Annals of Mathematical Sciences and Applications
Volume 3, Number 2, 405–438, 2018

A unified Monte-Carlo jackknife for small area estimation after model selection

Jiming Jiang*, P. Lahiri*, and Thuan Nguyen*

We consider estimation of measure of uncertainty in small area estimation (SAE) when a procedure of model selection is involved prior to the estimation. A unified Monte-Carlo jackknife method, called McJack, is proposed for estimating the logarithm of the mean squared prediction error. We prove the second-order unbiasedness of McJack, and demonstrate the performance of McJack in assessing uncertainty in SAE after model selection through empirical investigations that include simulation studies and real-data analyses.

Keywords and phrases: Computer intensive, jackknife, log-MSPE, measure of uncertainty, model selection, Monte-Carlo, second-order unbiasedness, small area estimation.

1. Introduction

Small area estimation (SAE) has become a very active area of statistical research and applications. Here the term small area typically refers to a population for which reliable statistics of interest cannot be produced based on direct sampling from the population due to certain limitations of the available data. Examples of small areas include a geographical region (e.g., a state, county, municipality, etc.), a demographic group (e.g., a specific age × sex × race group), a demographic group within a geographic region, etc. See, for example, Rao and Molina (2015) for an updated, comprehensive account of various methods used in SAE. Statistical models, especially mixed effects models, have played key roles in improving small area estimates by borrowing strength from relevant sources. Therefore, it is not surprising

*The research of Jiming Jiang, Partha Lahiri, and Thuan Nguyen are partially supported by the NSF grant SES-1121794, SES-1534413, and SES-1118469, respectively. The research of Jiming Jiang and Thuan Nguyen are partially supported by the NIH grant R01-GM085205A1. The authors are grateful for a referee's comments that are very helpful in improving the presentation.

that model selection in SAE has received considerable attention in recent literature. See, for example, Jiang, Nguyen and Rao (2010), Datta, Hall and Mandal (2011), Pfeffermann (2013), Lahiri and Suntornchost (2014), and Rao and Molina (2015).

The errors from model selection are likely to affect the uncertainty measures in SAE estimates. To elaborate this point, let us consider a specific aspect of model selection—inclusion of small area specific random effects. Should one include area specific random effect in small area modeling? Such a component is a compromise between area specific fixed effects and no area effect and helps improving the properties of model-based estimators. For example, without such an area specific random effect, the model-based estimator may not be design-consistent, which may result in model-based estimate for an area with large sample size to deviate significantly from the corresponding design-based estimate, especially if area specific auxiliary variables fail to capture variation across the areas. A decision to exclude small area specific random effect may be based on a significance test. But such a decision is anything but perfect and depends very much on the subjective choice of the prespecified level of significance. A reasonable uncertainty measure estimator must incorporate the impact of model selection. However, most of the uncertainty measure estimators, with the exception of Molina, Rao and Datta (2015), do not attempt to capture the variation due to the model choice and there is no analytical study to examine the important second-order unbiasedness property of any of these estimators, including that of Molina et al. (2015).

In this paper, we propose a new uncertainty measure of any small area model-based estimator that incorporates errors due to model selection and a Monte-Carlo jackknife second-order unbiased estimator of the proposed uncertainty measure. We propose to use the logarithm of the mean squared prediction error (MSPE) as the uncertainty measure, where MSPE incorporates errors due to model selection. Our rationale behind using the log-MSPE comes from the means by which lack-of-fit measure of a typical model selection criterion is constructed. To elaborate on this point, consider the case of regression model selection with normal data. The well-known information criteria take the form of

$$n \log(\hat{\sigma}^2) + \lambda_n |M|, \tag{1}$$

where n is the sample size, $\hat{\sigma}^2$ is the standard estimator of the error variance, σ^2, $|M|$ is the dimension of the model, M, typically defined as the number of free parameters under M, and λ_n is a penalty function. Thus, in this

case, the measure of lack-of-fit is proportional (under a fixed sample size) to the logarithm of a variance estimator. Note that, typically, the variance is of the same scale as the MSPE. Therefore, it is reasonable to consider the logarithm of the MSPE as a measure of uncertainty in SAE when a model selection procedure, such as an information criterion, is involved.

Besides the intuitive link to model selection, there are other advantages of using the log-MSPE as a measure of uncertainty. In the SAE literature, MSPE estimates have been routinely used in assessing an improvement of the empirical best linear unbiased predictor (EBLUP) over the direct estimator. For such a purpose, one can equivalently use the log-MSPE, and report the improvement in the log-scale. An advantage of log-MSPE over MSPE occurs when it is desirable to model uncertainty measure estimators. This is because one can reasonably assume normality of the error term when log-MSPE estimators are considered. Zimmerman *et al.* (1999) emphasized the need to model log-MSPE in the context of a geo-spatial application. Gershunskaya and Dorfman (2013) considered modeling of logarithm of variances in an application related to Current Employment Statistics survey. In a small area context, such a model can provide a guideline for making important decisions on the choice of different design factors (e.g., sample size, number of clusters) for a future survey in achieving an approximate certain desired level of log-MSPE of the proposed predictor for different small areas. Also, the model can be used for quickly producing uncertainty measures when it is time consuming to compute such measures when dealing with big data as well as computational comlexity to meet a tight production deadline. Furthermore, when dealing with large observations (e.g., income), it takes space to report estimates and the associated MSPE estimates. For example, if estimates of average income for small areas in a country with high inflation are in billions, the order of MSPE will be in squared billions. It will be cumbersome to produce a large number of tables with such huge estimates and MSPE estimates. Creating such tables with log-MSPE may be more sensible in such situations.

In terms of statistical inference, it is easier to carry out hypothesis testing when considering log-MSPE. For example, suppose that one wishes to compare MSPE_1 with MSPE_2, which may correspond to two different methods of SAE. If one has second-order unbiased estimators of the log-MSPEs, say, $\hat{l}_j$ for $l_j = \log(\mathrm{MSPE}_j)$, $j = 1, 2$, it is possible to construct a z-test, or t-test, by assuming (approximately) that $\hat{l}_j = l_j + e_j, j = 1, 2$, where e_j is normal with mean zero and constant variance.

Another issue related to the second-order unbiased MSPE estimation is that, in practice, square roots of MSPE estimates are usually reported

in official publications. However, second-order unbiased MSPE estimators do not automatically generate second-order unbiased estimators of square roots of MSPE. On the other hand, in a log-scale this issue does not arise, as second-order estimator of log-$\sqrt{\text{MSPE}}$ can be obtained as half of the second-order estimator of log-MSPE.

