Special Proceedings: ISBN #: 978-81-950383-5-0
26th Annual Conference, 26-28 February 2024; pp 207-220

V. K. Gupta Endowment Award Lecture 2024: Modernizing Linear Mixed Model Prediction

J. Sunil Rao
Division of Biostatistics, University of Minnesota, Twin Cities, USA

Received: 30 April 2024; Revised: 18 August 2024; Accepted: 20 August 2024

Abstract

Mixed models have widespread appeal in many areas of statistical modeling from biostatistics to small area estimation. Here we review a variety of recent approaches for modernizing linear mixed model prediction including robust prediction via the observed best predictor (OBP) to prediction for new test data using a classified random effect, namely classified mixed model prediction (CMMP). Finally, a brief mention will be made to a proposal for using mixed model prediction to project outside of the range of the training data using classified mixed model projections.

Key words: Mixed model selection; GIC; Fence method; Small area estimation; Subareas.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

It is now well accepted that it is necessary to *borrow strength* from relevant domains and resources to increase the efficiency of direct estimates and that linear mixed models provide a pathway to do so. In this regard, the empirical best linear unbiased prediction, or EBLUP, method has had a dominant influence (*e.g.*, Rao 2003; Jiang and Lahiri 2006). The method utilizes a linear mixed effects model (*e.g.*, Jiang 2007) in order to borrow strength. The standard procedure of computing the EBLUP is the following. First one derives the best predictor (BP) of the mixed effects of interests, such as the small area means. Then, one replaces the vector of the fixed effects by its maximum likelihood estimator (MLE), assuming that the variance components are known (up to this stage one obtains the best linear unbiased predictor, or BLUP). Finally, one replaces the unknown variance components by their ML or REML estimators. It follows that the EBLUP is the BP, in which the unknown fixed parameters, including the fixed effects and variance components, are estimated either by ML or REML. The latter are known to be asymptotically optimal under estimation considerations (*e.g.*, Jiang 2007). However, in many cases, such as in SAE, the problem of main interest is prediction, rather than estimation. The implication is that the EBLUP may be regarded as a hybrid of optimal prediction (*i.e.*, BP) and optimal estimation (*e.g.*, ML). Nevertheless, if prediction is of main interest, it would be more natural to have a purely

Corresponding Author: J. Sunil Rao
Email: js-rao@umn.edu

predictive procedure, in which both the predictor and estimator are derived from predictive considerations.

2. A general framework for the observed best predictor (OBP)

First, consider a general mixed model prediction problem (*e.g.*, Robinson 1991). The assumed model is $y = X\beta + Zv + e$, where X, Z are known matrices; β is a vector of fixed effects; v, e are vectors random effects and errors, respectively, such that $v \sim N(0, G)$, $e \sim N(0, \Sigma)$, and v, e are uncorrelated. Suppose that the true underlying model is $y = \mu + Zv + e$, where $\mu = \mathrm{E}(y)$. Here, again, E without subscript represents expectation with respect to the true distribution, which may be unknown but is not model-dependent. Following Jiang *et al.* (2011), our interest is prediction of a vector of mixed effects that can be expressed as $\theta = F'\mu + R'v$, where F, R are known matrices. Suppose that G, Σ are known. Then, the best predictor (BP) of θ, in the sense of minimum MSPE, under the assumed model is given by $\mathrm{E}_{\mathrm{a}}(\theta|y) = F'\mu + R'\mathrm{E}_{\mathrm{a}}(v|y) = F'X\beta + R'GZ'V^{-1}(y - X\beta)$, where E_{a} denotes expectation under the assumed model, $V = \mathrm{Var}(y) = \Sigma + ZGZ'$ and β is the true vector of fixed effects, under the assumed model. If we write $B = R'GZ'V^{-1}$ and $\Gamma = F' - B$, then the BP can be expressed as

$$\mathrm{E}_{\mathrm{a}}(\theta|y) \quad = \quad F'y - \Gamma(y - X\beta). \tag{1}$$

Now let $\check{\theta}$ denote the right side of (1) with a fixed, but arbitrary β. Then, it can be shown that $\mathrm{MSPE}(\check{\theta}) = \mathrm{E}(I_1 - 2I_2 + (y - X\beta)'\Gamma'\Gamma(y - X\beta)\}$, where I_1, I_2 do not depend on β. Thus, the best predictive estimator (BPE) (Jiang *et al.* 2011) of β is obtained by minimizing the expression inside the expectation, that is, $\tilde{\beta} = (X'\Gamma'\Gamma X)^{-1}X'\Gamma'\Gamma y$, assuming that $\Gamma'\Gamma$ is nonsingular and X is full rank. Once the BPE is obtained, the OBP of θ (Jiang *et al.* 2011), is given by the right side of (1) with β replaced by $\tilde{\beta}$. On the other hand, the BLUP of θ is given by the right side of (1) with β replaced by $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$, which is the MLE of β. To compare the predictive performance of the OBP and BLUP, let us consider a class of empirical best predictors (EBPs) that can be expressed as

$$\check{\theta} \quad = \quad F'y - \Gamma(y - X\check{\beta}), \tag{2}$$

where $\check{\beta}$ is a weighted least squares (WLS) estimator of β expressed as $\check{\beta} = (X'WX)^{-1}X'Wy$ and W is a positive definite weighting matrix. Note that (2) is the BP (1) with β replaced by $\check{\beta}$ (and hence explains the name EBP). Also note that the BPE and MLE are special cases of the WLS, hence the OBP and BLUP are special cases of the EBP.

2.1. Special Case 1: The Fay-Herriott Model

The Fay-Herriot model (Fay and Herriot 1979) is widely used in small area estimation (SAE). It was proposed to estimate the per-capita income of small places with population size less than 1,000. The model can be expressed as a mixed effects model:

$$y_i \quad = \quad x_i'\beta + v_i + e_i, \quad i = 1, \ldots, m, \tag{3}$$

where x_i is a vector of known covariates, β is a vector of unknown regression coefficients, v_i's are area-specific random effects and e_i's are sampling errors. It is assumed that the v_i's and e_i's are independent with $v_i \sim N(0, A)$ and $e_i \sim N(0, D_i)$. The variance A is unknown, but the sampling variances D_i's are assumed known. The problem of interest is estimation of the small area means, which, at least by a higher-order approximation, are equal to the mixed effects $\theta_i = x_i'\beta + v_i, i = 1, \dots, m$. Thus, without loss of generality, we treat the θ_i's as the small area means, so the problem is prediction of the mixed effects.

