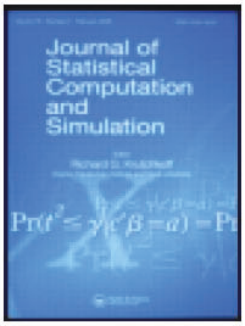

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Assessing uncertainty for classified mixed model prediction

Thuan Nguyen, Jiming Jiang & J. Sunil Rao

To cite this article: Thuan Nguyen, Jiming Jiang & J. Sunil Rao (2021): Assessing uncertainty for classified mixed model prediction, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2021.1955885

To link to this article: https://doi.org/10.1080/00949655.2021.1955885

Published online: 26 Jul 2021.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
https://doi.org/10.1080/00949655.2021.1955885

Assessing uncertainty for classified mixed model prediction

Thuan Nguyen[a], Jiming Jiang[b] and J. Sunil Rao[c]

[a]Oregon Health & Science University, Portland, OR, USA; [b]University of California Davis, Davis, CA, USA; [c]University of Miami, Coral Gables, FL, USA

ABSTRACT

Classified mixed model prediction (CMMP) is a new method that has embedded the traditional mixed model prediction (MMP) with a modern flavour. The basic idea is to first identify a class among the training data that matches the potential class corresponding to the new observations, whose associated mixed effect is of interest for prediction. Once such a matching is established, the MMP method can be utilized to make more accurate prediction that takes into account the subject-level differences. In this paper, we consider estimation of the mean squared prediction error (MSPE) of CMMP. A recently proposed Sumca method is implemented. Sumca combines analytic and Monte-Carlo approaches, leading to a second-order unbiased estimator of the MSPE. The performance of Sumca is investigated via simulation studies and comparisons are made with alternative methods. The simulation study shows that a brute-force bootstrap method performs almost as well as Sumca, while a naive approach and a Prasad-Rao estimator at the matched index are significantly inferior to Sumca. A real-data application is considered. Remarks and recommendation are offered.

ARTICLE HISTORY
Received 2 December 2020
Accepted 12 July 2021

KEYWORDS
CMMP; measure of uncertainty; MSPE; Sumca

1. Introduction

Classified mixed model prediction (CMMP [1]) is a recently proposed, modernized version of the traditional mixed model prediction (MMP; e.g. [2, section 2.3]). Such modern problems occur when interest is at subject level, such as in precision medicine, or (small) sub-population level (e.g. county, age by gender by race groups), as in precision public health, rather than at large population level. In such cases, it is possible to make substantial gains in prediction accuracy by identifying a class that a new subject belongs to. This idea was recently developed in CMMP, where a match between a random effect associated with the new data and one associated with the training data is built. Once the match is established, the well-developed MMP method can be utilized to take advantage of the available (massive) training data to make accurate prediction about characteristics of interest. Sun et al. [3] extended the CMMP method to classified mixed logistic model prediction; Sun, Luan and Jiang [4] proposed a variation of CMMP that incorporates covariate information in the prediction.

CONTACT Jiming Jiang jimjiang@ucdavis.edu University of California, Davis, CA 95616-5270, USA

An important issue in statistical prediction is measure of uncertainty. This was not addressed in the original CMMP method [1]. The measure of uncertainty in MMP, especially in terms of the mean squared prediction error (MSPE), has been well studied, in the context of small area estimation (SAE; e.g. [5]). However, CMMP is more complicated than MMP in that there is a matching procedure prior to the prediction. In a way, this is similar to a post-model-selection (PMS) prediction problem. Jiang and Torabi [6] proposed a Sumca method for MSPE estimation that applies to PMS prediction problems (Sumca is an abbreviation of 'simple, unified, Monte-Carlo assisted'). The method, in principle, is applicable to estimating the MSPE of CMMP. Although this was explored in Sun et al. (2020), the empirical evaluation is rather limited that did not show the nice theoretical property, namely, the second-order unbiasedness, of Sumca established in Jiang and Torabi [6]. As shown below, correct implementation of Sumca is not entirely straightforward, which is another point that we wish to make.

The Sumca method is carefully implemented to fit into the special feature of CMMP. Furthermore, we study the performance of Sumca in estimating the MSPE of CMMP in simulation studies. We also compare Sumca with several alternative approaches, including a brute-force bootstrap method (Boots), a naive method (Naive), and a Prasad-Rao method evaluated at the matched (via CMMP) index (Prami). Our simulation study shows that Boots performs nearly as well as Sumca while Naive and Prami perform considerably worse than Sumca. We also demonstrate real-data applicability of the CMMP/Sumca method. The computer code, carefully developed for the simulation studies in this paper, can serve as benchmarks for checking a future software package that implements the method.

In Section 2, a brief overview of Sumca as well as its implementation to CMMP is provided. In Section 3, we present results of Monte-Carlo simulation that evaluate finite-sample performance of Sumca as well as its comparison with the alternative methods. A real-data application is considered in Section 4. Concluding remarks and a recommendation are offered in Section 5.

2. Sumca: overview and implementation

2.1. Overview of Sumca

Let θ denote a mixed effect of interest, and $\hat{\theta}$ a predictor of θ. Then, the MSPE is defined as $\mathrm{MSPE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2$. The problem of estimating the MSPE of a possibly complex predictor, $\hat{\theta}$, has been extensively studied in the context of SAE. For example, a standard method for obtaining a second-order unbiased MSPE estimator, in the sense that $\mathrm{E}(\widehat{\mathrm{MSPE}}) = \mathrm{MSPE} + o(m^{-1})$, is the Prasad-Rao linearization method (PR; [7]). Here, m represents the number of (independent) clusters, or groups, in the (training) data. The PR method requires differentiability in order to use the Taylor series expansion for approximation. In the case of CMMP, however, the procedure involves a non-differentiable process, that is, the selection of i over a discrete space, $i \in \{1, \ldots, m\}$; as a result, the PR method may not apply.

