The Canadian Journal of Statistics
Vol. 00, No. 00, 2021, Pages 1–24
La revue canadienne de statistique

Classified Generalized Linear Mixed Model Prediction Incorporating Pseudo-prior Information

Haiqiang Ma[1] and Jiming Jiang[2]*

[1]*Jiangxi University of Finance and Economics, Nanchang, China*

[2]*University of California, Davis, USA*

Key words and phrases: Consistency; CMMP; generalized linear mixed models; matching; network communities ; prediction ; random effects.

MSC 2010: Primary 62F15; secondary 62G20

Abstract: We develop a method of classified mixed model prediction based on generalized linear mixed models that incorporates pseudo-prior information to improve prediction accuracy. We establish consistency of the proposed method both in terms of prediction of the true mixed effect of interest and in terms of correctly identifying the potential class corresponding to the new observations, if such a class matching one of the training data classes exists. Empirical results, including simulation studies and real-data validation, fully support the theoretical findings. *The Canadian Journal of Statistics* 00: 1–24; 2021 © 2021 Statistical Society of Canada

Résumé: Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* 00: 1–24; 2021 © 2021 Société statistique du Canada

1. INTRODUCTION

Prediction has had a long history in statistics. New and challenging problems are now emerging from such fields as precision medicine, public health and business, in which one may substantially improve prediction accuracy regarding a charac-

* *Author to whom correspondence may be addressed.*
E-mail: Insert your email address here only after your paper has been accepted

© 2021 Statistical Society of Canada / Société statistique du Canada

teristic of interest at the subject level, if one could identify a class that the subject is associated with. Here, a subject may be an individual, or a group of individuals sharing similar characteristics. This new type of prediction problem was first considered by Jiang et al. (2018), who proposed a method called classified mixed model prediction (CMMP). Under the CMMP setting, the class mentioned above corresponds to a random effect associated with some (massive) training data under a mixed effects model. 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 traditional mixed model prediction method (MMP; e.g., Jiang & Nguyen 2021, sec. 2.3) can be utilized to make accurate prediction that takes into account the subject-level differences.

The CMMP method of Jiang et al. (2018) was developed for linear mixed models (LMM; e.g., Jiang & Nguyen 2021), and applies to continuous responses only. There are many problems of practical interest involving discrete or categorical responses, for which the CMMP concept is potentially useful, if an extension can be made to generalized linear mixed models (GLMMs; e.g., Jiang & Nguyen 2021). A special case of such an extension was considered by Sun et al. (2018) in the case of binary responses. There are, of course, other types of non-normal responses, for which prediction problems are of interest.

In standard regression prediction, a mean response is estimated as a linear function of observed covariates and the estimated mean function is used to predict a future observation. However, given the values of the covariates, there may still be substantial variation not explained by the covariates. In regression analysis, the unexplained variation is treated as the "regression errors", and is typically assumed to be homoscedastic across the whole population. The latter assumption is often violated in practice. There is a rich literature on capturing heteroscedasticity within the population to improve prediction accuracy. This is known as

MMP (e.g., Jiang, 2007, sec. 2.3). Under a mixed effects model, the population is divided into subpopulations, and there is a random effect (possibly vector-valued) corresponding to each subpopulation. The random effect is estimated, or predicted, using the best predictor (BP), in the sense of minimum mean squared prediction error (MSPE).

More specifically, the matching strategy of Jiang et al. (2018) and Sun et al. (2018) is to minimize the distance between an observed characteristic of the subject and a predicted one under the mixed effects model, assuming that the subject belongs to a given class. The minimization is over all of the training data classes. It should be noted that, although the CMMP procedure involves a procedure of matching or classifying the new observations, the primary interest is not classification. It is the prediction of the mixed effect associated with the new observations that is of main interest. Such a mixed effect may correspond to a mean or probability of interest regarding the new observations. The matching or classification is merely a tool to achieve better prediction of the mixed effect. In fact, as shown by Jiang et al. (2018), even if the class of the new observations is misspecified, such a "false" classification still helps to improve the prediction accuracy. The rationale is that, even if the matching is not exactly correct, it still finds a class, among the training data classes, that is close to the true class of the new observations in terms of the corresponding mixed effect.

Sun, Luan & Jiang (2020) proposed a new matching strategy that incorporates covariate information. It is noted (Jiang et al., 2018; remark following Theorem 2 therein) that such a matching is not consistent in terms of identifying the true class index with a probability tending to one as the number of classes, m, increases. Nevertheless, as noted above, CMMP is consistent in terms of predicting a mixed effect of interest, even if the class index is mismatched, provided that $m \to \infty$. However, when m is relatively small, or only moderately large, as in many practical situations, the precision of CMMP can be substantially improved

with an improved method of matching.

For example, an important measure of progression of the 2019 coronavirus disease (Covid-19) is the daily number of newly infected persons in certain geographic areas or subpopulations. Specifically, we were able to obtain cumulative numbers of infected persons for every city in the 30 provinces and municipalities in China, along with auxiliary variables that potentially contribute to the disease infection. Because the responses are counts, it is natural to consider a Poisson log-linear mixed model with random effects at the province level, for the training data. Now suppose that one wishes to predict the cumulative number of infected persons in a new city, given the auxiliary information. The prediction accuracy may be perhaps improved significantly if one can identify the province that hosts the city. See below for details.

In general, the geographic area may be a country, or a region (e.g., province, state, or county) within a country; the subpopulation may be characterized by demographic variables such as sex, ethnicity and age. The latter may be of interest to health officials or researchers studying the disease. It should be noted that the daily numbers refer to the true, unobserved totals of newly infected persons on a certain day, not the ones reported by the news media or health officials based on the reported cases. It is known that community network plays an important role in the spread of infectious disease; in fact, such network information has been used in improving prediction accuracy in precision epidemiology for infectious disease control (e.g., Keeling & Eames, 2005; Ladner et al., 2019). Recent advances in statistical analysis of network data (e.g., Bickel & Chen 2009; McAuley & Leskovec, 2012; Bickel & Sarkar 2016; Li, Shen & Pan, 2019) have set up a basis for inference about the networks.

A key idea of this paper is to utilize such network information to improve the precision of matching. As we shall demonstrate, this idea leads to a new method which not only applies more broadly to non-continuous responses, it

also has better performance than the existing methods. In particular, we establish consistency of class matching when m, the number of classes in the training data, is either bounded or increasing, and consistency in terms of prediction of the mixed effect when m is bounded. We also obtain the rate of convergence for the prediction of the mixed effect. Not all of these theoretical results were available in the previous work of Jiang et al. (2018) and Sun et al. (2020).

In Section 2, a new matching strategy is proposed based on a pseudo prior. This is proposed under the framework of GLMM, so the resulting CMMP applies, in particular, to situations of binary responses or counts. We call the new method network-classified mixed model prediction, or NCMMP, because the method can be motivated naturally by networks. A real-data example of Covid-19 data is introduced and revisited later. In Section 3, we develop asymptotic theory for NCMMP, including its consistency properties both in terms of prediction of the mixed effect and in terms of the class matching. The consistency of class-matching holds, of course, only in the matched case, that is, when there exists a match between the class of the new observations and a class in the training data, but the consistency in mixed-effect prediction holds regardless of the true matching status (matched or unmatched). In Section 4, we investigate finite-sample performance of NCMMP via Monte-Carlo simulation, and demonstrate its advantage over existing methods. The real-data example is revisited, and used as validation. Proofs and additional details are deferred to the supplementary material.