One should note that the true small area means should not depend on the assumed model. In fact, it is easy to see that under the weak assumption

$$y_i = \mu_i + v_i + e_i, \ \text{ we have } \ \theta_i = \mathrm{E}(y_i|v_i) = \mathrm{E}(y_i) + v_i, \quad i = 1, \dots, m, \tag{4}$$

where $\mu_i = \mathrm{E}(y_i)$. The advantage of expressions (4) is that they are not model-dependent, which is a key to our approach. Here, E denotes the expectation with respect to the true distribution of y_i, which may be unknown, but not model-dependent. The most popular precision measure of a predictor is its mean squared prediction error (MSPE; *e.g.*, Prasad & Rao 1990, Das *et al.* 2004). Write $\theta = (\theta_i)_{1 \le i \le m}$ and let $\tilde{\theta} = (\tilde{\theta}_i)_{1 \le i \le m}$ be a predictor of θ. Then, the (overall) MSPE of $\tilde{\theta}$ is given by

$$\mathrm{MSPE}(\tilde{\theta}) = \mathrm{E}(|\tilde{\theta} - \theta|^2) = \sum_{i=1}^{m} \mathrm{E}(\tilde{\theta}_i - \theta_i)^2. \tag{5}$$

Once again, the expectation in (5) is with respect to the true underlying distribution (of whatever random quantities that are involved), which is unknown but <u>not</u> model-dependent. Under the assumed Fay-Herriot model, and given the parameters $\psi = (\beta', A)'$, the BP is given by

$$\tilde{\theta}(\psi) \ = \ \mathrm{E}_{\mathrm{m},\psi}(\theta|y) \ = \ \left[x_i'\beta + \frac{A}{A + D_i}(y_i - x_i'\beta)\right]_{1 \le i \le m}, \tag{6}$$

or $\tilde{\theta}(\psi)_i = x_i'\beta + B_i(y_i - x_i'\beta), 1 \le i \le m$, where $B_i = A/(A + D_i)$, and $\mathrm{E}_{\mathrm{m},\psi}$ represents (conditional) expectation under the assumed model with ψ being the true parameter vector. Note that $\mathrm{E}_{\mathrm{m},\psi}$ is different from E unless the model is correct, and ψ is the true parameter vector. For simplicity, let us assume, for now, that A is known. Then, the precision of $\tilde{\theta}(\psi)$, which is now denoted by $\tilde{\theta}(\beta)$ because A is no longer a parameter, is measured by

$$\mathrm{MSPE}\{\tilde{\theta}(\beta)\} = \sum_{i=1}^{m} \mathrm{E}\{B_i y_i - \theta_i + x_i'\beta(1 - B_i)\}^2 = I_1 + 2I_2 + I_3, \tag{7}$$

where $I_1 = \sum_{i=1}^m \mathrm{E}(B_i y_i - \theta_i)^2$, $I_2 = \sum_{i=1}^m x_i'\beta(1 - B_i)\mathrm{E}(B_i y_i - \theta_i)$, and $I_3 = \sum_{i=1}^m (x_i'\beta)^2(1 - B_i)^2$. Note that I_1 does not depend on β. As for I_2, we have $\mathrm{E}(B_i y_i - \theta_i) = (B_i - 1)\mathrm{E}(y_i)$. Thus, we have $I_2 = -\sum_{i=1}^m (1 - B_i)^2 x_i'\beta \mathrm{E}(y_i)$. It follows that the left side of (7) can be expressed as

$$\mathrm{MSPE}\{\tilde{\theta}(\beta)\} = \mathrm{E}\left\{I_1 + \sum_{i=1}^{m}(1 - B_i)^2(x_i'\beta)^2 - 2\sum_{i=1}^{m}(1 - B_i)^2 x_i'\beta y_i\right\}. \tag{8}$$

The right side of (8) suggests a natural estimator of β, by minimizing the expression inside the expectation, which is equivalent to minimizing $Q(\beta) = \sum_{i=1}^m (1 - B_i)^2(x_i'\beta)^2 - 2\sum_{i=1}^m (1 -$

$B_i)^2 x_i'\beta y_i = \beta' X'\Gamma^2 X\beta - 2y'\Gamma^2 X\beta$, where $X = (x_i')_{1\le i\le m}$, $y = (y_i)_{1\le i\le m}$ and $\Gamma = \text{diag}(1 - B_i, 1 \le i \le m)$. A closed-form solution is given by

$$\tilde{\beta} = (X'\Gamma^2 X)^{-1} X'\Gamma^2 y = \left\{\sum_{i=1}^{m} (1 - B_i)^2 x_i x_i'\right\}^{-1} \sum_{i=1}^{m} (1 - B_i)^2 x_i y_i. \tag{9}$$

Here we assume, without loss of generality, that X is of full column Note that $\tilde{\beta}$ is different from the MLE of β,

$$\hat{\beta} = (X'V^{-1}X)^{-1} X'V^{-1}y = \left(\sum_{i=1}^{m} \frac{x_i x_i'}{A + D_i}\right)^{-1} \sum_{i=1}^{m} \frac{x_i y_i}{A + D_i}, \tag{10}$$

where $V = \text{diag}(A + D_i, 1 \le i \le m) = \text{Var}(y)$. While $\hat{\beta}$ maximizes the likelihood function, $\tilde{\beta}$ minimizes the "observed" MSPE which is the expression inside the expectation on the right side of (8). We call $\tilde{\beta}$ given by (9) the *best predictive estimator*, or BPE, of β. Note that the BPE has the property that its expected value,

$$\text{E}(\tilde{\beta}) \;=\; (X'\Gamma^2 X)^{-1} X'\Gamma^2 \text{E}(y), \tag{11}$$

is the β that minimizes $\text{MSPE}\{\tilde{\theta}(\beta)\}$. However, the expression (11) is not computable.

A predictor of the mixed effects θ is then obtained by replacing β in the BP (6) by its BPE. We call this predictor the *observed best predictor*, or OBP. The reason is that the BPE is the minimizer of the observed MSPE. If the observed MSPE were the true MSPE, the BPE would give us the BP. However, because, instead, we are dealing with the observed MSPE, the corresponding predictor (obtained by the same procedure with the MSPE replaced by the observed MSPE) should be called the observed BP.

