A New Classified Mixed Model Predictor

HANMEI SUN[1], YIHUI LUAN[1] AND JIMING JIANG[2]
Shandong University, China[1] *and University of California, Davis*, USA[2]

We develop a new method for classified mixed model prediction (CMMP). The original CMMP method (Jiang *et al.* 2018) does not incorporate covariate information in matching the class between the new observations and the training data. As a result, the method may not outperform the mixed model prediction (MMP) method in terms of predictive performance. The new CMMP method that we develop utilizes covariate information, and therefore is more accurate in terms of the matching. We show that the new CMMP method outperform the MMP in terms of the predictive performance. Furthermore, we develop a second-order unbiased estimator of the mean squared prediction error (MSPE) for the new CMMP, which was previously not available for the original CMMP. Theoretical and empirical properties of the proposed new CMMP method as well as the MSPE estimator are studied. A real data application is considered.

Key Words. CMMP, MSPE, Sumca

1 Introduction

Classified mixed model prediction (Jiang *et al.* 2018) is new method of prediction based an idea of matching a random effect associated with the new observations and one of the random effects associated with the training data. It was shown that CMMP im-

proves predictive performance over the tradition regression-based prediction substantially. The method has potential application in such fields as precision medicine, where the primary interests are at the subject-level, and business, where prediction of business values at customer-level is of interest.

On the other hand, the current CMMP method does not utilize covariate information in its matching procedure; in other words, only the observed mean response is used in the matching. As a result, the probability of correct match is low, even though the predictive performance of the CMMP may still be satisfactory. But, the performance can be improved, if the precision of the matching improves. In practice, there are often covariates at the group or cluster level, which are associated with the group-specific random effects. For example, consider the following nested-error regression (NER) model (Battese, Fuller & Harter 1988): $y_{ij} = x'_{ij}\beta + v_i + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where y_{ij} is the jth observed (or sampled) response in the ith cluster, x_{ij} is a vector of covariates, β is a vector of unknown regression coefficients (the fixed effects), v_i is a cluster-specific random effect, and e_{ij} is an additional error. In practice, the random effect v_i is often used to "capture the un-captured", that is, variation not captured by the mean function, $x'_{ij}\beta$, at the cluster level. On the other hand, some components of x_{ij} may be also at the cluster level, that is, they depend on i but not j. It is natural to think that there may be association between v_i and some of the cluster-level components of x_{ij}; however, we do not know what kind of association it is except that it must be nonlinear (because, otherwise, it would be captured

by $x'_{ij}\beta$). Nevertheless, if such covariate information can be utilized, the precision of the matching of random effects in CMMP, hence the (predictive) performance of CMMP can be improved.

Furthermore, the current CMMP method does not provide an uncertainty measure for the predictor, such as the mean squared prediction error (MSPE). The MSPE is extensively used in mixed model prediction (e.g., Jiang & Lahiri 2006, Rao & Malina 2015). However, when it comes to CMMP, because the latter involves a matching procedure which is non-differentiable, the traditional methods of deriving MSPE estimators, especially second-order unbiased MSPE estimators, do not apply.

The goal of this paper is two-fold. First, we are going to implement the idea described above regarding incorporating the covariate information in the CMMP matching. Second, we develop a simple, unified, Monte-Carlo assisted (Sumca) method for estimating the MSPE of CMMP, as noted above. Theoretical and empirical performance of the proposed new CMMP method and the Sumca MSPE estimator are carefully studied. Our results show that the new CMMP method can improve performance of CMMP substantially; furthermore, the Sumca estimator is second-order unbiased in estimating the MSPE of CMMP.

In Section 2 we introduce a new CMMP procedure that incorporates covariate information in matching the random effects. Estimation of MSPE of CMMP is considered in Section 3. Section 4 presents empirical results regarding performance of the new CMMP as well as that of the proposed MSPE estimator. A real data application is considered in

Section 5. Proofs of theoretical results are deferred to the Appendix.

2 A new CMMP procedure

Suppose that the underlying model can be expressed as

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

$i = 1, \ldots, m, j = 1, \ldots, n_i$, where w_i corresponds to the cluster-specific covatiates, and γ is the corresponding vector of regression coefficients. Here, α_i is used to capture whatever is not captured by $x'_{ij}\beta + w'_i\gamma$. Suppose that there is a one-to-one correspondence between i and w_i so that $w_i = w_{i'}$ implies $i = i'$. Also assume that a similar one-to-one correspondence holds between i and α_i. Then, to match α_i, one only needs to match w_i. This leads to the following simple identifier: Let the new observation satisfy

$$y_{\rm n} = x'_{\rm n}\beta + w'_{\rm n}\gamma + \alpha_I + \epsilon_{\rm n}, \tag{2}$$

where $x_{\rm n}, w_{\rm n}$ are observed, α_I is the new random effect, and $\epsilon_{\rm n}$ the new error. Let us, for now, focus on the matched case (Jiang *et al.* 2017), that is, there is an actual match between α_I and one of the α_is. Our goal is to predict the mixed effect

$$\theta = x'_{\rm n}\beta + w'_{\rm n}\gamma + \alpha_I. \tag{3}$$

First assume that all of the parameters are known. Suppose that there is a prior distribution, π, for I over $\{1, \ldots, m\}$ that is independent with the training data. A key idea is that $w_{\rm n}$ is

supposed to be related to the true class, I, that is,

$$w_{\mathrm{n}} \;=\; w_I. \tag{4}$$

Thus, given that $I = i$, (3) becomes $\theta = x_{\mathrm{n}}'\beta + w_i'\gamma + \alpha_i$. Therefore, the best predictor (BP), in the sense of minimizing the MSPE, of θ is

$$\begin{aligned}
\tilde{\theta}_{(i)} &= \mathrm{E}(x_{\mathrm{n}}'\beta + w_i'\gamma + \alpha_i|y) \\
&= \mathrm{E}(x_{\mathrm{n}}'\beta + w_i'\gamma + \alpha_i|y_i) \\
&= x_{\mathrm{n}}'\beta + w_i'\gamma + \mathrm{E}(\alpha_i|y_i).
\end{aligned} \tag{5}$$

