Statistica Sinica Preprint No: SS-2023-0082	
Title	Asymptotic Analysis of Mis-Classified Linear Mixed Models
Manuscript ID	SS-2023-0082
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0082
Complete List of Authors	Haiqiang Ma and Jiming Jiang
Corresponding Authors	Jiming Jiang
E-mails	jimjiang@ucdavis.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

ASYMPTOTIC ANALYSIS OF MIS-CLASSIFIED LINEAR MIXED MODELS

Haiqiang Ma and Jiming Jiang

Jiangxi University of Finance and Economics,

and University of California, Davis

Abstract: **We study impact of class misspecification on the analysis of linear mixed models. Here, the misclassification means that some of the classes or groups associated with the random effects are mismatched. Such misclassification problems are becoming increasingly common in modern data science, including intentional and unintentional misclassifications. One important case of intentional misspecification is related to differential privacy; while a case of unintentional misspecification arises in classified mixed model prediction. Our study shows that standard asymptotic properties of the maximum likelihood and restricted maximum likelihood estimators, including consistency and asymptotic normality, remain valid under the misclassification provided that the proportion of the misclassified group numbers is asymptotically negligible in a suitable sense. Empirical results of simulation studies fully support our theoretical findings. A real-data example is considered.**

Key words and phrases: Asymptotic behavior, differential privacy, linear mixed models,

misclassification, random effects, robustness

1. Introduction

Mixed effects models (e.g., Jiang and Nguyen (2021)) are widely used in practice. These models explore heteroscedasticity within the population, such as subpopulations or groups. These groups, characterized by the random effects, are of fundamental importance, and a main reason for the broad application of mixed effects models. It should be noted that the groups depend on the model we define, namely the mixed effects model—the data itself is not necessarily grouped, or grouped according to the mixed effects model.

A classical mixed effects model assumes that the group classifications are correctly specified. For example, there are 58 counties in the state of California, USA. Thus, if data are clustered according to those counties, numbered from 1 to 58, it is assumed that the county number is correctly specified for each data record. Specifically, consider a linear mixed model (LMM) with county-level random effects, expressed as $y_{ij} = x'_{ij}\beta + \alpha_i + \epsilon_{ij}$, $i = 1, \cdots, 58$, $j = 1, \cdots, n_i$, where y_{ij} is jth value of the response variable from county i, x_{ij} is an associated vector of covariates, β is an unknown vector of fixed effects, α_i is a county-level random effect, and ϵ_{ij} is an additional error. The number n_i corresponds to the group size, that is, the number of observations for the ith group. By correct

group specification it means that each data record, (y_{ij}, x_{ij}), is associated with the random effect α_i, not $\alpha_{i'}$ for some $i' \neq i$.

There are, however, increasingly many situations in modern data science, in which some of the group classifications are incorrect or mismatched. Some of these mismatches are unintentional, such as recording errors or matching errors (see below), while the others are intentional. An important case of the latter situation is *differential privacy* (DP). This is because, the classifications, or some components of the classifications, are related to privacy issues such as location, race, gender, and age groups. For example, Table 1 shows a portion of data from the 2010 Census of the United States for population totals of the 58 counties of the state of California. Only data from the first county (in alphabetical order of county name) are reported. Column 2 reports population totals for the 18 age-groups of column 1, and column 3 are the corresponding DP "contaminated" fuzzy versions produced by a U. S. Census DP algorithm (Abowd *et al.* (2022)). The fuzzy versions were created due to concerns of privacy protection so that various classifications, such as race, gender and age groups, have been altered. As a result, the fuzzy population totals do not match the census population totals for all of the age groups. Nevertheless, it is observed that the differences are very small compared to the subpopulation sizes, indicating that the impact of the contamination is minor. In fact, a main objective of the DP is to maintain

primary content of the information and, at the same time, protect privacy.

Table 1: **Age-adjusted Population Totals and Their Fuzzy Counterparts**: Alameda County, California; Source: United States Census Bureau 2010 (Fuzzy Population Totals Produced in March 2022; Diff. = Difference between Pops and Fuzzy Pops)

Age Groups	Pops	Fuzzy Pops	Diff	Age Groups	Pops	Fuzzy Pops	Diff
00-04 years	97,652	97,676	24	45-49 years	114,111	114,155	44
05-09 years	94,546	94,547	1	50-54 years	108,506	108,457	-49
10-14 years	91,070	91,058	-12	55-59 years	94,648	94,659	11
15-19 years	100,394	100,256	-138	60-64 years	78,854	78,861	7
20-24 years	107,049	106,977	-72	65-69 years	52,663	52,719	56
25-29 years	113,597	113,643	46	70-74 years	37,774	37,733	-41
30-34 years	114,607	114,651	44	75-79 years	29,185	29,267	82
35-39 years	115,275	115,271	-4	80-84 years	23,391	23,386	-5
40-44 years	112,216	112,201	-15	85+ years	24,733	24,753	20

A case of unintentional misspecification is classified mixed model prediction (CMMP; Jiang *et al.* (2018); also see Ma and Jiang (2022)). The latter may be regarded as a modernized version of the traditional mixed model prediction (MMP; e.g., Rao and Molina (2015), Jiang and Nguyen (2021)). Basically, in many practical problems there are available "training" data, for which the group numbers are correctly specified, and one wishes to make prediction about certain characteristics associated with some new data. The new data are un-

classified in the sense that the group numbers for the new data are unknown. Therefore, as a first step, CMMP tries to identify the group numbers of the new data. Once the group numbers are identified, the well-developed MMP methods can be utilized to improve prediction accuracy. There are situations in which the group classifications within the training data are also unknown, in which case a procedure is needed to identify the training data groups before using CMMP (e.g., Rao, Li and Jiang (2023)). In all of these situations, there are mismatched groups within the training data, or expanded training data. Specifically, in the former case, it is quite possible that the group number for the new data is misidentified; thus, when the new data is combined with the training data (to be used as future training data), some of the group numbers are mismatched; in the latter case, there are mismatched groups within the training data as a result of the group-identifying process.

In situations like the above, a question of practical interest is to what extent the misclassification impacts the existing methods of mixed model analysis. Typically, one would expect such an impact to be "minor", but how minor is minor, and in what sense? We need guidelines from a theoretical standpoint. Given the wide-ranging applications of LMMs, which by far are the most popular type of mixed effects models, we shall focus on LMM in this paper.

The current state-of-the-art methods for LMM analysis are maximum like-

lihood (ML) and restricted maximum likelihood (REML); see, for example, Chapter 1 of Jiang and Nguyen (2021). When groups numbers are correctly classified, asymptotic behavior of the ML and REML estimators have been well established. See, for example, Sections 2.1 and 2.2 of Jiang (2017). In particular, Miller (1977) established asymptotic normality of the ML estimators of the fixed effects and variance components under a mixed ANOVA model; Jiang (1996) gives sufficient conditions for the consistency and asymptotic normality of the REML and ML estimators, respectively, under a mixed ANOVA model that are also necessary in the case of balanced data.

The main goal of our current asymptotic analysis is to investigate the conditions, under which these standard asymptotic behaviors of the ML and REML continue to hold under misclassification of the group numbers. Before introducing the general settings and results, in Section 2 we first consider a special case, which is conceptually simple for illustration purposes. In this case, the standard asymptotic properties of ML and REML estimators, namely, consistency and asymptotic normality, can be established provided that the proportion of the misclassified group numbers is asymptotically negligible in a suitable sense.

Here, the word "negligible" should be defined more precisely to avoid confusion. It is in the sense that the large-sample behaviors of the parameter estimators, namely, consistency and asymptotic normality, are preserved. In some

cases, such as the special case mentioned above, the condition for the negligibility can be expressed more explicitly; in some more complex situation, the condition is not explicit, but the general concept still applies.