2.2. Special Case 2: The nested error regression model

Consider sampling from finite subpopulations $P_i = \{Y_{ik}, k = 1, \ldots, N_i\}, i = 1, \ldots, m$. Suppose that auxiliary data $X_{ikl}, k = 1, \ldots, N_i, l = 1, \ldots, p$ are available for each P_i, and a super-population nested-error regression model (Battese *et al.* 1988) hold for all subpopulations:

$$Y_{ik} \;=\; X_{ik}'\beta + v_i + e_{ik}, \quad k = 1, \ldots, N_i, \tag{12}$$

where $X_{ik} = (X_{ikl})_{1\le l\le p}$, the v_i's are small-area specific random effects, and e_{ik}'s are additional errors, such that the random effects and errors are independent with $v_i \sim N(0, \sigma_v^2)$ and $e_{ik} \sim N(0, \sigma_e^2)$. The small area mean for P_i is then $\mu_i = N_i^{-1}\sum_{k=1}^{N_i} Y_{ik}$.

Now suppose that $y_{ij}, j = 1, \ldots, n_i$ are observed for the ith subpopulation, $i = 1, \ldots, m$. Let the corresponding auxiliary data be $x_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, m$. Write $y_i = (y_{ij})_{1\le j\le n_i}$, $y = (y_i)_{1\le i\le m}$, $\bar{y}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$ and $\bar{x}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} x_{ij}$. Let $\psi = (\beta', \sigma_v^2, \sigma_e^2)'$ denote the vector of parameters under the nested-error regression model (12). Under this model with ψ being the true parameter vector, the BP for μ_i is

$$\begin{aligned}\tilde{\mu}_i(\psi) \;&=\; \text{E}_{\text{m},\psi}(\mu_i|y) = \frac{1}{N_i}\left\{\sum_{j=1}^{n_i} y_{ij} + \sum_{k\notin I_i} \text{E}_{\text{m},\psi}(Y_{i,k}|y_i)\right\} \\ &=\; \bar{X}_i'\beta + \left\{\frac{n_i}{N_i} + \left(1 - \frac{n_i}{N_i}\right)\frac{n_i\sigma_v^2}{\sigma_e^2 + n_i\sigma_v^2}\right\}(\bar{y}_{i\cdot} - \bar{x}_{i\cdot}'\beta),\end{aligned} \tag{13}$$

where $\mathrm{E}_{\mathrm{m},\psi}$ denotes the model-based conditional expectation given that ψ is the true parameter vector, I_i is the set of sampled indexes such that Y_{ik} is in the sample iff $k \in I_i$, and $\bar{X}_i = N_i^{-1}\sum_{k=1}^{N_i} X_{ik}$ is the subpopulation mean of the X_{ik}'s for the ith subpopulation (which is known). Note that (13) is a model-based BP. The performance of the model-based BP is then evaluated by the design-based MSPE. This is because the latter is almost free of model assumptions, and therefore robust to model misspecifications [we could not do this under the Fay-Herriot model, however, because the sampling data were not available at the unit level; instead, we considered a model under very weak assumptions]. The design-based MSPE is given by $\mathrm{MSPE}\{\tilde{\mu}(\psi)\} = \mathrm{E}_{\mathrm{d}}\{|\tilde{\mu}(\psi) - \mu|^2\} \quad = \quad \sum_{i=1}^{m} \mathrm{E}_{\mathrm{d}}\{\tilde{\mu}_i(\psi) - \mu_i\}^2$, where $\tilde{\mu}(\psi) = [\tilde{\mu}_i(\psi)]_{1\leq i\leq m}$, $\mu = (\mu_i)_{1\leq i\leq m}$ and E_{d} denotes the design-based expectation. Assume, for simplicity, simple random sampling within each subpopulation P_i. Then, it can be shown that

$$\mathrm{MSPE} \;=\; \mathrm{E}_{\mathrm{d}}\left[\sum_{i=1}^{m}\{\tilde{\mu}_i^2(\psi) - 2a_i(\sigma_v^2,\sigma_e^2)\bar{X}_i'\beta\bar{y}_{i\cdot} + b_i(\sigma_v^2,\sigma_e^2)\hat{\mu}_i^2\}\right], \tag{14}$$

where $a_i(\sigma_v^2,\sigma_e^2) = (1-n_i/N_i)\sigma_e^2/(\sigma_e^2+n_i\sigma_v^2)$ and $b_i(\sigma_v^2,\sigma_e^2) = 1-2[n_i/N_i+(1-n_i/N_i)\{n_i\sigma_v^2/(\sigma_e^2+n_i\sigma_v^2)\}]$. Thus, the BPE of ψ is obtained by minimizing $Q(\psi) = \sum_{i=1}^{m}\{\tilde{\mu}_i^2(\psi) - 2a_i(\sigma_v^2,\sigma_e^2)\bar{X}_i'\beta\bar{y}_{i\cdot} + b_i(\sigma_v^2,\sigma_e^2)\hat{\mu}_i^2\}$, which is the expression inside the expectation in (14). Here $\hat{\mu}_i^2$ is a design-based unbiased estimator of μ_i^2, given by $\hat{\mu}_i^2 = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}^2 - (N_i-1)\{N_i(n_i-1)\}^{-1}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$. A similar numerical procedure can be developed to compute the BPE. Given the BPE $\tilde{\psi} = (\tilde{\beta}', \tilde{\sigma}_v^2, \tilde{\sigma}_e^2)'$, the OBP of μ_i is given by $\tilde{\mu}_i = \tilde{\mu}_i(\tilde{\psi})$, $1 \leq i \leq m$, where $\tilde{\mu}_i(\psi)$ is given by (13).

2.3. Estimation of MSPE of OBP

Obtaining a measure of uncertainty for OBP is particularly challenging. This is because the OBP is derived by taking into account of the potential model misspecification. Therefore, to derive a measure of uncertainty, the potential model misspecification also needs to be taken into consideration when considering measures of uncertainty. More importantly, it is desirable to evaluate uncertainty due to the potential model misspecification.