Note that, in a way, the choice of the class index for the new observations may be viewed as a model selection problem so that the selected model connects the training data and new data in a specified way. From this viewpoint, the Sumca method, which applies to PMS

prediction problems, is applicable, at least in principle. Write the MSPE of $\hat{\theta}_n$ as

$$\text{MSPE} = \text{E}(\hat{\theta} - \theta)^2 = \text{E}\left[\text{E}\{(\hat{\theta} - \theta)^2 | Y\}\right], \tag{1}$$

where $Y = (y, y_n)$, y denotes the training data and y_n the new observations. The conditional expectation inside the outer expectation on the right side of (1) is a function of Y and ψ, a vector of unknown parameters associated with the distribution of Y, that is,

$$\begin{aligned} a(Y, \psi) &= \text{E}\{(\hat{\theta} - \theta)^2 | Y\} \\ &= \hat{\theta}^2 - 2\hat{\theta}\text{E}(\theta | Y) + \text{E}(\theta^2 | Y) \\ &= \hat{\theta}^2 - 2\hat{\theta} a_1(Y, \psi) + a_2(Y, \psi), \end{aligned} \tag{2}$$

where $a_s(Y, \psi) = \text{E}(\theta^s | Y), s = 1, 2$. Note that $\hat{\theta}$ does not depend on ψ. If we replace the ψ in (2) by $\hat{\psi}$, a consistent estimator, the result is a first-order unbiased estimator, that is,

$$\text{E}\{a(Y, \hat{\psi}) - a(Y, \psi)\} = O(m^{-1}). \tag{3}$$

On the other hand, both $\text{MSPE} = \text{E}\{a(Y, \psi)\}$ [by (1), (2)] and $\text{E}\{a(Y, \hat{\psi})\}$ are functions of ψ. Let $b(\psi) = \text{E}\{a(Y, \psi)\}$, $c(\psi) = \text{E}\{a(Y, \hat{\psi})\}$ and $d(\psi) = b(\psi) - c(\psi)$. Then, (3) implies that $d(\psi) = O(m^{-1})$; thus, if we replace, again, ψ by $\hat{\psi}$ in $d(\psi)$, the difference is a lower-order term, that is $d(\hat{\psi}) - d(\psi) = o(m^{-1})$ (in a suitable sense; e.g. in probability). Now consider the following estimator:

$$\widehat{\text{MSPE}} = a(Y, \hat{\psi}) + b(\hat{\psi}) - c(\hat{\psi}). \tag{4}$$

We have, by combining the above arguments, that $\text{E}(\widehat{\text{MSPE}}) = \text{E}\{a(Y, \psi)\} + \text{E}\{a(Y, \hat{\psi}) - a(Y, \psi)\} + \text{E}\{d(\hat{\psi})\} = \text{MSPE} + \text{E}\{d(\hat{\psi}) - d(\psi)\} = \text{MSPE} + o(m^{-1})$. The above arguments can be made rigorous [6].

The following alternative expression of $a(Y, \psi)$ is sometimes more convenient:

$$a(Y, \psi) = \{\hat{\theta} - \text{E}_\psi(\theta | Y)\}^2 + \text{var}_\psi(\theta | Y). \tag{5}$$

Note that, in (4), $a(Y, \hat{\psi})$ is the leading term which is typically $O(1)$; the remaining term, $d(\hat{\psi}) = b(\hat{\psi}) - c(\hat{\psi})$ is typically $O(m^{-1})$. However, the remaining term is usually much more difficult to evaluate than the leading term. Jiang and Torabi [6] propose to evaluate this term via a Monte-Carlo method. Let P_ψ denote the distribution of Y with ψ being the true parameter vector. Given ψ, one can generate Y under P_ψ. Let $Y_{[b]}$ denote Y generated under the bth Monte-Carlo sample, $b = 1, \ldots, B$. Then, we have

$$b(\psi) - c(\psi) \approx \frac{1}{B} \sum_{b=1}^{B} \left\{a(Y_{[b]}, \psi) - a(Y_{[b]}, \hat{\psi}_{[b]})\right\}, \tag{6}$$

where $\hat{\psi}_{[b]}$ denotes $\hat{\psi}$ based on $Y_{[b]}$. Jiang and Torabi argued that a reasonable choice for the Monte-Carlo sample size, B, is $B = m$. Thus, we can replace the term $b(\hat{\psi}) - c(\hat{\psi})$

in (4) by the right side of (6) with ψ replaced by $\hat{\psi}$, leading to the Sumca MSPE estimator:

$$\widehat{\mathrm{MSPE}}_B = a(Y, \hat{\psi}) + \frac{1}{B}\sum_{b=1}^{B}\left\{a(Y_{[b]}, \hat{\psi}) - a(Y_{[b]}, \hat{\psi}_{[b]})\right\}, \tag{7}$$

where $Y_{[b]}, b = 1, \ldots, B$ are generated as described above (6), with $\psi = \hat{\psi}$, and $\hat{\psi}_{[b]}$ is the estimator of ψ based on $Y_{[b]}$.