It follows, by (3) and (5), that under the standard normality assumption that the α_is and ϵ_{ij}s are independent with $\alpha_i \sim N(0, G)$ and $\epsilon_{ij} \sim N(0, R)$, that the MSPE of $\tilde{\theta}_{(i)}$ is

$$\begin{aligned}
\mathrm{MSPE}\{\tilde{\theta}_{(i)}\} &= \mathrm{E}\{\mathrm{E}(\alpha_i|y_i) - \alpha_i\}^2 \\
&= \mathrm{E}\left(\frac{n_i G}{R + n_i G}\bar{\epsilon}_{i\cdot} - \frac{R}{R + n_i G}\alpha_i\right)^2 \\
&= \frac{GR}{R + n_i G}.
\end{aligned} \tag{6}$$

On the other hand, given that $I \neq i$, α_I is independent with $\mathrm{E}(\alpha_i|y_i)$. However, suppose that one does not know this, and therefore still assumes $I = i$, then the expression of $\tilde{\theta}_{(i)}$,

by (5), does not change. Therefore, one has

$$\begin{aligned}
\mathrm{MSPE}\{\tilde{\theta}_{(i)}\} &= \mathrm{E}\{(w_i - w_\mathrm{n})'\gamma + \mathrm{E}(\alpha_i|y_i) - \alpha_I\}^2 \\
&= \{(w_i - w_\mathrm{n})'\gamma\}^2 + \mathrm{E}\{\mathrm{E}(\alpha_i|y_i) - \alpha_I\}^2 \\
&= \{(w_i - w_\mathrm{n})'\gamma\}^2 + \mathrm{E}\{\mathrm{E}(\alpha_i|y_i)\}^2 + G \\
&= \{(w_i - w_\mathrm{n})'\gamma\}^2 + G\left(1 + \frac{n_i G}{R + n_i G}\right). \qquad (7)
\end{aligned}$$

If we combine the above results, we have

$$\begin{aligned}
\mathrm{MSPE}\{\tilde{\theta}_{(i)}\} &= \mathrm{E}(\mathrm{E}[\{\tilde{\theta}_{(i)} - \theta\}^2|I]) \\
&= \pi(I = i)\mathrm{E}[\{\tilde{\theta}_{(i)} - \theta\}^2|I = i] + \sum_{i' \neq i} \pi(I = i')\mathrm{E}[\{\tilde{\theta}_{(i)} - \theta\}^2|I = i'] \\
&= \pi(I = i)\frac{GR}{R + n_i G} \\
&\quad + \sum_{i' \neq i} \pi(I = i')\left[\{(w_i - w_\mathrm{n})'\gamma\}^2 + G\left(1 + \frac{n_i G}{R + n_i G}\right)\right] \\
&= \{1 - \pi(I = i)\}\{(w_i - w_\mathrm{n})'\gamma\}^2 + G\left(1 + \frac{n_i G}{R + n_i G}\right) \\
&\quad - \pi(I = i)\frac{2n_i G^2}{R + n_i G}. \qquad (8)
\end{aligned}$$

If $\pi(I = i)$ is known, one can identify I as the minimizer of the right side of (8), that is,

$$\begin{aligned}
\tilde{I} &= \mathrm{argmin}_{1 \leq i \leq m}\left[\{1 - \pi(I = i)\}\{(w_i - w_\mathrm{n})'\gamma\}^2 + G\left(1 + \frac{n_i G}{R + n_i G}\right)\right. \\
&\quad \left. - \pi(I = i)\frac{2n_i G^2}{R + n_i G}\right]. \qquad (9)
\end{aligned}$$

For example, if $\pi(I = i) = 1/m, 1 \leq i \leq m$, then (9) reduces to

$$\tilde{I} = \text{argmin}_{1\leq i\leq m}\left[\frac{m-1}{m}\{(w_i - w_{\text{n}})'\gamma\}^2 + G\left(1 + \frac{n_i G}{R + n_i G}\right) - \frac{2n_i G^2}{m(R + n_i G)}\right]. \tag{10}$$

(10) may be interpreted as the minimizer of MSPE under the non-informative prior for I. In particular, if n_i does not depend on i, then (10) is equivalent to

$$\tilde{I} = \text{argmin}_{1\leq i\leq m}\{(w_i - w_{\text{n}})'\gamma\}^2. \tag{11}$$

In practice, when the parameters are unknown, they are replaced by their consistent estimators, say, the REML estimators, $\hat{\gamma}, \hat{G}, \hat{R}$, leading to

$$\hat{I} = \text{argmin}_{1\leq i\leq m}\left[\frac{m-1}{m}\{(w_i - w_{\text{n}})'\hat{\gamma}\}^2 + \hat{G}\left(1 + \frac{n_i \hat{G}}{\hat{R} + n_i \hat{G}}\right) - \frac{2n_i \hat{G}^2}{m(\hat{R} + n_i \hat{G})}\right] \tag{12}$$

for the unequal n_i case, and $\hat{I} = \text{argmin}_{1\leq i\leq m}\{(w_i - w_{\text{n}})'\hat{\gamma}\}^2$ for the equal n_i case.

One concern about the above procedure is that the choice of prior (e.g., uniform) is a bit subjective. An alternative that does not depend on the choice of the prior is to focus on the case of misspecification only. From (7), it is seen that, if $I \neq i$, the MSPE of $\tilde{\theta}_{(i)}$ is given by the right side of (7). This leads to a modified procedure:

$$\hat{I} = \text{argmin}_{1\leq i\leq m}\left[\{(w_i - w_{\text{n}})'\hat{\gamma}\}^2 + \hat{G}\left(1 + \frac{n_i \hat{G}}{\hat{R} + n_i \hat{G}}\right)\right], \tag{13}$$

where $\hat{\gamma}, \hat{G}, \hat{R}$ are the same as in (12).

A major difference between (13) and the original CMMP procedure of Jiang *et al.* (2017) is that the covariate information in w is incorporated in the identification of I. Using a similar argument as in Jiang *et al.* (2017), consistency of the new CMMP procedure can be established under reasonable assumptions. Empirical performance of the new CMMP procedure will be evaluated in Section 4.