Finally, a desirable property for an MSPE estimator is that it needs to be positive. If the property is combined with the second-order unbiasedness property, it turns out that it is very difficult to produce an estimator that has both of these properties. Typically, it is relatively easy to obtain a positive MSPE estimator that is first-order unbiased. To achieve the second-order unbiasedness, either analytical (e.g., Prasad and Rao 1990) or resampling (e.g., Jiang, Lahiri and Wan 2002, Hall and Maiti 2006) methods are used. However, with very few exceptions (Prasad and Rao 1990, Chen and Lahiri 2011), these techniques do not produce MSPE estimators that are guaranteed positive, in spite of achieving the second-order unbiasedness. To ensure that the MSPE estimator is positive, some modification of the (second-order unbiased) MSPE estimator is often made. For example, Hall and Maiti (2006) suggested the following strategy. Let $\widehat{\text{MSPE}}_1$ and $\widehat{\text{MSPE}}_2$ be two estimators of the same MSPE, for example, the former being an MSPE estimator with an additive bias-correction, and the latter one with a multiplicative bias-correction. Both MSPE estimators have some types of problems. For example, $\widehat{\text{MSPE}}_1$ can take negative values, and $\widehat{\text{MSPE}}_2$ can be unreliable (Hall and Maiti 2006). The idea is to combine the two estimators by letting $\widehat{\text{MSPE}} = \widehat{\text{MSPE}}_1$ if something happens, and $\widehat{\text{MSPE}} = \widehat{\text{MSPE}}_2$ otherwise. This strategy takes care of the positivity issue, but it does not necessarily preserve the second-order unbiasedness, even if $\widehat{\text{MSPE}}_1$ and $\widehat{\text{MSPE}}_2$ are both second-order unbiased. In fact, no rigorous proof has even been given that such a combined MSPE estimator is both positive and second-order unbiased. In contrast, there is no requirement that log-MSPE needs to be positive. Therefore, for log-MSPE, one can simply focus on the second-order unbiasedness of its estimator. Question is: How to obtain such an estimator?

In the context of MSPE estimation, a standard approach is Prasad-Rao (P-R) linearization (Prasad and Rao 1990). However, the approach is not feasible to handle our current problem, which is much more complicated. More specifically, we are interested in estimating the log-MSPE when the small area predictor is obtained after a model-selection procedure. The existing literature on inference after model selection has mainly focused on the case of independent observations (e.g., Rao and Wu 2001, sec. 12 and

the references therein, Leeb 2009, Berk, Brown and Zhao 2010). In particular, the potential impact of model selection on MSPE has never been rigorously addressed in the SAE literature. Intuitively, there is an additional uncertainty involved in the model-selection process, that needs to be taken into account in the MSPE estimation. The P-R linearization method requires differentiability of the underlying operation. This usually holds for standard estimation and prediction procedures, but not for model selection. For example, the information criteria, such as AIC (Akaike 1973) and BIC (Schwarz 1978), or the fence methods (see Jiang 2014 for a review), select models from a discrete space of candidate models. Even the shrinkage methods (e.g., Tibshirani 1996, Fan and Li 2001) involve continuous but non-differentiable penalty functions, such as the L^1 norm. See Müller, Scealy and Welsh (2013) for a review. Even if it is possible to develop a P-R type method, the derivation is tedious, and the final analytic expression is likely to be complicated. More importantly, errors often occur in the process of derivations as well as computer programming based on the lengthy expressions.

In this paper, we develop a unified jackknife approach that is assisted by Monte-Carlo simulations for the estimation of log-MSPE. As will be seen, the approach is applicable not just to the current problem of SAE after model selection, but to a much broader class of problems to obtain nearly unbiased estimators of quantities that can be obtained via Monte-Carlo simulation, if one knows the parameters that are involved. The method is especially attractive if the quantity of interest does not carry a constraint, such as non-negativity. This will be the case for the log-MSPE. Furthermore, the Monte-Carlo jackknife method, called *McJack*, is "one-formula-for-all", which means that one needs not re-derive the formula, as in P-R type methods, every time there is a new problem.

The rest of the paper is organized as follows. The McJack is introduced in Section 2 by first considering a special case for ease of illustration. A simple example is used to demonstrate numerical performance of McJack before a general theory is established. In Section 3 we offer a critical review of Jiang, Lahiri and Wan (2002; hereafter, JLW). We point out some undesirable features of JLW, and make two important observations that motivate McJack. We also note some major differences between McJack and a jackknife-after-bootstrap method. For those who are more interested in the motivation before learning about the method, the orders of Section 2 and Section 3 may be reversed. In Section 4, we carry out further simulation studies on the performance of McJack, and compare it with alternative approaches. A real data application is considered in Section 5. Proofs of the theorems are given in Section 6.

2. Monte-Carlo jackknife

2.1. A special case

We first illustrate the method using an example of EBLUP under a Fay-Herriot model (Fay and Herriot 1979), where the BIC (Schwarz 1978) is used to select the fixed covariates and the area-specific random effects. The model can be expressed in a way more convenient for the model selection problem:

$$y_i = x_i'\beta + \sqrt{A}\xi_i + e_i, \tag{2}$$

$i = 1, \ldots, m$, where the components of x_i are to be selected from a set of candidate covariates; $\xi_i \sim N(0,1)$; if $A > 0$, the random effects are included in the model; if $A = 0$, the random effects are excluded from the model; $e_i \sim N(0, D_i)$, where $D_i, 1 \le i \le m$ are known; and the ξ_i's and e_i's are independent. Note that there have been further considerations regarding the choice of the random effects; see, for example, Datta *et al.* (2011), but here we focus on a simpler situation. Let $M_{\rm f}$ denote a full model, under which x_i is the vector that includes all of the candidate covariates, and $A \ge 0$. Denote the x_i under $M_{\rm f}$ by $x_{{\rm f},i}$, and the corresponding β by $\beta_{\rm f}$. Let $\psi = (\beta_{\rm f}', A)'$. It is easy to see that $M_{\rm f}$ is, at least, a correct model, which means that (2) holds with x_i replaced by $x_{{\rm f},i}$, β replaced by $\beta_{\rm f}$, and the range of A being $[0, \infty)$. Of course, the reason for the model selection is that some of the components of $\beta_{\rm f}$ may be zero, in case that the full model can be simplified, and the true A may be zero. But this does not change the fact $M_{\rm f}$ is a correct model. In particular, the true small-area mean, θ_i, can be expressed as

$$\theta_i = x_{{\rm f},i}'\beta_{\rm f} + \sqrt{A}\xi_i. \tag{3}$$

On the other hand, under a candidate model, M, which corresponds to (2), the EBLUP of θ_i can be expressed as

$$\tilde{\theta}_i = \frac{\hat{A}}{\hat{A} + D_i} y_i + \frac{D_i}{\hat{A} + D_i} x_i'\hat{\beta}, \tag{4}$$