2.2. Implementation to CMMP

Following Jiang et al. [1], we assume that the training data satisfy the following nested-error regression (NER; [8]) model:

$$y_{ij} = x'_{ij}\beta + \alpha_i + \epsilon_{ij}, \tag{8}$$

$i = 1, \ldots, m, j = 1, \ldots, k_i$, where i represents the group index corresponding to the training data, k_i is the group size for group i; y_{ij} is the outcome of interest, x_{ij} is a vector of associated covariates, β is an unknown vector of regression coefficients (the fixed effects), α_i is a group-specific random effect, and ϵ_{ij} is an error. In practice, the group index i may correspond to a subject, or group subject sharing similar characteristics, such as in precision medicine. In small area estimation [5], the index i may correspond to a small geographical area or subpopulation. It is assumed that the random effects and errors are independent with $\alpha_i \sim N(0, G)$ and $\epsilon_{ij} \sim N(0, R)$, where $G > 0$, $R > 0$ are unknown variances. Furthermore, suppose that the outcomes of interest corresponding to a new subject, $y_{\mathrm{n}j}, 1 \le j \le k_{\mathrm{n}}$, satisfy

$$y_{\mathrm{n}j} = x'_{\mathrm{n}j}\beta + \alpha_I + \epsilon_{\mathrm{n}j}, \tag{9}$$

$1 \le j \le k_{\mathrm{n}}$, where $x_{\mathrm{n}j}$ is the corresponding vector of covariates; I is an unknown group index that is thought to be one of the $1, \ldots, m$ corresponding to the training data groups, although, in reality, this may or may not be true; and $\epsilon_{\mathrm{n}j}$s are the new errors that are independent and distributed as $N(0, R)$, and are independent with the α_is and ϵ_{ij}s.

To apply the Sumca method, note that, here, $\psi = (\beta', G, R, I)'$, where I is the true group index treated as another unknown parameter, and $\theta = x'_{\mathrm{n}}\beta + \alpha_I$ is the mixed effect of interest that one wishes to predict. By the normal theory, it can be derived that

$$\mathrm{E}_\psi(\theta|Y) = x'_{\mathrm{n}}\beta + \frac{(k_I + k_{\mathrm{n}})G}{R + (k_I + k_{\mathrm{n}})G}\left(\bar{y}_{I\cup\mathrm{n}\cdot} - \bar{x}'_{I\cup\mathrm{n}\cdot}\beta\right), \tag{10}$$

$$\mathrm{var}_\psi(\theta|Y) = \frac{GR}{R + (k_I + k_{\mathrm{n}})G}, \tag{11}$$

where $\bar{y}_{I\cup\mathrm{n}\cdot} = (k_I + k_{\mathrm{n}})^{-1}(\sum_{j=1}^{k_I} y_{Ij} + \sum_{j=1}^{k_{\mathrm{n}}} y_{\mathrm{n}j})$ and $\bar{x}_{I\cup\mathrm{n}\cdot}$ is defined similarly. Note the similarity, and difference, between (10) and the empirical best predictor (EBP) based on the training data (only), which is used in CMMP [1]:

$$\hat{\theta}_{(i)} = x'_{\mathrm{n}}\hat{\beta} + \frac{k_i\hat{G}}{\hat{R} + k_i\hat{G}}\left(\bar{y}_{i\cdot} - \bar{x}'_{i\cdot}\beta\right), \tag{12}$$

where $\bar{y}_{i\cdot} = k_i^{-1}\sum_{j=1}^{k_i} y_{ij}$, $\bar{x}_{i\cdot} = k_i^{-1}\sum_{j=1}^{k_i} x_{ij}$, and $\hat{\beta}, \hat{G}, \hat{R}$ are also based on the training data. The Sumca estimator of the MSPE of $\hat{\theta}$ is then obtained by (7) with $a(Y, \psi)$ given

by (5), (10), (11) and $\hat{\psi} = (\hat{\beta}', \hat{G}, \hat{R}, \hat{I})'$, where $\hat{\beta}, \hat{G}, \hat{R}$ are the (consistent) estimators mentioned, and $\hat{I}$ is obtained from CMMP by minimizing the distance between $\hat{\theta}_{(i)}$ and $\bar{y}_{\mathrm{n}} = k_{\mathrm{n}}^{-1} \sum_{j=1}^{k_{\mathrm{n}}} y_{\mathrm{n}j}$ [1, p. 278]. Specifically, we have

$$
\begin{aligned}
a(Y, \hat{\psi}) &= \{\hat{\theta} - \mathrm{E}_{\hat{\psi}}(\theta|Y)\}^2 + \mathrm{var}_{\hat{\psi}}(\theta|Y) \\
&= \left\{ \frac{k_{\hat{I}}\hat{G}}{\hat{R} + k_{\hat{I}}\hat{G}} (\bar{y}_{\hat{I}\cdot} - \bar{x}'_{\hat{I}\cdot}\hat{\beta}) - \frac{(k_{\hat{I}} + k_{\mathrm{n}})\hat{G}}{\hat{R} + (k_{\hat{I}} + k_{\mathrm{n}})\hat{G}} (\bar{y}_{\hat{I}\cup\mathrm{n}\cdot} - \bar{x}'_{\hat{I}\cup\mathrm{n}\cdot}\hat{\beta}) \right\}^2 \\
&\quad + \frac{\hat{G}\hat{R}}{\hat{R} + (k_{\hat{I}} + k_{\mathrm{n}})\hat{G}}; \qquad (13) \\
a(Y_{[b]}, \hat{\psi}) &= \{\hat{\theta}_{[b]} - \mathrm{E}_{\hat{\psi}}(\theta_{[b]}|Y_{[b]})\}^2 + \mathrm{var}_{\hat{\psi}}(\theta_{[b]}|Y_{[b]}) \\
&= \left\{ x'_{\mathrm{n}}(\hat{\beta}_{[b]} - \hat{\beta}) + \frac{k_{\hat{I}_{[b]}}\hat{G}_{[b]}}{\hat{R}_{[b]} + k_{\hat{I}_{[b]}}\hat{G}_{[b]}} (y_{[b]\hat{I}_{[b]}\cdot} - x'_{[b]\hat{I}_{[b]}\cdot}\hat{\beta}_{[b]}) \right. \\
&\quad \left. - \frac{(k_{\hat{I}} + k_{\mathrm{n}})\hat{G}}{\hat{R} + (k_{\hat{I}} + k_{\mathrm{n}})\hat{G}} (\bar{y}_{[b]\hat{I}\cup\mathrm{n}\cdot} - \bar{x}'_{[b]\hat{I}\cup\mathrm{n}\cdot}\hat{\beta}) \right\}^2 \\
&\quad + \frac{\hat{G}\hat{R}}{\hat{R} + (k_{\hat{I}} + k_{\mathrm{n}})\hat{G}}; \qquad (14) \\
a(Y_{[b]}, \hat{\psi}_{[b]}) &= \{\hat{\theta}_{[b]} - \mathrm{E}_{\hat{\psi}_{[b]}}(\theta_{[b]}|Y_{[b]})\}^2 + \mathrm{var}_{\hat{\psi}_{[b]}}(\theta_{[b]}|Y_{[b]}) \\
&= \left\{ \frac{k_{\hat{I}_{[b]}}\hat{G}_{[b]}}{\hat{R}_{[b]} + k_{\hat{I}_{[b]}}\hat{G}_{[b]}} (\bar{y}_{[b]\hat{I}_{[b]}\cdot} - \bar{x}'_{[b]\hat{I}_{[b]}\cdot}\hat{\beta}_{[b]}) \right. \\
&\quad \left. - \frac{(k_{\hat{I}_{[b]}} + k_{\mathrm{n}})\hat{G}_{[b]}}{\hat{R}_{[b]} + (k_{\hat{I}_{[b]}} + k_{\mathrm{n}})\hat{G}_{[b]}} (\bar{y}_{[b]\hat{I}_{[b]}\cup\mathrm{n}\cdot} - \bar{x}'_{[b]\hat{I}_{[b]}\cup\mathrm{n}\cdot}\hat{\beta}_{[b]}) \right\}^2 \\
&\quad + \frac{\hat{G}_{[b]}\hat{R}_{[b]}}{\hat{R}_{[b]} + (k_{\hat{I}_{[b]}} + k_{\mathrm{n}})\hat{G}_{[b]}}. \qquad (15)
\end{aligned}
$$