The more general settings are considered in Section 3, in which we extend the results of the special case to two types of LMMs. According to Jiang and Nguyen (2021) (sec. 1.2), there are two main types of LMMs, namely, the mixed ANOVA model and longitudinal model. Due to the space limit, here in this paper we focus on the mixed ANOVA models. We have obtained similar results for the longitudinal models, which will be published elsewhere. Some empirical results, including simulation studies and a real-data example, are presented in Section 4. The technical lemmas and detailed proofs of theoretical results for the special and general cases are deferred to the Supplementary Material.

2. A special case

Before introducing the general results, we take a look at a special case, in which the general result can be expressed in an explicit form. Consider a balanced one-way random effects model (e.g., Jiang and Nguyen (2021)), which can be expressed as

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{2.1}$$

$i = 1, \ldots, m$, $j = 1, \ldots, n$, where i represents the group (e.g., subject, community), n is the group size, that is, the number of observations in the training data that belong to group i, which is assumed to be the same for difference groups, hence explaining the term "balanced". Furthermore, y_{ij} is the outcome of interest, μ is an unknown mean, α_i is a group-specific random effect, and ϵ_{ij} is an error. It is assumed that the random effects and errors are independent with $\alpha_i \sim N(0, \tau^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$, where $\sigma^2, \tau^2 > 0$ are unknown variances. The model can be expressed in a vector-matrix form:

$$y = X\mu + Z\alpha + \epsilon, \tag{2.2}$$

where $y = (y_i)_{1 \le i \le m}$ with $y_i = (y_{ij})_{1 \le j \le n}$; $X = 1_m \otimes 1_n$; $Z = I_m \otimes 1_n$, $I_n, 1_n$ denote the n-dimensional identity matrix and vector of 1's, respectively, and $\otimes$ denotes the Kronecker product; $\alpha = (\alpha_i)_{1 \le i \le m}$, and ϵ is defined in a similar way as y with y_{ij} replaced by ϵ_{ij}.

The model, (2.1) or (2.2), is the assumed model. In practice, such a model may be mis-classified so that, in reality, one has

$$y_{ij} = \mu + \alpha_{\gamma_{ij}} + \epsilon_{ij}, \tag{2.3}$$

$i = 1, \ldots, m$, $j = 1, \ldots, n$, where $\gamma_{ij} \in \{1, ..., m\}$ denotes the true index of the random effect that y_{ij} is associated with. Let $\tilde{z}'_{ij}$ be the $1 \times m$ vector whose uth component is $1_{(\gamma_{ij}=u)}$, $1 \le u \le m$. Note that one of the components of

z_{ij} is 1 and the rest are 0. Also, let $\tilde{Z}_i$ be the $n \times m$ matrix whose jth row is $\tilde{z}'_{ij}$, $1 \le j \le n$, that is, $\tilde{Z}_i = (\tilde{z}'_{ij})_{1 \le j \le n}$. Then, let $\tilde{Z}$ be the matrix of stacking $\tilde{Z}_i, i = 1, \ldots, m$, that is, $\tilde{Z} = (\tilde{Z}_i)_{1 \le i \le m}$. Note that $\tilde{Z}$ is an $N \times m$ matrix with $N = mn$. Then, the true model, (2.3), can be expressed as

$$y = X\mu + \tilde{Z}\alpha + \epsilon, \tag{2.4}$$

where X is the same as in (2.2).

Let $\hat{\theta} = (\hat{\tau}^2, \hat{\sigma}^2)'$ be the REML estimator of $\theta = (\tau^2, \sigma^2)'$, and $\theta_0 = (\tau_0^2, \sigma_0^2)'$ be the true θ. Similarly, let $\tilde{\psi} = (\tilde{\mu}, \tilde{\tau}^2, \tilde{\sigma}^2)'$ be the ML estimator of $\psi = (\mu, \tau^2, \sigma^2)'$, and $\psi_0 = (\mu_0, \tau_0^2, \sigma_0^2)'$ be the true ψ. We assume that γ_{ij} are independent, and independent with the random effects and errors, such that $\mathrm{P}(\gamma_{ij} \neq i) = p$ and $\mathrm{P}(\gamma_{ij} = k)$ does not depend on k for $k \neq i$. Intuitively, p is the probability that the group index is misclassified, that is, the true group index is not the same as it is thought to be. It follows that $\mathrm{P}(\gamma_{ij} = i) = 1 - p$ and $\mathrm{P}(\gamma_{ij} = k) = p/(m-1)$ for $k \neq i$. Note that there are different ways of misclassification. For example, one could have the group labels of two groups completely swapped while the group labels of the other groups correctly intact, or the group labels of some members in each group misclassified. The asymptotic theory, to be established below, does not distinguish these two scenarios, as long as p is asymptotically the same.

The following results, which are special cases of the general theory to be es-

tablished, state conditions, especially regarding the magnitude of p, under which the standard asymptotic theory for the REML and ML estimators holds. Here, the REML equations refers to setting the system of equations that set the derivatives of the restricted log-likelihood function to zero, and the ML equations are defined similarly. The notation $m \sim n$ means that $\limsup(m/n) < \infty$ and $\liminf(m/n) > 0$.

Theorem 1 (Consistency of REML and ML estimators; special case). Suppose that the above distributional assumptions about the random effects, errors, and γ_{ij} hold. Furthermore, suppose that σ_0^2 and τ_0^2 are positive and that, as $m, n \to \infty$, we have $m \sim n$ and $p = O(1/\sqrt{mn})$. Then, the following hold:

(I) With probability tending to 1, the REML equation has a solution, $\hat{\theta} = (\hat{\tau}^2, \hat{\sigma}^2)$, such that $[\sqrt{m}(\hat{\tau}^2 - \tau_0^2), \sqrt{mn}(\hat{\sigma}^2 - \sigma_0^2)]'$ is bounded in probability. Thus, in particular, we have $\hat{\theta} \xrightarrow{\mathrm{P}} \theta_0$, that is, $\hat{\theta}$ is consistent.

(II) With probability tending to 1, the ML equation has a solution, $\tilde{\psi} = (\tilde{\mu}, \tilde{\tau}^2, \tilde{\sigma}^2)'$, such that $[\sqrt{m}(\tilde{\mu} - \mu_0), \sqrt{m}(\tilde{\tau}^2 - \tau_0^2), \sqrt{mn}(\tilde{\sigma}^2 - \sigma_0^2)]'$ is bounded in probability. Thus, in particular, we have $\tilde{\psi} \xrightarrow{\mathrm{P}} \psi_0$, that is, $\tilde{\psi}$ is consistent.

(III) In fact, the result for $\tilde{\mu}$ holds without any restriction on p.

Note 1. It is interesting to note that, in the standard asymptotic theory, consistency of the REML estimator only requires $m \to \infty$, as long as $n > 1$, in this special case (e.g., Jiang (1996), th. 4.2). Now, because some of the group

labels are misclassified, it requires that, in addition, $n \to \infty$. Such a requirement is similar to that for the consistency of CMMP (e.g., Jiang *et al.* (2018), Ma and Jiang (2022)), and there is an intuitive explanation. Basically, in each group there are some "misters right", that is, those whose group indexes are correctly classified, and some "misters wrong", that is, those whose group indexes are misclassified. The condition $n \to \infty$, together with the order of p, ensure that, in each group, the misters right dominate the misters wrong so that the overall structure of the LMM is essentially unchanged. In a similar way, the next result can be interpreted.

Note 2. It is also interesting to note that the restriction on p is needed only for the estimation of the variance components—no restriction on p is needed for the estimation of μ. In fact, in this case, the MLE of μ is simply the overall sample mean, which is not affected by the classifications. A similar note also applies to the next result.

