Pseudo-Bayesian Classified Mixed Model Prediction

HAIQIANG MA AND JIMING JIANG
Jiangxi University of Finance and Economics, China and
University of California, Davis, USA

We propose a new classified mixed model prediction (CMMP) procedure, called pseudo-Bayesian CMMP, that utilizes network information in matching the group index between the training data and new data, whose characteristics of interest one wishes to predict. The current CMMP procedures (Jiang *et al.* 2018; Sun *et al.* 2018) do not incorporate such information; as a result, the methods are not consistent in terms of matching the group index. Although, as the number of training data groups increases, the current CMMP method can predict the mixed effects of interest consistently, its accuracy is not guaranteed when the number of groups is moderate, as is the case in many potential applications. The proposed pseudo-Bayesian CMMP procedure assumes a flexible working probability model for the group index of the new observation to match the index of a training data group, which may be viewed as a pseudo prior. We show that, given any working model satisfying mild conditions, the pseudo-Bayesian CMMP procedure is consistent and asymptotically optimal both in term of matching the group index and in terms of predicting the mixed effect of interest associated with the new observations. The theoretical results are fully supported by results of empirical studies, including Monte-Carlo simulations and real-data validation.

Key Words. Asymptotic optimality, classified mixed model prediction, consistency, matching, network communities, prediction, pseudo-Bayesian, random effects

1 Introduction

Classified mixed model prediction (Jiang *et al.* 2018) is a new method developed out of the traditional mixed model prediction that is particularly suitable for subject-level inference, such as in precision medicine and public health. For example, the National Research Council of the United States in 2014 defined precision medicine 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".

In spite of being a new idea in prediction with many potential applications, a key step in CMMP, that is, the matching between a class among the training data and an unknown class associated with the new/future observations, is based on a crude sample mean of the responses observed for the unknown class. In fact, as noted in Jiang *et al.* (2018, p. 273), "as m, the number of classes increases, the probability of identifying the true index I decreases, even in the matched case; thus, there is no consistency in terms of estimating I, even in the matched case." Here, I denotes the class index for the unknown class. However, in terms of prediction of the mixed effect of interest, CMMP is still consistent as m goes to infinity. This is because, as m increases, class-matching becomes less important in terms of approximation to the mixed effect associated with the unknown class. In other words, even if one does not have the exact match, there is a high chance, as m increases, that one will find an approximate match in the sense that the corresponding mixed effect is close to the one associated with the unknown class. Thus, if prediction of the mixed effect is of primary interest, CMMP will perform well as long as m is large.

On the other hand, there are many practical situations where m, the number of classes, is not very large. This happens, for example, in the analysis of network data, where the number of communities identified within the network is typically not large, or only moderately large (see below). Empirical studies have shown (e.g., Sections 3 and 4) that CMMP may perform poorly in such situations, which is consistent with the statement of Jiang *et al.* (2018) noted earlier. It turns out that, in the case that m is small or moderately large, precision in matching is critically important. Having realized the importance of the precision of matching in CMMP when m is relatively small, Sun *et al.* (2018) proposed to incorporate covariate information in the matching. The authors assumed that there exist

some class-level covariates that can be used in the matching. The modified CMMP procedure then focuses on matching the class-level covariates to those of a training data classes. However, for the latter method to work one needs a (nearly) one-to-one correspondence between the class-level covariates and random effects, which may not hold in practice. In fact, by definition, the random effects are supposed to be orthogonal to the covariates in a mixed effects model, so the strategy may also lack theoretical justification.

A main motivation of the current work comes from the analysis of network data, which has generated substantial interest in both research and applications. See, for example, Bickel and Chen (2009), McAuley and Leskovec (2012), Bickel and Sarkar (2016), Ma, Su and Zhang (2018), and Li, Shen and Pan (2019). One of the extensively studied topics in network analysis is community detection in networks. Such work and results are potentially useful in identifying the classes that we are concerned about. It should be noted that our primary interest is prediction of mixed effects, or future observations, by utilizing the class information. These classes are potentially closely related to the communities detected in the network. Thus, it is natural to consider utilizing the existing work in community detection in CMMP. In fact, community information within a network has been used in improving prediction accuracy in precision epidemiology for infectious disease control (e.g., Keeling and Eames 2005, Ladner *et al.* 2019). As another example, it is known that social networks have effects on economics (e.g., Bailey *et al.* 2018); therefore, it may be possible to utilize network information in making prediction of characteristics of economic interest.

One feature of the network data is that the number of communities within the network is typically not very large. For example, in the Jazz musician network (Gleiser and Danon 2003; data available at www.redhotjazz.com), the number of communities known to exist is 3. In the political books network (Newman 2006; data available at www.orgnet.com), the number of known communities is also 3. In the Facebook friendship network (available at www.snap.stanford.edu), 11 communities have been identified (Ma *et al.* 2018). With such a small or moderate number of communities, the original CMMP method of Jiang *et*

al. (2008) would not apply, if the communities are treated as the classes.

Furthermore, there is also a concern on whether the communities in a network and the classes in CMMP should match exactly. Typically, the classes in CMMP correspond to random effects under a mixed effects model. The random effects are associated with the means or proportions of some characteristics of interest. It is unlikely that these characteristics are solely determined by the community membership, even though the two may be associated. Therefore, uncertainty, or discrepancy, should be allowed in matching the communities to the classes. Due to such a consideration, we consider the following model for the data, which includes a working probability model for the class membership.

Suppose that we have a network with known communities. As mentioned, there has been extensive work on identification of the network communities (e.g., Bickel and Sarkar 2016, Ma *et al.* 2018), so we can assume that the network communities are known. Suppose that the training data associated with the network satisfy a nested-error regression (NER; Battese, Harter and Fuller 1988) model:

$$y_{ij} \quad = \quad x'_{ij}\beta + \alpha_i + \epsilon_{ij}, \tag{1}$$

$i = 1, \ldots, m$, $j = 1, \ldots, k_i$, where i represents the community, k_i is the number of subjects in the training data that belong to community 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 community-specific random effect, and ϵ_{ij} is an error. It is assumed that the random effects and errors are independent with $\alpha_i \sim N(0, \sigma^2)$ and $\epsilon_{ij} \sim N(0, \tau^2)$, where $\sigma^2 > 0, \tau^2 > 0$ are unknown variances.

We are interested in prediction of a mixed effect associated with a new subject. The mixed effect may be a conditional mean, proportion, or other characteristic, given the random effect associated with the subject. Suppose that the new subject belongs to a known community $c_{\rm n}$. Here the subscript n stands for "new". The random effect associated with the new subject, however, is not entirely determined by $c_{\rm n}$—it is subject to some uncertainty. This happens, for example, when the training data were collected from a previous

time period, a network that has grown bigger, or smaller, or a network that is not exactly the same as the one relevant to the new subject. Consider the following working probability model. Let $\gamma_{\rm n}$ denote the true class index of the new subject. Note that the words "class" and "community" do not necessarily mean the same in that the former refers to that associated with the random effect while the latter to that associated with the network. For the training data, however, the classes match the communities, by assumption, but this is not necessarily true for the new subject. For now, let us assume that $\gamma_{\rm n}$ is an unknown integer between 1 and m. This is called a matched case. Later we also consider the case that $\gamma_{\rm n}$ does not match any of the integers between 1 and m. This is called an unmatched case (Jiang *et al.* 2018). We assume that there is a working probability model for $\gamma_{\rm n}$:

$$\pi(\gamma_{\rm n} = i), \quad 1 \le i \le m, \tag{2}$$

where $\pi(\cdot)$ is a known probability distribution, which is not necessarily the true distribution of $\gamma_{\rm n}$. For example, in connection with utilizing the network information in the matching, one may consider the following working model:

$$\pi(\gamma_{\rm n} = i) \;\; = \;\; p^{1_{(i=c_{\rm n})}} \left(\frac{1-p}{m-1} \right)^{1_{(i \ne c_{\rm n})}}, \tag{3}$$

where p is a given probability (see below); in other words, $\pi(\gamma_{\rm n} = i) = p$ if $i = c_{\rm n}$, and $\pi(\gamma_{\rm n} = i) = (1-p)/(m-1)$ if $i \ne c_{\rm n}$. It is easy to verify that (3) is a probability distribution on $\{1, \ldots, m\}$. The p in (3) may be treated as a tuning parameter, which has an intuitive interpretation: It has to do with one's belief to what extent $c_{\rm n}$ determines $\gamma_{\rm n}$. Large sample theory, established later in this paper, shows that, as long as there are sufficient data information, it does not really matter what $\pi(\cdot)$ is used as the working model, or, in particular, what p is chosen in the special case of (3). We call $\pi(\cdot)$ a *pseudo-prior* due to a role it plays in deriving our method in the sequel.

Although (3) is a special case of (2), it is an important special case that motivates our method. In this special case, the community index to which the new subject belongs, $c_{\rm n}$, is known. However, it is not necessarily equal to the true class index, $\gamma_{\rm n}$. The pseudo-prior

probability that $\gamma_{\rm n} = c_{\rm n}$ is p. Because, in a way, $c_{\rm n}$ may be viewed as part of the data, the pseudo prior (3) may be viewed as a conditional probability. In general, the pseudo-prior (2) could be conditioning on certain data other than y, the combined responses of the training data and new data (see below). However, for notation simplicity, the conditioning notation is suppressed [e.g., $\pi(\gamma_{\rm n} = i)$ rather than $\pi(\gamma_{\rm n} = i|c_{\rm n})$ in (3)]. Similarly, the pseudo-posterior derived below may be viewed as conditioning on y and $c_{\rm n}$, in case of (3), although the $c_{\rm n}$ part is suppressed in notation [e.g., $\mathrm{P}_\pi(\gamma_{\rm n} = i|y)$ instead of $\mathrm{P}_\pi(\gamma_{\rm n} = i|y, c_{\rm n})$]. The asymptotic results, established later in this paper, should be regarded as conditional on $c_{\rm n}$. Also note that there is no randomness in $c_{\rm n}$; in other words, $c_{\rm n}$ is a known constant.