2. CLASSIFIED GLMM PREDICTION WITH A PSEUDO PRIOR

2.1. Method

Under the GLMM, it is assumed that, conditional on vectors of class-specific random effects, $\alpha_i = (\alpha_{ij})_{1\le j\le q}$, $1 \le i \le m$, responses $y_{ij}, 1 \le j \le k_i, 1 \le i \le$

m are independent with conditional density

$$f(y_{ij}|\alpha) = f(y_{ij}|\alpha_i) = \exp\left[\left(\frac{a_{ij}}{\phi}\right)\{y_{ij}\xi_{ij} - b(\xi_{ij})\} + c\left(y_{ij}, \frac{\phi}{a_{ij}}\right)\right], \quad (1)$$

where $\alpha = (\alpha_i)_{1\le i\le m}$, $b(\cdot)$ and $c(\cdot,\cdot)$ are functions associated with the exponential family (McCullagh & Nelder, 1989, ch. 2), ϕ is a dispersion parameter, a_{ij} is a weight such that $a_{ij} = 1$ for ungrouped data; $a_{ij} = l_{ij}$ for grouped data when the average is considered as response, and l_{ij} is the group size; and $a_{ij} = l_{ij}^{-1}$ when the sum of individual responses is considered. Furthermore, ξ_{ij} is associated with a linear function, $\eta_{ij} = \mathbf{x}_{ij}'\beta + \mathbf{z}_{ij}'\alpha_i$, through a link function $g(\cdot)$; that is, $g(\xi_{ij}) = \eta_{ij}$, or $\xi_{ij} = h(\eta_{ij})$, where $h = g^{-1}$. Here $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are known vectors, and β is a vector of unknown fixed effects. For simplicity, we focus on the case of a canonical link, that is, $\xi_{ij} = \eta_{ij}$. Finally, suppose that $\alpha_1, \ldots, \alpha_m$ are independent and distributed as $N(0, \mathbf{G})$, where the covariance matrix $\mathbf{G}$ depends on a vector γ of variance components, that is, $\mathbf{G} = \mathbf{G}(\gamma)$. Let $\psi = (\beta', \gamma')'$, and $\vartheta = (\psi', \phi)$. Note that in some cases, such as binomial or Poisson distributions, the dispersion parameter ϕ is known, so ψ is the vector of unknown parameters.

We assume that the above GLMM holds for the training data, $y_{ij}, \mathbf{x}_{ij}, 1 \le j \le k_i, 1 \le i \le m$, where the classes $1, \ldots, m$ correspond to known network communities. There have been extensive studies on community detection in networks; see, for example, Bickel & Chen (2009), Bickel & Sarkar (2016), and Ma, Su & Zhang (2018). Thus, without loss of generality, we assume that the network communities are known for the training data.

Furthermore, suppose that there are observations, $y_{\mathrm{n}j}, \mathbf{x}_{\mathrm{n}j}, 1 \le j \le k_{\mathrm{n}}$ that correspond to a new subject. The new observations are assumed to satisfy a similar GLMM, that is, $y_{\mathrm{n}j}, 1 \le j \le k_{\mathrm{n}}$ are conditionally independent given α_I with conditional density

$$f(y_{\mathrm{n}j}|\alpha_I) = \exp\left[\left(\frac{a_{\mathrm{n}j}}{\phi}\right)\{y_{\mathrm{n}j}\xi_{\mathrm{n}j} - b(\xi_{\mathrm{n}j})\} + c\left(y_{\mathrm{n}j}, \frac{\phi}{a_{\mathrm{n}j}}\right)\right], \quad (2)$$

where $\xi_{\mathrm{n}j} = \mathbf{x}'_{\mathrm{n}j}\beta + \mathbf{z}'_{\mathrm{n}j}\alpha_I$. Here, the subscript n stands for "new", and I represents an unknown index, which may or may not belong to $\{1, \ldots, m\}$. As in CMMP, a first step is to identify the index I.

Suppose that the new subject belongs to a known community c_{n}. The true index I, however, is not entirely determined by c_{n}. This may happen, for example, when the training data are well studied; therefore, one is certain about the classes in the training data, but the data corresponding to the new subject are "new" so there is uncertainty about the class index, I, even though c_{n} is known. Sometimes, the training data were collected from a past period of time; the network has since grown bigger, or smaller. It is also possible that the training network is not exactly the same as the one relevant to the new subject. We illustrate below using a real-data example.

Due to such concerns, we consider a working probability model described as follows. First, we allow the words "class" and "community" to not necessarily mean the same in that the former corresponds to the random effect while the latter to the network. For the training data, however, the classes match the communities, by our assumption, but this is not necessarily the case for the new subject. For now, let us assume that $I \in \{1, \ldots, m\}$. This is called the matched case. Later we also consider the case that $I \notin \{1, \ldots, m\}$, called the unmatched case (Jiang et al., 2018). Consider the following working probability model:

$$\mathrm{P}(I = i) = p^{1(i=c_{\mathrm{n}})} \left(\frac{1-p}{m-1}\right)^{1_{(i\neq c_{\mathrm{n}})}}, \tag{3}$$

where p is a given probability (see below); in other words, we have

$$\mathrm{P}(I = i) = p \text{ if } i = c_{\mathrm{n}}, \quad \mathrm{P}(I = i) = \frac{1-p}{m-1} \text{ if } i \neq c_{\mathrm{n}}.$$

It is easy to verify that (3) is a probability distribution on $\{1, \ldots, m\}$. We call (3) a working model because, unlike the GLMM, model (3) does not have to hold. In particular, the p in (3) is 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 I. We call working model (3) a *pseudo prior* due to its similarity to the Bayesian prior. Large sample theory, established later in this paper, shows that, as long as there is sufficient data information, it does not really matter what p is chosen, or whether or not (3) holds.

Let $I = i$ for some $1 \leq i \leq m$. Then, one can combine the new data with the ith class of the training data so that the following groups are independent: $\mathbf{y}_1, \ldots, \mathbf{y}_{i-1}, (\mathbf{y}_i, \mathbf{y}_{\rm n}), \mathbf{y}_{i+1}, \ldots, \mathbf{y}_m$, where $\mathbf{y}_i = (y_{ij})_{1 \leq j \leq k_i}$ and $\mathbf{y}_{\rm n} = (y_{{\rm n}j})_{1 \leq j \leq k_{\rm n}}$. The conditional pdf of $\mathbf{y}_u$ $(u \neq i)$ is given by

$$\begin{aligned} f(\mathbf{y}_u|\alpha) &= f(\mathbf{y}_u|\alpha_u) \\ &= \prod_{j=1}^{k_u} \exp\left[\left(\frac{a_{uj}}{\phi}\right)\{y_{uj}\xi_{uj} - b(\xi_{uj})\} + c\left(y_{uj}, \frac{\phi}{a_{uj}}\right)\right]. \end{aligned} \tag{4}$$

Similarly, given $I = i$, the joint conditional pdf of $(\mathbf{y}_i, \mathbf{y}_{\rm n})$ is given by

$$\begin{aligned} f(\mathbf{y}_i, \mathbf{y}_{\rm n}|\alpha) &= f(\mathbf{y}_i, \mathbf{y}_{\rm n}|\alpha_i) \\ &= f(\mathbf{y}_i|\alpha_i) f(\mathbf{y}_{\rm n}|\alpha_i) = \left\{\prod_{j=1}^{k_i} f(y_{ij}|\alpha_i)\right\}\left\{\prod_{j=1}^{k_{\rm n}} f(\mathbf{y}_{{\rm n}j}|\alpha_i)\right\}. \end{aligned} \tag{5}$$