Theorem 2 (Asymptotic normality of REML and ML estimators; special case). Under the conditions of Theorem 1, if the big O for p is replaced by

small o, then the REML and ML estimators are asymptotically normal in that

$$\begin{bmatrix} \sqrt{m}(\hat{\tau}^2 - \tau_0^2) \\ \sqrt{mn}(\hat{\sigma}^2 - \sigma_0^2) \end{bmatrix} \xrightarrow{\text{d}} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2\tau_0^4 & 0 \\ 0 & 2\sigma_0^4 \end{pmatrix} \right];$$
$$\begin{bmatrix} \sqrt{m}(\tilde{\tau}^2 - \tau_0^2) \\ \sqrt{mn}(\tilde{\sigma}^2 - \sigma_0^2) \end{bmatrix} \xrightarrow{\text{d}} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2\tau_0^4 & 0 \\ 0 & 2\sigma_0^4 \end{pmatrix} \right].$$

Also, we have $\sqrt{m}(\tilde{\mu} - \mu_0) \xrightarrow{\text{d}} N(0, \tau_0^2)$, which holds without any restriction on p.

Note 3. In modern data science, it has become increasingly possible to obtain massive training data in the sense that both m, n are large(e.g.,Kramlinger, Krivobokova and Sperlich (2022)), and n is even (much) larger than m. One special case of the latter is in the analysis of network data (e.g., Bickel and Chen (2009)). For example, community network plays an important role in precision epidemiology for infectious disease control (e.g., Ladner et al. (2019)). Typically, the number of distinct communities is relatively small compared to the number of individuals within each community. There are certainly correlations within the community but different communities may be treated as independent, or approximately independent. Also see Lyu and Welsh (2021) for another recent work under this kind of asymptotic framework.

The proofs of Theorem 1 and Theorem 2 are given in Supplementary Material for the REML parts. The proof for the ML parts is similar (actually, simpler),

and therefore omitted. Some remarks about the key differences between ML and REML, as well as similarity in the proofs for ML to that for REML, are provided in the Supplementary Material.

3. Mixed ANOVA model

A general mixed ANOVA model (e.g., Jiang and Nguyen (2021), sec. 1.2.1.1) can be expressed as

$$y_{ij} = x'_{ij}\beta + u'_{ij1}\alpha_{i1} + \cdots + u'_{ijs}\alpha_{is} + \epsilon_{ij}, \tag{3.1}$$

$i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where i is a group index, determined by possibly multiple levels of factors, n_i is the number of "replicates" within group i; y_{ij} is the outcome variable for the jth replicate within the ith group, x_{ij} is a q-vector of associated covariates, $\beta = (\beta_1, \cdots, \beta_q)'$ is an unknown vector of fixed effects, u_{ijl} are known q_l-vectors, α_{il} are q_l-vectors of random effects, and ϵ_{ij} is a random error. It is assumed that the random effects and errors are independent with $\alpha_{il} \sim N(0, \tau_l^2 I_{q_l}), 1 \le l \le s$ and $\epsilon_{ij} \sim N(0, \sigma^2)$, where $\sigma^2, \tau_l^2, 1 \le l \le s$ are unknown variances. The model can be written as

$$y_i = X_i\beta + U_{i1}\alpha_1 + \cdots + U_{is}\alpha_s + \epsilon_i, \tag{3.2}$$

where $y_i = (y_{ij})_{1 \le j \le n_i}$, $X_i = (x'_{ij})_{1 \le j \le n_i}$, $U_{il} = \mathrm{diag}(u'_{i1l}, \ldots, u'_{in_il})(Z_i \otimes I_{q_l})$ with $Z_i = e'_i \otimes 1_{n_i}$, e_i is the $m \times 1$ vector whose i-th component is 1 and other

components are 0, $\alpha_l = (\alpha'_{1l}, \ldots, \alpha'_{ml})'$, $1 \le l \le s$, and $\epsilon_i = (\epsilon_{ij})_{1\le j\le n_i}$. Furthermore, let $X = (X_i)_{1\le i\le m}$, $U_l = (U_{il})_{1\le i\le m}$, $1 \le l \le s$, $y = (y_i)_{1\le i\le m}$, and $\epsilon = (\epsilon_i)_{1\le i\le m}$. Then, the mixed ANOVA model can be rewritten as

$$y = X\beta + U_1\alpha_1 + \cdots + U_s\alpha_s + \epsilon. \tag{3.3}$$

The mixed ANOVA model (3.1), (3.2) or (3.3), is the assumed model. In practice, such a model may be misclassified in that, in reality, one has

$$y_{ij} = x'_{ij}\beta + u'_{ij1}\alpha_{\gamma_{ij1}} + \cdots + u'_{ijs}\alpha_{\gamma_{ijs}} + \epsilon_{ij}, \tag{3.4}$$

where $\gamma_{ijl} \in \{1, ..., m\}$ denotes the true group index for the random effects that y_{ij} is associated with. Let $\tilde{z}'_{ijl}$ be the $1 \times m$ vector whose uth component is $1_{(\gamma_{ijl}=u)}$, $1 \le u \le m$, $\tilde{Z}_{il} = (\tilde{z}_{i1l}, \cdots, \tilde{z}_{in_il})'$, $\tilde{Z}_l = (\tilde{Z}_{il})_{1\le i\le m}$, $X = (X_i)_{1\le i\le m}$,

$$\tilde{U}_{il} = \mathrm{diag}(u'_{i1l}, \cdots, u'_{in_il})(\tilde{Z}_{il} \otimes I_{q_l}),$$

$\tilde{U}_l = (\tilde{U}_{il})_{1\le i\le m}$, $1 \le l \le s$. Then, the true mixed ANOVA model, (3.4), can be expressed as

$$y = X\beta + \tilde{U}_1\alpha_1 + \cdots + \tilde{U}_s\alpha_s + \epsilon. \tag{3.5}$$

Comparing (3.5) with (3.3), the difference is apparent, that is, in the design matrices of the random effect vectors. Note that while $U_1, \ldots, U_s$ are known, $\tilde{U}_1, \ldots, \tilde{U}_s$ are unknown.

The standard methods for estimating the unknown coefficient parameters β, $\theta = (\tau_1^2, \cdots, \tau_s^2, \sigma^2)$ are ML and restricted maximum likelihood (REML) based on the assumed model, (3.3), because the true design matrices for the random effects are unknown. See, for example, Searle, Casella and McCulloch (1992), Jiang and Nguyen (2021). Let $N = \sum_{i=1}^{m} n_i$ be the total number of observations in the training data, and Φ be an $N \times (N-q)$ matrix satisfying rank$(\Phi) = N-q$ and $\Phi' X = 0$. Under (3.3), we have $y \sim N(X\beta, \sum_{l=1}^{s} U_l U_l' \tau_l^2 + I_N \sigma^2)$. Thus, the joint pdf of $\tilde{y} = \Phi' y$ is given by

$$f_\theta(\tilde{y}) = \frac{1}{(2\pi)^{(N-q)/2} |V(\Phi, \theta)|^{1/2}} \exp\left\{ -\frac{1}{2} \tilde{y}' V^{-1}(\Phi, \theta) \tilde{y} \right\},$$

where $V(\Phi, \theta) = \sigma^2 \Phi'\Phi + \sum_{l=1}^{s} \tau_l^2 \Phi' U_l U_l' \Phi$, and $|A|$ denotes the determinant of matrix A. The REML estimator of θ can be obtained by maximizing $\log f_\theta(\tilde{y})$. We first state the results regarding asymptotic behavior of the REML estimator under the misclassification.