Furthermore, suppose that the outcomes of interest corresponding to the new subject, $y_{{\rm n}j}, 1 \le j \le k_{\rm n}$, satisfy a similar NER model to (1), that is,

$$y_{{\rm n}j} \;\; = \;\; x'_{{\rm n}j}\beta + \alpha_{\gamma_{\rm n}} + \epsilon_{{\rm n}j}, \tag{4}$$

$1 \le j \le k_{\rm n}$, where $x_{{\rm n}j}$ is the corresponding vector of covariates, and $\epsilon_{{\rm n}j}$s are the new errors that are independent and distributed as $N(0, \tau^2)$, and are independent with $\alpha_{\gamma_{\rm n}}$ and the α_is and ϵ_{ij}s associated with the training data. Note that, given $\gamma_{\rm n} = i$, (4) becomes $y_{{\rm n}j} = x'_{{\rm n}j}\beta + \alpha_i + \epsilon_{{\rm n}j}$, $1 \le j \le k_{\rm n}$. This means that one can combine the training data and new data into m independent groups:

$$y_1, \ldots, y_{i-1}, (y_i, y_{\rm n}), y_{i+1}, \ldots, y_m,$$

where $y_i = (y_{ij})_{1 \le j \le k_i}$ and $y_{\rm n} = (y_{{\rm n}j})_{1 \le j \le k_{\rm n}}$. The pdf of y_u ($u \ne i$) is given by

$$f(y_u) \;\; = \;\; \frac{1}{(2\pi)^{k_u/2}|V_u|^{1/2}} \exp\left\{-\frac{1}{2}(y_u - X_u\beta)'V_u^{-1}(y_u - X_u\beta)\right\}, \tag{5}$$

where $X_u = (x'_{uj})_{1 \le j \le k_u}$ and $V_u = \tau^2 I_{k_u} + \sigma^2 J_{k_u}$ (I_k, J_k denote the $k \times k$ identity matrix and matrix of 1s, respectively). Similarly, given $\gamma_n = i$, the joint pdf of $(y_i, y_{\rm n})$ is given by

$$\begin{aligned} & f(y_i, y_{\rm n}|\gamma_n = i) \\ = \;\; & \frac{1}{(2\pi)^{(k_i+k_{\rm n})/2}|V_{i,{\rm n}}|^{1/2}} \exp\left\{-\frac{1}{2}\begin{pmatrix} y_i - X_i\beta \\ y_{\rm n} - X_{\rm n}\beta \end{pmatrix}' V_{i,n}^{-1} \begin{pmatrix} y_i - X_i\beta \\ y_{\rm n} - X_{\rm n}\beta \end{pmatrix}\right\}, \end{aligned} \tag{6}$$

where $X_{\rm n} = (x'_{{\rm n}j})_{1\le j\le k_{\rm n}}$ and $V_{i,n} = \tau^2 I_{k_i+k_{\rm n}} + \sigma^2 J_{k_i+k_{\rm n}}$. Combining the above results, and with some reorganization of terms, we obtain

$$\begin{aligned} f(y|\gamma_{\rm n}=i) &= f(y_i, y_{\rm n}|\gamma_{\rm n}=i)\prod_{u\neq i} f(y_u) \\ &= (2\pi)^{(k_\cdot+k_{\rm n})/2}(\tau^2)^{(k_\cdot+k_{\rm n}-m)/2}\{\tau^2+(k_i+k_{\rm n})\sigma^2\}^{1/2}\prod_{u\neq i}(\tau^2+k_u\sigma^2)^{1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}\begin{pmatrix} y_i - X_i\beta \\ y_{\rm n} - X_{\rm n}\beta \end{pmatrix}' V_{i,n}^{-1} \begin{pmatrix} y_i - X_i\beta \\ y_{\rm n} - X_{\rm n}\beta \end{pmatrix}\right. \\ &\quad \left. -\frac{1}{2}\sum_{u\neq i}(y_u - X_u\beta)'V_u^{-1}(y_u - X_u\beta)\right\}, \end{aligned} \tag{7}$$

where $k_\cdot = \sum_{u=1}^m k_u$, and $y = (y'_1, \ldots, y'_m, y'_{\rm n})'$.

From (2) and (7), we obtain the pseudo-posterior distribution of γ_n:

$$\mathrm{P}_\pi(\gamma_{\rm n}=i|y) = \frac{\pi(\gamma_{\rm n}=i)f(y|\gamma_{\rm n}=i)}{\sum_{j=1}^m \pi(\gamma_{\rm n}=j)f(y|\gamma_{\rm n}=j)}. \tag{8}$$

The match of γ_n to the training data groups is chosen as the pseudo-posterior mode, that is,

$$\begin{aligned} \hat\gamma_{\rm n} &= \mathrm{argmax}_{1\le i\le m}\mathrm{P}_\pi(\gamma_{\rm n}=i|y) \\ &= \mathrm{argmax}_{1\le i\le m}\{\pi(\gamma_{\rm n}=i)f(y|\gamma_n=i)\}. \end{aligned} \tag{9}$$

Note that the denominator in (8) is not needed in obtaining $\hat\gamma_{\rm n}$. In case that the maximizer of (8) or (9) is not unique, choose the one with the lowest index number.

Note. Although the procedure clearly resembles the *maximum posterior*, or Bayesian classification (e.g., Nurty and Devi 2011), the set-up is not Bayesian. In fact, the distribution π in (2) is treated as a working model rather than a prior. In other words, there is no underlying assumption that the working model is the true distribution of $\gamma_{\rm n}$. In this regard, the method is similar to Henderson's original derivation of BLUP—best linear unbiased predictor (e.g., Jiang, Jia and Chen 2001, p.98). On the other hand, whenever there is additional information regarding the class index, $\gamma_{\rm n}$, such as network information, it should be incorporated into the working model. As will be demonstrated, the choice of the working

model does not make a difference asymptotically, but it may make a difference in terms of the finite-sample performance. Due to its similarity to the Bayesian classifier, we call the proposed procedure of matching maximum pseudo-posterior matching (MPPM).

Once $\hat{\gamma}_{\rm n}$ is determined, the prediction of the new mixed effect is carried out as in CMMP (Jiang *et al.* 2018). Namely, given $\gamma_{\rm n} = i$, the best predictor (BP), in the sense of minimum mean squared prediction error (MSPE), of $\theta_{{\rm n}j} = x'_{{\rm n}j}\beta + \alpha_{\gamma_{\rm n}} = x'_{{\rm n}j}\beta + \alpha_i$ [see (4)] is

$$\mathrm{E}(\theta_{{\rm n}j}|y) \;\; = \;\; x'_{{\rm n}j}\beta + \frac{k_i\sigma^2}{\tau^2 + k_i\sigma^2}(\bar{y}_{i\cdot} - \bar{x}'_{i\cdot}\beta), \tag{10}$$

where $\bar{y}_{i\cdot} = k_i^{-1}\sum_{j=1}^{k_i} y_{ij}$ and $\bar{x}_{i\cdot} = k_i^{-1}\sum_{j=1}^{k_i} x_{ij}$. The classified mixed effect predictor (CMEP) of $\theta_{{\rm n}j}$, denoted by $\hat{\theta}_{{\rm n}j}$, is given by the right-hand side of (10) with i replaced by $\hat{\gamma}_{\rm n}$, and β, σ^2, τ^2 replaced by their estimators (e.g., REML estimators; e.g., Jiang 2007, sec. 1.3.2) based on the training data. We call the new procedure pseudo-Bayesian CMMP, or PBCMMP, due to its connection to both MPPM and CMMP.

The difference between PBCMMP and CMMP is in their strategies of matching the class index, $\gamma_{\rm n}$, which is a key component of CMMP. The CMMP of Jiang *et al.* (2018) used a strategy that matched the sample mean of the new observations and the empirical BP (EBP), given the class index, of the mixed effect associated with the new observations (see Jiang *et al.* 2018, 3rd paragraph in sec. 7.2). The new PBCMMP uses the maximum pseudo-posterior idea, as noted above. As will be seen, the new matching strategy of PBCMMP, that is, MPPM, has both asymptotic and empirical superiority over the CMMP matching strategy. The superior matching strategy of PBCMMP does lead to (significantly) better predictive performance over CMMP, as we show both theoretically and empirically.

One can also make prediction of a future observation, say, $y_{{\rm n}j}$, if it is unobserved. By (4), we have $y_{{\rm n}j} = \theta_{{\rm n}j} + \epsilon_{{\rm n}j}$. Because the new error, $\epsilon_{{\rm n}j}$, is independent of the training data, the BP of $y_{{\rm n}j}$ is $\mathrm{E}(y_{{\rm n}j}|y) = \mathrm{E}(\theta_{{\rm n}j}|y) + \mathrm{E}(\epsilon_{{\rm n}j}|y) = \mathrm{E}(\theta_{{\rm n}j}|y)$. This shows that the BP of $y_{{\rm n}j}$ is the same as the BP of $\theta_{{\rm n}j}$; therefore, naturally, the PBCMMP of $y_{{\rm n}j}$, denoted by $\hat{y}_{{\rm n}j}$, is the same as $\hat{\theta}_{{\rm n}j}$, the CMEP of $\theta_{{\rm n}j}$ based on MPPM.

In Section 2, we establish the asymptotic superiority of MPPM mentioned above, as

well as consistency of MPPM in terms of the class-index matching. In Section 3, we establish consistency and asymptotic optimality of the CMEP based on MPPM. It should be noted that the consistency and asymptotic optimality of MPPM and CMEP based on MPPM hold for any working model π subject to some regularity conditions. In Section 4, we carry out extensive simulation studies to investigate finite-sample performance of MPPM and CMEP based on MPPM, and their comparisons with existing methods. The simulation results demonstrate, in particular, the superiority of MPPM and CMEP based on MPPM over the CMMP of Jiang *et al.* (2018), as predicted by the theory. In Section 5, we discuss measure of uncertainty for the proposed classified predictor. In Section 6, we consider an application to Facebook network data, and provide a real-data validation that further demonstrate the advantage of our new methods. Conclusion and discussion are offered in Section 7. Proofs and additional tables are deferred to Supplementary Material.

2 Consistency and asymptotic optimality of MPPM

We first consider a simpler, but also realistic situation in some cases (e.g., network data), where m, the number of classes in the training data, is bounded. Later we extend the result to the case that m increases with the sample sizes at an appropriate rate. Throughout this section, we assume a match case, which means that $\gamma_{\rm n}$ matches one of the indexes $1 \leq i \leq m$. This is reasonable because, otherwise, there is, of course, no consistency.

2.1 Consistency and asymptotic optimality when m is bounded

We assume that a consistent estimator of β, $\hat{\beta}$, is available; however, for the σ^2, τ^2, we only assume that some estimators, $\hat{\sigma}^2, \hat{\tau}^2$, are available, which satisfy

$$0 < a \leq \hat{\sigma}^2, \hat{\tau}^2 \leq A < \infty, \tag{11}$$

where a, A are some known constants. Note that (11) can always be met by truncating the estimators of σ^2 and τ^2. For example, one may choose a as a small positive number (e.g.,

10^{-6}) and A a large positive number (e.g., 10^6), and re-define the value of $\hat{\sigma}^2$ as a, if it is less than a, and A, if it is greater than A; and similarly for τ^2.