Combining the above results, we obtain that, given $I = i$,

$$f(\mathbf{y}|\alpha) = f(\mathbf{y}_i, \mathbf{y}_{\rm n}|\alpha) \prod_{u \neq i} f(\mathbf{y}_u|\alpha) = f(\mathbf{y}_i, \mathbf{y}_{\rm n}|\alpha_i) \prod_{u \neq i} f(\mathbf{y}_u|\alpha_u)$$

with $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_m', \mathbf{y}_{\rm n}')'$. The above expression may be viewed as $f(\mathbf{y}|\alpha, I = i)$. Suppose that I and α are independent so that $f(\alpha|I = i) = f(\alpha)$, hence

$$\begin{aligned} & f(\mathbf{y}|I = i) \\ &= \int_{-\infty}^{\infty} f(\mathbf{y}|\alpha, I = i) f(\alpha) d\alpha \\ &= \int_{-\infty}^{\infty} f(\mathbf{y}_i, \mathbf{y}_{\rm n}|\alpha_i) f(\alpha_i) d\alpha_i \left\{\prod_{u \neq i} \int_{-\infty}^{\infty} f(\mathbf{y}_u|\alpha_u) f(\alpha_u) d\alpha_u\right\}. \end{aligned} \tag{6}$$

From (3), (6), we obtain the "posterior" distribution of I:

$$\mathrm{P}(I=i|\mathbf{y}) = \frac{\mathrm{P}(I=i)f(\mathbf{y}|I=i)}{\sum_{v=1}^{m}\mathrm{P}(I=v)f(\mathbf{y}|I=v)}. \tag{7}$$

The matching of I to the training data class is chosen as the "posterior mode":

$$\begin{aligned}\hat{I} &= \mathrm{argmax}_{1\le i\le m}\mathrm{P}(I=i|\mathbf{y}) \\ &= \mathrm{argmax}_{1\le i\le m}\left\{\mathrm{P}(I=i)f(\mathbf{y}|I=i)\right\} \\ &= \mathrm{argmax}_{1\le i\le m}\left\{\mathrm{P}(I=i)\frac{\int_{-\infty}^{\infty}f(\mathbf{y}_i,\mathbf{y}_{\mathrm{n}}|\alpha_i)f(\alpha_i)d\alpha_i}{\int_{-\infty}^{\infty}f(\mathbf{y}_i|\alpha_i)f(\alpha_i)d\alpha_i}\right\},\end{aligned} \tag{8}$$

using (6). Note that some factors are not needed in obtaining $\hat{I}$. Also, we have $f(\alpha_i) = \{(2\pi)^{\frac{q}{2}}|\mathbf{G}(\gamma)|^{\frac{1}{2}}\}^{-1}\exp\{-(1/2)\alpha_i'\mathbf{G}^{-1}(\gamma)\alpha_i\}$, and, by the conditional exponential family assumption, we have

$$\begin{aligned}f(\mathbf{y}_i|\alpha_i) = \exp&\left[\frac{1}{\phi}\sum_{j=1}^{k_i}a_{ij}\{y_{ij}(\mathbf{x}_{ij}'\beta+\mathbf{z}_{ij}'\alpha_i)-b(\mathbf{x}_{ij}'\beta+\mathbf{z}_{ij}'\alpha_i)\}\right] \\ &\times\exp\left\{\sum_{j=1}^{k_i}c\left(y_{ij},\frac{\phi}{a_{ij}}\right)\right\},\end{aligned} \tag{9}$$

$$\begin{aligned}f(\mathbf{y}_i,\mathbf{y}_n|\alpha_i) = \exp&\left[\frac{1}{\phi}\sum_{j=1}^{k_i}a_{ij}\{y_{ij}(\mathbf{x}_{ij}'\beta+\mathbf{z}_{ij}'\alpha_i)-b(\mathbf{x}_{ij}'\beta+\mathbf{z}_{ij}'\alpha_i)\}\right] \\ &\times\exp\left[\frac{1}{\phi}\sum_{j=1}^{k_{\mathrm{n}}}a_{nj}\{\mathbf{y}_{\mathrm{n}j}(\mathbf{x}_{\mathrm{n}j}'\beta+\mathbf{z}_{\mathrm{n}j}'\alpha_i)-b(\mathbf{x}_{\mathrm{n}j}'\beta+\mathbf{z}_{\mathrm{n}j}'\alpha_i)\}\right] \\ &\times\exp\left\{\sum_{j=1}^{k_i}c\left(y_{ij},\frac{\phi}{a_{ij}}\right)+\sum_{j=1}^{k_{\mathrm{n}}}c\left(y_{\mathrm{n}j},\frac{\phi}{a_{\mathrm{n}j}}\right)\right\}.\end{aligned} \tag{10}$$

Intuitively, the matching procedure proposed above may be viewed as maximum a posteriori. It should be noted that, although the matching procedure is derived assuming that $I \in \{1,\ldots,m\}$, it applies to any case in practice regardless of the true matching status (matched or unmatched), which is unknown.

Once $\hat{I}$ is determined, the prediction of the new mixed effect is carried out. Consider prediction of a mixed effect associated with a new observation, $y_{\rm n}$, that can be expressed as $\theta_{\rm n} = {\rm E}(y_{\rm n}|\alpha_I) = b'(x'_{\rm n}\beta + z'_{\rm n}\alpha_I)$. Given $I = i$, the BP (see Section 1, end of third paragraph) of $\theta_{\rm n}$ is given by

$$\begin{aligned}
&{\rm E}(\theta_{\rm n}|\mathbf{y}) = {\rm E}\{b'(\mathbf{x}'_{\rm n}\beta + \mathbf{z}'_{\rm n}\alpha_i)|\mathbf{y}\} \\
&= {\rm E}\{b'(\mathbf{x}'_{\rm n}\beta + \mathbf{z}'_{\rm n}\alpha_i)|\mathbf{y}_i\} = \int_{-\infty}^{\infty} b'(\mathbf{x}'_{\rm n}\beta + \mathbf{z}'_{\rm n}\alpha_i) f(\alpha_i|\mathbf{y}_i) d\alpha_i \\
&= \frac{\int_{-\infty}^{\infty} b'(\mathbf{x}'_{\rm n}\beta + \mathbf{z}'_{\rm n}\alpha_i) f(\mathbf{y}_i|\alpha_i) f(\alpha_i) d\alpha_i}{\int_{-\infty}^{\infty} f(\mathbf{y}_i|\alpha_i) f(\alpha_i) d\alpha_i}. \qquad (11)
\end{aligned}$$

Let $\mathbf{v}_i = \mathbf{G}^{-1/2}\alpha_i$, which is distributed as $N(\mathbf{0}, \mathbf{I}_q)$, $\mathbf{I}_q$ being the q-dimensional identity matrix, $\pi(\mathbf{v})$ denote the pdf of $N(\mathbf{0}, \mathbf{I}_q)$, and

$$\begin{aligned}
&s_i(\mathbf{v}_i) = s_i(\mathbf{y}_i, \alpha_i, \beta) = s_i(\mathbf{y}_i, \mathbf{G}^{1/2}\mathbf{v}_i, \beta) \\
&= -\frac{1}{k_i\phi} \sum_{j=1}^{k_i} a_{ij}\{y_{ij}(\mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\alpha_i) - b(\mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\alpha_i)\} \\
&= -\frac{1}{k_i\phi} \sum_{j=1}^{k_i} a_{ij}\{y_{ij}(\mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{G}^{1/2}\mathbf{v}_i) - b(\mathbf{x}'_{ij}\beta + \mathbf{z}'_{ij}\mathbf{G}^{1/2}\mathbf{v}_i)\}.
\end{aligned}$$