3. Simulation studies

We begin with an initial demonstration of the performance of Sumca as different factors such as the sample sizes and variances of the random effects and errors vary. Later we extend the simulation and make comparison with some alternative methods.

3.1. Initial demonstration

As an initial demonstration, we consider the following simple NER model:

$$
y_{ij} = \beta_0 + \beta_1 x_{ij} + v_i + e_{ij}, \qquad (16)
$$

Table 1. Simulated %RB of Sumca Estimator: M, Matched; UM, Unmatched.

β_0	β_1	G	R	m	k	k_n	M	UM
5.0	1.0	1.0	1.0	100	5	10	0.52	−8.59
5.0	1.0	1.0	1.0	100	5	1	9.11	−6.18
5.0	1.0	1.0	1.0	25	5	10	8.88	−9.16
5.0	1.0	1.0	1.0	100	20	10	5.09	2.69
5.0	1.0	0.1	1.0	100	5	10	−1.48	3.74
5.0	1.0	1.0	0.5	100	5	10	−0.15	−4.67

$i = 1, \ldots, m, j = 1, \ldots, k$; the new observations are generated as

$$y_{nj} = \theta + e_{nj}, \quad j = 1, \ldots, k_n, \tag{17}$$

where $\theta = \beta_0 + \beta_1 x_n + v_I$, where x_{ij}, x_n are generated from $N(0, 1)$, and fixed within each of the simulation scenarios corresponding to the rows of Table 1; $I = 1$ in the matched case, and α_I is generated independently in the unmatched case. We used $B = 100 \vee m$ for computing the Sumca estimator, as suggested by Jiang and Torabi [6]. We carried out $L = 500$ simulations runs for each combination of the true parameters and sample sizes listed in Table 1. In each simulation run, we generate the training data and new data according to the matched or unmatched scenarios, then compute $s = (\hat{\theta} - \theta)^2$ and the Sumca estimator of MSPE($\hat{\theta}$), denoted by S, as described above. Then, we have MSPE $\approx \bar{s} = L^{-1} \sum_{l=1}^{L} s_l$ and $\mathrm{E}(\widehat{\mathrm{MSPE}}) \approx \bar{S} = L^{-1} \sum_{l=1}^{L} S_l$, where s_l and S_l are s and S based on the lth simulation run. The percentage relative bias (%RB) is defined as

$$\%\mathrm{RB} = 100 \times \left\{ \frac{\mathrm{E}(\widehat{\mathrm{MSPE}}) - \mathrm{MSPE}}{\mathrm{MSPE}} \right\}.$$

The %RB are reported in Table 1. We start with the basic scenario in the first row and vary one component each time to generate a total of six scenarios for the matched and unmatched cases. It is seen that all %RBs are single digits (in absolute value), which are generally considered very good performance. Overall, the performance of Sumca is somewhat better in the matched case than in the unmatched case, which seems reasonable. However, the matched cases do not always outperform the unmatched cases. This may be due to the fact that CMMP does not have matching-consistency, that is, consistency in terms of identifying the true class, even if it exists [1]. As noted by the latter authors, the consistency of CMMP is in terms of estimation, or prediction, of the mixed effect associated with the new observations, not in term of correctly identifying the true class index corresponding to the new observations.

3.2. Extended simulation

We extend the simulation in the previous section by considering the same simulation setting in Jiang et al. [1]. The underlying model can be expressed as

$$y_{ij} = 1 + 2x_{ij,1} + 3x_{ij,2} + \alpha_i + \epsilon_{ij}, \tag{18}$$

$i = 1, \ldots, m$, $j = 1, \ldots, k$. The new observations satisfy (17) with $\theta_n = 1 + 2x_{n,1} + 3x_{n,2} + \alpha_I$, where $\alpha_i, \epsilon_{ij}, \alpha_I, \epsilon_{n,j}$ are the same as in the previous subsection, with G varies

among 0.25, 1.0, 2.0 and R fixed at 1.0. Three combinations of sample sizes are considered: $m = 50$, $k = 5$; $m = 50$, $k = 25$, and $m = 100$, $k = 25$; $k_n = 10$ in all cases. Again, we consider the matched and unmatched scenarios, as described in the previous subsection.