Also note that, for the consistency of $\hat{\beta}$, one does not need $m \to \infty$. In fact, $m \to \infty$ is necessary for the consistency of $\hat{\sigma}^2$ but not for that of $\hat{\beta}$ and $\hat{\tau}^2$. However, it is known that consistent estimator of β is available given any "working" estimators of σ^2 and τ^2, which need not to be consistent. For example, such a result was established earlier in the context of generalized estimating equations (GEE; Liang and Zeger 1986), and later in the context of generalized linear mixed models (GLMM; e.g., Jiang 1999, Jiang *et al.* 2001). Define $a_i = k_i/(k_i + k_{\rm n}), 1 \le i \le m$. To distinguish MPPM based on different working models, let us denote the MPPM, (9), by $\hat{\gamma}_{{\rm n},\pi}$, where π corresponds to the working model (2). Let $\gamma_{\rm n}^*$ denote the $\hat{\gamma}_{{\rm n},\pi}$ when π is the true distribution of $\gamma_{\rm n}$. The latter is not practically available, of course, but we can still get it involved in theoretical studies. Define

$$\tilde{\gamma}_{\rm n} \quad = \quad {\rm argmax}_{1 \le i \le m} f(y|\gamma_{\rm n} = i). \tag{12}$$

Note that $\tilde{\gamma}_{\rm n}$ does not depend on π (therefore no index of π is needed).

Theorem 1 (consistency of MPPM and more). Suppose that the following hold:
(i) $m > 1$, $\min_{1 \le i \le m} k_i \to \infty$, $k_{\rm n} \to \infty$, and

$$\min_{1 \le i \le m} a_i \ge b \ \text{ for some constant } b > 0; \tag{13}$$

(ii) $\bar{x}_{i\cdot} = k_i^{-1} \sum_{j=1}^{k_i} x_{ij}, 1 \le i \le m$ and $\bar{x}_{{\rm n}\cdot} = k_{\rm n}^{-1} \sum_{j=1}^{k_{\rm n}} x_{{\rm n}j}$ are bounded, so are

$$b_{ij} \equiv \log\{\pi(\gamma_{\rm n} = j)\} - \log\{\pi(\gamma_{\rm n} = i)\}, \quad 1 \le j \ne i \le m;$$

(iii) $\hat{\beta}$ is consistent, and (11) holds. Then, we have the following conclusions:
(I) ${\rm P}(\hat{\gamma}_{{\rm n},\pi} = \tilde{\gamma}_{\rm n}) \to 1$ for any working model π, including the true distribution of $\gamma_{\rm n}$.
(II) MPPM is consistent for any working model π, that is, ${\rm P}(\hat{\gamma}_{{\rm n},\pi} \ne \gamma_n) \to 0$.

The proof of Theorem 1 is given in the Supplementary Material. If we apply the result to the special case that π is the true distribution of $\gamma_{\rm n}$, we have the following result.

Corollary 1. Under the conditions of Theorem 1, we have, for any working model π,

$$\mathrm{P}(\hat{\gamma}_{\mathrm{n},\pi} = \gamma_{\mathrm{n}}^*) \;\longrightarrow\; 1. \tag{14}$$

Proof: Using conclusion (I) of Theorem 1, we have

$$\begin{aligned} \mathrm{P}(\hat{\gamma}_{\mathrm{n},\pi} = \gamma_{\mathrm{n}}^*) &\geq \mathrm{P}(\hat{\gamma}_{\mathrm{n},\pi} = \tilde{\gamma}_{\mathrm{n}}, \gamma_{\mathrm{n}}^* = \tilde{\gamma}_{\mathrm{n}}) \\ &\geq 1 - \mathrm{P}(\hat{\gamma}_{\mathrm{n},\pi} \neq \tilde{\gamma}_{\mathrm{n}}) - \mathrm{P}(\gamma_{\mathrm{n}}^* \neq \tilde{\gamma}_{\mathrm{n}}) \to 1. \quad \square \end{aligned}$$

Next, we establish asymptotic optimality of MPPM. To do so we first introduce a lemma that states the (exact) optimality of γ_{n}^*, which we call maximum posterior matching (MPM), even although it is not practically available. The optimality is in terms of minimizing the probability of mismatch, that is, $\mathrm{P}(\check{\gamma}_{\mathrm{n}} \neq \gamma_{\mathrm{n}})$, where $\check{\gamma}_{\mathrm{n}}$ is any class matcher of γ_{n}.

Lemma 1 (Optimality of MPM). $\mathrm{P}(\gamma_{\mathrm{n}}^* \neq \gamma_{\mathrm{n}}) \leq \mathrm{P}(\check{\gamma}_{\mathrm{n}} \neq \gamma_{\mathrm{n}})$ for any $\check{\gamma}_{\mathrm{n}}$.

Although such a result is well known in the literature of Bayesian classifier (e.g., Nurty and Devi 2011), we were unable to find a simple proof. Thus, a proof is given below.

Proof: Note that, when π is the true distribution of γ_{n}, the right side of (8) is equal to $\mathrm{P}(\gamma_{\mathrm{n}} = i|y)$, the (true) conditional probability. Thus, by definition, γ_{n}^* has the property that

$$\mathrm{P}(\gamma_{\mathrm{n}} = i|y) = \max_{1 \leq j \leq m} \mathrm{P}(\gamma_{\mathrm{n}} = j|y) \;\text{ on }\; \{\gamma_{\mathrm{n}}^* = i\}, \;\; 1 \leq i \leq m. \tag{15}$$

Now, for any other class matcher, $\check{\gamma}_{\mathrm{n}}$, we have

$$\begin{aligned} \mathrm{P}(\check{\gamma}_{\mathrm{n}} \neq \gamma_{\mathrm{n}}) &= \sum_{i=1}^{m} \mathrm{P}(\check{\gamma}_{\mathrm{n}} = i, \gamma_{\mathrm{n}} \neq i) \\ &= \sum_{i=1}^{m} \mathrm{E}\left[1_{(\check{\gamma}_{\mathrm{n}} = i)}\{1 - \mathrm{P}(\gamma_{\mathrm{n}} = i|y)\}\right] \\ &\geq \sum_{i=1}^{m} \mathrm{E}\left[1_{(\check{\gamma}_{\mathrm{n}} = i)}\left\{1 - \max_{1 \leq j \leq m} \mathrm{P}(\gamma_{\mathrm{n}} = j|y)\right\}\right] \\ &= 1 - \mathrm{E}\left\{\max_{1 \leq j \leq m} \mathrm{P}(\gamma_{\mathrm{n}} = j|y)\right\}. \end{aligned} \tag{16}$$

On the other hand, by replacing $\check{\gamma}_{\mathrm{n}}$ with γ_{n}^*, and using property (15), it is seen that the same arguments of (16) hold with the inequality in the third line replaced by equality. Therefore, the right side of (16) is equal to $\mathrm{P}(\gamma_{\mathrm{n}}^* \neq \gamma_{\mathrm{n}})$. $\square$

Because the MPM, $\gamma_{\rm n}^*$, is optimal, in view of (14), it is not surprising that MPPM is asymptotically optimal. Let N denote the different sample sizes associated with y, such as $k_i, 1 \le i \le m$ and $k_{\rm n}$. Let a_N be a sequence of positive numbers such that

$$a_N \to \infty, \;\; a_N\{{\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}) \wedge {\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}^*)\} \to 0. \tag{17}$$

By Theorem 1 and Corollary 1, we have ${\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}) \to 0$ and ${\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}^*) \to 0$. Thus, for example, for any $0 < \delta < 1$, $a_N = \{{\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}) \wedge {\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}^*)\}^{-\delta}$ is a sequence that satisfies (17). We have the following result.

Theorem 2 (Asymptotic optimality of MPPM). Suppose that the conditions of Theorem 1 are satisfied. Let $\check{\gamma}_{\rm n}$ be any other class matcher. Then, we have

$$\limsup\{a_N{\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n})\} \;\; \le \;\; \limsup\{a_N{\rm P}(\check{\gamma}_{\rm n} \neq \gamma_{\rm n})\} \tag{18}$$

for any sequence a_N that satisfies (17).

Proof: We have, using Lemma 1,

$$\begin{aligned}
{\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}) &= {\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}, \hat{\gamma}_{{\rm n},\pi} = \gamma_{\rm n}^*) + {\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}, \hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}^*) \\
&\le {\rm P}(\gamma_{\rm n}^* \neq \gamma_{\rm n}) + {\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}) \wedge {\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}^*) \\
&\le {\rm P}(\check{\gamma}_{\rm n} \neq \gamma_{\rm n}) + {\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}) \wedge {\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}^*).
\end{aligned}$$

Thus, by multiplying a_N on both sides, we obtain

$$a_N{\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}) \le a_N{\rm P}(\check{\gamma}_{\rm n} \neq \gamma_{\rm n}) + a_N\{{\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}) \wedge {\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}^*)\}.$$

The result then follows by taking the $\limsup$ on both sides. □

Now we know that MPM is optimal and MPPM is asymptotically optimal. A question is how much is the difference between the two. The next result gives an upper bound on this difference in terms of the probability of match. To simplify the notation, write $P(i) = {\rm P}(\gamma_{\rm n} = i|y)$ and $P_\pi(i) = {\rm P}_\pi(\gamma_{\rm n} = i|y)$, $1 \le i \le m$.

Theorem 3 (Error probability bounds). The following inequalities hold:

$$\begin{aligned}
0 &\le {\rm P}(\hat{\gamma}_{{\rm n},\pi} \neq \gamma_{\rm n}) - {\rm P}(\gamma_{\rm n}^* \neq \gamma_{\rm n}) \\
&\le {\rm E}\left\{\max_{1\le j\le m} P_\pi(j) - P_\pi(\gamma_{\rm n}^*)\right\} + {\rm E}\left\{\max_{1\le j\le m} P(j) - P(\hat{\gamma}_{{\rm n},\pi})\right\}.
\end{aligned} \tag{19}$$

Intuitively, the expression inside the first expectation on the right side of (19) is the distance between the maximum pseudo posterior and pseudo posterior at the MPM; similarly, the expression inside the second expectation on the right side of (19) is the distance between the maximum posterior and posterior at the MPPM. An implication is that what matters is how close the MPPM is to maximizing the posterior, and vice versa (how close the MPM is to maximizing the pseudo posterior). For example, the pseudo posterior does not need to be close to the posterior everywhere, as long as it is close to the posterior where it peaks; in other words, the posterior and pseudo posterior peak around the same place. The proof of Theorem 3 is given in the Supplementary Material.

2.2 Consistency and asymptotic optimality when $m \to \infty$

First note that some results from the previous subsection, namely, Lemma 1 and Theorem 3, are not asymptotic, which means that they hold under fixed sample size. These results extend without change when $m \to \infty$. The theorem below extends all of the asymptotic results, namely, Theorem 1 and Theorem 2, in that it allows m to increase at a suitable rate. Define $k_* = \min_{1 \le i \le m} k_i$ and $k^* = \max_{1 \le i \le m} k_i$.