Then, we have the following expression:

$$\begin{aligned}
&\int_{-\infty}^{\infty} f(\mathbf{y}_i|\alpha_i) f(\alpha_i) d\alpha_i \\
&= \exp\left\{\sum_{j=1}^{k_i} c\left(y_{ij}, \frac{\phi}{a_{ij}}\right)\right\} \int_{-\infty}^{\infty} \exp\{-k_i s_i(\mathbf{v})\}\pi(\mathbf{v}) d\mathbf{v}, \qquad (12)
\end{aligned}$$

and a similar expression for the numerator in (11). Thus, combining (11) and (12), we have the following expression for the BP:

$$\begin{aligned}
{\rm E}(\theta_{\rm n}|\mathbf{y}) &= \frac{\int_{-\infty}^{\infty} b'(\mathbf{x}'_{\rm n}\beta + \mathbf{z}'_{\rm n}\mathbf{G}^{1/2}\mathbf{u}) \exp\{-k_i s_i(\mathbf{u})\}\pi(\mathbf{u}) d\mathbf{u}}{\int_{-\infty}^{\infty} \exp\{-k_i s_i(\mathbf{u})\}\pi(\mathbf{u}) d\mathbf{u}} \\
&= \frac{{\rm E}[b'(\mathbf{x}'_{\rm n}\beta + \mathbf{z}'_{\rm n}\mathbf{G}^{1/2}\mathbf{v}) \exp\{-k_i s_i(\mathbf{v})\}]}{{\rm E}[\exp\{-k_i s_i(\mathbf{v})\}]}, \qquad (13)
\end{aligned}$$

where the expectations are with respect to $\mathbf{v} \sim N(\mathbf{0}, \mathbf{I}_q)$.

In (13), the parameters $\beta, \mathbf{G}, \phi$ are understood as the true parameters, which are unknown in practice. If we replace these parameters in (13) by their consistent estimators, for example, the maximum likelihood estimators (MLEs; e.g., Jiang, 2013) based on the training data, we obtain the network-classified mixed model predictor (NCMMP) of $\theta_{\rm n}$, denoted by $\hat{\theta}_{\rm n}$. The word "network" is used because of the involvement of the working probability model, (3), which is motivated by the network association. We use a real-data example to illustrate.

2.2. Real-data example

Since the first case reported in Wuhan, China in December 2019, Covid-19 has emerged as a global outburst of a public health incident with a rapid increase in cases and deaths. In this study, we were able to obtain cumulative numbers of infected persons for every city in the 30 provinces and municipalities in China, recorded for the period of time from December 8, 2019 to August 1, 2020. Because Xizang (Tibet) Autonomous Region has its specific geographical environment, and had not been effected by Covid-19, it is excluded. The rest of the provinces and municipalities are listed in column 1 of Table 3 and Table 4; all except Beijing, Chongqing, Shanghai, and Tianjin are provinces. The total number of cities in each region are given in column 2 of the table.

For the auxiliary variables, we have collected information about population density, real per capita GDP, distance between every city and Wuhan, general public budgets, proportion of the population age at 60 and over, the number of Grade Three A Hospitals, the mean air quality index and the annual lowest temperature. Population density and proportion of the population age at 60 and over are intended to characterize demographic information for each city. General public budgets and real per capita GDP are intended to reflect the level of economic development. The number of Grade Three A Hospitals can reflect the medical quality. The annual lowest temperature and distance between every city

and Wuhan are included to reflect the physical environment. The mean air quality index stands for the air environmental quality of each city. It should be noted that the current quarantine time in China is fixed as 14 days, and this is true for all provinces and municipalities; therefore, quarantine time is not a factor that makes a difference. See, for example, Guan *et al.* (2020) and Li *et al.* (2020).

The economic data are obtained from the National Bureau of Statistics (http://www.stats.gov.cn/tjsj/). The meteorological data are obtained from China Meteorological Administration (http://www.cma.gov.cn/). The population data are obtained from the Statistical Almanac of each city. The COVID-19 real data and medical data are obtained from National Health Commission of the People's Republic of China (http://www.nhc.gov.cn/).

Surprisingly, the province to which a city belongs is not necessarily known with certainty in practice, due to political, economic, or other concerns. For example, there may be a privacy issue for revealing the truth identities of the cities/provinces. The observed identities of the cities may be contaminated due to intentional errors, as in differential privacy (e.g., Dwork 2006), or unintentional errors, such as mistakes in data entry. However, it can be demonstrated (see Section 4.2) that, even if the city/province information for the new data may be contaminated for whatever reasons, our method can still help to identify the true province of the city associated with the new data, hence improve accuracy in predicting the associated mixed effects.

3. ASYMPTOTIC THEORY

In this section, we study asymptotic behavior of the NCMMP under two scenarios. The first is when m, the number of classes in the training data, is bounded; the second is when m increases with the sample sizes. Practically, $m \to \infty$ means that the number of classes in the training data is large. Under each scenario, we obtain consistency results both in terms of class matching, that is, in

the sense that $\mathrm{P}(\hat{I} \neq I) \to 0$, and in terms of prediction of the mixed effect, that is, in the sense that $\hat{\theta}_{\mathrm{n}} - \theta_{\mathrm{n}} \to 0$ in probability, where I is the true class index associated with the new observations, and $\hat{\theta}_{\mathrm{n}}$ is the NCMMP of θ_{n}, a mixed effect of interest associated with the new observations. Note that the consistency result in terms of the class matching is not available in Jiang et al. (2018) or Sun et al. (2018); neither is consistency of $\hat{\theta}_{\mathrm{n}}$ when m is bounded in those previous work.

The asymptotic theory developed below extends without difficulty to any fixed dimension q of α_i and α_I, any positive constants $a_{ik}, a_{\mathrm{n}k}$ that are bounded from above, and bounded below from zero, and any bounded numbers $z_{ik}, z_{\mathrm{n}k}$. Thus, for notation simplicity and without loss of generality, we assume that $q = 1$, $a_{ik} = a_{\mathrm{n}k} = 1$, and $z_{ik} = z_{\mathrm{n}k} = 1$.

3.1. Asymptotic behavior when m is bounded

First consider consistency of the proposed class-matching procedure.

Theorem 1 (consistency of class matching). *Suppose that the following hold:*
(i) $m > 1$ *and is bounded, and* $\min_{1\le i\le m} k_i \to \infty$, $k_{\mathrm{n}} \to \infty$*;*
(ii) $|\mathbf{x}_{ij}|, 1 \le j \le k_i, 1 \le i \le m$ *and* $|\mathbf{x}_{\mathrm{n}j}|, 1 \le j \le k_{\mathrm{n}}$ *are bounded;*
(iii) $b''(\cdot)$ *is positive and continuous; and*
(iv) $\hat{\beta}$, $\hat{G}$ *and* $\hat{\phi}$ *are consistent.*
Then, for any fixed $0 < p < 1$ *in (3),* $\hat{I}$ *is consistent, that is, we have* $\mathrm{P}(\hat{I} \neq I) \to 0$.