3 Estimation of MSPE

As noted, a standard uncertainty measure for a predictor is the MSPE. A "gold standard" for the MSPE estimation is to produce a second-order unbiased MSPE estimator, that is, the order of bias of the MSPE estimator is $o(m^{-1})$, where m is the total number of clusters in the training data. Typically, the $o(m^{-1})$ term is, in fact, $O(m^{-2})$, but this difference is usually ignored. For the most part, there have been two approaches for producing a second-order unbiased MSPE estimator. The first is the Prasad-Rao linearization method (Prasad & Rao 1990). The approach uses Taylor series expansion to obtain a second-order approximation to the MSPE, then corrects the bias, again to the second-order, to produce an MSPE estimator whose bias is $o(m^{-1})$. Various extensions of the Prasad-Rao method have been developed; see, for example, Datta & Lahiri (2000), Jiang & Lahiri (2001), Das, Jiang & Rao (2004), and Datta, Rao & Smith (2005). Although the method often leads to an analytic expression of the MSPE estimator, the derivation is tedious, and the final expression

is likely to be complicated. More importantly, errors often occur in the process of analytic derivations as well as computer programming based on the lengthy expressions. Furthermore, the linearization method does not apply to situations where a non-differentiable operation is involved in obtaining the predictor, such as shrinkage estimation (e.g., Tibshirani 1996), CMMP (Jiang *et al.* 2017), as well as the new CMMP developed in Section 2.

The second approach to second-order unbiased MSPE estimation is resampling methods. Jiang, Lahiri & Wan (2002; hereafter JLW) proposed a jackknife method to estimate the MSPE of an empirical best predictor (EBP). The method avoids tedious derivations of the Prasad-Rao method, and is "one formula for all". On the other hand, there are restrictions on the class of predictors to which JLW applies. Namely, JLW only applies to empirical best predictor (EBP), that is, predictor obtained by replacing the parameters involved in the best predictor (BP), which is the conditional expectation, by their (consistent) estimators. The CMMP predictor, however, is not an EBP, because it involves a matching process. Jiang, Lahiri & Nguyen (2017) proposed a Monte-Carlo jackknife method, call McJack, which potentially applies to CMMP; however, the method is computationally very expensive (see below). Another resampling-based approach is double bootstrapping (DB; Hall & Maiti 2006a,b). Although DB is capable of producing a second-order unbiased MSPE estimator, it is, perhaps, computationally even more intensive than the McJack. It is also unclear whether DB can be extended to CMMP.

In a way, the method to be proposed below may be viewed as a hybrid of the lineariza-

tion method and resampling method, by combining the best part of each method. In short, we use a simple, analytic approach to obtain the leading term of our MSPE estimator, and a Monte-Carlo method to take care a remaining, lower-order term. The computational cost for the Monte-Carlo part is much lesser compared to McJack. For example, the computational burden of our method is about $1/m^3$ to $1/m^2$ of that of McJack. More importantly, the method provides a unified, conceptually easy solution to a difficult problem, that is, obtaining a second-order unbiased MSPE estimator for CMMP (either that of Jiang *et al.* 2017 or the new CMMP proposed in Section 2).

Let θ be the mixed effect corresponding to the new observations, and $\hat{\theta}$ the CMMP predictor of θ. The MSPE of $\hat{\theta}$ can be expressed as

$$\mathrm{MSPE} = \mathrm{E}(\hat{\theta} - \theta)^2 = \mathrm{E}\left[\mathrm{E}\{(\hat{\theta} - \theta)^2 | y\}\right], \tag{14}$$

where y represents the available data. Suppose that the underlying distribution of y depends on a vector of unknown parameters, ϕ. Then, the conditional expectation inside the expectation on the right side of (14) is a function of y and ϕ, which can be written as

$$a(y,\phi) = \mathrm{E}\{(\hat{\theta} - \theta)^2 | y\} = \hat{\theta}^2 - 2\hat{\theta}\mathrm{E}(\theta|y) + \mathrm{E}(\theta^2|y) = \hat{\theta}^2 - 2\hat{\theta}a_1(y,\phi) + a_2(y,\phi), \tag{15}$$

where $a_j(y,\phi) = \mathrm{E}(\theta^j|y), j = 1, 2$. If we replace the ϕ in (15) by $\hat{\phi}$, a consistent estimator of ϕ, the result is a first-order unbiased estimator, that is, we have

$$\mathrm{E}\{a(y,\hat{\phi}) - a(y,\phi)\} \;\; = \;\; O(m^{-1}). \tag{16}$$

On the other hand, both $\mathrm{MSPE} = \mathrm{E}\{a(y,\phi)\}$ [by (14), (15)] and $\mathrm{E}\{a(y,\hat{\phi})\}$ are functions of ϕ, denoted by $b(\phi)$ and $c(\phi)$, respectively. By (16), we have $d(\phi) = b(\phi) - c(\phi) = O(m^{-1})$; thus, if we replace, again, ϕ by $\hat{\phi}$ in $d(\phi)$, the difference is a lower-order term, that is, we have

$$d(\hat{\phi}) - d(\phi) \;\; = \;\; o_{\mathrm{P}}(m^{-1}) \tag{17}$$

[see, e.g., Jiang 2010, sec. 3.4 for notation like o_{P} and O_{P}]. Now consider the estimator

$$\widehat{\mathrm{MSPE}} = a(y,\hat{\phi}) + d(\hat{\phi}) = a(y,\hat{\phi}) + b(\hat{\phi}) - c(\hat{\phi}). \tag{18}$$

We have, by (14)–(18), $\mathrm{E}(\widehat{\mathrm{MSPE}}) = \mathrm{E}\{a(y,\phi)\} + \mathrm{E}\{a(y,\hat{\phi}) - a(y,\phi)\} + \mathrm{E}\{d(\hat{\phi})\} = \mathrm{MSPE} + \mathrm{E}\{d(\hat{\phi}) - d(\phi)\} = \mathrm{MSPE} + o(m^{-1})$. Essentially, this one-line, heuristic derivation shows the second-order unbiasedness of the proposed MSPE estimator, (18), provided that the terms involved can be evaluated. A rigorous justification is given in Appendix.