Theorem 4 (Consistency and asymptotic optimality of MPPM). Suppose that (a) $m \to \infty$ and there is $d > 4$ such that $m^{2d}/k_* = O(1)$, $m^d/k_{\mathrm{n}} = o(1)$,

$$\max_{1 \le i \ne j \le m} \left| \log \left\{ \frac{\pi(\gamma_{\mathrm{n}} = i)}{\pi(\gamma_{\mathrm{n}} = j)} \right\} \right| = O\left(\frac{k_{\mathrm{n}}}{m^d} \right), \tag{20}$$

and (13) holds; (b) $\bar{x}_{i\cdot}, 1 \le i \le m$ and $\bar{x}_{\mathrm{n}\cdot}$ are bounded; and (c) $m^d(\hat{\beta} - \beta) \xrightarrow{\mathrm{P}} 0$ for the same d, and (11) holds. Then, we have the following conclusions:

(I) $\mathrm{P}(\hat{\gamma}_{\mathrm{n},\pi} = \gamma_{\mathrm{n}}^*) \to 1$ and $\mathrm{P}(\hat{\gamma}_{\mathrm{n},\pi} = \gamma_{\mathrm{n}}) \to 1$.

(II) (18) holds for any other class matcher $\check{\gamma}_{\mathrm{n}}$ and sequence a_N satisfying (17).

Assumption (a) states, in particular, that the rate that m increases is relatively slower than that of k_{n}, and even slower than those of $k_i, 1 \le i \le m$. This is reasonable because, so far as this paper is concerned, the cases we are concerned with are such that m is much

smaller than the k_i's or $k_{\rm n}$. The key ideas of the proof are similar to those of Theorem 1, but with more careful evaluation of the bound for the log-ratio of selection probability and the lower bound for the distance between different random effects, taking into account that now $m \to \infty$. Detail of the proof is given in the Supplementary Material.

To conclude this section, we have shown (Theorem 2 and Theorem 4) that MPPM is asymptotically superior than any other class matcher, including the CMMP class matcher of Jiang *et al.* (2018) [see the new paragraph below (10)], in terms of smaller probability of mismatch under a suitable asymptotic framework. This theoretical result is fully supported by our empirical studies presented in Section 4 and Section 6. Furthermore, this is the very reason for the superior performance of PBCMMP over CMMP, which we demonstrate both theoretically and empirically in the sequel.

3 Consistency and asymptotic optimality of $\hat{\theta}_{{\rm n},\pi}$

We now switch attention to asymptotic behavior of the CMEP based on MPPM. Let $\theta_{\rm n}$ denote a mixed effect associated with some new observations that satisfy (4). The mixed effect can be expressed as $\theta_{\rm n} = x_{\rm n}'\beta + \alpha_{\gamma_{\rm n}}$. For example, $x_{\rm n}$ may be one of the observed $x_{{\rm n}j}$, in which case $\theta_{\rm n} = \theta_{{\rm n}j}$, defined above (10), but $x_{\rm n}$ can also be a value not among the new observations $y_{{\rm n}j}, x_{{\rm n}j}, 1 \le j \le k_{\rm n}$. The CMEP of $\theta_{\rm n}$, denoted by $\hat{\theta}_{{\rm n},\pi}$, is given by (10) with $x_{{\rm n}j}$ replaced by $x_{\rm n}$, i replaced by $\hat{\gamma}_{\rm n}$, the MPPM defined by (9), and β, σ^2, τ^2 replaced by $\hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2$, respectively. We first establish consistency of the CMEP based on MPPM. Later, we study asymptotic optimality of the CMEP based on MPPM under a suitable framework.

3.1 Consistency when m is bounded

Although in Jiang *et al.* (2018), the authors also established consistency of CMEP based on their proposed class matcher, the result was proved under the assumption that $m \to \infty$ (see assumption *A5* in the supplementary material of Jiang *et al.* 2018). Here, we consider

consistency of CMEP based on MPPM when m is bounded, which is also practical, for example, in some cases of network data.

Theorem 5 (consistency of CMEP when m is bounded). Suppose that $x_{\rm n}$ is bounded. Then, under the assumptions of Theorem 1, the CMEP based on MPPM is consistent, that is, $\hat{\theta}_{{\rm n},\pi} - \theta_{\rm n} \xrightarrow{\rm P} 0$. The result holds regardless of the choice of the working model π.

The proof of Theorem 5 is given in the Supplementary Material. It should be noted that, although we have established consistency of the CMEP of $\theta_{\rm n}$, there is no consistency of the corresponding classified predictor for $y_{\rm n}$, a future observation associated with $\theta_{\rm n}$. This is because $y_{\rm n}$ is subject to a new error, which has a non-vanishing variance; in other words, $y_{\rm n} = \theta_{\rm n} + \epsilon_{\rm n}$, where $\epsilon_{\rm n} \sim N(0, \tau^2)$ with $\tau^2 > 0$. Therefore, there is no way to estimate, or predict, $y_{\rm n}$ consistently due to the non-vanishing $\epsilon_{\rm n}$.

3.2 Consistency when $m \to \infty$

In this case, there are two scenarios, the matched case and unmatched case (Jiang *et al.* 2018). In the matched case, it is assumed that the true class number of the new observation, $\gamma_{\rm n}$, belongs to $\{1, \ldots, m\}$, the set of indexes associated with the training data classes. This is the case that we have considered, so far, in studying the asymptotic behaviors. Note that consistency in terms of the class matching is only possible in the matched case. In the unmatched case, $\gamma_{\rm n}$ does not belong to the above index set. There is no matching consistency, of course, in the unmatched case. Nevertheless, we can still establish consistency of the CMEP in the unmatched case. The latter result was also obtained in Jiang *et al.* (2018).

Theorem 6 (consistency of CMEP in matched case). Suppose that $x_{\rm n}$ is bounded. Then, under the assumptions of Theorem 4, the CMEP based on MPPM is consistent, that is, $\hat{\theta}_{{\rm n},\pi} - \theta_{\rm n} \xrightarrow{\rm P} 0$, regardless of the choice of the working model π.

The proof of Theorem 6 is given in the Supplementary Material.

We now consider the case that $\gamma_{\rm n} \notin \{1, \ldots, m\}$. This means that the random effect corresponding to the new observations does not match one of the random effects associated

with the training data. Such a case was considered in Jiang *et al.* (2018) and Sun *et al.* (2018). Of course, in this case, there is no consistency in terms of matching the class index; however, it was shown (e.g., Jiang *et al.* 2018) that, as long as $m \to \infty$, the CMEP of $\theta_{\rm n}$ (based on the mismatched class index) is still consistent. The rationale is that, although there is no exact match of the class index, since m is large, there is alway some α_i that comes close to $\alpha_{\gamma_{\rm n}}$, which is all that matters, so far as consistency of CMEP is concerned. The following theorem states that a similar result holds for the CMEP based on the MPPM.

Theorem 7 (consistency of CMEP in unmatched case). Suppose that $\alpha_{\gamma_{\rm n}}$ is independent with $\alpha_i, 1 \le i \le m$, and $\alpha_{\gamma_{\rm n}} = O_{\rm P}(\log m)$. Then, under the assumptions of Theorem 6, we have $\hat{\theta}_{{\rm n},\pi} - \theta_{\rm n} \xrightarrow{\rm P} 0$, that is, the CMEP based on the MPPM is consistent, regardless of the choice of the working model π.

Note that, because $\gamma_{\rm n}$ does not match any of the indexes $1 \le i \le m$, it is reasonable to assume that $\alpha_{\gamma_{\rm n}}$ is independent with $\alpha_i, 1 \le i \le m$. The additional assumption, $\alpha_{\gamma_{\rm n}} = O_{\rm P}(\log m)$, takes into account the fact that $\gamma_{\rm n}$ is considered as a random index. Note that if, instead, $\gamma_{\rm n}$ is a fixed index, then we have $\alpha_{\gamma_{\rm n}} = O_{\rm P}(1)$ provided that $\mathrm{E}(|\alpha_i|) < \infty$ for every i. To see a case where $\gamma_{\rm n}$ is random, consider the following example.

Example 1. Suppose that $\mathrm{E}(|\alpha_i|) \le B, i = 1, 2, \ldots$ for some constant $B > 0$, and that $\gamma_{\rm n} \in \{m+1, m+2, \ldots\}$ such that, given $\gamma_{\rm n} = i$, $\alpha_{\gamma_{\rm n}}$ is distributed as α_i. Then, we have $\alpha_{\gamma_{\rm n}} = O_{\rm P}(1)$. To show this, note that, for any $M > 0$, we have

$$\mathrm{P}(|\alpha_{\gamma_{\rm n}}| > M) \;=\; \sum_{i=m+1}^{\infty} \mathrm{P}(\gamma_{\rm n} = i)\mathrm{P}(|\alpha_{\gamma_{\rm n}}| > M | \gamma_{\rm n} = i), \tag{21}$$

and $\mathrm{P}(|\alpha_{\gamma_{\rm n}}| > M | \gamma_{\rm n} = i) = \mathrm{P}(|\alpha_i| > M) \le B/M$ for every $i > m$. Thus, the right side of (21) is bonded by $(B/M)\sum_{i=m+1}^{\infty} \mathrm{P}(\gamma_{\rm n} = i) = B/M$, which is arbitrarily small if M is sufficiently large. Thus, by definition (e.g., Jiang 2010, sec. 3.4), $\alpha_{\gamma_{\rm n}} = O_{\rm P}(1)$.

The proof of Theorem 7 is given in the Supplementary Material.

3.3 Asymptotic optimality

Let us focus on the matched case with $m \to \infty$. To simplify the notation, write $\gamma = \gamma_{\rm n}$, $\theta = \theta_{\rm n}$, and $\hat{\theta}_\pi = \hat{\theta}_{{\rm n},\pi}$. Our approach is to first establish a lower bound for a normalized limit of the MSPE of a classified predictor, that is, a predictor of θ, $\hat{\theta}$, based on a class matcher $\hat{\gamma}$. We then show that the normalized MSPE of $\hat{\theta}_\pi$ attains the asymptotic lower bound. In a way, the approach is similar to the Cramér-Rao lower bound and asymptotic optimality of the maximum likelihood or Bayes estimators (e.g., Lehmann and Casella 1998, ch. 6). Recall the notation b_{ij} defined in Theorem 1. Recall $\bar{y}_{i\cdot}, \bar{x}_{i\cdot}$ are defined below (10), $\bar{x}_{{\rm n}\cdot}$ is defined below (13). Also let $\bar{y}_{{\rm n}\cdot} = k_{\rm n}^{-1} \sum_{j=1}^{k_{\rm n}} y_{{\rm n}j}$, $B_{\max} = \max_{1 \le i \ne j \le m} b_{ij}$.