Next, we consider consistency of the NCMMP.

Theorem 2 (consistency of NCMMP). *Suppose that* $|\mathbf{x}_{\mathrm{n}}|$ *is bounded. Then, under the assumptions of Theorem 1, the NCMMP is consistent, that is,* $\hat{\theta}_{\mathrm{n}} - \theta_{\mathrm{n}} \xrightarrow{\mathrm{P}} 0$. *The result holds regardless of the choice of the tuning parameter,* p *in (3), as long as* $0 < p < 1$.

The proofs of Theorems 1 and 2 are given in the supplementary material.

3.2. Asymptotic behavior when $m \to \infty$

Now let us consider the asymptotic behavior of $\hat{I}$ and $\hat{\theta}_{\mathrm{n}}$ as m, the number of classes in the training data, increases with other parts of the sample size. As in Jiang et al. (2018), we consider both the matched and unmatched scenarios (see Section 1). In the matched case, we assume that the true class number of the new observation, I, belongs to $\{1, \ldots, m\}$, the set of indexes associated with the training data classes. In the unmatched case, I does not belong to the above index set. It makes sense to consider consistency of $\hat{I}$, as in the previous subsection, in the matched case, but there is no matching consistency, of course, in the unmatched case. Nevertheless, we can still establish consistency of the NCMMP in the unmatched case. Note that Jiang et al. (2018) also established consistency of the CMMP of the mixed effect in the unmatched case; however, consistency of class matching has not been previously obtained, even in the matched case.

1. Matched case. First introduce the following notation: $k_* = \min_{1 \le i \le m} k_i$,

$$B_m = \sup_{|u| \le 2 \log m} |b'(u)|, \quad D_m = \inf_{|u| \le 2 \log m} b''(u), \quad H_m = \sup_{|u| \le 2 \log m} b''(u).$$

Define $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

Theorem 3 (consistency of class matching). *Suppose that assumption (ii) of Theorem 1 is satisfied, $m > 1$, and $\hat{\beta}, \hat{G}, \hat{\phi}$ are $\sqrt{k_*}$-consistent, that is, $\sqrt{k_*}(\hat{\beta} - \beta, \hat{G} - G, \hat{\phi} - \phi) = O_{\mathrm{P}}(1)$. In addition, suppose that $k_* \wedge k_{\mathrm{n}} \to \infty$ and there are $d > 0$, $0 < \gamma < 1/2$ and $\eta > 2$ such that the following hold: $\log m / k_*^{1/2-\gamma} = O(1)$, $k_*^{\gamma} \log m / \sqrt{k_{\mathrm{n}}} m^d = O(1)$, $B_m \vee H_m = O(m^d)$ and $(k_*^{\gamma}/m^{d+\eta}) D_m \to \infty$. Then, for any fixed $0 < p < 1$ in (3), $\hat{I}$ is consistent, that is, we have $\mathrm{P}(\hat{I} \ne I) \to 0$.*

Remark 1: *The assumptions of Theorem 3, in particular, set restriction on how fast m increases, relative to k_* and k_{n}. This is reasonable in most network related applications, where the k_i's are often much larger than m.*

Remark 2: *The following special cases deserve some attention. For the case of LMM, we have* $b'(u) = u$ *and* $b''(u) = 1$*; thus, we have* $B_m = 2\log m$*, and* $D_m = H_m = 1$*. For the mixed logistic model, we have* $b'(u) = e^u/(1+e^u)$*,* $b''(u) = e^u/(1+e^u)^2$*; thus, we have* $B_m \leq 1$*,* $D_m \geq (1/4)m^{-2}$*, and* $H_m \leq 1/4$*. For the Poisson log-linear mixed model, we have* $b'(u) = b''(u) = e^u$*; hence, we have* $B_m = H_m = m^2$*, and* $D_m = m^{-2}$*.*

The proof of Theorem 3 is given in the supplementary material.

The next result is regarding consistency of the NCMMP of $\theta_{\rm n}$.

Theorem 4 (consistency of NCMMP). *Suppose that* $b''(\cdot)$ *is continuous and* $|\mathbf{x}_{\rm n}|$ *is bounded. Then, under the assumptions of Theorem 3, we have* $\hat{\theta}_{\rm n} - \theta_{\rm n} = O_{\rm P}(k_*^{-\gamma})$*, where* $0 < \gamma < 1/2$ *is the same as in Theorem 3. The result holds regardless of the choice of the tuning parameter,* p *in (3), as long as* $0 < p < 1$*.*

Note that the conclusion of Theorem 4 is stronger than consistency in that there is a rate of convergence in probability. Again, the proof is given in the supplementary material.

2. Unmatched case. Now consider the case that $I \notin \{1, \ldots, m\}$. 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 CMMP 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, but since m is large, there is always some α_i that comes close to α_I, which is all that matters, so far as prediction of the mixed effect is concerned. We now extend such a result to NCMMP under a GLMM. Even more, we obtain the rate of convergence in probability, which was not previously obtained for CMMP.

First introduce the following definition. The function $b(\cdot)$ is called regular if for any $A > 0$, there are constants $U_A, \delta_A > 0$ such that

$$D(a,u) = b(a+u) - b(a) - b'(a)u \geq \delta_A, \ \ \forall\, |a| \leq A \text{ and } |u| > U_A. \quad (14)$$

Some examples are discussed in Section A.6 of the supplementary material, in which condition (14) is verified.

Theorem 5 (consistency of NCMMP). *Suppose that the following hold:*
(i) $b''(\cdot)$ *is continuous and positive, and* $b(\cdot)$ *is regular;*
(ii) $|\mathbf{x}_{ij}|, 1 \leq j \leq k_i, 1 \leq i \leq m$, $|\mathbf{x}_{\mathrm{n}j}|, 1 \leq j \leq k_{\mathrm{n}}$ *and* $|\mathbf{x}_{\mathrm{n}}|$ *are bounded;*
(iii) $\hat{\beta}, \hat{G}, \hat{\phi}$ *are* $\sqrt{k_*}$*-consistent;*
(iv) α_I *is independent with* $\alpha_i, 1 \leq i \leq m$*; and*
(v) $m > 1$, $\log m/\sqrt{k_* \wedge k_{\mathrm{n}}} \to 0$*, and* $B_m/k_*^{\gamma} \to 0$ *for some* $0 < \gamma < 1/2$*.*
Then, $\hat{\theta}_{\mathrm{n}}$ *is consistent and has the convergence rate*

$$\hat{\theta}_{\mathrm{n}} - \theta_{\mathrm{n}} = O_{\mathrm{P}}\left[m^{-\nu} + \frac{\sqrt{\log m}}{(k_* \wedge k_{\mathrm{n}})^{1/4}} + \frac{\sqrt{B_m}}{k_*^{\gamma/2}}\right] \quad (15)$$

for any $0 < \nu < 1$*. The result holds regardless of the choice of the tuning parameter,* p *in (3), as long as* $0 < p < 1$*.*

The proof of Theorem 5 is given in the supplementary material.

4. EMPIRICAL STUDIES

We carry out a series of empirical studies, including simulation studies and real-data validation, on the finite-sample performance of the proposed NCMMP method. For the Monte Carlo simulation studies, we compare the predictive performance of NCMMP to existing methods. Here we present a simulation study under a mixed logistic model. An additional simulation study under a Poisson log-linear mixed model is presented in Section B of the supplementary material. For the real-data validation, we consider a problem of predicting the number of infected persons of Covid-19.