In addition to the Sumca estimator, we consider three alternative methods of MSPE estimation. The first is what we call brute-force (parametric) bootstrap (Boots). What one does is to treat β, G, R and I, the unknown group index for the new observations, as unknown parameters. Then, from CMMP, one obtains estimates of β, G, R (e.g. REML), say, $\hat{\beta}, \hat{G}, \hat{R}$, respectively, and $\hat{I}$, the matched group index. One then treats $\hat{\beta}, \hat{G}, \hat{R}, \hat{I}$ as the true parameters to regenerate both the training data and the new data. Although this seems to be a straightforward application of Efron's plug-in principle [9], it is worth noting that, typically, the application of the bootstrap requires consistency of the parameter estimation. Here, $\hat{\beta}, \hat{G}, \hat{R}$ are consistent estimators, but the story about $\hat{I}$ is different. As noted by Jiang et al. [1], $\hat{I}$ may not be consistent as $m \to \infty$. Nevertheless, the latter authors showed that the CMMP of θ is consistent, which is all that matters. This is the rationale behind Boots. Note that, for the bootstrapped data, one knows the value of θ, the mixed effect associated with the new data; one also has the CMMP, $\hat{\theta}$, of θ. Therefore, one can evaluate the empirical MSPE of the CMMP based on the bootstrap replications. The number of bootstrap replicates, B, is chosen as the same as the B for Sumca, which is 100.

The second alternative is to simply use the leading term in (5) or (7), that is, $a(Y, \hat{\psi})$. This is called the naive estimator (Naive).

The third alternative is the PR estimator (see Section 2.1) at matched index (Prami). Note that, given the group index i, the PR MSPE estimator, with REML estimator of the variance components, can be computed as

$$\widehat{\mathrm{MSPE}}_{\mathrm{PR},i} = g_{1i}(\hat{\sigma}^2) + g_{2i}(\hat{\sigma}^2) + 2g_{3i}(\hat{\sigma}^2), \tag{19}$$

where $\hat{\sigma}^2 = (\hat{G}, \hat{R})'$ is the REML estimator of $\sigma^2 = (G, R)'$. Datta and Lahiri [10] gave more specific expressions of the terms on the right side of (19). The Prami estimator is given by (19) with $i = \hat{I}$, the matched index by CMMP.

We carry out $L = 500$ simulation runs that compare the performance of the four methods, Sumca, Boots, Naive and Prami, in terms of %RB. The results are presented in Table 2. It is seen that both Sumca and Boots significantly outperform Naive and Prami in terms of %RB. In fact, the %RB for Sumca and Boots are mostly single-digit, or slightly over single-digit. There seems to be no obvious patterns between different scenarios, matched or unmatched. The %RBs of Naive are consistently negative, while those of Prami are consistently positive when k is smaller, and consistently negative when k is larger.

The picture is less clear, however, regarding the comparison between Sumca and Boots. It appears that Sumca performs slightly better than Boots in terms of the number of double-digit %RB cases (3 vs. 5) and in terms of the maximum absolute value of %RB (14.72 vs. 18.60). To help with the comparison, we compute key summary statistics including the mean (Mean), standard deviation (Sd), minimum (Min), first quartile (Q1), median (Median), third quartile (Q3) and maximum (Max) for both %RB and percentage absolute relative bias (%|RB|; that is, the absolute value of %RB) over the 18 scenarios considered in Table 2. The results are presented in Table 3. The figures confirm that Sumca, indeed, performs slightly better than Boots in overall performance; in terms of the median %|RB|, Sumca performs moderately better than Boots (3.65 vs. 6.11).

Table 2. Simulated %RB of Sumca, Boots, Naive and Prami Estimators: M, Matched; UM, Unmatched.

			Matched				Unmatched			
m	*k*	*G*	Sumca	Boots	Naive	Prami	Sumca	Boots	Naive	Prami
50	5	0.25	1.53	6.51	−34.98	37.22	−2.88	−10.89	−37.48	30.44
50	5	1.0	−5.23	−1.94	−36.30	69.19	−9.80	−9.74	−37.73	47.53
50	5	2.0	−10.20	−4.11	−40.03	57.70	−1.68	−18.60	−30.08	62.45
50	25	0.25	−0.80	0.75	−71.00	−58.67	−1.06	−1.66	−70.97	−60.88
50	25	1.0	6.98	0.83	−69.19	−57.58	−3.24	−2.58	−72.13	−58.84
50	25	2.0	5.39	1.16	−69.95	−56.68	−14.72	−9.57	−74.65	−67.01
100	25	0.25	−3.22	−13.83	−72.96	−63.69	−0.45	−1.67	−72.15	−62.68
100	25	1.0	−9.04	−11.74	−74.34	−62.80	−11.55	−5.70	−74.21	−65.52
100	25	2.0	4.05	−12.26	−70.52	−58.77	0.32	−7.80	−71.44	−59.56

Table 3. Summary Statistics of %RB and %|RB| for Sumca and Boots.

%RB	Mean	Sd	Min	Q1	Median	Q3	Max
Sumca	−3.09	6.03	−14.72	−8.09	−2.28	0.13	6.98
Boots	−5.71	6.48	−18.60	−10.60	−4.91	−1.66	6.51
%|RB|	Mean	Sd	Min	Q1	Median	Q3	Max
Sumca	5.12	4.32	0.32	1.57	3.65	8.53	14.72
Boots	6.74	5.33	0.75	1.74	6.11	10.60	18.60

Note that, theoretically speaking, Sumca is known to be second-order unbiased [6], while bootstrap without double-bootstrap bias correction is typically first-order unbiased (e.g. [11]; however, the double-bootstrap method is computationally too intensive). On the other hand, Boots has some advantage of its own. First, it is guaranteed positive, a desirable property that Sumca is not known to possess; in fact, out all of the 15,000 Sumca estimates computed in our simulation studies of the current and previous subsections (18 scenarios in the current subsection and 12 scenarios in the previous one, with 500 simulation runs under each scenario), we observed a single case in which the Sumca estimate is non-positive (the value is -6.58×10^{-5}).