3.1 Asymptotic behavior of REML estimator under misclassification

Let $l_\theta(\tilde{y}) = \log |V(\Phi, \theta)| + \tilde{y}' V^{-1}(\Phi, \theta) \tilde{y}$. As noted in Jiang (1996), a key condition for the consistency and asymptotic normality properties to hold for the REML estimator is that the LMM is asymptotically identifiable and infinitely informative under the invariant class. This condition is abbreviated as AI[4]. Define, for any matrices A, B, $\text{cor}(A, B) = \text{tr}(A'B)/\|A\|_2 \|B\|_2$, where

$\|A\|_2 = \sqrt{\mathrm{tr}(A'A)}$. Let

$$V_l = \sigma^2 V(\Phi,\theta)^{-1/2}\Phi' U_l U_l' \Phi V(\Phi,\theta)^{-1/2}, \;\; 1 \le l \le s,$$

and $V_{s+1} = I_{N-q}/\sigma^2$. Then, define

$$\mathrm{Cor}(V_1,\ldots,V_{s+1}) = [\mathrm{cor}(V_k, V_l)]_{1\le k,l\le s+1}.$$

θ_0 is said to be asymptotically identifiable under the invariant class (AI^2) if

$$\liminf \lambda_{\min}(\mathrm{Cor}(V_1,\ldots,V_{s+1})) > 0$$

[$\lambda_{\min}(A)$ denotes the smallest eigenvalue of symmetric matrix A]. The model is infinitely informative under the invariant class (I^3) about θ_0 if

$$\lim \|V_l\|_2 = \infty, \;\; 1 \le l \le s+1.$$

Then, the AI^4 condition holds provided that θ_0 is AI^2, about which the model is I^3. As noted in Jiang (1996), the standard limiting process in mixed effects model is not as simple as $N \to \infty$. Multiple numbers associated with the sample size, such as m, $q_1,\ldots,q_s$ and $n_i, 1 \le i \le m$, may increase. In the sequel, the notation $\lim$ (without subscript) simply represents such a complex limiting process. We also use the w. p. $\to 1$ for "with probability tending to one" under such a limiting process.

We further introduce the following notation. For $i, j, l = 1, \cdots, s+1$, define

$$\begin{aligned}
\Psi_{1i}(\theta_0) &= V^{-1}(\Phi, \theta_0)\frac{\partial V(\Phi, \theta_0)}{\partial \theta_i}, \\
\Psi_{2ij}(\theta_0) &= V^{-1}(\Phi, \theta_0)\frac{\partial V(\Phi, \theta_0)}{\partial \theta_i}V^{-1}(\Phi, \theta_0)\frac{\partial V(\Phi, \theta_0)}{\partial \theta_j}, \\
\Psi_{3ijl}(\theta) &= V^{-1}(\Phi, \theta)\frac{\partial V(\Phi, \theta)}{\partial \theta_i}V^{-1}(\Phi, \theta)\frac{\partial V(\Phi, \theta)}{\partial \theta_j}V^{-1}(\Phi, \theta)\frac{\partial V(\Phi, \theta)}{\partial \theta_l}.
\end{aligned}$$

Also, let $b(\theta_0) = (\sigma_0 I_N, \tau_{10}\tilde{U}_1, \cdots, \tau_{s0}\tilde{U}_s)'$,

$$\begin{aligned}
A_{1i}(\theta_0) &= b(\theta_0)\Phi\Psi_{1i}(\theta_0)V^{-1}(\Phi, \theta_0)\Phi' b'(\theta_0), \\
A_{2ij}(\theta_0) &= b(\theta_0)\Phi\Psi_{2ij}(\theta_0)V^{-1}(\Phi, \theta_0)\Phi' b'(\theta_0), \\
A_{3ijl}(\theta) &= \Phi\Psi_{3ijl}(\theta)V^{-1}(\Phi, \theta)\Phi'.
\end{aligned}$$

Define $\Theta_N = \{\theta : \|\theta_l - \theta_{l0}\| < q_l(N), l = 1, \cdots, s+1\}$, where $q_l(N)$ are sequences of positive numbers such that $q_l(N) \to 0$ and $p_l(N)q_l(N) \to \infty$, $l = 1, \cdots, s+1$.

Theorem 3. Consider a general mixed ANOVA model (3.3). Let the true model be (3.5) and the variances, $\sigma^2, \tau_l^2, 1 \le l \le s$ be positive. Let $p_l(N)$ be any sequence such that

$$p_l(N) \sim \mathrm{E}_{\theta_0}\left\{\frac{\partial^2 l_\theta(\tilde{y})}{\partial \theta_l^2}\bigg|_{\theta=\theta_0}\right\}, \quad l = 1, \cdots, s+1.$$

Also assume that AI[4] and the following additional conditions hold:

(i) The class indexes γ_{ijl}'s are independent with the random effects and errors.

(ii) For $1 \le i \le s+1$, we have $\mathrm{E}_{\theta_0}\big[\mathrm{tr}\{A_{2ii}(\theta_0)\}\big] = \mathrm{tr}\{\Psi_{2ii}(\theta_0)\}\{1+o(1)\}$,

$$\mathrm{Var}_{\theta_0}[\mathrm{tr}\{A_{2ii}(\theta_0)\}] = p_i^4(N)o(1),\ \mathrm{E}_{\theta_0}\big[\mathrm{tr}\{A_{2ii}^2(\theta_0)\}\big] = p_i^4(N)o(1).$$

(iii) For $1 \le i, j \le s+1$, we have $\mathrm{E}_{\theta_0}\big[\mathrm{tr}\{A_{2ij}(\theta_0)\}\big] = \mathrm{tr}\{\Psi_{2ij}(\theta_0)\}\{1+o(1)\}$,

$$\mathrm{Var}_{\theta_0}[\mathrm{tr}\{A_{2ij}(\theta_0)\}] = p_i^2(N)p_j^2(N)o(1), \mathrm{E}_{\theta_0}\Big(\mathrm{tr}\{A_{2ij}^2(\theta_0)\}\Big) = p_i^2(N)p_j^2(N)o(1).$$

(iv) For $i, j, l = 1, \cdots, s+1$, we have

$$\sup_{\theta\in\Theta_N} |\mathrm{tr}\{\Psi_{3ijl}(\theta)\}| = p_i(N)p_j(N)p_l(N)o(1)$$

and $\{\sup_{\theta\in\Theta_N} \|A_{3ijl}(\theta)\|\}\mathrm{tr}\big[\mathrm{E}\{b'(\theta_0)b(\theta_0)\}\big] = p_i(N)p_j(N)p_l(N)o(1)$, where for any matrix, A, $\|A\| = \sqrt{\lambda_{\max}(A'A)}$ denotes its spectral norm.

(v) For $1 \le l \le s+1$, we have $\mathrm{E}_{\theta_0}\big[\mathrm{tr}\{A_{1i}(\theta_0)\}\big] = \mathrm{tr}\{\Psi_{1i}(\theta_0)\} + p_i(N)O(1)$,

$$\mathrm{Var}_{\theta_0}\big[\mathrm{tr}\{A_{1i}(\theta_0)\}\big] = p_i^2(N)O(1),\ \mathrm{E}_{\theta_0}\big[\mathrm{tr}\{A_{1i}^2(\theta_0)\}\big] = p_i^2(N)O(1).$$

Then, w. p. $\to 1$, the REML equation has a solution, $\hat\theta = (\hat\tau_1^2, \ldots, \hat\tau_s^2, \hat\sigma^2)'$, satisfying

$$[p_1(N)(\hat\tau_1^2 - \tau_{10}^2), \ldots, p_s(N)(\hat\tau_s^2 - \tau_{s0}^2), p_{s+1}(N)(\hat\sigma^2 - \sigma_0^2)]' = O_{\mathrm{P}}(1).$$

Thus, in particular, we have $\hat\theta \xrightarrow{\mathrm{P}} \theta_0$.

Conditions (ii)—(v) basically regularize the orders of certain expectations and variances, which become trivial in the special case considered in Theorem 1

and Theorem 2. Note that these conditions also implicitly set constraints on the degree of class misspecification, which once again becomes more clear in the case of Theorem 1 and Theorem 2.