Note that the leading term, $a(y,\hat{\phi})$, in (18) is guaranteed positive, a desirable property for an MSPE estimator. The lower-order term, $b(\hat{\phi}) - c(\hat{\phi})$, corresponds to a bias correction to the leading term. This term is typically much more difficult to evaluate than the leading term. We propose to approximate this term using a Monte-Carlo method. Let P_ϕ denote the distribution of y with ϕ being the true parameter vector. Given ϕ, one can generate y under P_ϕ. Let $y_{[k]}$ denote y generated under the kth Monte-Carlo sample, $k = 1, \ldots, K$. Then, by the law of large numbers, we have

$$b(\phi) - c(\phi) \;\; \approx \;\; \frac{1}{K}\sum_{k=1}^{K}\left\{a(y_{[k]},\phi) - a(y_{[k]},\hat{\phi}_{[k]})\right\}, \tag{19}$$

where $\hat{\phi}_{[k]}$ denotes $\hat{\phi}$ based on $y_{[k]}$. If K is sufficiently large, which one has control over during the Monte-Carlo simulation, the difference between the two sides of (19) is $o(m^{-1})$. Write the right side of (19) as $d_K(\phi)$ (note that $y_{[k]}, k = 1, \ldots, K$ also depend on ϕ). Then, a Monte-Carlo assisted MSPE estimator is given by

$$\widehat{\mathrm{MSPE}}_K = a(y, \hat{\phi}) + d_K(\hat{\phi}) = a(y, \hat{\phi}) + \frac{1}{K}\sum_{k=1}^{K}\left\{a(y_{[k]}, \hat{\phi}) - a(y_{[k]}, \hat{\phi}_{[k]})\right\}, \qquad (20)$$

where $y_{[k]}, k = 1, \ldots, K$ are generated as above with $\phi = \hat{\phi}$, and $\hat{\phi}_{[k]}$ is, again, the estimator of ϕ based on $y_{[k]}$. (20) is called the Sumca estimator of the MSPE of $\hat{\theta}$ (Sumca is abbreviation of "simple, unified, Monte-Caro assisted"). It is shown in Appendix that, under regularity conditions, the Sumca estimator is second-order unbiased. Empirical performance of the Sumca estimator will be evaluated in Section 4.

4 Simulation studies

4.1 Performance of new CMMP

In Jiang *et al.* (2017), the authors showed that CMMP significantly outperforms the standard regression prediction (RP) method. On the other hand, the latter authors have not compared CMMP with mixed model prediction (MMP; e.g., Jiang & Lahiri 2006, Rao & Molina 2015), which is known to outperform RP as well. In fact, some unpublished simulation results suggest that CMMP may not outperform MMP, depending on the situation.

Further investigation shows that most of the prediction errors by CMMP are due to mismatching the classes; in other words, if the classes are matched with high accuracy, CMMP is expected to outperform MMP.

In this subsection, we compare empirical performance of the new CMMP, developed in Section 2, with MMP under two scenarios. In each of the scenarios, there is no exact match between the new observation and a group in the training data. More specifically, under Scenario I, the training data satisfy

$$y_{ij} \quad = \quad \beta_0 + \beta_1 w_i + \alpha_i + \epsilon_{ij}, \tag{21}$$

$i = 1, \ldots, m, j = 1, \ldots, n_i$, where w_i is an observed, cluster-level covariate, α_i is a cluster-specific random effect, and ϵ_{ij} is an error. The random effects and errors are independent with $\alpha_i \sim N(0, G)$ and $\epsilon_{ij} \sim N(0, R)$. The new observation, on the other hand, satisfies

$$y_{\text{new}} \quad = \quad \beta_0 + \beta_1 w_1 + \alpha_1 + \delta + \epsilon_{\text{new}}, \tag{22}$$

where $\delta, \epsilon_{\text{new}}$ are independent with $\delta \sim N(0, D)$ and $\epsilon_{\text{new}} \sim N(0, R)$, and $(\delta, \epsilon_{\text{new}})$ are independent with the training data. It is seen that, because of δ, there is no exact match between the new random effect (which is $\alpha_1 + \delta$) and one of the random effects α_i associated with the training data; however, the value of D is small, $D = 10^{-4}$, hence there is an approximate match between the new random effect and α_1, the random effect associated with the first group in the training data.

Under Scenario 2, the training data satisfy (21) except that now $\alpha_i = w_i^3 + v_i$ with $v_i \sim$

$N(0, D)$, and v_i is independent with ϵ_{ij}. Under this scenario, there is a misspecification of the cluster-specific random effect in that the random effect represents a nonlinear function of the covariate, plus some random noise. This is motivated by the notion of "capture the un-captured" discussed in Section 1 (second paragraph). Similarly, the new observation satisfies (22) with $\alpha_1 + \delta$ replaced by $w_1^3 + v_{\text{new}}$, where $v_{\text{new}} \sim N(0, D)$ and is independent with ϵ_{new}, and $(v_{\text{new}}, \epsilon_{\text{new}})$ are independent with the training data.

We consider $m = 50$. The n_i are chosen according to one of the following four patterns: 1. $n_i = 5, 1 \leq i \leq m/2$; $n_i = 25, m/2 + 1 \leq i \leq m$; 2. $n_i = 50, 1 \leq i \leq m/2$; $n_i = 250, m/2 + 1 \leq i \leq m$; 3. $n_i = 25, 1 \leq i \leq m/2$; $n_i = 5, m/2 + 1 \leq i \leq m$; 4. $n_i = 250, 1 \leq i \leq m/2$; $n_i = 50, m/2 + 1 \leq i \leq m$. The consideration of the first two patterns is to see how results change when the cluster sizes of the training data get bigger; the consideration of the last two patterns, in comparison with the first two patterns, is to see if the apparent asymmetry due to the fact the the new random effect has an approximate match with the first half of the training data (i.e., α_1) affects the results. The true βs are $\beta_0 = 5$ and $\beta_1 = 1$. The results, based on 1000 simulation runs are presented in Tables 1 and 2. The numbers in the rows of CMMP and MMP are empirical MSPEs based on the simulation; and %Imp represents percentage improvement of CMMP over MMP, that is

$$\%\text{Imp} \quad = \quad 100\% \times \left(\frac{\text{MMP} - \text{CMMP}}{\text{CMMP}} \right).$$

It is seen that, under both scenarios, CMMP improves MMP substantially, as shown

Table 1: **Comparing CMMP with MMP: Scenario I**

		$R=1$					$G=1$				
		$G=.25$	$G=.5$	$G=1$	$G=2$	$G=4$	$R=.25$	$R=.5$	$R=1$	$R=2$	$R=4$
Pattern 1	CMMP	.123	.156	.171	.172	.176	.049	.091	.183	.294	.485
	MMP	.187	.342	.459	.590	.801	.190	.323	.476	.650	.836
	%Imp	52.2	118.8	168.4	243.3	355.3	285.9	255.5	160.5	121.0	72.4
Pattern 2	CMMP	.019	.018	.020	.020	.021	.005	.010	.019	.039	.070
	MMP	.178	.315	.494	.618	.780	.186	.321	.526	.633	.718
	%Imp	861.8	1630.8	2351.2	3023.6	3634.0	3483.5	3138.3	2701.5	1504.3	918.8
Pattern 3	CMMP	.034	.034	.035	.039	.041	.010	.019	.038	.084	.140
	MMP	.175	.283	.481	.642	.894	.199	.335	.476	.620	.749
	%Imp	418.4	725.9	1272.6	1562.1	2099.8	1890.1	1617.0	1155.1	638.7	434.4
Pattern 4	CMMP	.004	.004	.004	.004	.004	.001	.002	.004	.008	.016
	MMP	.181	.296	.466	.552	.809	.193	.308	.442	.631	.756
	%Imp	4235.2	7132.0	10918.6	12942.3	19873.0	16859.0	14875.6	10670.7	7481.8	4521.1