Theorem 8. Suppose that (11) holds and there is a constant $h > 1$ such that, as $m \to \infty$, the following are $O(1)$: $m^h {\rm E}(|\hat{\beta} - \beta|^2)$, $m^{h+1/2}/k_*$ and $m^h/k_{\rm n}$.

(I) For any classified predictor $\hat{\theta}$, and any sequence of positive numbers, b_N, satisfying $b_N \to \infty$ and $b_N/m^{(2h-1)/4} \to 0$, we have

$$\begin{aligned} & \liminf \{ b_N {\rm E}(\hat{\theta} - \theta)^2 \} \\ \ge \; & \liminf \left\{ b_N {\rm E} \left[\left(\min_{1 \le i \le m} d_i^2 \right) \left\{ 1 - \max_{1 \le j \le m} {\rm P}(\gamma = j | y) \right\} \right] \right\}, \end{aligned} \tag{22}$$

where $d_i = \bar{y}_{i\cdot} - \bar{y}_{{\rm n}\cdot} - (\bar{x}_{i\cdot} - \bar{x}_{{\rm n}\cdot})'\beta$, $1 \le i \le m$.

(II) If, in addition, we have $(k^*/k_{\rm n})^2 {\rm E}(|\hat{\beta} - \beta|^2) = O(m^{-g_1})$ for some $g_1 > (2h-1)/4$, and there is $g > 2h + 3$ such that $B_{\max}(m^g/k_{\rm n}) = o(1)$, $m^{g_2}/(k_* \wedge k_{\rm n}) = O(1)$ and $(k_{\rm n}/k_*) m^{g_3} = O(1)$ for some $g_2 > 2g + (5/2)h + 7/4$ and $g_3 > g + h/2 + 7/4$. Then, for any sequence b_N satisfying the conditions in (I), we have

$$\begin{aligned} & \liminf \{ b_N {\rm E}(\hat{\theta}_\pi - \theta)^2 \} \\ = \; & \liminf \left\{ b_N {\rm E} \left[\left(\min_{1 \le i \le m} d_i^2 \right) \left\{ 1 - \max_{1 \le j \le m} {\rm P}(\gamma = j | y) \right\} \right] \right\}. \end{aligned} \tag{23}$$

The result holds regardless of the choice of π.

Note that, if $\hat{\beta}$ is a (consistent) estimator of β based on the training data, its accuracy, or effective sample size, is typically much higher than the number of classes, m. In fact,

under some regularity conditions, we have $\mathrm{E}(|\hat{\beta} - \beta|^2) = O(n^{-1})$, where $n = \sum_{i=1}^{m} k_i$ is the total sample size. Therefore, the conditions involving $\mathrm{E}(|\hat{\beta} - \beta|^2)$ are expected to hold. The rest of the conditions are regarding the relative rates that m, k_* and k_{n} increase. For the most part, it requires that k_* and k_{n} increase at a (much) faster rate than m, and k_* increases at a faster rate than k_{n}. The conditions seem reasonable, at least, in applications to network data. The proof of Theorem 8 is given in the Supplementary Material.

4 Simulation studies

We carry out a series of Monte-Carlo simulation studies to investigate finite-sample performance of the proposed PBCMMP method and compare it with the existing methods, including the CMMP method of Jiang *et al.* (2018) and the standard regression prediction (RP) method. We begin with an example under the same simulation setting of Jiang *et al.* (2018). We then study a number of more complex situations, including more covariate predictors, different working models, and noise in random effects so that an exact match does not exist. The section is concluded with a summary.

4.1 An example under the setting of Jiang *et al.* (2018)

We consider an example studied by Jiang *et al.* (2018) based the following model:

$$y_{ij} \quad = \quad 1 + 2x_{1,ij} + 3x_{2,ij} + \alpha_i + \epsilon_{ij}, \tag{24}$$

$i = 1, \ldots, m, j = 1, \ldots, k$, where α_i's and ϵ_{ij}'s are independent with $\alpha_i \sim N(0, G)$, $\epsilon_{ij} \sim N(0, 1)$; $x_{k,ij}, k = 1, 2$ are generated from $N(0, 1)$, then fixed throughout the simulation.

Suppose that a new subject satisfies the same NER model:

$$y_{\mathrm{n}j} \quad = \quad 1 + 2x_{1,\mathrm{n}} + 3x_{2,\mathrm{n}} + \alpha_{\gamma_{\mathrm{n}}} + \epsilon_{nj}, \tag{25}$$

$j = 1, \ldots, k_{\mathrm{n}}$, where $x_{k,\mathrm{n}}, k = 1, 2$ are generated from $N(0, 1)$, and fixed in the simulation.

Let $c_{\rm n} = 1$, that is, the new subject is thought to belong to the first community ($i = 1$), but there is a chance that this may be wrong. The true index, $\gamma_{\rm n}$, satisfies (3), where the true value of p is 0.85. However, we pretend that this is unknown, and two proposed values of p are considered: $0.75, 0.9$. The following combinations of sample sizes are considered: $m = 10, k = 10, 50, 100$; $m = 20, k = 50, 100, 200, 400$. For each of the combinations, we consider $G = 0.1, 1$; $k_{\rm n} = 1, 10$ for the cases of $m = 10$, and $k_{\rm n} = 1, 10, 50$ for the cases of $m = 20$, resulting a total of 36 combinations of $m, k, G, k_{\rm n}$.

There are two objectives of interest: Identification of the true index, $\gamma_{\rm n}$, and prediction of the true mixed effect, $\theta_{\rm n} = 1 + 2x_{1,{\rm n}} + 3x_{2,{\rm n}} + \alpha_{\gamma_{\rm n}}$. As in Jiang *et al.* (2018), unknown parameters are replaced by their REML estimators based on the training data. We run 100 simulations under each combination of $m, k, G, k_{\rm n}$, and p values specified above, and report (i) the empirical MSPE: $\mathrm{E}(\hat{\theta}_{\rm n} - \theta_{\rm n})^2$, where $\hat{\theta}_{\rm n}$ corresponds to PBCMMP, CMMP, or RP, based on the simulation runs; (ii) ratio of the empirical MSPEs, that is, $\mathrm{R}_{\rm c/p}$, which is the empirical MSPE of CMMP divided by that of PBCMMP, and $\mathrm{R}_{\rm r/p}$, which is the empirical MSPE of RP divided by that of PBCMMP; and (iii) empirical probability of correct matching, that is, proportion of times that the class index is matched correctly, for PBCMMP ($\mathrm{PCM}_{\rm p}$) and CMMP ($\mathrm{PCM}_{\rm c}$). The results for $p = 0.75$ are presented in Table 1; the results for $p = 0.90$ are deferred to the Supplementary Material. The numbers in the parentheses are empirical standard deviations for the empirical MSPE.

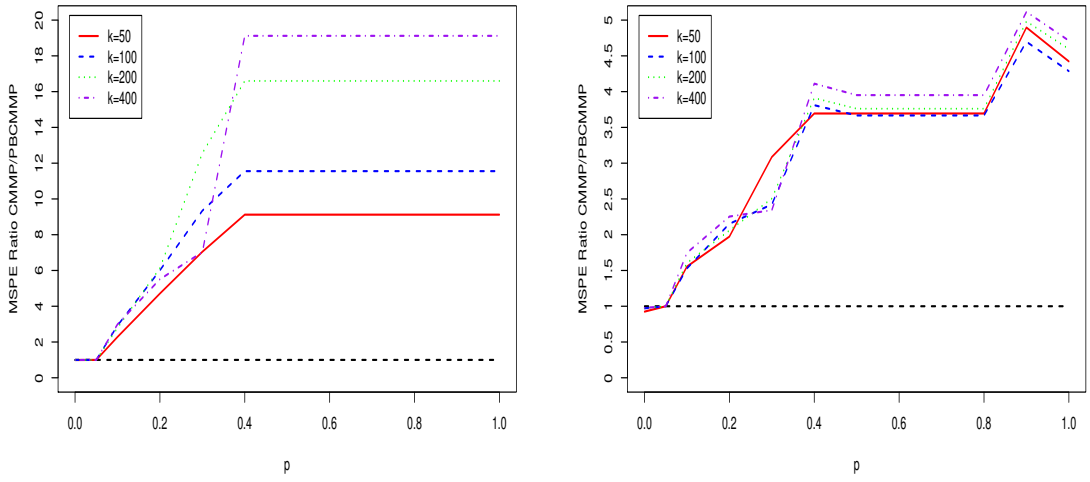
The ratio of MSPE is a measure of relative efficiency comparing two predictors, with the ratio greater than 1 indicating the denominator method (i.e., PBCMMP) is more efficient than the numerator method (CMMP or RP). It is seen that the majority of the ratios for CMMP are greater than 1, some much greater than 1; and all of the ratios for RP are greater than 1, some much greater than 1. This suggests that, overall, PBCMMP performs significantly better than CMMP and RP in terms of the predictive performance. It is interesting to note that CMMP does not always perform better than RP, while PBCMMP does. It is also seen that, in terms of probability of correct matching, PBCMMP performs much

Table 1: Empirical MSPE, Ratio of MSPE, and Probability of Correct Matching (p=0.75)

