908 and DP240100143. Thank you to Nickson Xu Ning for useful discussions.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2025.108254.

References

Agresti, A., Caffo, B., Ohman-Strickland, P., 2004. Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. Comput. Stat. Data Anal. 47, 639–653.
Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1.
Booth, J.G., Hobert, J.P., 1998. Standard errors of prediction in generalized linear mixed models. J. Am. Stat. Assoc. 93, 262–272.
Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. 88, 9–25.
Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Maechler, M., Bolker, B.M., 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J. 9, 378–400.
Cantoni, E., Mills Flemming, J., Welsh, A., 2017. A random-effects hurdle model for predicting bycatch of endangered marine species. Ann. Appl. Stat. 11, 2178–2199.
Carpenter, J.R., Goldstein, H., Rasbash, J., 2003. A novel bootstrap procedure for assessing the relationship between class size and achievement. J. R. Stat. Soc., Ser. C, Appl. Stat. 52, 431–443.
Chambers, R., Tzavidis, N., 2006. M-quantile models for small area estimation. Biometrika 93, 255–268.
Chatterjee, S., Lahiri, P., Li, H., 2008. Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. Ann. Stat. 36, 1221–1245.
Das, K., Jiang, J., Rao, J., 2004. Mean squared error of empirical predictor. Ann. Stat. 32, 818–840.
Erciulescu, A.L., Opsomer, J.D., 2022. A model-based approach to predict employee compensation components. J. R. Stat. Soc., Ser. C, Appl. Stat. 71, 1503–1520.
Flores-Agreda, D., Cantoni, E., 2019. Bootstrap estimation of uncertainty in prediction for generalized linear mixed models. Comput. Stat. Data Anal. 130, 1–17.
Gee, J.M., Warwick, R.M., Schaanning, M., Berge, J.A., Ambrose Jr, W.G., 1985. Effects of organic enrichment on meiofaunal abundance and community structure in sublittoral soft sediments. J. Exp. Mar. Biol. Ecol. 91, 247–262.
Ha, I.D., Sylvester, R., Legrand, C., MacKenzie, G., 2011. Frailty modelling for survival data from multi-centre clinical trials. Stat. Med. 30, 2144–2159.
Heagerty, P.J., Kurland, B.F., 2001. Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika 88, 973–985.
Ho, H.J., Lin, T.-I., 2010. Robust linear mixed models using the skew t distribution with application to schizophrenia data. Biom. J. 52, 449–469.
Hui, F.K.C., Bondell, H.D., 2021. A shared parameter mixture model for longitudinal income data with missing responses and zero rounding. Aust. N. Z. J. Stat. 63, 221–240.
Hui, F.K.C., Warton, D.I., Foster, S.D., 2015. Order selection in finite mixture models: complete or observed likelihood information criteria? Biometrika 102, 724–730.
Hui, F.K.C., Müller, S., Welsh, A.H., 2021. Random effects misspecification can have severe consequences for random effects inference in linear mixed models. Int. Stat. Rev. 89, 186–206.
Hui, F.K.C., Hill, N.A., Welsh, A.H., 2022. Assuming independence in spatial latent variable models: consequences and implications of misspecification. Biometrics 78, 85–99.
Jiang, J., 2003. Empirical best prediction for small-area inference based on generalized linear mixed models. J. Stat. Plan. Inference 111, 117–127.
Jiang, J., Nguyen, T., 2007. Linear and Generalized Linear Mixed Models and Their Applications. Springer.
Kackar, R.N., Harville, D.A., 1984. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. J. Am. Stat. Assoc. 79, 853–862.
Komárek, A., Lesaffre, E., 2008. Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. Comput. Stat. Data Anal. 52, 3441–3458.
Korhonen, P., Hui, F.K.C., Niku, J., Taskinen, S., 2023. Fast and universal estimation of latent variable models using extended variational approximations. Stat. Comput. 33, 26.
Lee, Y., Jang, M., Lee, W., 2011. Prediction interval for disease mapping using hierarchical likelihood. Comput. Stat. 26, 159–179.
Litiere, S., Alonso, A., Molenberghs, G., 2008. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. Stat. Med. 27, 3125–3144.
Liu, L., Yu, Z., 2008. A likelihood reformulation method in non-normal random effects models. Stat. Med. 27, 3105–3124.
Lyu, Z., Welsh, A., 2022. Asymptotics for EBLUPs: nested error regression models. J. Am. Stat. Assoc. 117, 2028–2042.
Manole, T., Khalili, A., 2021. Estimating the number of components in finite mixture models via the group-sort-fuse procedure. Ann. Stat. 49, 3043–3069.
Marino, M.F., Ranalli, M.G., Salvati, N., Alfo, M., 2019. Semiparametric empirical best prediction for small area estimation of unemployment indicators. Ann. Appl. Stat. 13, 1166–1197.
McCulloch, C., Searle, S., Neuhaus, J., 2011. Generalized, Linear, and Mixed Models. Wiley.
McCulloch, C.E., Neuhaus, J.M., 2011a. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. Stat. Sci. 26, 388–402.