A standard measure of uncertainty is the MSPE. Let us first consider this under a Fay-Herriot model. In proposing the OBP, Jiang *et al.* (2011) also proposed an MSPE estimator under potential model misspecification, which we call the JNR estimator in the sequel. The authors showed that the JNR estimator is second-order unbiased, that is, its bias is $o(m^{-1})$. However, the estimator is known to have large variation. To see why, note that the leading term of the JNR estimator has the expression

$$(\hat{\theta}_i - y_i)^2 + D_i(2\hat{B}_i - 1), \tag{15}$$

where $\hat{\theta}_i$ is the OBP of θ_i, $B_i = A/(A + D_i$, and $\hat{B}_i$ is B_i with A replaced by $\hat{A}$, the BPE of A. The direct estimator, y_i, is involved in (15), which has large variation compared to, for example, $\hat{A}$. The latter is an estimator based on all of the data, $y_i, x_i, 1 \leq i \leq m$, which has relatively small variation. In particular, the JNR estimator has a significantly nonzero chance of taking negative values.

In addition to the JNR estimator, Jiang *et al.* (2011) also proposed a bootstrap MSPE estimator. Although the bootstrap estimator is gauranteed non-negative, its bias was shown to be significantly larger than the JNR estimator. The method also seemed lack of theoretical justification, in which the bootstrap samples were drown independently under the model $y_i^* \sim \hat{\theta}_i + e_i^*$, where $\hat{\theta}_i$ is the OBP and $e_i^* \sim N(0, D_i)$, $1 \leq i \leq m$.

Liu *et al.* (2022a) proposed a OBOR estimator for the MSPE of OBP. Here, OBOR is an abbreviation of one-bring-one-route. It is called OBOR because the estimator consists of averages of terms, where each term involves y_i, plus one other y_j for $j \neq i$. The average is over $m - 1$ such y_j's for $j \neq i$. The idea can be generalized to one-bring-two, one-bring-three, etc., but the computational burden mounts as this moves on. In this regard, the JNR estimator may also be viewed as a special case of one-bring-none. Although the OBOR estimator reduces the variation over the JNR estimator, the result was not all satisfactory, compared to a much better estimator found later.

A well-known method for obtaining a second-order unbiased MSPE estimator is the Prasad-Rao (PR) linearization method (Prasad and Rao 1990). The method is developed under the assumption that the underlying model is correct. In fact, the assumed model is substantially used in the derivation of the P-R MSPE estimator. Given that, it would be surprising to learn that, in spite of the model misspecification, the PR MSPE estimator for OBP is, still, mostly correct. In fact, Liu *et al.* (2022b) found that the PR MSPE estimator remains first-order unbiased in the sense that the bias of the estimator is $O(m^{-1})$, even if the underlying model is misspecified in its mean function. Furthermore, the same authors showed the PR MSPE estimator can be modified to achieve the second-order unbiasedness, again under the potential model misspecification in the mean function.

3. Classified mixed model prediction (CMMP)

The world has been witnessing an information explosion in many areas of society from medicine to economics and business to social media for instance. The rapid increase in the unprecedented amount of data has resulted in many new important shifts of interest in the types of questions that can be potentially answered. These new shifts are focusing more and more attention on knowledge at individual or subject levels. One of the currently "hot" areas is *precision medicine*. The National Research Council of the United States in 2014 defined the latter as the "ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those disease they may develop, or in their response to a specific treatment. Preventive or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not". Another area, in economic studies, is *family economics*, which applies basic economic concepts to families, which are viewed as (small) firms or companies. For example, China Household Finance Survey, the largest non-governmental household panel survey since 2009, has so far collected massive financial and economical data at household level. The latest wave, conducted in the summer of 2017, had more than 40,000 nationally and provincially representative households. More than 10,000 registered users worldwide are using the data for their studies about China. In particular, the data provide important information about household finance, which is a driving force of China's national economy (*e.g.*, Zhang *et al.* 2014, Gan, Yin and Tan 2016).

The target of classical statistical inference is a (large) population, from which data are collected, and to be used to make inference about the same population parameters, such as the mean, proportion, and regression coefficients. However, in each of the above subject matter disciplines, the primary interest is inference at the subject levels. For example, in precision medicine, the subject may be a patient, or small group of patients sharing similar characteristics; in family economics, the subject is typically a family, whose definition may vary depending on factors such as culture or interest.

Nevertheless, it should be noted that making inference about a specific subject does not mean that the inference is based only on data collected from the subject, which is key. The idea can be best explained using a mixed effects model (MEM; *e.g.*, Jiang 2007). There are fixed effects and random effects in a MEM. The fixed effects are parameters that are common for all of the subjects; the random effects are typically subject-specific. The fixed effects are estimated using all of the data, that is, combining all subjects, but what about the random effects? This question has direct implications on how one predicts accurately from such models (*e.g.* Welham *et al.* 2004). In mixed model prediction (MMP), the best predictor for a characteristic of interest associated with a specific subject is derived, which depends on the subject-specific data as well as the fixed effects and variance components, which are population parameters. Thus, through MMP, inference about the subject-specific characteristic borrows strength from data from other subjects, as well as information from other sources.

In a significant recent development toward potentially much broader, modern-time applications, Jiang *et al.* (2018) proposed a method called classified mixed model prediction (CMMP) for two types of prediction problems - predicting the mixed effect associated with a set of new observations and predicting future values associated with new sets of covariates. The basic idea is to create a "match" between a group or cluster in the population, for which one wishes to make prediction, and a (massive) training data, with known groups or clusters. Once such a match is built, the traditional MMP method can be utilized to make accurate predictions. Even more interestingly, it can handle the situation where a real match may not exist.

To illustrate the CMMP method, let us focus on prediction of a mixed effect associated with new observations. Suppose that we have a set of training data, $y_{ij}, i = 1, \dots, m, j = 1, \dots, n_i$ in the sense that their classifications are known, that is, one knows which group, i, that y_{ij} belongs to. The assumed model is a linear mixed model (LMM; *e.g.*, Jiang 2007):

$$y_i = \mathrm{E}(y_i|\alpha) + \epsilon_i = X_i\beta + Z_i\alpha_i + \epsilon_i, \tag{16}$$

where $y_i = (y_{ij})_{1\le j\le n_i}$, $X_i = (x'_{ij})_{1\le j\le n_i}$ is a matrix of known covariates, β is a vector of unknown regression coefficients (the fixed effects), Z_i is a known $n_i \times q$ matrix, α_i is a $q \times 1$ vector of group-specific random effects, ϵ_i is an $n_i \times 1$ vector of errors, and $\alpha = (\alpha_i)_{1\le i\le m}$. It is assumed that the α_i's and ϵ_i's are independent, with $\alpha_i \sim N(0, G)$ and $\epsilon_i \sim N(0, R_i)$, where the covariance matrices G and R_i depend on a vector ψ of dispersion parameters, or variance components. Our goal is to make a classified prediction for a mixed effect associated with a new observation, y_{n}. Suppose that

$$y_{\mathrm{n}} = \mathrm{E}(y_{\mathrm{n}}|\alpha) + \epsilon_{\mathrm{n}} = x'_{\mathrm{n}}\beta + z'_{\mathrm{n}}\alpha_{\mathrm{I}} + \epsilon_{\mathrm{n}}, \tag{17}$$

where $x_{\mathrm{n}}, z_{\mathrm{n}}$ are known vectors, $I \in \{1, \dots, m\}$ but one does not know which element i,