Table 2: **Comparing CMMP with MMP: Scenario II**

		$R = 0.25$	$R = 0.5$	$R = 1$	$R = 2$	$R = 4$
Pattern 1	CMMP	0.049	0.104	0.216	0.498	1.004
	MMP	0.264	0.539	1.273	2.736	5.357
	%Imp	444.1	416.8	489.1	449.0	433.5
Pattern 2	CMMP	0.005	0.010	0.020	0.040	0.079
	MMP	0.292	0.602	1.250	2.786	5.151
	%Imp	5656.1	5742.3	5998.8	6815.5	6457.1
Pattern 3	CMMP	0.010	0.020	0.045	0.087	0.177
	MMP	0.260	0.553	1.176	2.977	5.550
	%Imp	2554.1	2613.1	2510.5	3307.8	3027.4
Pattern 4	CMMP	0.001	0.002	0.004	0.007	0.017
	MMP	0.281	0.593	1.167	2.598	5.564
	%Imp	23012.9	27150.1	27283.9	35431.9	33316.0

by, for example, %Imp. Under Scenario I, the MSPE of both CMMP and MMP increases with G and with R when the other variance component is held fix. In term of %Imp, it increases with G but decreases with R. This makes sense because large G makes identification of the group (effect) more important, hence increasing the advantage of CMMP; on the other hand, larger R makes it less accurate to estimate the group (or cluster) effect, hence decreasing the advantage of CMMP. The amount of improvement by CMMP increases substantially, as shown by %Imp, as the cluster sizes increase. Comparatively, the advantage of CMMP over MMP is more significant, some above 10,000, under Patterns 3–4 than under Patterns 1–2. The explanation is that, under Patterns 3–4, the cluster sizes for the matching group (i.e., group 1) is relatively larger, making estimation of the cluster effect more accurate, hence increasing the advantage of CMMP. Under Scenario II, the improvement of CMMP over MMP is even more substantial, with some %Imp over 30,000. The MSPEs follow the same pattern, but %Imp does not show a clear trend, with the maximum occurring near the mid-range of R. Note that there is no G under Scenario II because the random effect is misspecified in this situation.

4.2 Performance of Sumca estimator

In this subsection, we investigate empirical performance of the Sumca MSPE estimator, developed in Section 3. We consider, again, model (21), but under two scenarios with different accuracy in terms of matching the groups. More specifically, under the first sce-

nario, one has (21) with $\beta_0 = 5$ and $\beta_1 = 1$; $w_i = -2.5 + 0.1i$; $\alpha_i = w_i^3 + v_i$ with $v_i \sim N(0, 10^{-4})$. The new observation satisfies (22). Furthermore, we consider two cases of sample size configurations for n_i, namely,

$$\text{Case 1}: \; n_i = \begin{cases} 5, & 1 \le i \le m/2 \\ 25, & m/2+1 \le i \le m \end{cases} \qquad \text{Case 2}: \; n_i = \begin{cases} 50, & 1 \le i \le m/2 \\ 250, & m/2+1 \le i \le m \end{cases}$$

The results, based on 1000 simulation runs, are presented in Table 3, where the percentage relative bias is defined as

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

Here, MSPE is the MSPE of the CMMP $\hat{\theta}$ of $\theta = \beta_0 + \beta_1 w_1 + \alpha_I$ based on the simulation runs, and $\widehat{\text{MSPE}}$ is the mean of the Sumca estimator over the simulation runs. As a comparison, we also consider what we call the 1st-term estimator which is the first term on the right side of (20). It is seen that the Sumca estimator improves the 1st-term estimator, as indicated by the %RB (note that these numbers are percentages), but the improvement is not significant, especially when m is larger. One possible reason is that, under this scenario, the percentage of correct matching (% match, that is, the group number identified by the CMMP procedure is the same as the true group number, $I = 1$) is 100%.

The improvement of Sumca over 1st-term is more significant, however, under the second scenario. In this case, w_i is the same as above, but $\alpha_i = 0.5(w_i^2 - 1) + v_i$ with $v_i \sim N(0, 0.01)$. In a way, the "signal", in terms of w_i, for identifying α_i is weaker, but the "noise", in terms of v_i, is stronger, compared to the first scenario. Furthermore, the cluster

Table 3: **MSPE Estimation:** First Scenario

	R	0.25	0.5	1	2	4
Case 1	MSPE	0.051	0.107	0.220	0.462	1.011
	Sumca	0.050	0.099	0.197	0.382	0.730
	%RB	-0.66	-7.17	-10.14	-17.28	-27.83
	1st-term	0.050	0.099	0.196	0.377	0.711
	%RB	-0.83	-7.49	-10.80	-18.46	-29.74
Case 2	MSPE	0.0055	0.0104	0.0209	0.0396	0.0858
	Sumca	0.0050	0.0100	0.0199	0.0398	0.0790
	%RB	-8.45	-4.02	-4.38	0.56	-7.92
	1st-term	0.0050	0.0100	0.0199	0.0397	0.0787
	%RB	-8.46	-4.05	-4.45	0.41	-8.19

Table 4: **MSPE Estimation:** Second Scenario

R	0.25	0.5	1	2	4
MSPE	0.064	0.149	0.307	0.651	1.143
Sumca	0.083	0.118	0.174	0.284	0.437
%RB	29.43	-20.94	-43.37	-56.32	-61.79
1st-term	0.025	0.054	0.108	0.198	0.323
%RB	-61.03	-63.76	-64.93	-69.57	-71.72
%match	78.2	59.1	49.8	40.0	34.6

sizes are no longer fixed; instead, the n_is are randomly selected from the integers between 5 and 25, and the selection changes with each simulation. As a result, the % match is no longer 100%, ranging from 34.6% to 78.2%. The results, again based on 1000 simulation runs, are presented in Table 4. It is seen that the improvement of Sumca over 1st-term, in terms of %RB, is more significant now, especially for smaller values of R.