where $\hat{\beta} = \{\sum_{i=1}^m (\hat{A} + D_i)^{-1} x_i x_i'\}^{-1} \sum_{i=1}^m (\hat{A} + D_i)^{-1} x_i y_i$, and $\hat{A}$ is a consistent estimator of A obtained using a certain method (e.g., P-R, ML, REML; see Rao and Molina 2015). The BIC procedure chooses the model, M, by minimizing

$$\mathrm{BIC}(M) = -2\hat{l} + |M| \log(m), \tag{5}$$

where $\hat{l}$ is the maximized log-likelihood under M; $|M| = \dim(\beta) + 1$ if M includes the random effects, and $|M| = \dim(\beta)$ if M excludes the random effects. Here, for simplicity, we assume that $X = (x_i')_{1 \le i \le m}$ is full rank under any M. Let the minimizer of (5) be $\hat{M}$. We then compute the EBLUP (4) under $M = \hat{M}$, that is,

$$\hat{\theta}_i = \frac{\hat{A}_{\hat{M}}}{\hat{A}_{\hat{M}} + D_i} y_i + \frac{D_i}{\hat{A}_{\hat{M}} + D_i} x'_{\hat{M},i} \hat{\beta}_{\hat{M}}, \tag{6}$$

where $\hat{\beta}_{\hat{M}}$, $\hat{A}_{\hat{M}}$ are the $\hat{\beta}$, $\hat{A}$ obtained under $\hat{M}$, respectively. The MSPE of interest is

$$\mathrm{MSPE}(\hat{\theta}_i) = \mathrm{E}(\hat{\theta}_i - \theta_i)^2, \tag{7}$$

where θ_i is given by (3). It is clear that the joint distribution of $(\theta_i, y_i), 1 \le i \le m$ depends only on $\psi = (\beta_{\mathrm{f}}', A)$. Thus, (7) is a function of ψ and so is its logarithm. Let

$$b(\psi) = \log\{\mathrm{MSPE}(\hat{\theta}_i)\}. \tag{8}$$

Note that, in the context of SAE, $b(\psi)$ typically depends on i, but for notational simplicity the subscript i is dropped when considering a fixed i. Given ψ, for the kth Monte-Carlo simulation, one first generates θ_i by (3) with ξ_i replaced by $\xi_i^{(k)}$, $1 \le i \le m$, generated independently from $N(0,1)$. Denote the generated θ_i by $\theta_i^{(k)}$. Next, let $y_i^{(k)} = \theta_i^{(k)} + e_i^{(k)}$, $1 \le i \le m$, where $e_i^{(k)} \sim N(0, D_i), 1 \le i \le m$, generated independently and independent with $\xi_i^{(k)}$'s. The Monte-Carlo approximation to $b(\psi)$ is

$$\tilde{b}(\psi) = \log\left[\frac{1}{K}\sum_{k=1}^{K}\left\{\hat{\theta}_i^{(k)} - \theta_i^{(k)}\right\}^2\right], \tag{9}$$

where $\hat{\theta}_i^{(k)}$ is obtained the same way as the $\hat{\theta}_i$ of (6) except with y_i replaced by $y_i^{(k)}$, $1 \le i \le m$. Write the above procedure as a function, say, $\tilde{b}(\psi) = \mathbf{mcjack}(\psi)$, that computes (9) for every given ψ. Now suppose that $\hat{\psi}$ is an M-estimator of ψ. For example, $\hat{A}$ is the P-R estimator (Prasad and Rao 1990; truncated at zero if the expression turns out to be negative), and $\hat{\beta}_{\mathrm{f}}$ is given below (4) with $x_i = x_{\mathrm{f},i}, 1 \le i \le m$. Let $\hat{\psi}_{-j}$ be the delete-j version of $\hat{\psi}$. The McJack estimator of (8) is then given by

$$\widehat{b(\psi)} = \tilde{b}(\hat{\psi}) - \frac{m-1}{m}\sum_{j=1}^{m}\{\tilde{b}(\hat{\psi}_{-j}) - \tilde{b}(\hat{\psi})\}. \tag{10}$$

The motivation of (10), including its connection to JLW, will be discussed in the next section. Before we present the McJack under a general framework and develop a related theory, we would like to demonstrate its numerical performance using a simulated example.

2.2. Numerical demonstration

Let us consider a very simple situation, which may be viewed as a special case of the Fay-Herriot model,

$$y_i = x_i'\beta + v_i + e_i, \quad i = 1, \ldots, m, \tag{11}$$

where the components of x_i consist of an intercept, a group indicator, $x_{1,i}$, which is 0 if $1 \le i \le m_1 = m/2$, and 1 if $m_1 + 1 \le i \le m$, and potentially a third component, $x_{2,i}$, which is generated from the $N(0, 1)$ distribution, and fixed throughout the simulation. There are two candidate models: Model 1, which includes $x_{2,i}$, and Model 2: which does not include $x_{2,i}$. The model selection is carried out by BIC (Schwarz 1978).

For this demonstration, we consider a special case that the variance of the random effects, v_i, is known to be zero, that is, $A = 0$. There have been considerations of such situations in SAE (e.g., Datta *et al.* 2011). The variance of e_i, D_i, is equal to 1 for $1 \le i \le m_1$, and a for $m_1 + 1 \le i \le m$, where the value of a is either 4 or 16. Because $A = 0$, the small area mean, θ_i, under a given model, is equal to $x_i'\beta$. The corresponding EBLUP is $\hat{\theta}_i = x_i'\hat{\beta}$, where $\hat{\beta} = (X'D^{-1}X)^{-1}X'D^{-1}y$, with $X = (x_i')_{1\le i\le m}$ and $D = \mathrm{diag}(D_i, 1 \le i \le m)$, is the best linear unbiased estimator (BLUE) of β (e.g., Jiang 2007, sec. 2.3), under the given model. Due to the unbiasedness of the BLUE, the MSPE of the EBLUP is equal to its variance, that is,

$$\mathrm{MSPE}(\hat{\theta}_i) = \mathrm{var}(\hat{\theta}_i) = x_i'(X'D^{-1}X)^{-1}x_i, \quad 1 \le i \le m, \tag{12}$$

which are known under the given model. Now suppose that the EBLUP is obtained based on the model selected by the BIC. A naive estimator of the MSPE of $\hat{\theta}_i$, which ignores model selection, would be (12) computed under the selected model. The naive estimator of the log-MSPE is the logarithm of the naive MSPE estimator. We compare this estimator with two competitors. The first is what we call bootstrap MSPE estimator, which corresponds to the first term in (10), that is, without the jackknife bias correction, where $b(\cdot)$

is the log-MSPE function. The second is the McJack estimator given by (10). The bootstrap and McJack estimators are computed based on $K = 1000$ Monte-Carlo samples.