Second, Boots is relatively simpler to program than Sumca. The trickier part of Sumca programming is to note that the difference $a(Y_{[b]}, \hat{\psi}) - a(Y_{[b]}, \hat{\psi}_{[b]})$ in (7) involves three things: $Y_{[b]}$, $\hat{\psi}$, and $\hat{\psi}_{[b]}$; except for the places where $\hat{\psi}$ and $\hat{\psi}_{[b]}$ appear, everywhere else involves the same $Y_{[b]}$. This may be easier said than done and mistakes can occur. For example, the CMMP, $\hat{\theta}_{[b]}$, is the same in both $a(Y_{[b]}, \hat{\psi})$ and $a(Y_{[b]}, \hat{\psi}_{[b]})$, and so are the means of the training data groups for x and y that are involved [see (14) and (15)]. In contrast, the programming for Boots is fairly straightforward.

Also, note again that the matched cases do not always outperform the unmatched cases, and this is true for any given method. This may be explained similarly as in Section 3.1 (see the end of Section 3.1).

4. Real-data example

We illustrate the measures of uncertainty for CMMP with a real-data example on predicting Framingham Risk Score (FRS) using various cardiovascular risk factors based on data from the University of Miami Health System's electronic health record. The FRS is one of many different scoring systems that are frequently used to determine individual risk of

developing cardiovascular disease. It provides an estimate of the probability that a person will develop cardiovascular disease usually within the next 10 years; it also indicates who is most likely to benefit from prevention. The FRS is estimated using age, total cholesterol, smoking status, HDL and systolic blood pressure as inputs. Slight variations in cutoffs are used in the scoring system depending on gender of the individual. Scores of $< 10\%$ indicate low risk, 10%–20% indicate moderate risk and $> 20\%$ indicate high risk. However, these cutoffs are somewhat arbitrary [12,13].

Our original data as of June 2020 and contains 219,499 unique patient records with a potential cardiovascular indicated issue and 33 variables including information on age, gender, race/ethnicity, country of origin, language of preference, clinical diagnoses, blood pressure, hemoglobin A1C, body mass index (BMI), smoking status, and linked residential address to census area level covariates. Some patients have repeated scores over time; in those cases the average FRS over the repeated scores are used. We then take the logit transformation of the FRS, or average FRS, which is the response variable.

There were 471 zip codes associated with the records. We first select a subset of the data as our training data, as described below. We focus on all of the patient-level variables and there are 17 of them. After some initial cleaning that removed all of the records with NAs on at least one of the 19 variables, we have about 135 thousands of records remaining, covering 267 distinct zip codes. We focus on the first record in case there are multiple records from the same ID. We chose those zip codes that contain at least 10 unique IDs. There are 130 such zip codes. From each of those zip codes, we dram a 25% random sample of the IDs with their corresponding records. This gives us 2331 records, which are used as the training data. Thus the training data consist of $m = 130$ groups with the total sample size $n = 2331$. The group sizes in the training data range from 3 to 83.

We then randomly select 20 zips codes from the 267 zip codes. From each selected zip code, we randomly select 5 first records from those excluded from the training data (in case the zip code matches one of the zip code in the training data). This is used as the new data. This leaves up 59 ID/records, among which 30 have matched IDs and 29 are unmatched. There are no replicates in the covariate (which is 17-dimensional), so there are 59 different mixed effects, one for each new observation.

We fit a NER model of (8), which is expressed as

$$\text{logit}(y_{ij}) = x_{ij}'\beta + \alpha_i + \epsilon_{ij}, \tag{20}$$

$i = 1, \ldots, 130$, $j = 1, \ldots, n_i$, where y_{ij} is FRS, and x_{ij} the 17-dimensional covariate, as described below (8). Then, for each of the 59 new data records, we obtain the CMMP for its corresponding mixed effect as well as the associated Sumca and Boots MSPE estimates. Boxplots of the MSPE estimates are presented in Figure 1. It appears that the, overall, the Sumca estimates are slightly smaller, and less variable, than the Boots estimates. All of the Sumca estimates are positive (the Boots estimates are positive by definition).

Tables 4 and 5 present the values of the response variable for the new data [that is, $\text{logit}(y_{\text{new}})$ for each new data record], the corresponding CMMP, and square roots of the MSPE estimates, $\widehat{\text{MSPE}}_{\text{S}}$ for Sumca and $\widehat{\text{MSPE}}_{\text{B}}$ for Boots. The first 30 cases are matched (Table 4) and last 29 cases are unmatched (Table 5). For the matched, we also present their true group numbers as well as the matched group numbers. As can be seen, none of the group numbers is actually matched. However, this does not matter, as CMMP does not rely