4.1. Simulation study: mixed logistic model

First, we consider a case of mixed logistic model (MLM). For the training data, the model is expressed as $p_{ij} = \mathrm{P}(y_{ij} = 1|\alpha_i)$ and

$$\mathrm{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = 1 + x_{1ij} + x_{2ij} + \alpha_i, \tag{16}$$

$i = 1, \ldots, m, j = 1, \ldots, k$, where α_i's are generated independently from the $N(0, G)$ distribution with different values of G; $x_{rij}, r = 1, 2$ are generated from $N(0, 1)$, then fixed throughout the simulation.

A new subject satisfies $p_{\mathrm{n}j} = \mathrm{P}(y_{\mathrm{n}j} = 1|\alpha_I)$ and

$$\mathrm{logit}(p_{\mathrm{n}j}) = \log\left(\frac{p_{\mathrm{n}j}}{1-p_{\mathrm{n}j}}\right) = 1 + x_{1\mathrm{n}j} + x_{2\mathrm{n}j} + \alpha_I, \tag{17}$$

$j = 1, \ldots, k_{\mathrm{n}}$, where $x_{r\mathrm{n}j}, r = 1, 2$ are generated from $N(0, 1)$ and then fixed throughout; and I is the true class index for the new observations.

Let $c_{\mathrm{n}} = 1$, that is, the new subject belongs to the first community, but there is a chance that this is not the same as its true class index, I. The latter satisfies (3), where the true value of p is 0.85. However, we pretend that this is not known, and two proposed values of p are considered: $0.75, 0.95$. The following combinations of sample sizes are considered: $m = 10, k = 10$; $m = 10, k = 100$; $m = 10, k = 200$, $m = 50, k = 10$, $m = 50, k = 100$; $m = 50, k = 200$. For each of these combinations, we consider $G = 0.1, 1$; $k_{\mathrm{n}} = 1, 50$, resulting four pairs of combinations for each case of m, k.

There are two objectives of interest: identification of the true index, I, and prediction of the true mixed effect, $\theta_{\mathrm{n}} = 1 + x_{1\mathrm{n}} + x_{2\mathrm{n}} + \alpha_I$. It should be noted that the conditional probability $p_{\mathrm{n}} = \exp(\theta_{\mathrm{n}})/\{1 + \exp(\theta_{\mathrm{n}})\}$ is often of practical interest; however, because p_{n} is a smooth monotone function of θ_{n}, we focus on the latter for simplicity. The unknown parameters under the MLM are estimated by their MLEs based on the training data. We run 100 simulations under each combination of m, k, G, k_{n}, and p values specified above, and report (i) the

empirical MSPE: $\mathrm{E}(\hat{\theta}_{\mathrm{n}} - \theta_{\mathrm{n}})^2$, where $\hat{\theta}_{\mathrm{n}}$ corresponds to NCMMP, CMLMP, or GLM, based on the simulation runs; and (ii) empirical probability (i.e., proportion of times) that the class index is matched correctly, for NCMMP (EP_{N}) and CMLMP (EP_{C}). Here, CMLMP and GLM stand for the method of Sun et al. (2020) and the standard generalized linear model prediction, respectively. Note that GLM does not involve class matching.

The results for $p = 0.75$ and $p = 0.95$ are presented in Table 1 and Table 2, respectively. Reported are empirical MSPE (the number in the parentheses is the empirical standard deviation for the empirical MSPE) except for the last two columns, which are empirical probabilities of correct matching.

A remarkable observation is that the results are highly consistent across all proposed values of p. It turns out that the value of the tuning parameter, p in (3), does not really matter, so far as this simulation is concerned. Note that the true p used to generate the data is 0.85. This is consistent with our theoretical results, which show consistency of the class matching as well as that of NCMMP regardless of the value of p.

It is also seen that NCMMP performs substantially better than CMLMP when $k_{\mathrm{n}} = 1$; in fact, the difference is nearly 10 fold in some cases. When $k_{\mathrm{n}} = 50$, CMLMP in most cases performs slightly to moderately better than NCMMP. The trends are reasonable because NCMMP mainly relies on its much more accurate class matching while CMLMP relies on higher number of repeated observations. This is supported by the much higher empirical probability of correct matching in NCMMP; in fact, for CMLMP the empirical probability seems to get worse when the sample sizes, especially m, get larger. As noted by Jiang et al. (2018), CMMP does not have matching consistency, while we have established matching consistency for NCMMP. As for comparison with GLM, with the exception of three cases under $m = k = 10, G = 0.1$ (one with $p = 0.75$ and two with $p = 0.95$), NCMMP performs consistently (much) better than GLM; on the other

TABLE 1: Empirical MSPE & Probability of Correct Matching: MLM($p = 0.75$)

	G	k_{n}	NCMMP	CMLMP	GLM	$\mathrm{EP_N}$	$\mathrm{EP_C}$
m=10, k=10	0.1	1	.0108 (.0017)	.0112 (.0018)	.0097 (.0017)	.80	.29
	0.1	50	.0094 (.0014)	.0077 (.0013)	.0097 (.0017)	.79	.33
	1	1	.0329 (.0078)	.0366 (.0071)	.0389 (.0062)	.80	.13
	1	50	.0185 (.0032)	.0106 (.0017)	.0386 (.0062)	.80	.22
m=10, k=100	0.1	1	.0029 (.0006)	.0060 (.0013)	.0039 (.0006)	.80	.12
	0.1	50	.0029 (.0006)	.0020 (.0003)	.0039 (.0006)	.80	.24
	1	1	.0134 (.0042)	.0400 (.0075)	.0301 (.0048)	.80	.17
	1	50	.0049 (.0015)	.0043 (.0009)	.0301 (.0048)	.86	.35
m=10, k=200	0.1	1	.0024 (.0005)	.0074 (.0016)	.0037 (.0006)	.80	.14
	0.1	50	.0023 (.0004)	.0023 (.0005)	.0037 (.0006)	.81	.16
	1	1	.0135 (.0048)	.0399 (.0075)	.0306 (.0046)	.80	.18
	1	50	.0024 (.0009)	.0031 (.0006)	.0306 (.0046)	.84	.28
m=50, k=10	0.1	1	.0044 (.0007)	.0076 (.0015)	.0045 (.0007)	.80	.14
	0.1	50	.0044 (.0007)	.0030 (.0005)	.0045 (.0007)	.80	.16
	1	1	.0193 (.0041)	.0444 (.0068)	.0307 (.0047)	.80	.00
	1	50	.0183 (.0040)	.0053 (.0013)	.0307 (.0047)	.81	.07
m=50, k=100	0.1	1	.0014 (.0003)	.0105 (.0016)	.0030 (.0005)	.80	.01
	0.1	50	.0014 (.0003)	.0025 (.0005)	.0030 (.0005)	.80	.05
	1	1	.0083 (.0020)	.0647 (.0090)	.0275 (.0042)	.80	.01
	1	50	.0055 (.0015)	.0041 (.0007)	.0275 (.0042)	.80	.07
m=50, k=200	0.1	1	.0015 (.0003)	.0120 (.0018)	.0030 (.0005)	.80	.02
	0.1	50	.0016 (.0003)	.0029 (.0005)	.0030 (.0005)	.80	.04
	1	1	.0094 (.0025)	.0656 (.0090)	.0277 (.0042)	.80	.03
	1	50	.0029 (.0007)	.0039 (.0007)	.0277 (.0042)	.82	.08