A series of simulation studies were carried out with $m = 20$ and $\beta_0 = \beta_1 = 1$, where β_0 is the intercept and β_1 the slope of $x_{1,i}$, and under two different true underlying models. In the first scenario, Model 1 is the true underlying model with the slope of $x_{2,i}$, $\beta_2 = 0.5$. In the second scenario, Model 2 is the true underlying model (i.e., $\beta_2 = 0$). We present the simulated percentage relative bias (%RB), based on $N_{\text{sim}} = 1000$ simulation runs, in Figures 2 and 3, where, for a given area, the %RB is defined as

$$\%\text{RB} = \left[\frac{\text{E}\{\log(\widehat{\text{MSPE}})\} - \log(\text{MSPE})}{|\log(\text{MSPE})|}\right] \times 100\%, \tag{13}$$

MSPE is the true MSPE based on the simulations, and $\text{E}\{\log(\widehat{\text{MSPE}})\}$ is the mean of the estimated log-MSPE based on the simulations. It is seen that the naive estimator significantly underestimates the log-MSPE; in fact, when Model 1 is the true model, the %RB for one of the areas is 516% in the case of $a = 4$, and there is a similar case in the case of $a = 16$. More specifically, there are some interesting trends observed. Namely, when the true model is Model 1, all of the methods seem to underestimate the log-MSPE, but the bootstrap and McJack estimators are doing much better, with McJack offering significant improvement over the bootstrap. On the other hand, when the true model is Model 2, the naive estimator again underestimate the log-MSPE, but the bootstrap and McJack estimators seem to overestimate the log-MSPE, with McJack significantly improving the bootstrap. The amount of underestimation by the naive estimator is less dramatic when Model 2 is the true model compared to when Model 1 is the true model. One explanation is that the BIC is known to have the tendency to overpenalize larger models. This would have bigger impact when Model 1 is the true model, which is the full model. In other words, there is a higher chance of model misspecification by the BIC, which impacts the log-MSPE estimation. To have a closer look at the numbers, we present one set of the detailed results in Table 1.

2.3. A unified approach, and theory

Although the above illustration is based on the Fay-Herriot model, its general principle, namely, (7)–(10), applies to much broader cases. Using the result of JLW, we can justify the second-order unbiasedness of McJack un-

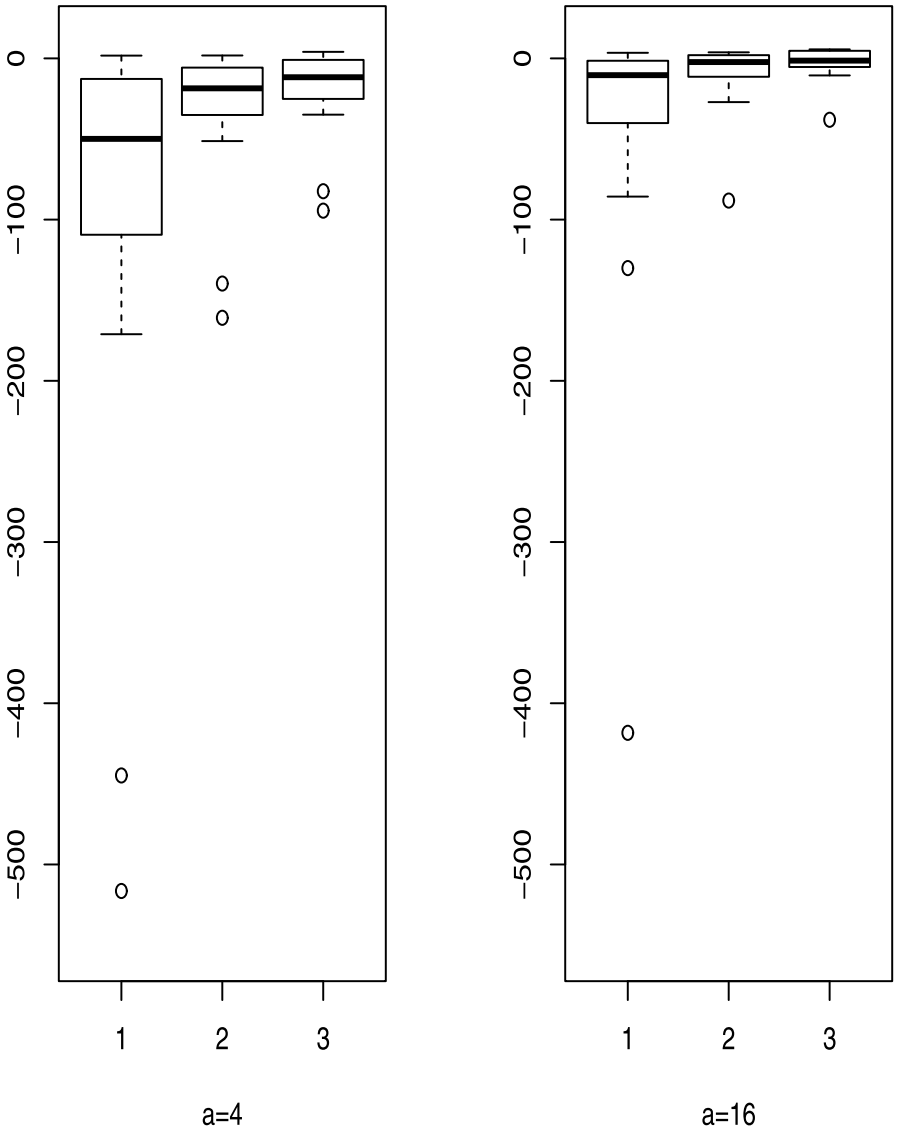

Figure 1: Boxplots of %RB when Model 1 is the true model. In each plot, from left to right: 1–Naive estimator, 2–bootstrap estimator, and 3–McJack estimator, of log-MSPE.

der the general framework. The justification also takes into account the effect of the Monte-Carlo errors. First note that, to establish a rigorous result about the unbiasedness, we need to make sure that the expectations