			Empirical MSPE			**Ratio of MSPE**		**Prob. of Correct Matching**	
	G	k_n	PBCMMP	CMMP	RP	$\mathrm{R_{c/p}}$	$\mathrm{Rr/p}$	$\mathrm{PCM_p}$	$\mathrm{PCM_c}$
m=10,k=10	0.1	1	0.132(0.022)	0.251(0.037)	0.164(0.021)	1.90	1.24	0.85	0.21
	0.1	10	0.131(0.021)	0.091(0.012)	0.164(0.021)	0.70	1.25	0.83	0.26
	1	1	0.387(0.082)	0.798(0.125)	1.140(0.149)	2.06	2.95	0.84	0.23
	1	10	0.193(0.031)	0.099(0.015)	1.140(0.149)	0.52	5.91	0.83	0.35
m=10,k=50	0.1	1	0.038(0.006)	0.257(0.036)	0.111(0.014)	6.79	2.93	0.85	0.13
	0.1	10	0.040(0.006)	0.064(0.009)	0.111(0.014)	1.61	2.80	0.84	0.20
	1	1	0.246(0.077)	0.857(0.126)	1.018(0.128)	3.48	4.14	0.85	0.19
	1	10	0.063(0.017)	0.089(0.014)	1.018(0.128)	1.40	16.10	0.86	0.36
m=10,k=100	0.1	1	0.027(0.006)	0.276(0.037)	0.103(0.013)	10.38	3.86	0.85	0.10
	0.1	10	0.030(0.007)	0.070(0.009)	0.103(0.013)	2.34	3.43	0.84	0.18
	1	1	0.219(0.074)	0.868(0.130)	0.989(0.124)	3.97	4.52	0.85	0.19
	1	10	0.052(0.014)	0.087(0.015)	0.989(0.124)	1.65	18.86	0.85	0.40
m=20,k=50	0.1	1	0.033(0.007)	0.303(0.039)	0.124(0.016)	9.12	3.72	0.85	0.07
	0.1	10	0.037(0.008)	0.070(0.010)	0.124(0.016)	1.88	3.31	0.84	0.13
	0.1	50	0.029(0.004)	0.021(0.003)	0.124(0.016)	0.73	4.32	0.85	0.19
	1	1	0.199(0.073)	0.734(0.114)	1.147(0.142)	3.69	5.78	0.85	0.14
	1	10	0.065(0.015)	0.093(0.013)	1.147(0.142)	1.43	17.68	0.86	0.27
	1	50	0.029(0.005)	0.025(0.004)	1.147(0.142)	0.87	40.18	0.88	0.33
m=20,k=100	0.1	1	0.027(0.005)	0.308(0.038)	0.113(0.013)	11.55	4.23	0.85	0.09
	0.1	10	0.025(0.005)	0.073(0.011)	0.113(0.013)	2.96	4.56	0.85	0.16
	0.1	50	0.020(0.003)	0.017(0.002)	0.113(0.013)	0.85	5.73	0.86	0.24
	1	1	0.199(0.072)	0.729(0.112)	1.134(0.137)	3.67	5.71	0.85	0.16
	1	10	0.066(0.017)	0.095(0.014)	1.134(0.137)	1.45	17.29	0.85	0.26
	1	50	0.014(0.002)	0.023(0.005)	1.134(0.137)	1.67	80.69	0.91	0.45
m=20,k=200	0.1	1	0.021(0.005)	0.344(0.042)	0.117(0.014)	16.60	5.65	0.85	0.09
	0.1	10	0.019(0.005)	0.075(0.011)	0.117(0.014)	3.82	6.00	0.85	0.16
	0.1	50	0.013(0.003)	0.015(0.003)	0.117(0.014)	1.22	9.31	0.86	0.24
	1	1	0.203(0.079)	0.764(0.120)	1.154(0.138)	3.76	5.69	0.85	0.13
	1	10	0.043(0.013)	0.084(0.011)	1.154(0.138)	1.97	27.08	0.86	0.29
	1	50	0.008(0.001)	0.020(0.003)	1.154(0.138)	2.46	141.52	0.93	0.45
m=20,k=400	0.1	1	0.018(0.005)	0.338(0.042)	0.116(0.014)	19.12	6.55	0.85	0.06
	0.1	10	0.016(0.005)	0.071(0.009)	0.116(0.014)	4.32	7.02	0.85	0.13
	0.1	50	0.009(0.002)	0.019(0.003)	0.116(0.014)	2.16	12.95	0.86	0.27
	1	1	0.195(0.076)	0.770(0.117)	1.159(0.141)	3.95	5.94	0.85	0.10
	1	10	0.038(0.013)	0.091(0.012)	1.159(0.141)	2.41	30.77	0.86	0.26
	1	50	0.004(0.001)	0.017(0.003)	1.159(0.141)	3.77	259.27	0.88	0.42

better than CMMP, which may explain the better predictive performance of PBCMMP.

Table 1 of the Supplementary Material shows very much the same pattern for $p = 0.90$. In fact, similar pattern is found much more broadly, not just for the values of p presented here. For example, Figure 1 presents two plots of the ratio of the empirical MSPE of CMMP over that of PBCMMP, that is, $\mathrm{R_{c/p}}$, for the cases of $m = 20, k_{\mathrm{n}} = 1$. It is seen that, as long as p is not very small, the raio is (well) above 1, especially when k is large.



(a) (b)

Figure 1: Ratio of empirical MSPE of CMMP / Empirical MSPE of PBCMMP: (a) $m = 20, k_{\mathrm{n}} = 1, G = 0.1$; (b) $m = 20, k_{\mathrm{n}} = 1, G = 1$

4.2 More covariate predictors

In this subsection, we extend the model in the previous subsection by adding more covariates. The extended model can be expressed as

$$y_{ij} \quad = \quad 1 + x_{1,ij} + x_{2,ij} + 2x_{3,ij} + 4x_{4,ij} + \alpha_i + \epsilon_{ij}, \tag{26}$$

$i = 1, \ldots, m, j = 1, \ldots, k$, where $x_{k,ij}, k = 1, 2$ are generated from $N(0, 1)$, $x_{3,ij}$ are generated from $\mathrm{Bernoulli}(0.5)$, and $x_{4,ij}$ are generated from $U(0, 1)$. The x's are then fixed throughout the simulation. The rest of the settings are the same as in the previous subsection. The results are presented in Table 2 and Table 3 of the Supplementary Material. It is seen that the observed patterns are very much the same as those of Table 1.

4.3 Different working model

In this subsection, we consider a different type of working model than the one considered earlier. We compare PBCMMP with CMMP as well as an ideal matching strategy, in which the working model is the true distribution of γ_{n}. The latter is not available, of course, in practice, but the comparison would give us some idea on the relative efficiency of a working model compared with the true distribution (see Theorem 3).

Specifically, we let $m = 20$; the true γ_{n} is generated from the $1 + \mathrm{Binomial}(9, 0.01)$ distribution while the working model is $\pi(\gamma_{\mathrm{n}}) = 0.8$ if $\gamma_{\mathrm{n}} = 1$; $\pi(\gamma_{\mathrm{n}}) = 0.1$ if $\gamma_{\mathrm{n}} = 2$; and $\pi(\gamma_{\mathrm{n}}) = 1/180$ if $\gamma_n = 3, \cdots, 20$. Other settings are the same as in Section 4.1. The results are reported in Table 2, where BCMMP represents PBCMMP using the true distribution of γ_{n} as the working model, and $\mathrm{P_b}$, $\mathrm{P_p}$ and $\mathrm{P_c}$ denote the empirical probabilities of correct matching for BCMMP, PBCMMP and CMMP, respectively.

The big picture is quite similar to what have been observed in the previous two subsections. It is also seen that the matching probability under the true distribution of γ_{n}, which corresponds to γ_{n}^* in, say, Lemma 1, is the highest in most cases. This is consistent with Lemma 1. It is also observed that the matching probability under the working model, which corresponds to $\hat{\gamma}_{\mathrm{n},\pi}$ in, say, Theorem 2, comes quite close to that of γ_{n}^*, and the two probabilities, $\mathrm{P_b}$ and $\mathrm{P_p}$, are both much higher than $\mathrm{P_c}$. These are consistent with the theory we have established, namely, Theorem 2 and Theorem 4.

Table 2: Empirical MSPE, Ratio of MSPE, and Probability of Correct Matching

k	G	k_{n}	Empirical MSPE				Ratio of MSPE		Prob. of Correct Matching		
			BCMMP	PBCMMP	CMMP	RP	$\mathrm{R_{c/p}}$	$\mathrm{R_{r/p}}$	$\mathrm{P_b}$	$\mathrm{P_p}$	$\mathrm{P_c}$
50	0.1	1	.044(.008)	.038(.006)	.306(.039)	.114(.015)	8.11	3.01	.83	.83	.06
	0.1	10	.035(.005)	.037(.006)	.068(.010)	.114(.015)	1.86	3.10	.83	.82	.12
	0.1	50	.026(.003)	.026(.003)	.020(.003)	.114(.015)	0.76	4.34	.88	.88	.19
	1	1	.304(.100)	.259(.091)	.763(.129)	1.06(.141)	2.95	4.08	.85	.86	.14
	1	10	.037(.010)	.050(.013)	.091(.013)	1.06(.141)	1.80	20.95	.93	.90	.23
	1	50	.029(.004)	.029(.004)	.022(.003)	1.06(.141)	0.77	36.96	.93	.92	.29
100	0.1	1	.037(.007)	.035(.007)	.315(.038)	.104(.014)	8.97	2.96	.83	.83	.06
	0.1	10	.034(.007)	.034(.007)	.074(.010)	.104(.014)	2.19	3.06	.84	.84	.15
	0.1	50	.016(.002)	.016(.002)	.017(.002)	.104(.014)	1.04	6.53	.91	.91	.22
	1	1	.287(.093)	.287(.093)	.769(.124)	1.05(.138)	2.68	3.66	.85	.85	.11
	1	10	.030(.010)	.030(.010)	.089(.013)	1.05(.138)	2.95	34.73	.93	.93	.27
	1	50	.013(.002)	.013(.002)	.024(.005)	1.05(.138)	1.78	78.98	.96	.96	.41
200	0.1	1	.037(.010)	.037(.010)	.353(.042)	.111(.014)	9.60	3.00	.83	.83	.06
	0.1	10	.031(.009)	.035(.009)	.076(.011)	.111(.014)	2.19	3.17	.85	.84	.14
	0.1	50	.009(.002)	.009(.002)	.016(.003)	.111(.014)	1.79	12.51	.92	.92	.21
	1	1	.264(.092)	.302(.103)	.812(.135)	1.08(.139)	2.69	3.56	.85	.85	.09
	1	10	.021(.009)	.036(.014)	.080(.011)	1.08(.139)	2.21	29.64	.94	.92	.28
	1	50	.008(.002)	.008(.002)	.020(.003)	1.08(.139)	2.48	135.50	.97	.97	.43
400	0.1	1	.031(.009)	.031(.009)	.347(.042)	.107(.014)	11.22	3.47	.83	.83	.04
	0.1	10	.020(.006)	.029(.009)	.072(.009)	.107(.014)	2.48	3.69	.86	.84	.13
	0.1	50	.006(.002)	.006(.002)	.019(.003)	.107(.014)	3.20	17.73	.91	.90	.28
	1	1	.250(.088)	.292(.101)	.815(.133)	1.07(.142)	2.79	3.67	.85	.85	.06
	1	10	.018(.009)	.033(.014)	.091(.013)	1.07(.142)	2.81	33.00	.94	.92	.26
	1	50	.006(.002)	.005(.002)	.017(.003)	1.07(.142)	3.27	206.59	.96	.96	.45

4.4 Noise in random effect

Finally, we consider a situation where there is no exact match between the random effect associated with the new observations and a random effect associated with the training data. Specifically, a noise is added to the random effect of the new observations that an exact match does not exist. This is practical in some situations. The simulation setting is the same as that of Section 4.1 except that the true random effect of the new observations is $\alpha_{\gamma_{\mathrm{n}}} + e_{\mathrm{n}}$, where e_{n} is an additional noise that is distributed as $N(0, G/10)$. It follows that the standard deviation of the noise is about about $1/3$ of that of the true random effect. The results for $p = 0.9$ are presented in Table 3; those for $p = 0.75$ are deferred to Table 4 of the Supplementary Material. Although there is no exact match of the random effect in this case, we can still obtain empirical probability of matching the main part of the random effect, $\alpha_{\gamma_{\mathrm{n}}}$, and call it probability of approximate matching, denoted by $\mathrm{PAM_p}$ and $\mathrm{PAM_c}$, respectively, for PBCMMP and CMMP.