$1 \leq i \leq m$, is equal to I. Furthermore, $\epsilon_{\rm n}$ is a new error that is independent of $y_i, 1 \leq i \leq m$, and has mean zero. The mixed effect of interest is $\theta = {\rm E}(y_{\rm n}|\alpha) = y_{\rm n} - \epsilon_{\rm n}$. From the training data, one can estimate the parameters, β and ψ. Thus, we can assume that estimators $\hat{\beta}, \hat{\psi}$ are available for β, ψ, respectively. Suppose that $I = i$. Then, it can be shown that ${\rm E}(\theta|y_1, \ldots, y_m) = {\rm E}(\theta|y_i)$ and, according to the normal theory,

$${\rm E}(\theta|y_i) = x_{\rm n}'\beta + z_{\rm n}'GZ_i'(R_i + Z_iGZ_i')^{-1}(y_i - X_i\beta). \tag{18}$$

The right side of (18) is the BP under the assumed LMM, if the true parameters, β and ψ, are known. Because the latter are unknown, we replace them by $\hat{\beta}$ and $\hat{\psi}$, respectively. The result is called an empirical best predictor (EBP), noted by $\tilde{\theta}_{(i)}$. In practice, however, I is unknown. In order to identify I, we consider the MSPE of predicting θ by the BP when I is classified as i, that is ${\rm MSPE}_i = {\rm E}\{\tilde{\theta}_{(i)} - \theta\}^2 = {\rm E}\{\tilde{\theta}_{(i)}^2\} - 2{\rm E}\{\tilde{\theta}_{(i)}\theta\} + {\rm E}(\theta^2)$. Using the expression $\theta = y_{\rm n} - \epsilon_{\rm n}$, we have ${\rm E}\{\tilde{\theta}_{(i)}\theta\} = {\rm E}\{\tilde{\theta}_{(i)}y_{\rm n}\} - {\rm E}\{\tilde{\theta}_{(i)}\epsilon_{\rm n}\} = {\rm E}\{\tilde{\theta}_{(i)}y_{\rm n}\}$. Thus, we have

$${\rm MSPE}_i = {\rm E}\{\tilde{\theta}_{(i)}^2 - 2\tilde{\theta}_{(i)}y_{\rm n} + \theta^2\}. \tag{19}$$

Note that the E in (19) denotes the true expectation, which may be unknown; nevertheless, the observed MSPE corresponding to (4) is the expression inside the expectation. Therefore, a natural idea is to identify I as the index i that minimizes the observed MSPE. Because θ^2 does not depend on i, this is equivalent to

$$I = {\rm argmin}_i\{\tilde{\theta}_{(i)}^2 - 2\tilde{\theta}_{(i)}y_{\rm n}\}. \tag{20}$$

Denote the I identified by (20) by $\hat{I}$. Then, the classified mixed-effect predictor (CMMP) of θ is given by $\hat{\theta} = \tilde{\theta}_{(\hat{I})}$.

The basic idea of CMMP has been extended to multiple new observations from the same, unknown group, and to prediction of a future observation. See Jiang *et al.* (2018) for details. An important concept being exploited by CMMP is the idea that it *captures what is not captured by the fixed effect (the uncaptured)* through the classified random effect. It is important to note that the primary interest is not to identify the correct “match”, I. In fact, in many applications such a match may not exist, that is, there is no group among the training that matches exactly the group corresponding to the new observations. Even if the exact match does exist, as the number of training data groups, m, increases, the probability of identifying the correct group, that is, ${\rm P}(\hat{I} = I)$, goes to zero (as opposed to going to one, as one might expect). But, regardless, the CMMP of the mixed effect, θ, is consistent (in fact, converges in L^2 to the true mixed effect), which is all we care about. For example, it was demonstrated that CMMP significantly outperform the traditional regression prediction whether or not the true match exists. The rationale behind the mismatch-led-correct-prediction is because, as m increases, the difference between different groups becomes smaller and smaller; thus, even though there is no exact match, there is a “close match” between one of the training data groups and the new group, of which CMMP is able to take advantage. This important, and interesting, feature makes the CMMP idea practically more attractive because, in practice, an exact match may not exist but a close resemblance may well be expected.

Following the initial work of CMMP, Sun *et al.* (2018) extended the idea to classified prediction of mixed effects associated with binary outcomes, such as conditional probabilities associated with a group of new observations, and demonstrated similar properties to CMMP for the resulting classified predictor. A number of recent results have been derived to extend the idea of CMMP to topics like functional data analysis (Xiu & Jiang (2024)) and a psuedo-Bayesian version of CMMP (Ma & Jiang (2023)).

3.1. Estimation of MSPE for CMMP

A standard uncertainty measure for a predictor is the MSPE. A "gold standard" for the MSPE estimation is to produce a second-order unbiased MSPE estimator, that is, the order of bias of the MSPE estimator is $o(m^{-1})$, where m is the total number of clusters in the training data. Typically, the $o(m^{-1})$ term is, in fact, $O(m^{-2})$, but this difference is usually ignored. For the most part, there have been two approaches for producing a second-order unbiased MSPE estimator. The first is the Prasad-Rao linearization method (Prasad & Rao 1990). The approach uses Taylor series expansion to obtain a second-order approximation to the MSPE, then corrects the bias, again to the second-order, to produce an MSPE estimator whose bias is $o(m^{-1})$. Various extensions of the Prasad-Rao method have been developed; see, for example, Datta & Lahiri (2000), Jiang & Lahiri (2001), Das, Jiang & Rao (2004), and Datta, Rao & Smith (2005). Although the method often leads to an analytic expression of the MSPE estimator, the derivation is tedious, and the final expression is likely to be complicated. More importantly, errors often occur in the process of analytic derivations as well as computer programming based on the lengthy expressions. Furthermore, the linearization method does not apply to situations where a non-differentiable operation is involved in obtaining the predictor, such as shrinkage estimation (*e.g.*, Tibshirani 1996), CMMP (Jiang *et al.* 2018), as well as the CMMP described in what follows in the next section.