Next, define $I_N(\theta_0) = [I_{Nij}(\theta_0)]_{1\le i,j\le s+1}$, $\Sigma(\theta_0) = [\Sigma_{ij}(\theta_0)]_{1\le i,j\le s+1}$ with

$$I_{Nij}(\theta_0) = \frac{\mathrm{tr}\{\Psi_{2ij}(\theta_0)\}}{p_i(N)p_j(N)},\ \Sigma_{ij}(\theta_0) = 2\lim \frac{\mathrm{tr}[\mathrm{E}\{A_{1i}(\theta_0)\}\mathrm{E}\{A_{1j}(\theta_0)\}]}{p_i(N)p_j(N)},$$

$1 \le i, j \le s+1$. For random matrices $M_i, 1 \le i \le k$, their trace covariance matrix is defined as

$$\mathrm{Tc}(M_i, 1 \le i \le k) = (\mathrm{tr}[\{M_i - \mathrm{E}(M_i)\}\{M_j - \mathrm{E}(M_j)\}])_{1\le i,j\le k}. \qquad (3.6)$$

Theorem 4. Suppose that the conditions of Theorem 3 hold with condition (v) replaced by the following: (v+) $\mathrm{E}_{\theta_0}[\mathrm{tr}\{A_{1i}(\theta_0)\}] = \mathrm{tr}\{\Psi_{1i}(\theta_0)\} + p_i(N)o(1)$, $\mathrm{Var}_{\theta_0}[\mathrm{tr}\{A_{1i}(\theta_0)\}] = p_i^2(N)o(1)$, $1 \le i \le s+1$. Furthermore, suppose that (vi) $\mathrm{E}\{\mathrm{Tc}(A_{1i}(\theta_0)/p_i(N), 1 \le i \le s+1)\} = o(1)$. Then, the $\hat{\theta}$ in Theorem 3 satisfies

$$\Sigma^{-\frac{1}{2}}(\theta_0) I_N(\theta_0)[p_1(N)(\hat{\tau}_1^2 - \tau_{10}^2), \ldots, p_{s+1}(N)(\hat{\sigma}^2 - \sigma_0^2)]' \xrightarrow{\mathrm{d}} N(0, I_{s+1}).$$

Conditions (v+) and (vi) specify the orders of some additional expectation and variance. Note that $\mathrm{E}(\mathrm{Tc}(\cdots)\}$ is similar to a correlation matrix.

Note 4. Unlike Theorems 1 and 2, in the last two theorems the restriction on the extent of misclassification is not explicit due to the generality of the results;

however, the restriction is embedded in the limiting behavior of the quantities controlling the asymptotic behavior, namely, conditions (i)–(v) in Theorem 3, and (i)–(vi) in Theorem 4. A good thing about these general conditions is that they are not restricted to a specific form of misclassification, which may or may not be known in practice. In the special case of the balanced one-way random effects model, those conditions reduce to simply $m/n = O(1), p = O(1/\sqrt{mn})$ for Theorem 1, and $m/n = o(1), p = o(1/\sqrt{mn})$ for Theorem 2.

The proofs of Theorem 3 and Theorem 4 are given in the Supplementary Material.

3.2 Asymptotic behavior of ML estimator under misclassification

We now state the results on asymptotic behavior of the ML estimators. To avoid introducing too many notation, which a reader may lose track of, we maintain similar notation used for REML estimation; this may also help to link the ML results to the corresponding REML results, which may actually help understand the results on both ends. Just keep in mind that all the notation used in this subsection stay within the subsection.

Let $\theta = (\theta_1', \theta_2, \ldots, \theta_{s+2})' = (\beta', \tau_1^2, \ldots, \tau_s^2, \sigma^2)'$. Similarly, let $\theta_0 = (\beta_0', \tau_{10}^2, \cdots, \tau_{s0}^2, \sigma_0^2)'$ be the true θ. Note that $\theta_1 = \beta$ is a p-dimensional. Let

$$l_\theta(y) = \log[|V(\tau^2, \sigma^2)|] + (y - X\beta)' V^{-1}(\tau^2, \sigma^2)(y - X\beta),$$

where $V(\tau^2,\sigma^2)=\sigma^2 I_N+\sum_{l=1}^{s}\tau_l^2U_lU_l'$. For $i,j,l=1,\cdots,s+2$, define

$$\begin{aligned}
\Psi_{1i}(\theta_0) &= V^{-1}(\tau_0^2,\sigma_0^2)\frac{\partial V(\tau_0^2,\sigma_0^2)}{\partial\theta_i},\\
\Psi_{2ij}(\theta_0) &= V^{-1}(\tau_0^2,\sigma_0^2)\frac{\partial V(\tau_0^2,\sigma_0^2)}{\partial\theta_i}V^{-1}(\tau_0^2,\sigma_0^2)\frac{\partial V(\tau_0^2,\sigma_0^2)}{\partial\theta_j},\\
\Psi_{3ijl}(\theta_0) &= V^{-1}(\tau_0^2,\sigma_0^2)\frac{\partial V(\tau_0^2,\sigma_0^2)}{\partial\theta_i}V^{-1}(\tau_0^2,\sigma_0^2)\frac{\partial V(\tau_0^2,\sigma_0^2)}{\partial\theta_j}\\
&\quad\times V^{-1}(\tau_0^2,\sigma_0^2)\frac{\partial V(\tau_0^2,\sigma_0^2)}{\partial\theta_l};
\end{aligned}$$

and $b(\theta_0)=(\sigma_0 I_N,\tau_{10}\tilde{U}_1,\cdots,\tau_{s0}\tilde{U}_s)'$, $A_{1i}(\theta_0)=b(\theta_0)\Psi_{1i}(\theta_0)V^{-1}(\tau_0^2,\sigma_0^2)b'(\theta_0)$,

$A_{2ij}(\theta_0)=b(\theta_0)\Psi_{2ij}(\theta_0)V^{-1}(\tau_0^2,\sigma_0^2)b'(\theta_0)$, $A_{3ijl}(\theta_0)=\Psi_{3ijl}(\theta_0)V^{-1}(\tau_0^2,\sigma_0^2)$.

Also, define $q_l(N), l=1,\ldots,s+2$ as above Theorem 3.

Theorem 5. Consider the mixed ANOVA model (3.3) and let the true model be (3.5) and that the variances, $\sigma^2,\tau_l^2, 1\le l\le s$ be positive. Let $p_l(N)$ be any sequence such that $p_1^2(N)\sim \mathrm{tr}[\mathrm{E}_{\theta_0}\{\partial^2 l_\theta(y)/\partial\theta_1\partial\theta_1'|_{\theta=\theta_0}\}]$, $p_l^2(N)\sim \mathrm{E}_{\theta_0}\{\partial^2 l_\theta(y)/\partial\theta_l^2\}$, $l=2,\ldots,s+2$. Also assume that AI[4] and the following conditions hold:

(i) The class indexes γ_{ijl}'s are independent with the random effects and errors.