Once again, the results follow the same patterns that have been observed in the previous subsections; in particular, the probability of approximating match behaves similarly as the probability of correct matching that were observed previously.

4.5 Summary

To summarize this section, we focus on the comparison between PRCMMP and CMMP. There are a total of 7 cases of different scenarios, including those presented in the Supplementary Material. Denote the two cases in Section 4.1 corresponding to $p = 0.75$ and $p = 0.90$ by I-1 and I-2; the two cases in Section 4.2 corresponding to $p = 0.75$ and $p = 0.90$ by II-1 and II-2; the scenario of Section 4.3 by III; and the two cases in Section 4.4 corresponding to $p = 0.75$ and $p = 0.90$ by IV-1 and IV-2, respectively. Table 4 summarizes the mean probability of correct matching, or probability of approximate matching (results with *), and the six-number summary of the MSPE ratio, that is, the minimum, first quartile, median, mean, third quartile, and maximum of the empirical MSPE of CMMP di-

Table 3: Empirical MSPE, Ratio of MSPE, and Probability of Approximate Matching (p=0.90)

			Empirical MSPE			**Ratio of MSPE**		**Prob. of Approx. Matching**	
	G	k_n	PBCMMP	CMMP	RP	$\mathrm{R_{c/p}}$	$\mathrm{R_{r/p}}$	$\mathrm{PAM_p}$	$\mathrm{PAM_c}$
m=10,k=10	0.1	1	0.132(0.022)	0.215(0.031)	0.164(0.021)	1.63	1.24	0.85	0.21
	0.1	10	0.108(0.015)	0.103(0.014)	0.164(0.021)	0.96	1.51	0.82	0.20
	1	1	0.363(0.076)	0.729(0.098)	1.140(0.149)	2.01	3.14	0.85	0.27
	1	10	0.314(0.068)	0.274(0.046)	1.140(0.149)	0.87	3.63	0.78	0.26
m=10,k=50	0.1	1	0.038(0.006)	0.216(0.034)	0.111(0.014)	5.71	2.93	0.85	0.13
	0.1	10	0.038(0.006)	0.090(0.015)	0.111(0.014)	2.37	2.93	0.85	0.16
	1	1	0.220(0.065)	0.781(0.108)	1.018(0.128)	3.55	4.63	0.85	0.20
	1	10	0.186(0.052)	0.258(0.043)	1.018(0.128)	1.39	5.74	0.83	0.28
m=10,k=100	0.1	1	0.027(0.006)	0.226(0.029)	0.103(0.013)	8.51	3.86	0.85	0.15
	0.1	10	0.027(0.006)	0.099(0.014)	0.103(0.013)	3.72	3.86	0.85	0.22
	1	1	0.192(0.060)	0.805(0.406)	0.989(0.124)	4.19	5.15	0.85	0.19
	1	10	0.190(0.053)	0.267(0.043)	0.989(0.124)	1.40	5.20	0.83	0.31
m=20,k=50	0.1	1	0.033(0.007)	0.266(0.037)	0.124(0.016)	8.01	3.72	0.85	0.08
	0.1	10	0.033(0.007)	0.094(0.013)	0.124(0.016)	2.84	3.72	0.85	0.14
	0.1	50	0.029(0.004)	0.030(0.005)	0.124(0.016)	1.04	4.32	0.85	0.19
	1	1	0.166(0.053)	0.828(0.111)	1.147(0.141)	4.99	6.91	0.85	0.15
	1	10	0.110(0.038)	0.233(0.045)	1.147(0.141)	2.11	10.41	0.83	0.22
	1	50	0.065(0.014)	0.095(0.016)	1.147(0.141)	1.47	17.77	0.81	0.26
m=20,k=100	0.1	1	0.027(0.005)	0.269(0.033)	0.113(0.013)	10.11	4.23	0.85	0.09
	0.1	10	0.027(0.005)	0.096(0.013)	0.113(0.013)	3.60	4.23	0.85	0.17
	0.1	50	0.022(0.004)	0.029(0.004)	0.113(0.013)	1.29	5.08	0.86	0.24
	1	1	0.170(0.056)	0.824(0.108)	1.135(0.137)	4.85	6.67	0.85	0.13
	1	10	0.126(0.039)	0.239(0.044)	1.135(0.137)	1.90	9.02	0.83	0.20
	1	50	0.053(0.012)	0.094(0.014)	1.135(0.137)	1.78	21.51	0.84	0.28
m=20,k=200	0.1	1	0.021(0.005)	0.296(0.037)	0.117(0.014)	14.28	5.65	0.85	0.05
	0.1	10	0.021(0.005)	0.100(0.014)	0.117(0.014)	4.83	5.65	0.85	0.15
	0.1	50	0.014(0.004)	0.024(0.004)	0.117(0.014)	1.68	8.14	0.85	0.17
	1	1	0.166(0.059)	0.852(0.115)	1.155(0.138)	5.13	6.95	0.85	0.15
	1	10	0.139(0.048)	0.243(0.049)	1.155(0.138)	1.75	8.34	0.84	0.20
	1	50	0.048(0.012)	0.078(0.012)	1.155(0.138)	1.63	24.15	0.86	0.34
m=20,k=400	0.1	1	0.018(0.005)	0.287(0.036)	0.116(0.014)	16.24	6.55	0.85	0.07
	0.1	10	0.018(0.005)	0.095(0.013)	0.116(0.014)	5.37	6.55	0.85	0.13
	0.1	50	0.008(0.002)	0.026(0.005)	0.116(0.014)	3.20	14.24	0.86	0.21
	1	1	0.163(0.060)	0.800(0.108)	1.159(0.141)	4.89	7.09	0.85	0.16
	1	10	0.139(0.047)	0.247(0.048)	1.159(0.141)	1.78	8.37	0.84	0.19
	1	50	0.057(0.016)	0.092(0.016)	1.159(0.141)	1.62	20.39	0.82	0.28

Table 4: Mean Prob. of Matching and 6-number Summary of MSPE Ratio

	Mean Prob. of Matching		**6-number Summary of MSPE Ratio**					
Case	PBCMMP	CMMP	Min.	Q1	Median	Mean	Q3	Max.
I-1	0.86	0.22	0.52	1.57	2.38	3.95	3.85	19.12
I-2	0.86	0.22	0.43	1.40	2.09	3.96	4.75	19.12
II-1	0.86	0.22	0.60	1.46	2.49	3.95	3.89	19.05
II-2	0.86	0.22	0.57	1.31	2.45	3.96	4.88	19.05
III	0.89	0.19	0.76	1.85	2.58	3.44	3.01	11.22
IV-1	0.84*	0.19*	0.90	1.58	2.60	4.13	5.57	16.24
IV-2	0.84*	0.19*	0.87	1.63	2.61	3.97	4.92	16.24

vided by the empirical MSPE of PBCMMP.

It is seen that the minimum MSPE ratio is less than one, with the smallest about 0.43; after the first quartile (Q1), all of the summaries of the MSPE ratio are greater than one, with most of the maximum over 19. Overall, PBCMMP is seen to have significant advantage over CMMP both in terms of the probability of correct (or approximate) matching and in terms of the MSPE of the prediction.

5 Measure of uncertainty

A standard measure of uncertainty, in the context of prediction, is the MSPE. Jiang and Torabi (2020) proposed a Sumca method of MSPE estimation that is applicable to a broad range of problems involving complex predictors, including the current PBCMMP. In fact, in a similar application, Sun *et al.* (2018) has applied the Sumca method to classified mixed logistic model prediction. The method takes advantage from small area estimation (e.g., Rao and Molina 2015), where estimation of the area-specific MSPE has been extensively studied. Two of the main approaches are the Prasad-Rao linearization method (P-R; Prasad and Rao 1990) and resampling methods. See Rao and Molina (2015) for details. Sumca proposes to put together the best parts of the two approaches. Specifically, it uses analytic expression to compute a leading term of the MSPE estimator, which is similar to the P-R

method, and a Monte-Carlo method to evaluate a lower-order, bias correction term, which is similar to the resampling method. See Jiang and Torabi (2020) for details. An alternative method is double bootstrapping (DB; Hall and Maiti 2006), which is one of the resampling methods noted above. DB is known to be computationally very intensive. On the theoretical side, both Sumca and DB are known to produce second-order unbiased MSPE estimators, that is, the order of the bias is $o(m^{-1})$.

A simulation study is carried out to evaluate performance of the two MSPE estimators, Sumca and DB, under the setting of Section 4.1. We consider a case of relatively small sample sizes with $m = k = k_{\rm n} = 10$, and a case of relatively large sample sizes with $m = 20, k = 200$ and $k_{\rm n} = 50$. In each case $G = 0.1$ and $G = 1$ are considered. We use the Monte-Carlo sample size $K = 50$ for computing the Sumca estimator; for DB, we use $K_1 = K_2 = 50$ as the bootstrap sample sizes for the two stages of DB. As noted, DB is computationally very intensive. Although for Sumca we have no computational difficulties even with larger K, we intensionally keep K, K_1, K_2 the same so that the results are more comparable when computational costs are put aside (see below).

Table 5 presents the percentage relative bias defined as $\%\mathrm{RB} = 100 \times [\{\mathrm{E}(\widehat{\mathrm{MSPE}}) - \mathrm{MSPE}\}/\mathrm{MSPE}\}]$, where MSPE is the true simulated MSPE and $\mathrm{E}(\widehat{\mathrm{MSPE}})$ the simulated mean of the MSPE estimator, either Sumca or DB. The %RB is a standard measure of performance for MSPE estimation (e.g., Jiang and Torabi 2020). The results are based on 100 simulation runs.