The second approach to second-order unbiased MSPE estimation is resampling methods. Jiang, Lahiri & Wan (2002; hereafter JLW) proposed a jackknife method to estimate the MSPE of an empirical best predictor (EBP). The method avoids tedious derivations of the Prasad-Rao method, and is "one formula for all". On the other hand, there are restrictions on the class of predictors to which JLW applies. Namely, JLW only applies to empirical best predictor (EBP), that is, predictor obtained by replacing the parameters involved in the best predictor (BP), which is the conditional expectation, by their (consistent) estimators. The CMMP predictor, however, is not an EBP, because it involves a matching process. Jiang, Lahiri & Nguyen (2018) proposed a Monte-Carlo jackknife method, called McJack, which potentially applies to CMMP; however, the method is computationally very expensive. Another resampling-based approach is double bootstrapping (DB; Hall & Maiti 2006a,b). Although DB is capable of producing a second-order unbiased MSPE estimator, it is, perhaps, computationally even more intensive than the McJack. It is also unclear whether DB can be extended to CMMP.

In a way, the method to be proposed below may be viewed as a hybrid of the linearization method and resampling method, by combining the best part of each method. In short, we use a simple, analytic approach to obtain the leading term of our MSPE estimator, and a Monte-Carlo method to take care a remaining, lower-order term. The computational cost for the Monte-Carlo part is much lesser compared to McJack. For example, the computational burden of our method is about $1/m^3$ to $1/m^2$ of that for McJack. More importantly, the method provides a unified, conceptually easy solution to a difficult problem, that is, obtaining a second-order unbiased MSPE estimator for CMMP.

Let θ be the mixed effect corresponding to the new observations, and $\hat{\theta}$ the CMMP predictor of θ. The MSPE of $\hat{\theta}$ can be expressed as MSPE $= \mathrm{E}(\hat{\theta}-\theta)^2 = \mathrm{E}\left[\mathrm{E}\{(\hat{\theta}-\theta)^2|y\}\right]$, where y represents the available data. Suppose that the underlying distribution of y depends on a vector of unknown parameters, ϕ. Then, the conditional expectation inside the expectation is a function of y and ϕ, which can be written as $a(y,\phi) = \mathrm{E}\{(\hat{\theta}-\theta)^2|y\} = \hat{\theta}^2 - 2\hat{\theta}\mathrm{E}(\theta|y) + \mathrm{E}(\theta^2|y) = \hat{\theta}^2 - 2\hat{\theta}a_1(y,\phi) + a_2(y,\phi)$, where $a_j(y,\phi) = \mathrm{E}(\theta^j|y), j = 1,2$. If we replace the ϕ in $a(y,\psi)$ by $\hat{\phi}$, a consistent estimator of ϕ, the result is a first-order unbiased estimator, that is, we have $\mathrm{E}\{a(y,\hat{\phi}) - a(y,\phi)\} = O(m^{-1})$. On the other hand, both MSPE $= \mathrm{E}\{a(y,\phi)\}$ and $\mathrm{E}\{a(y,\hat{\phi})\}$ are functions of ϕ, denoted by $b(\phi)$ and $c(\phi)$, respectively. It follows that $d(\phi) = b(\phi) - c(\phi) = O(m^{-1})$; thus, if we replace, again, to replace ϕ by $\hat{\phi}$ in $d(\phi)$, the difference is a lower-order term, that is, $d(\hat{\phi}) - d(\phi) = o_{\mathrm{P}}(m^{-1})$ [see, *e.g.*, Jiang 2010, sec. 3.4 for notation like o_{P} and O_{P}]. Now consider the estimator

$$\widehat{\mathrm{MSPE}} = a(y,\hat{\phi}) + d(\hat{\phi}) = a(y,\hat{\phi}) + b(\hat{\phi}) - c(\hat{\phi}). \tag{21}$$

We have $\mathrm{E}(\widehat{\mathrm{MSPE}}) = \mathrm{E}\{a(y,\phi)\} + \mathrm{E}\{a(y,\hat{\phi}) - a(y,\phi)\} + \mathrm{E}\{d(\hat{\phi})\} = \mathrm{MSPE} + \mathrm{E}\{d(\hat{\phi}) - d(\phi)\} = \mathrm{MSPE} + o(m^{-1})$. Essentially, this one-line, heuristic derivation shows the second-order unbiasedness of the proposed MSPE estimator, (21), provided that the terms involved can be evaluated.

Note that the leading term, $a(y,\hat{\phi})$, in (21) is guaranteed positive, a desirable property for an MSPE estimator. The lower-order term, $b(\hat{\phi}) - c(\hat{\phi})$, corresponds to a bias correction to the leading term. This term is typically much more difficult to evaluate than the leading term. We propose to approximate this term using a Monte-Carlo method. Let P_ϕ denote the distribution of y with ϕ being the true parameter vector. Given ϕ, one can generate y under P_ϕ. Let $y_{[k]}$ denote y generated under the kth Monte-Carlo sample, $k = 1,\ldots,K$. Then, by the law of large numbers, we have $b(\phi) - c(\phi) \approx K^{-1}\sum_{k=1}^{K}\left\{a(y_{[k]},\phi) - a(y_{[k]},\hat{\phi}_{[k]})\right\} \equiv d_K(\phi)$, where $\hat{\phi}_{[k]}$ denotes $\hat{\phi}$ based on $y_{[k]}$. If K is sufficiently large, which one has control over during the Monte-Carlo simulation, the difference between the two sides of the approximation is $o(m^{-1})$. Note that $y_{[k]}, k = 1,\ldots,K$ also depend on ϕ. Then, a Monte-Carlo assisted MSPE estimator (Nguyuen *et al.* 2022), is given by

$$\widehat{\mathrm{MSPE}}_K = a(y,\hat{\phi}) + d_K(\hat{\phi}) = a(y,\hat{\phi}) + K^{-1}\sum_{k=1}^{K}\left\{a(y_{[k]},\hat{\phi}) - a(y_{[k]},\hat{\phi}_{[k]})\right\} \tag{22}$$

where $y_{[k]}, k = 1,\ldots,K$ are generated as above with $\phi = \hat{\phi}$, and $\hat{\phi}_{[k]}$ is, again, the estimator of ϕ based on $y_{[k]}$. (22) is called the Sumca estimator of the MSPE of $\hat{\theta}$ (Sumca is abbreviation of "simple, unified, Monte-Caro assisted").