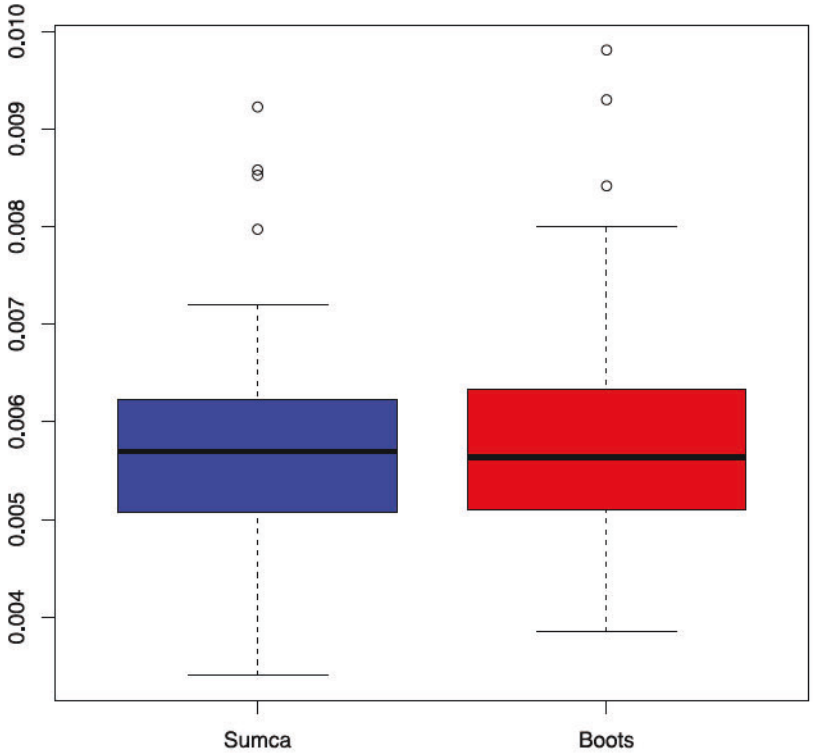

Figure 1. Boxplots of MSPE estimates.

on matching the group numbers. As shown in Jiang et al. [1], even if the group numbers are mismatched, or in an unmatched case where, of course, there is no actual match of the group number, CMMP still provides consistent prediction of the mixed effect of interest, which is what we are interested. Also note that CMMP does not intend to predict the response, that is, $\text{logit}(y_{\text{new}})$, but rather the mixed effect associated with the response, that is, $\theta_{\text{new}} = x'_{\text{new}}\beta + \alpha_I$ corresponding to (20), where I is the (unknown) group number. In terms of the MSPE estimates (or their square roots), it is seen that the Sumca and Boots estimates are quite close in most cases.

5. Conclusion, remarks and recommendation

We have implemented Sumca method for the estimation of the MSPE of CMMP, and studied its performance empirically in comparison with several alternative methods, including Boots, Naive and Prami. The empirical study shows that Boots performs almost as well as the Sumca, while Naive and Prami are significantly inferior than Sumca.

In our simulation study, we consider the NER model with one or two covariate variables. In Jiang et al. [1], CMMP is developed under two scenarios, the match case and unmatched case. Under the match scenario, it is assumed that the training data satisfy a linear mixed model that can be expressed as $y_i = X_i\beta + Z_i\alpha_i + \epsilon_i$, $i = 1, \ldots, m$, where y_i is $n_i \times 1$ vector representing the responses in group i; X_i is an $n_i \times p$ matrix of known covariates; Z_i is a known $n_i \times q$ matrix; α_i is a $q \times 1$ vector of group-specific random effects, and ϵ_i is an

Table 4. Response, CMMP and Square Root of MSPE Estimate (Matched).

Index	True Group #	Matched Group #	Response	CMMP	$\widehat{\text{MSPE}}_S^{1/2}$	$\widehat{\text{MSPE}}_B^{1/2}$
1	4	77	−6.454	−5.967	0.068	0.067
2	4	95	−0.929	−1.006	0.084	0.082
3	4	35	−1.306	−1.306	0.074	0.076
4	4	95	−3.546	−3.588	0.070	0.068
5	4	77	−6.529	−6.031	0.071	0.073
6	28	95	−1.504	−2.075	0.080	0.089
7	28	95	−2.194	−2.871	0.075	0.076
8	28	43	−3.690	−3.691	0.072	0.072
9	28	95	−4.261	−4.432	0.075	0.076
10	28	77	−3.214	−3.189	0.063	0.066
11	61	77	−0.541	−0.361	0.074	0.079
12	61	77	−2.284	−1.988	0.069	0.074
13	61	77	−1.691	−0.895	0.068	0.069
14	61	77	−5.336	−5.032	0.068	0.067
15	61	77	−3.657	−3.634	0.070	0.075
16	64	95	−1.095	−1.269	0.077	0.076
17	64	95	−0.382	−0.721	0.089	0.089
18	64	77	−0.594	1.5000	0.096	0.092
19	64	95	−1.817	−1.862	0.075	0.075
20	64	77	−3.800	−3.792	0.058	0.062
21	98	95	−5.180	−5.312	0.076	0.071
22	98	77	−5.079	−3.918	0.085	0.087
23	98	112	−1.113	−1.113	0.079	0.081
24	98	77	−6.974	−6.410	0.072	0.070
25	98	95	−1.580	−1.807	0.076	0.072
26	116	77	−0.737	−0.370	0.068	0.076
27	116	77	−2.310	−0.672	0.076	0.070
28	116	77	−2.020	−1.210	0.077	0.081
29	116	95	2.926	1.888	0.079	0.073
30	116	77	−0.176	−0.133	0.077	0.075

$n_i \times 1$ vector of additional errors. Furthermore, $\alpha_i, \epsilon_i, i = 1, \ldots, m$ are independent with $\alpha_i \sim N(0, G)$ and $\epsilon_i \sim N(0, R_i)$, where G, R_i are covariance matrices that depend on a vector γ of variance components. Note that the NER model is a special case of the above model with $q = 1$. Under the unmatched scenario, an NER model is assumed. In this paper, we focus on the NER model because it covers both the matched and unmatched cases. The Sumca estimator is expected to perform similarly under the NER models with more covariate predictors. Furthermore, extension to the linear mixed model considered by Jiang et al. [1] under the matched scenario is fairly straightforward. Another class of models, for which CMMP is developed, is the mixed logistic model with a structure similar to the NER model. It is expected that the Sumca method can be extended to the latter case as well. See Sun et al. [3].