McCulloch, C.E., Neuhaus, J.M., 2011b. Prediction of random effects in linear and generalized linear models under model misspecification. Biometrics 67, 270–279.
McLachlan, G.J., Rathnayake, S., 2014. On the number of components in a Gaussian mixture model. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 4, 341–355.
Meza, C., Osorio, F., De la Cruz, R., 2012. Estimation in nonlinear mixed-effects models using heavy-tailed distributions. Stat. Comput. 22, 121–139.
Nguyen, H.D., McLachlan, G., 2019. On approximations via convolution-defined mixture models. Commun. Stat., Theory Methods 48, 3945–3955.
Ning, X., Hui, F.K.C., Welsh, A.H., 2024. Asymptotic results for penalized quasi-likelihood estimation in generalized linear mixed models. Stat. Sin.
Ning, X., Hui, F.K.C., Welsh, A.H., 2025. Inferential procedures for random effects in generalized linear mixed models. PLoS ONE 20, e0320797.
Ormerod, J.T., Wand, M.P., 2012. Gaussian variational approximate inference for generalized linear mixed models. J. Comput. Graph. Stat. 21, 2–17.
Pan, L., Li, Y., He, K., Li, Y., Li, Y., 2020. Generalized linear mixed models with Gaussian mixture random effects: inference and application. J. Multivar. Anal. 175, 104555.
Pinheiro, J.C., Liu, C., Wu, Y.N., 2001. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. J. Comput. Graph. Stat. 10, 249–276.
Prasad, N.N., Rao, J.N., 1990. The estimation of the mean squared error of small-area estimators. J. Am. Stat. Assoc. 85, 163–171.
Rao, J.N.K., Molina, I., 2015. Small Area Estimation. John Wiley & Sons.
Singer, J.D., Willett, J.B., 2003. Applied Longitudinal Data Analysis. Oxford University Press.
Skrondal, A., Rabe-Hesketh, S., 2009. Prediction in multilevel generalized linear models. J. R. Stat. Soc., Ser. A, Stat. Soc. 172, 659–687.
Smith, A., Cullis, B.R., Thompson, R., 2005. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. J. Agric. Sci. 143, 449–462.
Verbeke, G., Lesaffre, E., 1996. A linear mixed-effects model with heterogeneity in the random-effects population. J. Am. Stat. Assoc. 91, 217–221.
Verbeke, G., Lesaffre, E., 1997. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. Comput. Stat. Data Anal. 23, 541–556.
Vidoni, P., 2006. Response prediction in mixed effects models. J. Stat. Plan. Inference 136, 3948–3966.
Wolfinger, R., 1993. Laplace's approximation for nonlinear mixed models. Biometrika 80, 791–795.
Zheng, N., Cadigan, N., 2021. Frequentist delta-variance approximations with mixed-effects models and tmb. Comput. Stat. Data Anal. 160, 107227.
Zheng, N., Cadigan, N., 2023. Frequentist conditional variance for nonlinear mixed-effects models. J. Stat. Theory Pract. 17, 3.