4. Classified mixed model projections

In many practical problems, there is interest in the estimation of mixed effect projections for new data that are outside the range of the training data. Examples include predicting extreme small area means for rare populations or making treatment decisions for patients who do not fit typical risk profiles. Standard methods have long been known to struggle with such problems since the training data may not provide enough information about potential model changes for these new data values (extrapolation bias). Rao *et al.* (2024) proposed a new framework called Prediction Using Random-effect Extrapolation (PURE) which involves constructing a generalized independent variable hull (gIVH) to isolate a minority training set which is "close" to the prediction space, followed by a regrouping of the minority data according to the response variable which results in a new (but misspecified) random effect distribution. This misspecification reflects "extrapolated random effects" which prove vital to capture information that is needed for accurate model projections. Projections were then made using classified mixed model prediction (CMMP) (Jiang et al. 2018) with the regrouped minority data. Let us assume that, for $i = 1, \dots, m$, $\boldsymbol{y}^{(k)}$ follow a mixed model as follows:

$$\boldsymbol{y}_i^{(k)} = X_i^{(k)}\boldsymbol{\beta}_k + Z_i^{(k)}\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{23}$$

where $\boldsymbol{y}_i^{(k)} = (y_{ij})_{1\le j\le n_i^{(k)}}$, $X_i^{(k)} = (x_{ij}^{(k)})^T_{1\le j\le n_i^{(k)}}$ is a matrix of known covariates, $Z_i^{(k)}$ is a matrix of known covariates, $\boldsymbol{\beta}_k$ is a p-vector of unknown regression coefficients (the fixed effects), $\boldsymbol{b}_i$ is q-vector of group-specific random effects, and $\boldsymbol{\varepsilon}_i$ is an vector of errors. *Notice the different notation for the random effects from the previous CMMP in order to distinguish the two methods.*

The subscript (k) denotes the population k, and $1 \le k \le K$. It is assumed that $\boldsymbol{b}_i \sim N(0, G)$, $\boldsymbol{\varepsilon}_i \sim N(0, R_i)$ and they are independent, and the covariance matrices G and R_i depend on a vector $\boldsymbol{\psi}$ of variance components. Note $\boldsymbol{\beta}_k$ is different for different population k, and the random effects $\boldsymbol{b}_i$ are the same across k populations. The total number of observations in each population is $n^{(k)} = \sum_{i=1}^m n_i^{(k)}$, and the overall total population $n = \sum_{i=1}^K n^{(k)}$. Note that $n = \sum_{i=1}^m n_i$ where n_i is the number of observations in the group i. If the data follows (23), people usually fit a one component mixed model that assumed only one set of fixed effects parameters when the true model information is unknown, which results in a convenient but "misspecified" model fit.

Assume new test observations, which follow:

$$y_{n,j} = x_n'\boldsymbol{\beta}_n + z_n'\boldsymbol{b}_I + \varepsilon_{n,j}, \quad 1 \le j \le n_{new}, \tag{24}$$

where x_n and z_n are known vectors, and I belongs to one of the m groups. The new errors $\varepsilon_{n,j}$ are independent with mean zero, and variance R_{new} and are assumed independent of the training data. Notice $\boldsymbol{\beta}_n \ne \boldsymbol{\beta}_k, 1 \le k \le K$. The mixed effect we wish to predict is $\theta_n = E(y_{n,j} \mid b_I) = x_n'\boldsymbol{\beta} + z_n'\boldsymbol{b}_I$ where $I \in \{1, \dots, m\}$ but we do not know which group I belongs to.

4.1. Generalized independent variable hull

Conn et al. (2015) proposed one possible definition of "the range of observation data" which turns to early works on outlier detection in simple linear regression analysis. Cook

(1979) referred that the smallest convex set containing all design points of a full-rank linear regression model as the independent variable hull (IVH). The IVH definition is based on linear model which require full rank of design matrix and i.i.d Gaussian error. Therefore, it can not be applied to generalized models such as binary response or random effects. Cook (1979) notes that design points with maximum prediction variance will be located on the boundary of IVH, then Conn et al. (2015) defined a generalized independent variable hull (gIVH) as a set of all predicted locations $\boldsymbol{S}_0$ for which

$$\mathrm{var}(\lambda_i) \leq \max(\mathrm{var}(\boldsymbol{\lambda}_{\boldsymbol{S}})), \tag{25}$$

where $i \in \boldsymbol{S}_0$, λ_i corresponds to the mean prediction at i, $\boldsymbol{S}$ denoted the set of locations where data are observed, and $\boldsymbol{\lambda}_{\boldsymbol{S}}$ denotes predictions at $\boldsymbol{S}$. Conn et al. (2015) proposed that the gIVH can be applied to determine whether predictions are interpolations (predictive design points lying inside the gIVH) or extrapolations (predictive design points lying outside the gIVH). This uses the generalization,

$$\boldsymbol{\mu} = X_{aug}\boldsymbol{\beta}_{aug}, \tag{26}$$

where X_{aug} is an augmented design matrix to accommodate the random effects design matrix Z and $\boldsymbol{\beta}_{aug}$ is the corresponding regression parameter vector. We can then write the prediction variance as,

$$\mathrm{var}(\hat{\boldsymbol{\lambda}}) = \mathrm{var}(\hat{\boldsymbol{\mu}}) = X_{aug}\mathrm{var}(\hat{\boldsymbol{\beta}}_{aug})X'_{aug}. \tag{27}$$

One possibility is to use a flexible generalized additive model (GAM) (Hastie and Tibshirani, 1990) and then estimate the appropriate form of var($\hat{\boldsymbol{\beta}}_{aug}$). If y is not on the linear predictor scale (*e.g.* generalized linear models outside of the normal model), then the delta method can be used to estimate var($\hat{\boldsymbol{\lambda}}$) (Conn et al. 2015). Outside of these situations, simulation based methods like bootstrapping can be used to estimate the variance.