5 Real data example

Datta, Lahiri, and Maiti (2002) considered a data set regarding median income of four-person families for the fifty states of U.S. and the District of Columbia using cross-sectional and time series modeling. The primary source of data is the annual supplement to the March Sample of the Current Population Survey (CPS), which provides individual annual income

data categorized into intervals of $2500. The direct survey estimates were obtained from the CPS using linear interpolation. Two secondary sources of data were also available. The first is the U.S. decennial censuses (Census) which produce median incomes for the 50 states and D.C. based on the "long form" filled out by approximately one-sixth of the U.S. population. These census median income estimates are believed to be free of sampling errors. The second is per-capita income estimates produced by the Bureau of Economic Analysis (BEA) division of the U.S. Department of Commerce. Since the per-capita income estimates are not based on any sampling techniques, they do not have any sampling errors associated with them. From the Census and BEA data, an adjusted census median income (adjusted Census) is obtained by multiplying the preceding census median income by the ratio of BEA per-capita income for the current year to that of the preceding census year.

The data are used to illustrate the new CMMP method as well as the Sumca MSPE estimation. The 51 states and D.C. are considered as 51 clusters or groups. The 1979–1988 (10 years) are used as the training data; the 1989 data are treated as the new data. Our goal is to predict the mixed effect associated with the new data for each of the 51 states and D.C. The following NER model is considered for the training data:

$$y_{ij} \quad = \quad \beta_0 + \beta_1 x_{ij} + \gamma w_i + \alpha_i + \epsilon_{ij}, \tag{23}$$

$i = 1, \ldots, 51$, $j = 1, \ldots, 10$, where y_{ij} is the direct survey estimate obtained from the CPS, and x_{ij}, w_i correspond to the adjusted Census and Census variables described above, respectively. Note that, here, w_i is a cluster-level covariate that can be used to help with the

matching, and α_i is a cluster-specific random effect that is used to capture the "un-captured" variation (by the covariates) at the cluster level. A similar model is assumed for the new data. We apply the new CMMP method of Section 2, more specifically, (13) to the data. Furthermore, we compute the associated Sumca MSPE estimates (see Section 3), and use the square roots of them as measures of uncertainty. The results are presented in Figure 1, where the (red) dot represents the value of CMMP predictor while the dash line represents the error marginal determined by plus/minus 2 times the square root of the corresponding Sumca estimate, for each of the 51 states. Note that the % match is 100% in this case.

Appendix: Second-order unbiasedness of Sumca estimator

A1 Second-order unbiasedness of (18)

We impose the following regularity conditions.

A1. $\mathrm{E}(\theta^2|y)$ is finite almost surely.

A2. The parameter space for ϕ, Φ, is compact, and $\hat{\phi} \in \Phi$.

A3. $\mathrm{E}(|\hat{\phi} - \phi|^4) = O(m^{-2})$.

A4. $\mathrm{E}\{\hat{\theta}(\partial a_1/\partial \phi')(\hat{\phi} - \phi)\} = O(m^{-1})$, $\mathrm{E}\{(\partial a_2/\partial \phi')(\hat{\phi} - \phi)\} = O(m^{-1})$.

A5. The second moments of the following are finite, where $\|\cdot\|$ denotes the spectral norm of a matrix:

$$|\hat{\theta}| \sup_{\phi \in \Phi} \left\| \frac{\partial^2 a_1}{\partial \phi \partial \phi'} \right\|, \quad \sup_{\phi \in \Phi} \left\| \frac{\partial^2 a_2}{\partial \phi \partial \phi'} \right\|.$$

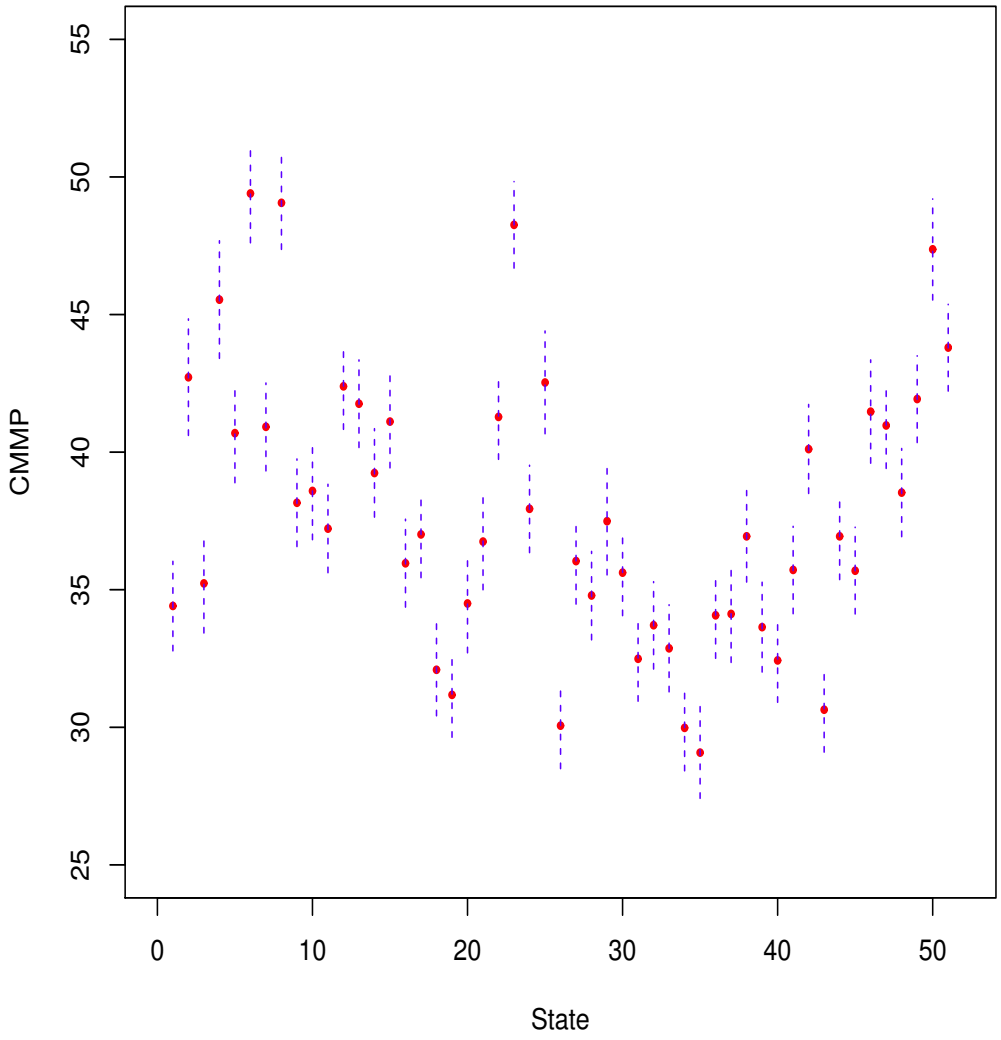

Figure 1: *Predicted Mixed Effects for 1989 via CMMP for Income Data: Dots indicate Predicted Values while Dash Lines Margins of Errors*