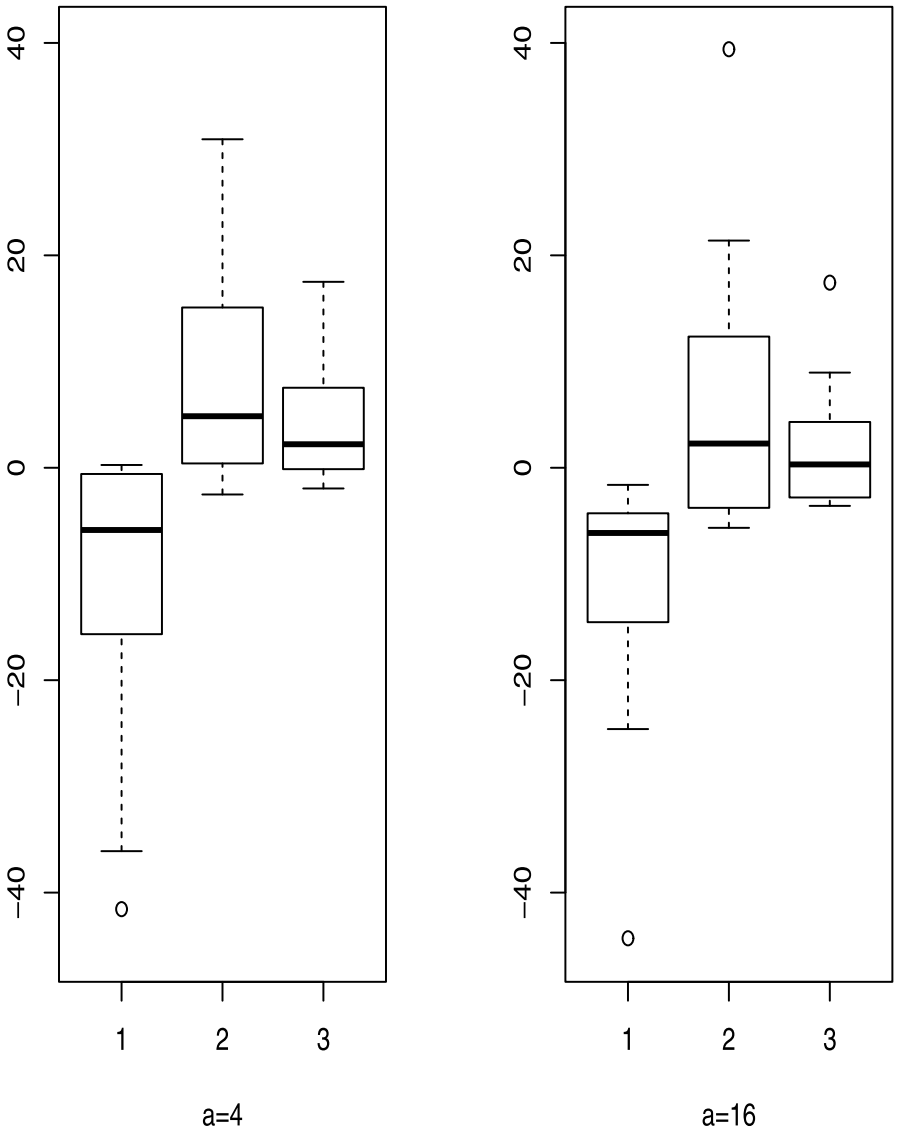

Figure 2: Boxplots of %RB when Model 2 is the true model. In each plot, from left to right: 1–Naive estimator, 2–bootstrap estimator, and 3–McJack estimator, of log-MSPE.

of $\tilde{b}(\hat{\psi}_{-j}), 0 \leq j \leq m$ exist. To avoid complicated technical conditions, we regularize these estimators (e.g., Jiang *et al.* 2002, Das *et al.* 2004). Let $\tilde{s}(\psi) = \exp\{\tilde{b}(\psi)\}$, and define

Table 1: **Log-MSPE estimation**: Model 2 is True Model; $a = 4$; %RB in ()s

Area	True log-MSPE	E(Naive Est.)	E(Bootstrap Est.)	E(McJack Est.)
1	-1.98	-2.26 (-14.0)	-1.79 (9.6)	-1.91 (3.3)
2	-1.62	-2.21 (-36.1)	-1.22 (25.0)	-1.41 (12.8)
3	-2.07	-2.27 (-9.8)	-1.95 (5.6)	-2.01 (2.9)
4	-2.20	-2.30 (-4.3)	-2.26 (-2.5)	-2.25 (-1.9)
5	-1.70	-2.22 (-30.4)	-1.33 (21.7)	-1.52 (10.7)
6	-2.05	-2.27 (-10.8)	-1.91 (6.7)	-1.97 (3.7)
7	-2.14	-2.29 (-6.9)	-2.11 (1.5)	-2.16 (-1.0)
8	-1.55	-2.20 (-41.6)	-1.11 (28.4)	-1.28 (17.5)
9	-2.19	-2.30 (-4.8)	-2.23 (-1.6)	-2.22 (-1.4)
10	-2.06	-2.27 (-10.2)	-1.94 (6.0)	-2.00 (3.2)
11	-0.91	-0.92 (-0.5)	-0.91 (0.6)	-0.91 (-0.0)
12	-0.91	-0.92 (-0.1)	-0.91 (0.2)	-0.92 (-0.1)
13	-0.76	-0.89 (-17.4)	-0.61 (19.8)	-0.69 (9.5)
14	-0.87	-0.90 (-3.7)	-0.78 (10.4)	-0.82 (5.6)
15	-0.92	-0.92 (0.3)	-0.92 (0.1)	-0.92 (-0.1)
16	-0.92	-0.92 (0.1)	-0.92 (0.1)	-0.92 (-0.1)
17	-0.74	-0.88 (-18.1)	-0.52 (30.9)	-0.62 (17.2)
18	-0.92	-0.91 (0.1)	-0.90 (2.1)	-0.91 (1.1)
19	-0.91	-0.92 (-0.6)	-0.90 (0.7)	-0.91 (-0.0)
20	-0.88	-0.91 (-3.6)	-0.84 (4.1)	-0.87 (1.5)

$$\hat{s}(\psi) \;=\; \begin{cases} e^{-\lambda m^{\rho}}, & \text{if } \tilde{s}(\psi) < e^{-\lambda m^{\rho}}, \\ \tilde{s}(\psi), & \text{if } e^{-\lambda m^{\rho}} \le \tilde{s}(\psi) \le e^{\lambda m^{\rho}}, \\ e^{\lambda m^{\rho}}, & \text{if } \tilde{s}(\psi) > e^{\lambda m^{\rho}}, \end{cases}$$

and $\hat{b}(\psi) = \log\{\hat{s}(\psi)\}$, where λ, ρ are given positive numbers. Let $s(\psi)$ denote $\mathrm{MSPE}(\hat{\theta}_i)$ when ψ is the true parameter vector. We truncate $s(\cdot)$ the same way as $\tilde{s}(\cdot)$, and let $b(\psi) = \log\{s(\psi)\}$. For notation convenience, write $\hat{\psi}_{-0} = \hat{\psi}$. Also, let $F_{-0}(\psi), F_{-j}(\psi)$ denote the left sides of (16) and (17), respectively. The M-estimators, $\hat{\psi}_{-j}, 0 \le j \le m$ are said to be consistent uniformly (c.u.) at rate m^{-d} if, for any $\delta > 0$, there is a constant c_δ such that