(ii) For $i=2,\ldots,s+2$, we have $\mathrm{E}_{\theta_0}[\mathrm{tr}\{A_{2ii}(\theta_0)\}]=\mathrm{tr}\{\Psi_{2ii}(\theta_0)\}\{1+o(1)\}$,

$$\begin{aligned}
&\mathrm{tr}\big[X'\Psi_{1i}(\theta_0)V^{-1}(\tau_0^2,\sigma_0^2)\mathrm{E}\{b'(\theta_0)b(\theta_0)\}\Psi_{1i}(\theta_0)V^{-1}(\tau_0^2,\sigma_0^2)X\big]\\
&=p_1^2(N)p_i^2(N)o(1);\\
&\mathrm{Var}_{\theta_0}\big[\mathrm{tr}\{A_{2ii}(\theta_0)\}\big]=p_i^4(N)o(1),\quad \mathrm{E}_{\theta_0}\big[\mathrm{tr}\{A_{2ii}^2(\theta_0)\}\big]=p_i^4(N)o(1).
\end{aligned}$$

(iii) For $i, j = 2, \cdots, s+2$, we have $\mathrm{E}_{\theta_0}[\mathrm{tr}\{A_{2ij}(\theta_0)\}] = \mathrm{tr}\{\Psi_{2ij}(\theta_0)\}\{1+o(1)\}$,

$$\mathrm{tr}\big[X' \sup_{\theta\in\Theta_N}\{\Psi_{1i}(\theta)V^{-1}(\tau^2,\sigma^2)\}X\big] = p_1^2(N)p_i(N)o(1),$$
$$\mathrm{Var}_{\theta_0}\big[\mathrm{tr}\{A_{2ij}(\theta_0)\}\big] = p_i^2(N)p_j^2(N)o(1),$$
$$\mathrm{E}_{\theta_0}\big[\mathrm{tr}\{A_{2ij}^2(\theta_0)\}\big] = p_i^2(N)p_j^2(N)o(1).$$

(iv) For $i, j, l = 2, \cdots, s+2$, we have

$$\sup_{\theta\in\Theta_N}|\mathrm{tr}\{\Psi_{3ijl}(\theta)\}| = p_i(N)p_j(N)p_l(N)o(1),$$
$$\mathrm{tr}\big[\sup_{\theta\in\Theta_N}\{A_{3ijl}(\theta)\}\mathrm{E}\{b'(\theta_0)b(\theta_0)\}\big] = p_i(N)p_j(N)p_l(N)o(1),$$
$$\sup_{\theta\in\Theta_N}\mathrm{tr}\Big[X'\{\Psi_{2ji}(\theta)+\Psi_{2ij}(\theta)\}V^{-1}(\tau^2,\sigma^2)\mathrm{E}\{b'(\theta_0)b(\theta_0)\}$$
$$V^{-1}(\tau^2,\sigma^2)\{\Psi_{2ji}(\theta)+\Psi_{2ij}(\theta)\}'X\Big] = p_1^2(N)p_i^2(N)p_j^2(N)o(1).$$

(v) For $i = 2, \cdots, s+2$, we have $\mathrm{E}_{\theta_0}[\mathrm{tr}\{A_{1i}(\theta_0)\}] = \mathrm{tr}\{\Psi_{1i}(\theta_0)\} + p_i(N)O(1)$,

$$\mathrm{tr}\big[X'V^{-1}(\tau_0^2,\sigma_0^2)\mathrm{E}\{b'(\theta_0)b(\theta_0)\}V^{-1}(\tau_0^2,\sigma_0^2)X\big] = p_1^2(N)O(1),$$

$$\mathrm{Var}_{\theta_0}\big[\mathrm{tr}\{A_{1i}(\theta_0)\}\big] = p_i^2(N)O(1),\ \ \mathrm{E}_{\theta_0}\big[\mathrm{tr}\{A_{1i}^2(\theta_0)\}\big] = p_i^2(N)O(1).$$

Then, w. p. $\to 1$, the ML equation has a solution, $\hat\theta = (\hat\beta', \hat\tau_1^2, \ldots, \hat\tau_s^2, \hat\sigma^2)'$, satisfying

$$\big[p_1(N)(\hat\beta-\beta_0)', p_2(N)(\hat\tau_1^2-\tau_{10}^2), \ldots, p_{s+1}(N)(\hat\tau_s^2-\tau_{s0}^2),$$
$$p_{s+2}(N)(\hat\sigma^2-\sigma_0^2)\big] = O_{\mathrm{P}}(1).$$

Thus, in particular, we have $\hat\theta \xrightarrow{\mathrm{P}} \theta_0$.

Conditions (ii)–(v) of Theorem 5, and conditions (vi)–(viii) of Theorem 6 below, have similar interpretations as the corresponding conditions in Theorem 3 and Theorem 4.

Next, we consider asymptotic normality of the ML estimator. First introduce some notation. Let $I_N(\theta_0) = \mathrm{diag}[I_{N1}(\theta_0), I_{N2}(\theta_0)]$, where $I_{N1}(\theta_0) = (2/N)XV^{-1}(\tau_0^2, \sigma_0^2)X'$, $I_{N2}(\theta_0) = \{I_{N2ij}(\theta_0)\}_{2\le i,j\le s+2}$ with $I_{N2ij}(\theta_0) = \mathrm{tr}\{\Psi_{2ij}(\theta_0)\}/p_i(N)p_j(N)$, and $\Sigma(\theta_0) = \big\{\Sigma_{ij}(\theta_0)\big\}_{2\le i,j\le s+2}$ with $\Sigma_{ij}(\theta_0) = \lim_{N\to\infty}\{2/p_i(N)p_j(N)\}\mathrm{tr}\big[\mathrm{E}\{A_{1i}(\theta_0)\}\mathrm{E}\{A_{1j}(\theta_0)\}\big]$, $i, j = 2, \cdots, s+2$. Also recall the trace covariance matrix defined via (3.6).

Theorem 6. Suppose that, in addition to the conditions of Theorem 5,

(vi) $\mathrm{tr}\big[X'V^{-1}(\tau_0^2, \sigma_0^2)\mathrm{Var}\{b'(\theta_0)\}V^{-1}(\tau_0^2, \sigma_0^2)X\big] = No(1)$ and

$\mathrm{tr}\Big(X'V^{-1}(\tau_0^2, \sigma_0^2)\big[\mathrm{E}\{b'(\theta_0)\}\mathrm{E}\{b(\theta_0)\} - V(\tau_0^2, \sigma_0^2)\big]V^{-1}(\tau_0^2, \sigma_0^2)X\Big) = No(1)$;

(vii) $\mathrm{E}_{\theta_0}\big[\mathrm{tr}\{A_{1i}(\theta_0)\}\big] = \mathrm{tr}\{\Psi_{1i}(\theta_0)\} + p_i(N)o(1)$ and

$\mathrm{Var}_{\theta_0}\big[\mathrm{tr}\{A_{1i}(\theta_0)\}\big] = p_i^2(N)o(1)$ for $i = 2, \cdots, s+2$;

(viii) $\mathrm{E}\big\{\mathrm{Tc}(A_{1i}(\theta_0)/p_i(N), 1 \le i \le s+2)\big\} = o(1)$.

Then, the ML estimator $\hat\theta$ of Theorem 5 satisfies

$$\sqrt{N}(\hat\beta - \beta_0) \xrightarrow{\mathcal{L}} N\Big[0, \Big\{\frac{1}{N}X'V^{-1}(\tau_0^2, \sigma_0^2)X\Big\}^{-1}\Big],$$

$$\Sigma^{-1/2}(\theta_0)I_{N2}(\theta_0)\Big[p_2(N)(\hat\tau_1^2 - \tau_{10}^2), \cdots, p_{s+2}(N)(\hat\sigma^2 - \sigma_0^2)\Big]'$$

$$\xrightarrow{\mathcal{L}} N(0, I_{s+1}).$$

The proofs of Theorem 5 and Theorem 6 for the asymptotic behavior of the ML estimator of β are given in the supplementary material; the proofs for the ML estimators of the variance components are similar to their REML counterparts, and therefore omitted.

Note 5. Like Theorems 3 and 4, the restriction on the extent of misclassification is not explicit in Theorems 5 and 6 due to the generality of the results; however, the restriction is embedded in the limiting behavior of the quantities controlling the asymptotic behavior. As noted, an advantage about these general conditions is that they are not restricted to a special kind of misclassification, which may or may not be known in practice. Basically, one needs is to verify that (i)–(v) in Theorem 1, and (vi)–(viii) in Theorem 2, are satisfied. In particular, for the special case considered in Section 2, those conditions reduce to the restriction on p.