TABLE 2: Empirical MSPE & Probability of Correct Matching: MLM($p = 0.95$)

	G	k_{n}	NCMMP	CMLMP	GLM	$\mathrm{EP_N}$	$\mathrm{EP_C}$
m=10, k=10	0.1	1	.0108 (.0017)	.0112 (.0018)	.0097 (.0017)	.80	.29
	0.1	50	.0107 (.0016)	.0077 (.0013)	.0097 (.0017)	.80	.33
	1	1	.0329 (.0078)	.0366 (.0071)	.0389 (.0062)	.80	.13
	1	50	.0205 (.0032)	.0106 (.0017)	.0386 (.0062)	.81	.22
m=10, k=100	0.1	1	.0029 (.0006)	.0060 (.0013)	.0039 (.0006)	.80	.12
	0.1	50	.0029 (.0006)	.0020(.0003)	.0039 (.0006)	.80	.24
	1	1	.0134 (.0042)	.0400 (.0075)	.0301 (.0048)	.80	.17
	1	50	.0049 (.0015)	.0043 (.0009)	.0301 (.0048)	.86	.35
m=10, k=200	0.1	1	.0024 (.0005)	.0074 (.0016)	.0037 (.0006)	.80	.14
	0.1	50	.0024 (.0005)	.0023 (.0005)	.0037 (.0006)	.80	.16
	1	1	.0135 (.0048)	.0399 (.0075)	.0306 (.0046)	.80	.18
	1	50	.0035 (.0012)	.0031 (.0006)	.0306 (.0046)	.85	.28
m=50, k=10	0.1	1	.0044 (.0007)	.0076 (.0014)	.0045 (.0007)	.80	.14
	0.1	50	.0044 (.0007)	.0030 (.0005)	.0045 (.0007)	.80	.16
	1	1	.0193 (.0041)	.0444 (.0068)	.0307 (.0047)	.80	.00
	1	50	.0192 (.0041)	.0053 (.0013)	.0307 (.0047)	.80	.07
m=50, k=100	0.1	1	.0014 (.0003)	.0105 (.0016)	.0030 (.0005)	.80	.01
	0.1	50	.0014 (.0003)	.0025 (.0005)	.0030 (.0005)	.80	.05
	1	1	.0083 (.0020)	.0647 (.0090)	.0275 (.0042)	.80	.01
	1	50	.0057 (.0015)	.0041 (.0007)	.0275 (.0042)	.80	.07
m=50, k=200	0.1	1	.0016 (.0003)	.0120 (.0018)	.0030 (.0005)	.80	.02
	0.1	50	.0016 (.0003)	.0029 (.0005)	.0030 (.0005)	.80	.04
	1	1	.0094 (.0025)	.0656 (.0090)	.0277 (.0042)	.80	.03
	1	50	.0037 (.0012)	.0039 (.0007)	.0277 (.0042)	.82	.08

hand, there are more cases across different scenarios where CMLMP does not perform better than GLM.

4.2. Real-data validation: Prediction with Covid-19 data

We now return to the Covid-19 example, introduced in Section 2.2. Note that the outcome variable, the cumulative number of infected persons, is a count; therefore, a Poisson log-linear mixed model (PLMM; see Section B of the supplementary material) is considered with regionally specific random effects. Specifically, we assume that, given the regional random effects $\alpha_1, \ldots, \alpha_{30}$, the cumulative numbers of infected persons, y_{ij}, are conditionally independent such that $y_{ij}|\alpha \sim \text{Poisson}(\mu_{ij})$, where

$$\log(\mu_{ij}) = \beta_0 + \sum_{l=1}^{8} \beta_k x_{ijl} + \alpha_i, \tag{18}$$

$i = 1, \cdots, 30,\ j = 1, \cdots, k_i,$ where $x_{ijl}, l = 1, \ldots, 8$ correspond to the eight auxiliary variables mentioned above. Here, k_i is the number of cities in the ith region; x_{ijr}, $\beta_l, l = 0, \ldots, 8$ are unknown fixed effects, and α_i is a region-specific random effect. The random effects are assumed to be independent and distributed as $N(0, G)$. Note that the classes in the training data are assumed to be exclusive. It is true that the new observations may belong to the border of two or more regions, in which case it may not be so clear which random effect is associated. Nevertheless, what NCMMP does is identify one random effect that (it thinks) is closest to the one associated with the new observations. This identified random effect may be a wrong classification, but the impact may be insignificant, as long as its value is close to the one associated with the new observations. This is a nice feature of CMMP (Jiang et al. 2018, Sun et al. 2018).

On the other hand, the fact that a false match can still help is in comparison to not making the match at all, as in regression or GLM predictions. But, there is still plenty of room to improve, if one can improve the precision of matching in CMMP. This has been demonstrated in the earlier sections.

Specifically, we use this real-data situation to test our NCMMP method, and compare it with other prediction methods. For that, we must know the "ground truth" in order to validate predictive performance. The ground truth is the observed cumulative number of infected persons, y_{ij}. Specifically, we take out the jth observation of the ith region, and consider it as a new observation; the rest of the data (including $y_{ij'}$ for $j' \neq j$ and $y_{i'j'}$ for $i' \neq i$) are used as the training data. We then use the NCMMP, CGLMMP (that is, CMMP applied to PLMM; see Section B of the supplementary material), and GLM methods to predict y_{ij}. Note that, if y_{ij} were unobserved, its BP based on the training data is the same as the BP of μ_{ij} based on the training data. This is because we can write $y_{ij} = \mu_{ij} + e_{ij}$, where e_{ij} is the new error, which is independent with the training data (note that only μ_{ij} is potentially correlated with the training data). Therefore, we have

$$\begin{aligned} \mathrm{E}(y_{ij}|\text{training data}) &= \mathrm{E}(\mu_{ij}|\text{training data}) + \mathrm{E}(e_{ij}|\text{training data}) \\ &= \mathrm{E}(\mu_{ij}|\text{training data}). \end{aligned}$$

Thus, one can use each of the three methods, NCMMP, CGLMMP and GLM, to obtain the corresponding predictor of μ_{ij}, which is a mixed effect, then use it as a predictor of y_{ij}, for which we know the truth. We do this for every j in region i. and every $1 \leq i \leq 30$.

To compare the performance of the three methods, we compute, for each $1 \leq i \leq 30$, the average squared prediction errors (ASPE) over $1 \leq j \leq k_i$. The results are reported in Table 3 and Table 4. Also reported (in the parentheses) are observed probability of correct matching, that is, proportion of times (over different j's) that the region corresponding to the new observation is correctly identified, for NCMMP and CGLMMP. For NCMMP we use $p = 0.75$, or $p = 0.95$, as in the simulation study of the previous subsection.

It is seen that the ASPE of NCMMP, with either $p = 0.75$ or $p = 0.95$, is less than or equal to that of CGLMMP for all but two regions (Henan and Shaanxi).