A6. The following expansion holds for $d(\phi)$:

$$d(\phi) \;=\; d_1(\phi)m^{-1} + r(\phi)$$

for some $d_1(\cdot)$, $r(\cdot)$ satisfying $\sup_{\phi\in\Phi}|r(\phi)| = o(m^{-1})$ and $\mathrm{E}\{|d_1(\hat{\phi}) - d_1(\phi)|\} = o(1)$.

A1, *A2* are satisfied in most practical situations. *A3* is satisfied, for example, for the consistent estimators truncated from above considered in Jiang *et al.* (2002) and Das *et al.* (2004). *A4* is expected to hold, for example, if the data are divided into independent clusters while the mixed effect of interest is associated with a finite number of clusters. Due to the compactness of Φ, *A5* is expected to hold. Finally, *A6* is suggested by the fact that $d(\phi) = O(m^{-1})$ [see below (16)], which can be shown to hold under *A1–A5*; that $\mathrm{E}\{|d_1(\hat{\phi}) - d_1(\phi)|\} = o(1)$ is suggested by the consistency of $\hat{\phi}$, if $d_1(\cdot)$ is continuous.

Theorem 1. Under Assumptions *A1–A6*, the MSPE estimator given by (18) is second-order unbiased, that is, $\mathrm{E}(\widehat{\mathrm{MSPE}}) = \mathrm{MSPE} + o(m^{-1})$.

Proof. Essentially, we need to verify that (16) and (17) hold, which are used in the heuristic derivation below (18) of the second-order unbiasedness. But first note that *A2* implies that all of the conditional expectations involved in (15) exist and are finite almost surely. To verify (16), note that, by Taylor series expansion, we have

$$a_2(y,\hat{\phi}) - a_2(y,\phi) = \frac{\partial a_2}{\partial \phi'}(\hat{\phi}-\phi) + \frac{1}{2}(\hat{\phi}-\phi)'\left(\frac{\partial^2 a_2}{\partial\phi\partial\phi'}\bigg|_{\tilde{\phi}}\right)(\hat{\phi}-\phi), \qquad \text{(A.1)}$$

where $\tilde{\phi}$ lies between ϕ and $\hat{\phi}$. The second term on the right side of (A.1) without the factor $1/2$ is bounded in absolute value by $\sup_{\phi\in\Phi}\|\partial^2 a_2/\partial\phi\partial\phi'\|\cdot|\hat{\phi}-\phi|^2$, whose expected value

is bounded by

$$\left\{ \mathrm{E}\left(\sup_{\phi\in\Phi} \left\| \frac{\partial^2 a_2}{\partial\phi\partial\phi'} \right\| \right)^2 \right\}^{1/2} \left\{ \mathrm{E}(|\hat{\phi}-\phi|^4) \right\}^{1/2},$$

by the Cauchy-Schwarz inequality. Thus, by *A3*, *A5*, the second term on the right side of (A.1) is $O(m^{-2})$ in $\mathrm{E}(|\cdots|)$. Furthermore, by *A4*, the first term on the right side of (A.1) is $O(m^{-1})$ in expectation. It follows that the left side of (A.1) is $O(m^{-1})$ in expectation. Similarly, we have, by Taylor expansion,

$$\hat{\theta}\{a_1(y,\hat{\phi}) - a_1(y,\phi)\} = \hat{\theta}\frac{\partial a_1}{\partial\phi'}(\hat{\phi}-\phi) + \frac{1}{2}(\hat{\phi}-\phi)\hat{\theta}\left(\frac{\partial^2 a_1}{\partial\phi\partial\phi'}\bigg|_{\tilde{\phi}} \right)(\hat{\phi}-\phi). \tag{A.2}$$

Thus, by *A3–A5*, and similar argument as above, the left side of (A.2) is $O(m^{-1})$ in expectation. (16) now follows by the proved results and expression

$$a(y,\hat{\phi}) - a(y,\phi) = a_2(y,\hat{\phi}) - a_2(y,\phi) - 2\hat{\theta}\{a_1(y,\hat{\phi}) - a_1(y,\phi)\}.$$

To verify (17), note that, by *A6*, we have

$$d(\hat{\phi}) - d(\phi) = \frac{d_1(\hat{\phi}) - d_1(\phi)}{m} + r(\hat{\phi}) - r(\phi). \tag{A.3}$$

We have $|r(\hat{\phi}) - r(\phi)| \le 2\sup_{\phi\in\Phi}|r(\phi)| = o(m^{-1})$, by *A2* and *A6*. It follows that (17) holds, again by *A6*. In fact, by *A6*, we have

$$|\mathrm{E}\{d(\hat{\phi}) - d(\phi)\}| \le \frac{\mathrm{E}(|d_1(\hat{\phi}) - d_1(\phi)|)}{m} + o(m^{-1}) = o(m^{-1}),$$

which is a key argument used in the heuristic derivation below (18).

A2 Second-order unbiasedness of Sumca estimator

We assume that the Monte-Carlo (MC) samples, under ϕ, are generated by first generating some standard [e.g., $N(0,1)$] random variables, say, ξ, that do not depend on ϕ. We then combine ξ with ϕ to produce the MC samples under ϕ. For example, y_{ij}'s are generated by first generating the ξ_i's and η_{ij}'s, which are independent $N(0,1)$, and then letting $y_{ij} = x'_{ij}\beta + \sqrt{G}\xi_i + \sqrt{R_i}\eta_i$, where G and R_i are known functions of ϕ.

Theorem 2. Suppose that ξ are independent of y, the original data. Then, under the assumptions of Theorem 1, we have $\mathrm{E}(\widehat{\mathrm{MSPE}}_K) = \mathrm{MSPE} + o(m^{-1})$, where the expectation is with respect to the joint distribution of the data and Monte-Carlo sampling.