Note 6. Unlike Theorem 1 and Theorem 2, it can be shown that, in general, the asymptotic behavior of β can be affected by p, the probability of misclassification, under the special form of misclassification considered in Section 2.

4. Empirical results

4.1 A simulation study

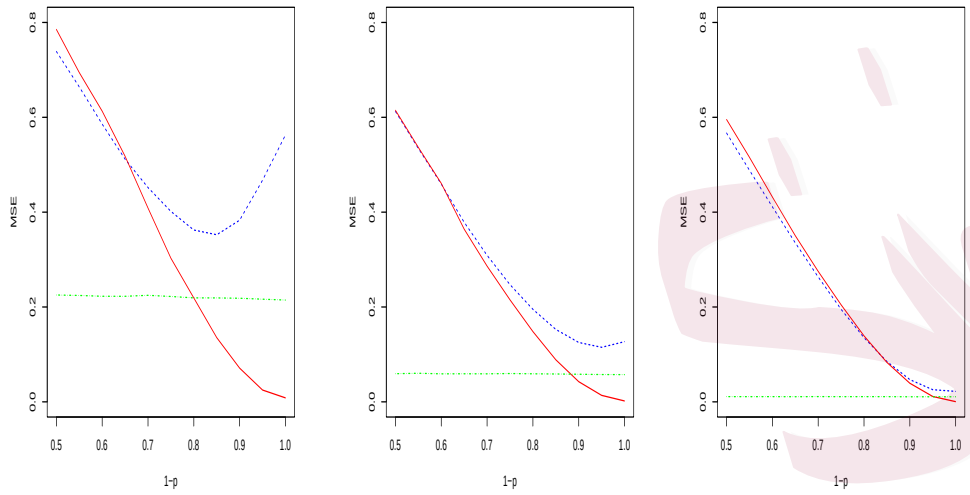
We carry out a simulation study on the finite-sample performance of the ML estimators under the mis-classified LMM. Specifically, we consider an example studied by Jiang *et al.* (2018) based the following assumed model: $y_{ij} = \beta_{10} + x_{1,ij}\beta_{20} + x_{2,ij}\beta_{30} + \alpha_i + \epsilon_{ij}$, where $\beta_0 = (\beta_{10}, \beta_{20}, \beta_{30}) = (1, 2, 3)$, $i = 1, \ldots, m, j = 1, \ldots, n$, α_i's and ϵ_{ij}'s are independent with $\alpha_i \sim N(0, \tau^2)$, $\epsilon_{ij} \sim N(0, \sigma^2)$, with $\tau^2 = \sigma^2 = 1$; $x_{k,ij}, k = 1, 2$ are generated from $N(0, 1)$, then fixed throughout the simulation. The true index γ_{ij} of the random effect that y_{ij} is associated with satisfies the model described below (2.4) with $1 - p \in \{0.5, 0.55, 0.6, \cdots, 1\}$. We consider $n = 50$ and $m = 5, 20, 100$.

We run 200 simulations under each combination of m, n, p values specified above, and report the empirical MSE of the MLE of the fixed effects β and variance components τ^2, σ^2. The empirical MSEs are presented in Figure 1. It can be seen that, when m is relatively small, the empirical MSE for τ^2 is relatively large, and does not converge to zero as p approaches zero. On the other hand, the empirical MSEs for σ^2 are very small, and rapidly converge to zero as p tends to zero. This is not surprising, because the convergence rate of the MLE of σ^2 is $\sqrt{mn}$, while that of τ^2 is $\sqrt{m}$ (e.g., Theorem 1).

When m is relatively large, the empirical MSEs for both τ^2 and σ^2 are very small, and approach to zero as p tends to zero. Furthermore, the convergence patterns of two variance estimators are very similar, with $\hat{\sigma}^2$ having better performance than $\hat{\tau}^2$.

In addition, regarding the empirical MSE of the MLE of β, it seems clear that, for fixed m, and different misclassification probability $p \in [0, 1]$, the empirical MSE of remain almost unchanged. Moreover, as m increases from 5 to 100, the empirical MSE of $\hat{\beta}$ tends to zero rapidly. This can be also explained by the convergence rate of $\hat{\beta}$ (e.g., Theorem 1).

In summary, the results seem to be consistent with the asymptotic behavior of the MLEs under the mis-classified LMM, as established by our theory.



(a) $m = 5, n = 50$ (b) $m = 20, n = 50$ (c) $m = 100, n = 50$

Figure 1: Trend of Empirical MSE as p decreases for MLE of β (dotted, green), τ^2 (dashed, blue), and σ^2 (solid, red)

4.2 A real data analysis

Here we analyze a more modern dataset, specifically, breast cancer data from The Cancer Genome Atlas (TCGA) (https://tcgabdata.nci.nih.gov/tcga). TCGA is a public repository data portal of high-quality pan-cancer tumor samples where clinical information, metadata, histopathology and molecular profiling information is available to researchers. For some cancers, non-tumor samples are also in the repository. It is a data repository that is commonly used to study underlying

genomic determinants of cancer.

This dataset has been studied by Jiang *et al.* (2018). Here, we re-analyzed breast cancer samples and their clinical information variables only to illustrate the robustness of maximum likelihood method under the misclassification. To facilitate this analysis, we focus on those patients who died and use a transformation of their survival time as our response of interest. This left 104 patients. Clinical variables included history of other malignancies, race, tumor status, surgical procedure performed, lymph nodes examined (yes/no), number of positive lymph nodes, stage of tumor based on AJCC staging criteria, estrogen receptor status, progesterone receptor status (both determined by an in situ hybridzation (ISH) assay) and age at diagnosis.

Since the group classifications within this real data are unknown, one does not know the truth in this application study. Thus, we need pre-perform a procedure to identify the training data groups. Similar to Jiang *et al.* (2018), we use the prediction analysis of microarrays (pam) algorithm of Tibshirani *et al.* (2002) to cluster survival time into different groups based on the clinical predictors and chose informative predictors based on this clustering. This left only the age and surgery variables. After removing observations with missing values in either of these two predictors, we had 95 patients for the robustness analysis.

In Jiang *et al.* (2018), they adopted an "elbow" point in scree plot to se-

lect the number of groups as 10 (note that the between cluster to within cluster variance ratio continued to decrease as the number of groups increased, but 10 represented an "elbow" point where larger group numbers gave diminishing improvements in fit). Given the fact that there is some uncertainty in determining the number of groups produced by the pam algorithm, therefore, this selection method has certain subjective and inaccuracy, and there may be some misclassifications in both the number of groups and the group index for every observation. Now, we assume the true number of groups as 10 (the "elbow" point in scree plot) and consider the number of groups as 8,9,10,11,12, respectively. Then, we use the pam algorithm to conduct the clustering analysis for this dataset. At last, we consider the following LMM: $y_{ij} = \beta_0 + x'_{ij}\beta + \alpha_i + e_{ij}$, where the covariate vector x_{ij} consists of age and surgery variables; the response, y_{ij} is the square root of of survival time, which give a suitable normalizing transformation, the group-specific random effects, α_i, and errors, e_{ij}, are assumed to be independent with $\alpha_i \sim N(0, \tau^2)$ and $e_{ij} \sim N(0, \sigma^2)$.

We report the averaged relative absolute bias (RAB), defined as

$$\begin{aligned}
\mathrm{RAB}_{\tau^2}(m) &= |\hat{\tau}^2(m) - \hat{\tau}^2(10)|/\hat{\tau}^2(10), \\
\mathrm{RAB}_{\sigma^2}(m) &= |\hat{\sigma}^2(m) - \hat{\sigma}^2(10)|/\hat{\sigma}^2(10), \\
\mathrm{RAB}_{\beta}(m) &= \frac{1}{3}\sum_{r=1}^{3} |\hat{\beta}_{r-1}(m) - \hat{\beta}_{r-1}(10)|/|\hat{\beta}_{r-1}(10)|,
\end{aligned}$$

where $\hat{\beta}_{r-1}(m), \hat{\tau}^2(m), \hat{\sigma}^2(m)$ are the MLEs of β_{r-1}, τ^2, σ^2, respectively, under the number of groups $m = 8, 9, 10, 11, 12$, and $\hat{\beta}_{r-1}(10), \hat{\tau}^2(10), \hat{\sigma}^2(10), r = 1, 2, 3$, are the MLEs for the "true" case, that is, $m = 10$.