$$\mathrm{P}(A_{j,\delta}^c) \;\le\; c_\delta m^{-d}, \quad 0 \le j \le m,$$

where $A_{j,\delta}$ is the event that $F_{-j}(\hat{\psi}_{-j}) = 0$ and $|\hat{\psi}_{-j} - \psi| \le \delta$, with ψ being the true parameter vector. Also, write $f_i = f_i(\psi, y_i)$, $g_i = \partial f_i/\partial\psi'$, $h_{i,k} = \partial^2 f_{i,k}/\partial\psi\partial\psi'$, where $f_{i,k}$ is the kth component of f_i. Furthermore, for

any function f of ψ, define

$$\|\Delta^3 f\|_w = \max_{1\le s,t,u\le r} \sup_{|\tilde{\psi}-\psi|\le w} \left| \frac{\partial^3 f(\tilde{\psi})}{\partial\psi_s \partial\psi_t \partial\psi_u} \right|,$$

where ψ is the true parameter vector, and $r = \dim(\psi)$. A similar definition is extended to $\|\Delta^4 f\|_w$. The spectral norm of a matrix, B, is defined as $\|B\| = \sqrt{\lambda_{\max}(B'B)}$, where $\lambda_{\max}$ denotes the largest eigenvalue. Also write $\Delta_j = a - a_{-j}$, where a, a_{-j} are the functions of ψ that appear in (16) and (17), respectively. We shall consider estimation of log-MSPE of $\hat{\theta}_i$, a predictor of θ_i after model selection, for a fixed i. Furthermore, we assume that the Monte-Carlo samples, under ψ, are generated by first generating some standard [e.g., $N(0,1)$] random variables and then plugging ψ. For example, under the full Fay-Herriot model of (2), y_i is generated by first generating the ξ_i's and η_i's, which are independent $N(0,1)$, and then letting $y_i = x'_{\mathrm{f},i}\beta_{\mathrm{f}} + \sqrt{A}\xi_i + \sqrt{D_i}\eta_i$, with $\psi = (\beta'_{\mathrm{f}}, A)'$. Let ξ denote the vector of the standard random variables. We first make the following general assumptions.

A1. There are $d > 2$ and $w > 0$ such that the $2d$th moments of $|f_i|$, $\|g_i\|$, $\|h_{i,k}\|$, $\|\Delta^3 f_{i,k}\|_w$, $1 \le i \le m, 1 \le k \le r$ are bounded for some $d > 2 + \rho$.

A2. For the same d and w in *A1*, a_{-j} and its up to third order partial derivatives, $0 \le j \le m$, as well as $\Delta_j, 1 \le j \le m$, all evaluated at $\tilde{\psi}$, are bounded uniformly for $|\tilde{\psi} - \psi| \le w$, where ψ is the true parameter vector, and $m^{\tau}(|\Delta_j| \vee \|\partial\Delta_j/\partial\psi\|), 1 \le j \le m$, evaluated at ψ, are bounded, where $\tau = (d-2)/(2d+1)$.

A3. The log-MSPE function $b(\cdot)$ of (8) is four-times continuously differentiable, and, for the same w in *A1*, $\|\Delta^4 b\|_w$ is bounded.

A4. $\limsup_{m\to\infty} \|\{\mathrm{E}(\bar{g})\}^{-1}\| < \infty$, where $\bar{g} = m^{-1}\sum_{j=1}^m g_j$, evaluated at the true ψ.

A5. $\hat{\psi}_{-j}, 0 \le j \le m$ are c.u. at rate m^{-d} for the same d in *A1*.

A6. $\sum_{j=1}^m \Delta_j = O(m^{-\nu})$ for some $\nu > 0$.

Recall the way that the Monte-Carlo samples are generated specified above *A1*. Under this assumption, $\theta_i^{(k)}, \hat{\theta}_i^{(k)}, 1 \le k \le K$, generated under $\tilde{\psi}$, are functions of $\tilde{\psi}$ and ξ. The additional assumptions below are regarding the Monte-Carlo sampling.

A7. ξ is independent with the data, y.

A8. Let ψ be the true parameter vector, and w be the same as in *A1*. There are constants $0 < c_1 < c_2$ such that $c_1 \le s(\tilde{\psi}) \le c_2$ for $|\tilde{\psi} - \psi| \le w$, and random variables $G_k, 1 \le k \le K$, which do not depend on $\tilde{\psi}$, such that $|\hat{\theta}_i^{(k)} - \theta_i^{(k)}| \le G_k$ and $\mathrm{E}(G_k^q)$ are bounded for some $q \ge 2\{2 + (\rho \vee 1)\}$.

A9. $m^2/K \to 0$, as $m \to \infty$.

Theorem 1. *Suppose that A1–A9 hold. Let* $\widehat{b(\psi)}$ *denote (10) with* $\tilde{b}$ *replaced by* $\hat{b}$*. Then, we have* $\mathrm{E}\{\widehat{b(\psi)} - b(\psi)\} = o(m^{-1})$*, where* ψ *is the true* ψ *[hence* $b(\psi)$ *is the true log-MSPE], and* E *is with respect to both* y *and* ξ*.*

The next result focuses on the special case of Fay-Herriot model.

Theorem 2. *Suppose that the true* $A > 0$*, and there are positive constants* $0 < c_1 < c_2$ *such that* $c_1 \le |x_{\mathrm{f},i}| \le c_2$*,* $c_1 \le D_i \le c_2, 1 \le i \le m$*. Furthermore, suppose that*

$$\limsup_{m\to\infty} \lambda_{\min}\left(\frac{1}{m}\sum_{i=1}^{m} x_{\mathrm{f},i}x'_{\mathrm{f},i}\right) > 0, \tag{14}$$

and A9 *holds. Then, the conclusion of Theorem 1 holds.*

The proofs of Theorem 1 and Theorem 2 are given in Section 6.