Also, in our simulation study we have considered moderate number of groups, namely, m between 25 and 100. We expect that the Sumca method will perform similarly when m is smaller, as long as B, the Monte-Carlo sample size is not too small [in this paper, we use $B = 100 \vee m$; see below (17)]. As for larger m, it is known that, theoretically, for consistency of CMMP, the group sizes need to increase with m. For example, if only m increases but the group sizes do not increase fast enough, neither CMMP nor Sumca are expected to perform well. Note that such a scenario is not unpractical in modern data science, when it is becoming more feasible to collect data at subject levels. Another factor

Table 5. Response, CMMP and Square Root of MSPE Estimate (Unmatched).

Index	Response	CMMP	$\widehat{\text{MSPE}}_S^{1/2}$	$\widehat{\text{MSPE}}_B^{1/2}$
31	−4.189	−5.621	0.077	0.075
32	−1.568	−1.568	0.076	0.080
33	−3.177	−2.035	0.092	0.096
34	−2.198	−2.846	0.077	0.072
35	−4.468	−5.315	0.069	0.065
36	−2.379	−2.798	0.070	0.068
37	−1.916	−1.928	0.073	0.070
38	−1.716	−2.106	0.082	0.083
39	−2.029	−1.730	0.093	0.099
40	−3.669	−3.159	0.069	0.075
41	−0.401	−1.177	0.077	0.069
42	−1.030	−1.031	0.082	0.078
43	1.737	1.310	0.080	0.075
44	−3.014	−2.984	0.076	0.075
45	−0.409	−0.835	0.075	0.077
46	−2.111	−2.590	0.075	0.078
47	−0.805	−0.769	0.072	0.075
48	−2.200	−1.612	0.081	0.080
49	−1.313	−1.991	0.084	0.086
50	−1.915	−1.794	0.071	0.072
51	−2.832	−2.698	0.071	0.072
52	1.126	0.609	0.076	0.076
53	−5.413	−5.497	0.072	0.064
54	−1.945	−1.735	0.074	0.072
55	−0.121	−0.210	0.078	0.076
56	−1.410	−0.864	0.073	0.069
57	−1.636	−0.755	0.074	0.079
58	−2.094	−2.334	0.081	0.080
59	−1.447	−1.662	0.079	0.081

that needs to be considered is computational cost. It is recommended [1] that $B = m$. Thus when m is very large, Sumca may become computationally intensive.

As noted in Sun et al. (2020), in practice, the random effect, α_i, is often used to 'capture the un-captured', that is, variation not captured by the mean function, $x'_{ij}\beta$, at the cluster or group level. On the other hand, some components of x_{ij} may be also at the group level, that is, they depend on i but not j. It is natural to think that there may be association between α_i and some of the group-level components of x_{ij}; however, we do not know what kind of association it is excepted that it must be nonlinear (because, otherwise, it would be captured by $x'_{ij}\beta$). Sun et al. (2020) explored this scenario and proposed an extension of CMMP by incorporating the covariate information in the matching procedure. The extension of Sumca to this case is fairly straightforward.

Regarding software development, computer code is carefully developed for the two simulated examples considered in this paper. These can serve as benchmarks for checking future software; namely, any future software package should produce the same or very similar results as our code does under those two examples.

As noted, Sumca is theoretically more attractive (second-order unbiased) than Boots (first-order unbiased); on the other hand, the implementation of Sumca is less straightforward than Boots (see the last paragraph of Section 3). A practical recommendation would be to compute both Sumca and Boots estimates, as outputs of the Monte-Carlo samples can be used for both Sumca and Boots (in other words, one does not need to generate different Monte-Carlo samples for computing Sumca and Boots). The Boots estimates can be

used as a check on whether the Sumca method is implemented correctly, in which case, the Sumca and Boots estimates should be mostly close. Also, in a (very) rare situation that the Sumca estimate is negative, the Boots estimate can be used as a back-up.

Disclosure statement

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

Funding

Thuan Nguyen's research is supported by NSF grants DMS-1914760; Jiming Jiang's research is supported by NSF grant DMS-1914465; J. Sunil Rao's research is supported by NSF grant DMS-1915976.

References

[1] Jiang J, Rao JS, Fan J, et al. Classified mixed model prediction. J Amer Statist Assoc. 2018;113:269–279.
[2] Jiang J. Linear and generalized linear mixed models and their applications. New York: Springer; 2007.
[3] Sun H, Nguyen T, Luan Y, et al. Classified mixed logistic model prediction. J Multivar Anal. 2018;168:63–74.
[4] Sun H, Luan Y, Jiang J. A new classified mixed model predictor. J Stat Plan Inference. 2020;207:45–54.
[5] Rao JNK, Molina I. Small area estimation. 2nd ed., New York: Wiley; 2015.
[6] Jiang J, Torabi M. Sumca: simple, unified, Monte-Carlo-assisted approach to second-order unbiased mean-squared prediction error estimation. J Roy Statist Soc Ser B. 2020;82:467–485.
[7] Prasad NGN, Rao JNK. The estimation of mean squared errors of small area estimators. J Amer Statist Assoc. 1990;85:163–171.
[8] Battese GE, Harter RM, Fuller WA. An error-components model for prediction of county crop areas using survey and satellite data. J Amer Statist Assoc. 1988;80:28–36.
[9] Efron B. Bootstrap method: another look at the jackknife. Ann Statist. 1979;7:1–26.
[10] Datta GS, Lahiri P. A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Statist Sinica. 2000;10:613–627.
[11] Hall P, Maiti T. On parametric bootstrap methods for small area prediction. J Roy Statist Soc Ser B. 2006b;68:221–238.
[12] Wilson PW, D'Agostino RB, Levy D, et al. Prediction of coronary artery heart disease using risk factor categories. Circulation. 1998;97:1837–1847.
[13] D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the framingham heart study. Circulation. 2008;117:743–753.