TABLE 3: Average Squared Prediction Error (Proportion of Correct Matching): Part I

Region	# of City	NCMMP ($p = 0.75$)	NCMMP ($p = 0.95$)	CGLMMP	GLM
Anhui	16	1058(0.6)	1058(0.6)	1366(0)	12449
Beijing	15	0(1)	0(1)	17.2(0.2)	55.3
Chongqing	37	168(0.10)	158(0.27)	181(0)	383
Fujian	9	0(1)	0(1)	25.4(0)	897
Gansu	14	0(1)	0(1)	6.24(0.5)	589
Guangdong	21	1840(0)	1840(0)	1848(0)	394
Guangxi	14	153(0.25)	79.2(0.75)	172(0)	455
Guizhou	9	0(1)	0(1)	12.5(0)	1764
Hainan	16	0(1)	0(1)	3.67(0.2)	45.6
Hebei	11	70.5(0.67)	0(1)	223(0)	460
Henan	18	14648(0.2)	14648(0.2)	12690(0)	38174
Heilongjiang	13	76.3(0.75)	76.3(0.75)	407(0)	1903
Hubei	17	59647(0.6)	59647(0.6)	72499(0)	77324
Hunan	14	495(0.5)	184(0.75)	709(0)	1365
Jilin	9	0(1)	0(1)	26(0)	2.78

The ASPE of NCMMP, with either $p = 0.75$ or $p = 0.95$, is less than or equal to that of GLM in all but two regions (Guangdong and Shandong). In fact, in most cases the ASPEs of NCMMP and CGLMMP are much smaller than that of GLM; and in many cases the ASPEs of NCMMP are much smaller than that of CGLMMP. As for comparison between the NCMMP with $p = 0.75$ and NCMMP with $p = 0.95$, the ASPE of the latter is less than or equal to that of the former, but the equality holds in, by far, most of the cases. In terms of the proportion of correct matching, the proportions for the two NCMMPs are greater than or equal to that of CGLMMP in all cases, with the equality holding only in two occasions (Guangdong and Tianjin); the proportion of NCMMP with $p = 0.95$

TABLE 4: Average Squared Prediction Error (Proportion of Correct Matching): Part II

Region	# of City	NCMMP ($p = 0.75$)	NCMMP ($p = 0.95$)	CGLMMP	GLM
Jiangsu	13	324(0.5)	284(0.75)	442(0)	1326
Jiangxi	11	1201(0.67)	1201(0.67)	2148(0)	2788
Liaoning	14	0(1)	0(1)	7.47(0)	57.8
Neimenggu	12	0(1)	0(1)	21.8(0)	28.9
Ningxia	5	86(0)	0(1)	85.9(0)	87.2
Qinghai	2	0(1)	0(1)	3.25(0)	249
Shandong	16	9374(0.6)	9374(0.6)	9401(0)	1612
Shanxi	11	0(1)	0(1)	34.8(0)	2929
Shaanxi	11	160(0.67)	160(0.67)	114(0)	2107
Shanghai	16	0(1)	0(1)	22.8(0.2)	149
Sichuang	21	0.15(0.83)	0.15(0.83)	109(0)	369
Tianjin	16	0(1)	0(1)	0(1)	235
Xinjiang	18	0(1)	0(1)	9.17(0)	92.8
Yunnan	16	0(1)	0(1)	19.9(0)	201
Zhejiang	11	167(0.33)	93(0.67)	1284(0)	655

is always greater than or equal that of NCMMP with $p = 0.75$ but in most cases the equality holds. Note that, although in our simulation study (see Section 4.1 and also Section B of the supplementary material), the predictive performance of NCMMP is almost the same whether $p = 0.75$ or $p = 0.95$, here we see some small discrepancy (e.g., Ningxia in Table 4). Our theoretical results state that, in large samples, the consistency of NCMMP, both in terms of class-matching and in terms of prediction, is not affected by p but in a finite-sample situation, such as the current example, there can be a nonignorable difference.

It should be noted that this is a situation of matched case, that is, the true index I corresponding to the new observation belongs to $\{1, \ldots, 30\}$. Considering

that m is relatively large ($m = 30$), the appropriate theorems that would apply are Theorem 3 and Theorem 4. Although some of the regularity conditions of Theorem 4 regarding the relative sizes of k_i and m might not hold, as some of the k_i's are quite small, the relative performance of the three methods, NCMMP, CGLMMP and GLM, still fully demonstrates the advantage of NCMMP.

5. CONCLUSION AND DISCUSSION

We extend the original CMMP method and its variations to GLMM so that the method can be applied, in particular, to cases of binary responses and counts. Unlike the previous CMMP methods, the extension is based on a new matching strategy that utilizes a pseudo prior to borrow strength across different classes of the training data, and the new data. The new method is shown to enjoy consistency both in terms of the class matching and in terms of the prediction of the mixed effect of interest associated with the new observations, the so-called *double consistency*. It should be noted that the previous CMMP methods are known to be consistent only in terms of the mixed effect prediction, when the number of classes in the training data, m, increases. The new CMMP method that we propose posses the double consistency, whether m increases or not.

It should be reminded that our primary interest is prediction of the mixed effect (associated with the new observations), while consistency of the class matching is used as a tool to improve the prediction accuracy, as we have demonstrated in our empirical studies. Nevertheless, the class-matching consistency is an important by-product, which should be useful practically in some cases.

ACKNOWLEDGEMENTS

Haiqiang Ma's research was supported by NNSF of China (No.12161042 and 11701235) , China Postdoctoral Science Foundation (No. 2019M662262). The research of Jiming Jiang is partially supported by the NSF grants DMS-1510219 and DMS-1713120. The authors are grateful to an Associate Editor and two

referees for their valuable comments that have led to improvement of the manuscript.

6. SUPPLEMENTARY MATERIAL

The Supplementary Material contains proofs of the theoretical results, additional simulation results, as well as details about computation of the maximum likelihood estimates under a GLMM.

BIBLIOGRAPHY

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

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

Bickel, P. J. & Sarkar, P. (2016). Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society. Series B (Methodological)*, 78, 253–273.

Dwork, C. (2006). Differential Privacy. *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, Springer, 4052, 1–12.

Gleiser, P. & Danon, L. (2003). Community structure in jazz. *Advances in Complex Systems*, 6, 565–573.

Guan, W. J., Ni, Z. Y., Hu, Y., Liang, W. H., Ou, C. Q., He, J. X., ... and Du, B. (2020). Clinical characteristics of 2019 novel coronavirus infection in China. *New England J. Med.*, 382, 1708–1720.

Jiang, J. (1999). Conditional inference about generalized linear mixed models. *The Annals of Statistics*, 27, 1974–2007.

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

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

Jiang, J. (2013). The subset argument and consistency of MLE in GLMM: Answer to an open problem and beyond. *The Annals of Statistics*, 41, 177–195.

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

Jiang, J. & Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 53, 217–243.

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

Jiang, J. & Torabi, M. (2020). Sumca: Simple, unified, Monte-Carlo-assisted approach to second-order unbiased mean-squared prediction error estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 82, 47–485.

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

Keeling, M. J. & Eames, K. T. D. (2005). Network and epidemic models. *Journal of the Royal Society Interface*, 2, 295–307.

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

Li, C., Shen, X., & Pan, W. (2019). Likelihood Ratio Tests for a Large Directed Acyclic Graph. *Journal of the American Statistical Association*, 115, 1304–1319.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., ... and Xing, X. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *New England J. Med.*, 382, 1199–1207.

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

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

McAuley, J. & Leskovec, J. (2012). Learning to discover social circles in ego networks. *Neural Information Processing Systems*, 1, 539–547.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman & Hall, London.

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

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

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

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

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

Sun, H., Luan, Y. and Jiang, J. (2020). A new classified mixed model predictor. *Journal of Statistical Planning and Inference*, 207, 45–54.

Received 9 July 2009

Accepted 8 July 2010