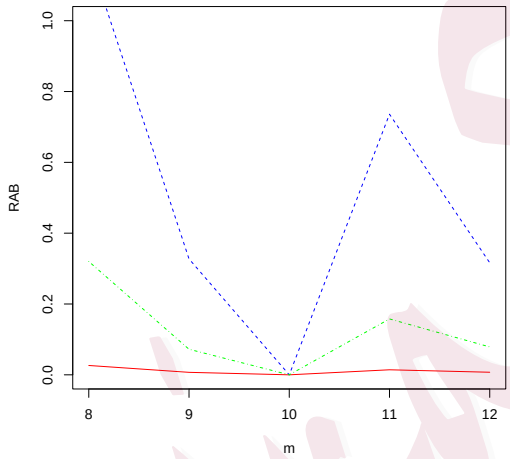


Figure 2: Trend of Empirical RAB as the number of groups m increases for MLE of β (dotted, green), τ^2 (dashed, blue) , and σ^2 (solid, red)

From Figure 2, it is clear that when the number of groups is deviated from the "true" value $m = 10$, the RAB increases for every parameter estimation. However, the increase rates and patterns are different for different parameter estimations. From a theoretical standpoint, the convergence rates for different parameter estimation are different (Theorem 5). Specifically, the convergence rate for σ^2 and the slope coefficients of β is $O(N^{-1/2})$ with $N = \sum_{i=1}^{m} n_i$, while the convergence rate for τ^2 and the intercept β_0 is $O(m^{-1/2})$. Note that the RAB

for β is defined as the average RAB for different β components; thus, the overall convergence rate for β is somewhere between $O(N^{-1/2})$ and $O(m^{-1/2})$. This is consistent with what we have observed in Figure 2, that is, in terms of the magnitudes of increase in RAB, τ^2 is the largest, followed by β, and σ^2 is the smallest.

5. Discussion

We have studied the impact of class misspecification on the asymptotic analysis of linear mixed models for the special and general cases of LLM.

For balanced data (Theorem 1 and Theorem 2), we consider the standard asymptotic theories for the REML and ML estimators under the misclassification, with restrictions on the extent of misclassification, that is, $m/n = O(1), p = O(1/\sqrt{mn})$ for Theorem 1, and $m/n = o(1), p = o(1/\sqrt{mn})$ for Theorem 2. For the general case when the data are not necessarily balanced (Theorems 3-6), although the conditions do not seem to put restrictions regarding the misclassification rate and the degree of data balancedness due to the generality of the results, the restrictions are embedded in the limiting behavior of the quantities controlling the asymptotic behavior, namely, conditions (i)–(v) in Theorem 3, (i)–(vi) in Theorem 4, (i)–(v) in Theorem 5, and (vi)–(viii) in Theorem 6. A good thing about these general conditions is that they are not restricted to a spe-

cific form of misclassification, which may or may not be known in practice.

On the other hand, from the results of the empirical studies, it can be seen that the data unbalancedness seems to impact estimation of different parameters differently. See the discussions of our simulation results and real-data analysis. We suspect that this has to do with the convergence rates for estimating different parameters, but not in a simple way. We shall investigate the impact of data unbalancedness theoretically in our future work.

As noted in the first paragraph of Section 1, the groups in a mixed effects model characterized by the random effects are of fundamental importance. Intuitively, one cannot maintain good asymptotic behavior of the estimators without restriction on the group misclassification. For example, suppose the one wishes to estimate the between-cluster variation. If the classes are substantially misclassified, the observed variation is not the actual between-cluster variation (for instance, if all the larger observations are misclassified into one group, and all the smaller observations to another, one might think there is a larger between-cluster variation, while, in fact, each group have both larger and smaller observations). In the special case of Theorem 1 and Theorem 2, the restrictions on p, the proportion of misclassified groups, is quite close to necessary, if not already necessary. However, for the general results of Theorems 3–6, it is possible to express the constraints in more explicit ways, although we do not believe that

they can be substantially weakened.

Supplementary Materials

The Supplementary Material contains some technical lemmas and detailed proofs of the theoretical results as well as additional empirical results.

Acknowledgements

Haiqiang Ma's research is supported by NNSF of China (No. 12161042), China Postdoctoral Science Foundation (No. 2019M662262), and Postdoctoral Foundation of Jiangxi Province (No. 2021KY18). The research of Jiming Jiang is supported by the NSF grants DMS-1510219 ahd DMS-713120.

References

Abowd, J. M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., and Zhuravlev, P. (2022). The 2020 Census Disclosure Avoidance System TopDown Algorithm, Tech. Report. https://www.census.gov/library/working-papers/2022/adrm /CED-WP-2022-002.html.

Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and Newman-Girvan and other modularities. *PNAS* **106**, 21068–21073.

Jiang, J. (1996). REML estimation: Asymptotic behavior and related topics. *Ann. Statist.* **24**, 255–286.

Jiang, J. (2022). *Large Sample Techniques for Statistics* (2nd ed.), Springer, New York.

REFERENCES

Jiang, J. (2017). *Asymptotic Analysis of Mixed Effects Models: Theory, Application, and Open Problems*, Chapman & Hall/CRC.

Jiang, J. and Nguyen, T. (2021). *Linear and Generalized Linear Mixed Models and Their Applications* (2nd ed.), Springer, New York.

Jiang, J., Jia, H. and Chen, H. (2001). Maximum posterior estimation of random effects in generalized linear mixed models. *Statist. Sinica* **11**, 97–120.

Jiang, J., Rao, J. S., Fan, J., and Nguyen, T. (2018). Classified mixed model prediction. *J. Amer. Statist. Assoc.* **113**, 269–279.

Johnson, N. L. (1962). The folded normal distribution: Accuracy of the estimation by maximum likelihood. *Technometrics* **4**, 249–256.

Kramlinger, P., Krivobokova, T., and Sperlich, S. (2022). Marginal and conditional multiple inference for linear mixed model predictors. *J. Amer. Statist. Assoc.*, in press.

Ladner, J. T., Grubaugh, N. D., Pybus, O. G., & Anderson, K. G. (2019). Precision epidemiology for infectious disease control. *Nature Medicine* **25**, 206–211.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lyu, Z. and Welsh, A. H. (2021). Asymptotics for EBLUPs: Nested error regression models. *J. Amer. Statist. Asssoc.*, https://doi.org/10.1080/01621459.2021.1895178.

Ma, H. and Jiang, J. (2022). Pseudo-Bayesian classified mixed model prediction. *J. mer. Statist. Assoc.*,

https://doi.org/10.1080/01621459.2021.2008944.

Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of analysis of variance. *Ann. Statist.* **5**, 746–762.

Nurty, M. N. and Devi, V. S. (2011). *Pattern Recognition: An Algorithmic Approach*. Springer-Verlag, London.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of mean squared errors of small area estimators. *J. Amer. Statist. Assoc.* **85**, 163-171.

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation* (2nd ed.), Wiley, New York.

Rao, J. S., Li, M. and Jiang, J. (2023). Classified mixed model projections. *J. Amer. Statist. Assoc.*, https://doi.org/10.1080/01621459.2023.2218041

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*. Wiley, New York.

Tibshirani, R., Hastie, T., Narasimhan, N. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS.* **99**, 6567–6572.