3. Review of JLW, motivation of McJack, and discussion

In the context of resampling methods for SAE, JLW proposed a jackknife method for estimating the MSPE of empirical best predictor (EBP) when the parameters of interest are estimated by M-estimators. Let ξ denote a mixed effect, for example, a small area mean. Let $\tilde{\xi}$ and $\hat{\xi}$ denote the best predictor (BP), defined as conditional expectation of ξ given the data, y, and EBP of ξ, respectively. Then, one has the decomposition

$$\mathrm{MSPE}(\hat{\xi}) = \mathrm{MSPE}(\tilde{\xi}) + \mathrm{E}\{(\hat{\xi} - \tilde{\xi})^2\}, \tag{15}$$

where $\mathrm{MSPE}(\hat{\xi})$ is defined as $\mathrm{E}\{(\hat{\xi} - \xi)^2\}$ and $\mathrm{MSPE}(\tilde{\xi})$ is defined similarly. The idea of JLW is to jackknife the two terms on the right side of (15) separately. For the first term, the authors assume that it is a function of ψ, a vector of parameters, that is, $\mathrm{MSPE}(\tilde{\xi}) = b(\psi)$, which can be computed analytically. The parameter vector ψ is then estimated by an M-estimator, defined as the solution, $\hat{\psi}$, to a system of equations of the following form:

$$\sum_{i=1}^{m} f_i(\psi, y_i) + a(\psi) = 0. \tag{16}$$

In (16), y_i is the data vector from the ith cluster (e.g., small area), and the clusters are assumed to be independent; $f_i(\cdot,\cdot)$ is a vector-valued function that satisfies $\mathrm{E}\{f_i(\psi, y_i)\} = 0, 1 \le i \le m$, if ψ is the true parameter vector;

and $a(\cdot)$ corresponds to a penalizer, which in some cases is the zero vector. The delete-j estimator, $\hat{\psi}_{-j}$, of ψ is defined as the solution to the following system of equations:

$$\sum_{i \neq j} f_i(\psi, y_i) + a_{-j}(\psi) \quad = \quad 0, \tag{17}$$

where $a_{-j}(\cdot)$ has a similar interpretation. Given the M-estimators, $b(\psi)$ is estimated by a plug-in estimator, minus a jackknife bias correction, that is,

$$b(\hat{\psi}) - \frac{m-1}{m} \sum_{j=1}^{m} \{b(\hat{\psi}_{-j}) - b(\hat{\psi})\}. \tag{18}$$

As for the second term on the right side of (15), it is estimated by a jackknife variance-type estimator that has the following expression:

$$\frac{m-1}{m} \sum_{j=1}^{m} (\hat{\xi}_{-j} - \hat{\xi})^2, \tag{19}$$

where $\hat{\xi}_{-j}$ is a delete-j version of $\hat{\xi}$, the EBP, defined in a certain way, which is not important for the current paper. JLW showed that, when the two terms, (18) and (19), are put together, the combined jackknife estimator of the MSPE of EBP is second-order unbiased. The work has had a significant impact in SAE, especially in the literature of resampling methods in SAE (e.g., Hall and Maiti 2006, Lohr and Rao 2009, Pfeffermann 2013, Rao and Molina 2015). On the other hand, we note the following undesirable features of JLW:

(a) JLW requires analytical computation of $b(\psi)$.

(b) JLW does not incorporate errors from model selection. In particular, the proof for the second-order unbiased property of (19) fails if a model selection procedure is involved prior to obtaining the EBP, such as in Datta *et al.* (2011).

(c) JLW does not guarantee a strictly positive MSPE estimator, in spite of its second-order unbiasedness. See our discussion in Section 1 (7th paragraph).

As far as this paper is concerned, what is most important is not the full JLW theory, but rather an intermediate result. In obtaining their theory, JLW showed, in particular, that (18) is a second-order unbiased estimator of $b(\psi)$, if the penalizers $a, a_{-j}, 1 \leq j \leq m$ in (16) and (17) satisfy certain mild conditions. In particular, those conditions are satisfied if the penalizers are zero (vectors), in which case the M-estimating equations are unbiased.

Having given the proof of the result, we realize the following two facts, both are critically important to the idea of the current paper.

(I) The fact that $b(\psi)$ is an MSPE is not used anywhere in the proof. In other words, as long as $b(\cdot)$ is a sufficiently smooth function, and ψ is estimated by the M-estimators, the second-order unbiased estimation of $b(\psi)$ by (18) holds. In particular, $b(\psi)$ can be log(MSPE), which is of primary interest here.

(II) More importantly, $b(\psi)$ does not have to have an analytic expression, as long as one knows how to compute it. An analytic expression would be nice, but, in the new era, the computation is typically implemented in some code and executed with a high-speed computer. In particular, suppose that, given ψ, $b(\psi)$ can be approximated by a Monte-Carlo method to an arbitrary degree of accuracy. Then, one can write programming codes, based on the Monte-Carlo, to compute $b(\cdot)$ as a function. Given this "computer-powered" function, all one needs to do is to plug the M-estimators, $\hat{\psi}, \hat{\psi}_{-j}, 1 \le j \le m$, into this function to obtain the second-order unbiased estimator of $b(\psi)$.

The importance of the above observations is that they apply to virtually any kind of situation, not just the EBP. In particular, the predictor, $\hat{\xi}$, can be much more complicated than the EBP, such as an EBP obtained following a model-selection procedure. Also, the decomposition (15), and jackknifing of the second term in the decomposition, (19), are altogether not needed to apply these observations. The McJack, proposed in the previous section, is based on these two important observations; it addresses all of the undesirable features of JLW noted above. Other complicated situations, to which our idea may apply, include (i) regression inference after variable selection (e.g., Leeb 2009); (ii) mixed model prediction with non-normal random effect distribution (e.g., Lahiri and Rao 1995); and (iii) shrinkage estimation/selection with data-driven choice of regularization parameter (e.g., Pang, Lin and Jiang 2016).

In the context of resampling methods, a well-known method is jackknife-after-bootstrap (JAB; Efron 1992). There are major differences between JAB and McJack. First, the objectives are different. The main purpose of JAB is to assess accuracy of the usual bootstrap estimates; while the objective of McJack is to estimate quantities of interest, such as measures of uncertainty for estimates based on the original data. Secondly, JAB works, for the most part, under the standard nonparametric bootstrap setting, to achieve efficient computation so that no additional bootstrap samples are needed; in other words, the JAB estimates are obtained from the original bootstrap samples. However, this is difficult to do under a parametric bootstrap setting. For example, although Efron (1992) has discussed JAB with parametric bootstrap using the idea of importance sampling, the approach does not

necessarily lead to a real gain in computation if the major computational burden is not due to sampling. On the other hand, standard nonparametric bootstrap procedures do not apply to SAE problems, in spite of some variations that have been developed. See, for example, Pfeffermann (2013), for a review. Finally, McJack does not have to be associated with bootstrap–any kind of Monte-Carlo method can be used to assist the computation. For example, JLW discussed an example in which the Monte-Carlo method used to compute the MSPE is not considered as bootstrapping.