4.2. Prediction Using Random-effects Extrapolation (PURE)

Suppose we have a set of training data and test data as in (23) and (24). Let π_k denote the percentage of the population that comes from the population k, and $\sum_{k=1}^{K} \pi_k = 1$. If $K = 2$, we have π_1 percent of the population comes from the minority and the rest $1 - \pi_1$ population comes from the majority. We define the following relevant features:

1. Extreme data: This is the test dataset which may or may not be outside of range of the training data. Both cases can be handled here.
2. Majority data: Notationally, we can concatenate all observations in the full training data as $\mathscr{L} = \{(x_l, y_l); l = 1, \ldots, (n_1 + n_2 + \ldots + n_m)\}$. Then define the majority dataset as those further away from the test data. Let ‡ denotes the majority, we have a distance measure $d_\ddagger = |\mathrm{median}(\mathrm{var}(\lambda_\ddagger)) - \max(\mathrm{var}(\boldsymbol{\lambda}_{\boldsymbol{S}}))|$ where $\mathrm{var}(\lambda_\ddagger) > \max(\mathrm{var}(\boldsymbol{\lambda}_{\boldsymbol{S}}))$ and $\lambda_\ddagger$ denotes the λ that calculated from the majority data. Similarly, $d_\dagger$ denotes the distance measure for the minority data and $d_\ddagger > d_\dagger$. Therefore:

$$\mathscr{L}^\ddagger = \{(x_l, y_l) | d_\ddagger > d_\dagger\}.$$

The original groupings are maintained so the majority data can be re-expressed according to the groupings.

3. Minority data: This portion of the data that is the complement of the majority data. This is found by a minority data decision rule to be described.

$$\mathscr{L}^{\dagger} = \{(x_l, y_l) | d_{\dagger} \leq d_{\ddagger}\}.$$

Again, the original groupings are maintained so the minority data can be re-expressed according to these groupings.

4. Re-grouped minority data $\mathscr{L}_R^{\dagger}$: For this, we take the minority data and re-group it according to a hierarchical clustering algorithm with respect to the responses $\boldsymbol{y}$ resulting in $m_r = m$ groupings with potentially revised memberships.

Rao *et al.* (2024) presented comprehensive simulation studies and analysis of data from the National Longitudinal Mortality Study (NLMS) which demonstrated superior predictive performance in these very challenging paradigms. An asymptotic analysis revealed why PURE resulted in more accurate projections.

References

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data, *Journal of American Statistical Association*, **80**, 28-36.

Conn, P. B., Johnson, D. S., and Boveng, P. L. (2015). On extrapolating past the range of the observed data when making statistical predictions in ecology, *PLoS One*, **10**, e0141416.

Cook, R. D. (1979). Influential observations in linear regression, *Journal of American Statistical Association*, **74**, 169-174.

Das, K., Jiang, J., and Rao, J. N. K. (2004). Mean squared error of emprical predictor, *Annals of Statistics*, **32**, 818-840.

Datta, G. S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statistica Sinica*, **10**, 613–627.

Datta, G. S. and Rao, J. N. K. and Smith, D. D. (2005). On measuring the variability of small area estimators under a basic area level model, *Biometrika*, **92**, 183-196.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data, *Journal of American Statistical Association*, **74**, 269-277.

Hall, P. and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models, *Annals of Statistics*, **34**, 1733-1750.

Hall, P. and Maiti, T. (2006b). On parametric bootstrap methods for small area prediction, *Journal of Royal Statistical Society Series B*, **68**, 221-238.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman & Hall/CRC, Boca Raton, Florida.

Gan, L., Yin, Z., and Tan, J. (2016). *Report on The Development of Household Finance in Rural China (2014)*, Springer, Singapore.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York.

Jiang, J. (2010). *Large Sample Techniques for Statistics*, Springer, New York.

Jiang, J. and Lahiri, P. (2001). Empirical best prediction for small area inference with binary data, *Annals of Institute of Statistics and Mathematics*, **53**, 217-243.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussion), *TEST*, **15**, 1-96.

Jiang, J., Lahiri, P., and Wan, S. (2002). A unified jackknife theory for empirical best prediction with M-estimation, *Annals of Statistics.*, **30**, 1782-1810.

Jiang, J., Lahiri, P., and Nguyen, T. (2018). A unified Monte-Carlo jackknife for small area estimation after model selection, *Annals of Mathematical Sciences and Applications*, **3**, 405-438.

Jiang, J., Nguyen, T., and Rao, J.S. (2011). Best predictive small area estimation, *Journal of American Statistical Association*, **106**.

Jiang, J., Rao, J. S., Fan, J., and Nguyen, T. (2018). Classified mixed model prediction, *Journal of American Statistical Association*, **113**, 269-279.

Liu, X., Ma, H., and Jiang, J. (2022b). That Prasad-Rao is robust: Estimation of mean squared prediction error of observed best preditor under potential model misspecification, *Statistica Sinica*, **32**, 2217-2240.

Liu, X. and Jiang, J. (2024). Classified functional mixed effects model prediction, *Statistics in Medicine*, **43**, 1329-1340.

Ma, H. and Jiang, J. (2023). Psuedo-Bayesian classified mixed model prediction, *Journal of American Statistical Association*, **118**, 1747-1759.

Nguyen, T., Jiang, J., and Rao, J. S. (2022). Assessing uncertainty for classified mixed model prediction, *Journal of Statistical Computation and Simulation*, **92**, 249-261.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error for small-area esitmators, *Journal of American Statistical Association*, **85**, 163-171.

Rao, J. N. K. (2003). *Small Area Estimation*, Wiley, New York.

Rao, J. S., Li, M., and Jiang, J. (2024). Classified mixed model projections, *Journal of American Statistical Association*, to appear.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion), *Statistical Science*, **6**, 15-51.

Sun, H., Nguyen, T., Luan, Y., and Jiang, J. (2018). Classified mixed logistic model prediction, *Journal of Multivariate Analysis*, **168**, 63-74.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso, *Journal of Royal Statistical Society Series B*, **58**, 267-288.

Welham, S., Cullis, B., Gogel, B., Gilmour, A., and Thompson, R. (2004). Prediction in linear mixed models, *Australian & New Zealand Journal of Statistics*, **46**, 325-347.

Zhang, J., Gan, L., Xu, L. C., and Yao, Y. (2014). Health shocks, village elections, and household income: Evidence from rural China, *China Economic Review*, **30**, 155-168.