Proof. First, let us re-clarify some concepts and notation introduced in Section 3. Recall $d(\phi) = b(\phi) - c(\phi)$, where $b(\phi) = \mathrm{E}\{a(y,\phi)\}, c(\phi) = \mathrm{E}\{a(y,\hat{\phi})\}$. Note that $b(\phi)$ is the MSPE when ϕ is the true parameter vector. Denote the right side of (19) by $\hat{d}(\psi)$, which is an approximation to $d(\psi)$. Note that, in (19), $y_{[k]}$ is y generated under ψ through the ξ introduced above, which does not depend on ϕ and is independent with $\hat{\phi}$, the estimator of ϕ based on the original data (by the assumption of Theorem 2). Furthermore, $\hat{\phi}_{[k]}$ is a function of $y_{[k]}$. Thus, the summand in (19) is a function of $\xi_{[k]}$, the copy of ξ generated in the kth Monte-Carlo simulation, and ϕ. Denote the summands by $\Delta(\xi_{[k]}, \phi), 1 \le k \le K$. Then, we have $d(\phi) = \mathrm{E_d}\{a(y,\phi) - a(y,\hat{\phi})\} = \mathrm{E_{mc}}\{\Delta(\xi,\phi)\}$, where $\mathrm{E_d}$ denotes expectation with respect to the data, and $\mathrm{E_{mc}}$ that with respect to the Monte-Carlo simulation. The two expectations are equal because y can be generated the same way as $y_{[k]}$, through ξ, given

the same ϕ. It follows that

$$\mathrm{E}_{\mathrm{mc}}\{\hat{d}(\phi)\} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{E}_{\mathrm{mc}}\{\Delta(\xi_{[k]},\phi)\} = \mathrm{E}_{\mathrm{mc}}\{\Delta(\xi,\phi)\} = d(\phi). \tag{A.4}$$

The Sumca estimator, (20), can now be expressed as

$$\widehat{\mathrm{MSPE}}_K = a(y,\hat{\phi}) + \hat{d}(\hat{\phi}), \tag{A.5}$$

where, in $\hat{d}(\hat{\phi})$, that is, the summand in (20), $y_{[k]}$ is generated via $\xi_{[k]}$ and $\phi = \hat{\phi}$ as described above. In other words, the summands in (20) are $\Delta(\xi_{[k]},\hat{\phi}), 1 \le k \le K$.

By the proof of Theorem 1, we have $\mathrm{E}_{\mathrm{d}}\{d(\hat{\phi}) - d(\phi)\} = o(m^{-1})$. Thus, we have

$$\begin{aligned}\mathrm{E}\{\hat{d}(\hat{\phi}) - d(\phi)\} &= \mathrm{E}\{\hat{d}(\hat{\phi}) - d(\hat{\phi})\} + \mathrm{E}\{d(\hat{\phi}) - d(\phi)\} \\ &= \mathrm{E}\{\hat{d}(\hat{\phi}) - d(\hat{\phi})\} + o(m^{-1}),\end{aligned} \tag{A.6}$$

where E denotes expectation with respect to both the data and Monte-Carlo simulation. Note that $d(\hat{\phi}) - d(\phi)$ depends only on y and not on ξ.

On the other hand, we have $\mathrm{E}\{\hat{d}(\hat{\phi}) - d(\hat{\phi})\} = \mathrm{E}[\mathrm{E}\{\hat{d}(\hat{\phi}) - d(\hat{\phi})|\hat{\phi}\}]$. For any given value of ϕ, we have, by the independence of ξ and y, and (A.4),

$$\mathrm{E}\{\hat{d}(\hat{\phi}) - d(\hat{\phi})|\hat{\phi} = \phi\} = \mathrm{E}\{\hat{d}(\phi) - d(\phi)|\hat{\phi} = \phi\} = \mathrm{E}_{\mathrm{mc}}\{\hat{d}(\phi) - d(\phi)\} = 0.$$

Note that $\hat{d}(\phi) - d(\phi)$ depends only on ξ and not on y. Thus, $\mathrm{E}\{\hat{d}(\hat{\phi}) - d(\hat{\phi})|\hat{\phi}\} = 0$, hence

$$\mathrm{E}\{\hat{d}(\hat{\phi}) - d(\hat{\phi})\} = 0. \tag{A.7}$$

Combining (A.6), (A.7), we have $\mathrm{E}\{\hat{d}(\hat{\phi}) - d(\phi)\} = o(m^{-1})$. Therefore, by (A.5), we have $\mathrm{E}(\widehat{\mathrm{MSPE}}_K) = \mathrm{E}\{a(y, \hat{\phi})\} + d(\phi) + \mathrm{E}\{\hat{d}(\hat{\phi}) - d(\phi)\} = c(\phi) + b(\phi) - c(\phi) + o(m^{-1}) = b(\phi) + o(m^{-1}) = \mathrm{MSPE} + o(m^{-1})$.

References

[1] Battese, G. E., Fuller, W. A. and Harter, R. M. (1988), An error-components model for prediction of county crop areas using survey and satellite data, *J. Amer. Statist. Assoc.* 80, 28–36.

[2] Das, K., Jiang, J. and Rao, J. N. K. (2004), Mean squared error of empirical predictor, *Ann. Statist.* 32, 818–840.

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

[4] Datta, G. S., Lahiri, P., and Maiti, T. (2002), Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data, *J. Statist. Planning Inference* 102, 83-97.

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

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

[7] Hall, P. and Maiti, T. (2006b), On parametric bootstrap methods for small area prediction, *J. Roy. Statist. Soc. Ser. B* 68, 221–238.

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

[9] Jiang, J. and Lahiri, P. (2001), Empirical best prediction for small area inference with binary data, *Ann. Inst. Statist. Math.* 53, 217–243.

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

[11] Jiang, J., Lahiri, P. and Nguyen, T. (2017), A unified Monte-Carlo jackknife for small area estimation after model selection, *Ann. Math. Sci. Apple.*, in press.

[12] Jiang, J., Lahiri, P. and Wan, S. (2002), A unified jackknife theory for empirical best prediction with M-estimation, *Ann. Statist.* 30, 1782–1810.

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

[14] Prasad, N. G. N. & Rao, J. N. K. (1990), The estimation of mean squared errors of small area estimators, *J. Amer. Statist. Assoc.* 85, 163–171.

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

[16] Tibshirani, R. J. (1996), Regression shrinkage and selection via the Lasso, *J. Roy. Statist. Soc. Ser. B* 58, 267–288.