Table 5: Empirical % RB of MSPE Estimation for PBCMMP

			$p = 0.75$		$p = 0.90$		$p = 0.95$	
	G	k_n	Sumca	DB	Sumca	DB	Sumca	DB
m=10, k=10	0.1	10	-3.11	-4.70	-7.49	-9.51	-6.90	-10.11
	1	10	-1.97	-1.36	-8.14	-5.59	-16.65	-8.85
m=20, k=200	0.1	50	-34.40	-20.65	-53.66	-57.12	-58.36	-61.93
	1	50	3.30	28.28	-5.39	22.03	-14.4	24.72

It is seen that the performance of both Sumca and DB improve as G increases. Also,

the performance of both Sumca and DB seem to get worse when m increases, especially when G is small. One explanation is that, when the sample sizes get larger, or when G gets smaller, the actual MSPE, which serves as the numerator of %RB, decreases. Therefore, it requires more accuracy in the MSPE estimation (the numerator) in order to keep the %RB small (in absolute value). As for the comparison between Sumca and DB, it is seen that Sumca performs better in most cases. One should also be reminded that DB is computationally much more expensive than Sumca. Roughly speaking, the computing time for DB is about 15 times that of Sumca for the current simulation study.

6 Prediction with Facebook network data

As a real-data validation, we apply our proposed metod to a large social network data regarding Facebook users (available at http://snap.stanford.edu/data/ego-Facebook.html). A node in the network represents a user and an edge a friendship between two users. From Facebook we obtained user profile information and network data from 10 ego-networks, consisting of 4039 nodes and 88234 edges. For each node, feature vectors have been provided and their interpretations obscured. For instance, where the original dataset may have contained a feature "political=Democratic Party", the new data would simply contain "political=anonymized feature 1". Therefore, by using the anonymized data it is possible to determine whether two users have the same political affiliations, but not what their individual political affiliations represent. In this dataset, features are '1' if the user has this property in the profile, and '0' otherwise.

The data have been analyzed by several authors; see McAuley and Leskovec (2012), Bickel and Sarkar (2016) and Ma *et al.* (2018), among others. However, the focuses were on problem of determining the number of communities or clusters, and on identifying users' social circles within the Facebook network. Here, we use the data to demonstrate the effectiveness of PBCMMP in predicting the number of friendships for a Facebook user, assuming, of course, that the latter is unknown. The PBCMMP is based on the working

model (3) with $p = 0.75$, or $p = 0.95$. Because the dimension of feature vectors for each ego-network is different, and the feature value is either 0 or 1, the proportion of features with the value being 1, that is, the number of features equal to 1 divided by the dimension of feature vector, is used as a covariate. The outcome of interest is the log-transformed number of friendships, that is, the number of edges for each node.

Table 6: Average Squared Prediction Error (Proportion of Correct Matching)

Community	Size	PBCMMP ($p = 0.75$)	PBCMMP ($p = 0.95$)	CMMP	RP
1	348	0 (1)	0 (1)	0.788 (0)	0.781
2	225	0.167 (0.957)	0 (1)	1.131 (0)	0.076
3	113	0 (1)	0 (1)	0.562 (0.083)	0.364
4	171	0 (1)	0 (1)	1.036 (0.059)	0.500
5	39	0 (1)	0 (1)	0.062 (0.250)	1.441
6	1016	0.142 (0.980)	0.071 (0.991)	0.510 (0.177)	0.082
7	749	0.229 (0.973)	0.126 (0.987)	0.636 (0.640)	0.479
8	776	0 (1)	0 (1)	0.708 (0.103)	0.021
9	543	0 (1)	0 (1)	0.908 (0.111)	0.392
10	59	0 (1)	0 (1)	0 (1)	1.499

To assess the predictive performance of PBCMMP, and its comparison with CMMP and RP, we randomly selected 10% of the data from each community (see below) as testing data; the remaining 90% of the data were used as training data. The testing data has size 404 and training data 3635. It is widely believed that there are 10 communities within the network associated with the Facebook data (e.g., McAuley and Leskovec 2012, Bickel and Sarkar 2016). Those 10 communities were used to divide the training data into $m = 10$ classes. For the testing data, however, the community information is known, but is only used as a "prior" according to the description of our PBCMMP method [see Section 1, above (2)]. Table 6 reports the average squared prediction errors, that is, the average of the squared prediction errors over the subset of testing data according to each community for three comparing methods, PBCMMP (with $p = 0.95$ and $p = 0.95$, respectively), CMMP and RP. Note that these are the true prediction errors, which is a more convincing measure of predictive performance than the empirical MSPE based on simulation studies

(see Section 4). The results clearly demonstrate the superiority of PBCMMP over CMMP and RP. It is remarkable that, in most cases, the average squared prediction error is zero (which means zero in every single case that was predicted). It is also seen that there is no difference, in most cases, under different values of p for PBCMMP, $p = 0.75$ or $p = 0.95$.

Also reported in Table 6 (in the parentheses) are proportion of correctly matched class index, for PBCMMP and CMMP. Note that, for the testing data, their class indexes are known, which are the same as their community indexes. However, we pretend that this is unknown to us. Instead, we use the working probability model, (3), with either $p = 0.75$ or $p = 0.95$ as the tuning parameter, to carry out the PBCMMP. In the end, we can find out exactly how many class indexes are correctly determined using either PBCMMP or CMMP. It is seen that the matching proportion of PBCMMP is nearly perfect; in contrast, the matching proportion of CMMP is relatively poor. This explains the difference in the average squared prediction errors between PBCMMP and CMMP, namely, PBCMMP predicts better by matching correctly. Again, there is almost no difference between different values of p for PBCMMP, $p = 0.75$ or $p = 0.95$.

It is also seen that, in terms of the predictive performance, PBCMMP, with $p = 0.95$, performs (much) better than RP for all communities; PBCMMP, with $p = 0.75$, performs (much) better than RP for most communities (8 out of 10). On the other hand, CMMP does not perform better than RP for most communities (8 out of 10). Therefore, this real-life-data example fully demonstrates the power of PBCMMP.

7 Conclusion and discussion

We have developed a pseudo-Bayes strategy called PBCMMP to significantly improve the efficiency of CMMP. The strategy has flexibility of using a working prior, typically chosen based on knowledge about the connection between classes in the training data and the potential class of the new data. The superiority of PBCMMP is demonstrated theoretically via the established theory of consistency and asymptotic optimality, both in terms of class-

matching and in terms of prediction of the mixed effect associated with the new data, and these results hold regardless of the choice of the working prior, subject to mild regularity conditions. It should be noted that such results of asymptotic analysis are rarely seen in the context of mixed effects models (e.g., Jiang 2007, McCulloch, Searle and Neuhaus 2008, Demidenko 2013, and Jiang 2017).

The theoretical results are fully supported by the results of extensive simulation studies, where we compare the finite-sample performance of PBCMMP and CMMP as well as the standard regression predictor (RP). A real-data application on the Facebook social network illustrates the striking difference in improvement of prediction accuracy via PBCMMP over CMMP and RP.

A major advantage of the new PBCMMP method over the existing CMMP methods is that it enjoys the double consistency, that is, consistency in terms of the class matching and that in terms of prediction of the mixed effect associated with the new data, and this property holds whether the number of classes in the training data, m, is bounded or not. In contrast, all of the previous CMMP methods only possess single consistency in terms of the mixed effect prediction, under the assumption that m increases. The double consistency of PBCMMP not only improves the predictive performance of CMMP, as a by-product it is also capable of correctly identifying the class index for the new observations, which in some cases may be of interest as well. Furthermore, in some cases, such as in case of network data, the number of classes, m, is fairly small, or at most moderate. The previous CMMP methods do not have guaranteed performance in such situations, while our new PBCMMP method has guaranteed performance, as we have demonstrated both theoretically and empirically. Furthermore, we have established asymptotic optimality of PBCMMP both in terms of the class matching and in terms of the prediction of the new mixed effect. This kind of theoretical results are rarely seen in the context of mixed model analysis. Basically, the asymptotic optimality assures that PBCMMP is the best class of classified predictors that one could possibly get.

One area that deserves further research, both theoretically and empirically, is regarding the MSPE estimation. Although the Sumca and DB methods are both known to be second-order unbiased, our simulation results have not found that their relative biases decrease when the sample sizes get larger. It is possible that the regularity conditions for the second-order unbiasedness of these estimators are not met in a CMMP situation; it is also possible that (much) larger number of simulation replicates are needed in order to evaluate the performance more accurately. We plan to research on such topics in future studies.

Acknowledgements. Haiqiang Ma's research is supported by NNSF of China (No. 11701235) and NSF of Jiangxi Province (No. 20171BAB211004). The research of Jiming Jiang is partially supported by the NSF grants DMS-1510219 and DMS-1713120.

References

[1] Bailey, M., Cao, R., Kuchler, T., and Stroebel, J. (2018), The economic effects of social networks: Evidence from the housing market, *J. Political Econ.* 126, 2224–2276.

[2] 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, *J. Amer. Statist. Assoc.* 80, 28–36.

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

[4] Bickel, P. J. and Sarkar, P. (2016), Hypothesis testing for automated community detection in networks, *J. R. Statist. Soc. B* 78, 253–273.

[5] Demidenko, E. (2013), *Mixed Models—Theory and Application with R*, 2nd ed., Wiley, New York.

[6] Gleiser, P. and Danon, L. (2003), Community structure in jazz, *Adv. Complex Syst.* 6, 565–573.

[7] Hall, P. and Maiti, T. (2006), Nonparametric estimation of mean-squared prediction error in nested-error regression models, *Ann. Stat.* 34, 1733-1750.

[8] Jiang, J. (1999), Conditional inference about generalized linear mixed models, *Ann. Statist.* 27, 1974–2007.

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

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

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

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

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

[14] Jiang, J. and Torabi, M. (2020), Sumca: Simple, unified, Monte-Carlo-assisted approach to second-order unbiased mean-squared prediction error estimation, *J. Roy. Statist. Soc. Ser. B* 82, 467–485.

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

[16] Keeling, M. J. and Eames, K. T. D. (2005), Network and epidemic models, *J. R. Soc. Interface* 2, 295–307.

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

[18] Lehmann, E. L. and Casella, G. (1998), *Theory of Point Estimation*, 2nd ed., Springer, New York.

[19] Li, C., Shen, X. and Pan, W. (2019), Likelihood inference for a large causal network, *J. Amer. Statist. Assoc.*, in press.

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

[21] Ma, S., Su, L. and Zhang, Y. (2018), Determining the number of communities in degree-corrected stochastic block models, *arXiv*: 1809.01028.

[22] McAuley, J. and Leskovec, J. (2012), Learning to discover social circles in ego networks, *NIPS* 1, 539–547.

[23] McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008), *Generalized, Linear, and Mixed Models*, 2nd ed., Wiley, Hoboken, NJ.

[24] Newman, M. E. J. (2006), Modularity and community structure in networks, *PNAS* 103, 8577–8582.

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

[26] 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.

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

[28] Sun, H., Nguyen, T., Luan, Y., and Jiang, J. (2018), Classified mixed logistic model prediction, *J. Multivariate Anal.* 168, 63–74.