4. More simulation studies

4.1. Testing the presence of random effects in a Fay-Herriot model

Datta *et al.* (2011) proposed a method of model selection by testing for the presence of the area-specific random effects, $v_i = \sqrt{A}\xi_i$, in the Fay-Herriot model (2). This is equivalent to testing the null hypothesis $\mathrm{H}_0 : A = 0$. The test statistic, $T = \sum_{i=1}^{m} D_i^{-1}(y_i - x_i'\hat{\beta})^2$, where $\hat{\beta}$ is the same as in Subsection 2.1, has a χ^2_{m-p} distribution, with $p = \mathrm{rank}(X)$, under H_0. If H_0 is rejected, the EBLUP is used to estimate the small area mean θ_i, where in this simulation A is estimated by the P-R estimator, and the corresponding MSPE estimator is the P-R MSPE estimator; if H_0 is accepted, the estimator $\hat{\theta}_i = x_i'\hat{\beta}$ is used to estimate θ_i, and the corresponding MSPE is given by (12). Thus, if the level of significance is chosen as 0.05, the proposed MSPE estimator, denoted by DHM, is the P-R MSPE estimator if $T > \chi^2_{m-p}(0.05)$, and (12) if $T <= \chi^2_{m-p}(0.05)$.

We run a simulation study to compare the performance of McJack with DHM. The simulation is under the full model considered in the previous subsection (hence $p = 3$), and three different true values of A: $A = 0$, $A = 0.5$, and $A = 1$. The boxplots of %RB for these three cases are presented in Figure 3, with the detailed numbers for DHM and McJack given in Table 2. It is seen that DHM works better for the case $A = 0$, which is not surprising because, under the null hypothesis, the DHM MSPE estimator is "right" 95% of the times. On the other hand, McJack works significantly better in those two cases of nonzero A. Simple simulations show that, in the latter cases, the probability of rejecting the null hypothesis is about 0.26 when $A = 0.5$, and 0.44 when $A = 1$. The worst scenario seems to be the case where A is not zero but closer to zero ($A = 0.5$). There are a few "blown-up" cases under this scenario where the %RB exceeds 1000% for DHM. It is also obvious that McJack improves bootstrap in every case.

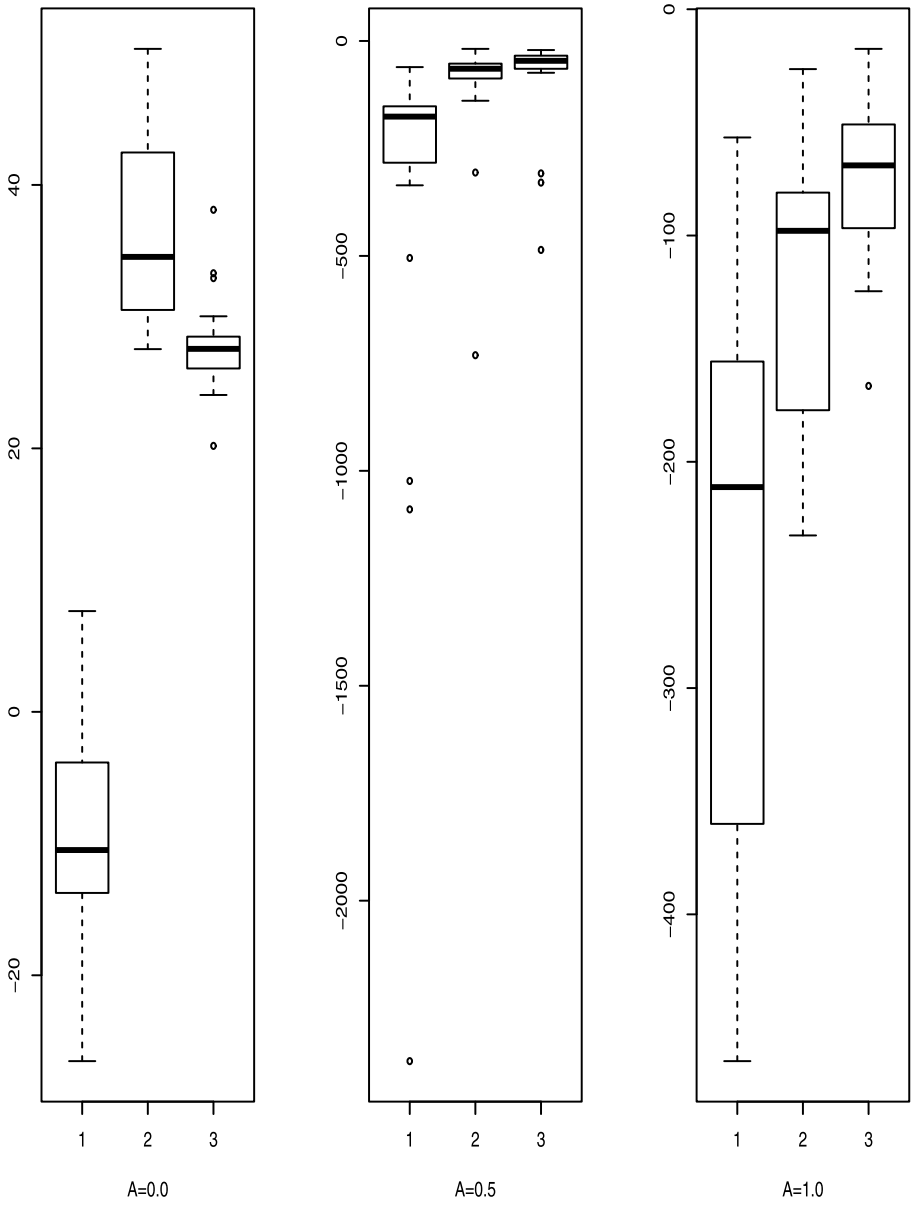

Figure 3: Boxplots of %RB. In each plot, from left to right: 1–DHM, 2–bootstrap, 3–McJack. Scales are different due to the huge difference in range.

Table 2: DHM vs McJack in %RB

Area	$A = 0.0$		$A = 0.5$		$A = 1.0$	
	DHM	McJack	DHM	McJack	DHM	McJack
1	-13.6	26.2	-216.8	-59.8	-342.0	-99.5
2	0.3	26.0	-131.0	-45.8	-343.8	-72.9
3	-3.9	27.5	-107.4	-21.1	-135.8	-30.3
4	1.1	28.2	-158.0	-37.7	-362.9	-77.6
5	-8.1	24.9	-178.9	-36.9	-191.3	-59.6
6	-6.0	30.0	-180.3	-50.5	-375.4	-166.5
7	-3.1	24.1	-210.0	-51.5	-395.5	-124.6
8	7.6	27.6	-135.3	-33.1	-464.9	-123.0
9	-10.9	27.4	-149.4	-43.0	-357.2	-108.1
10	-3.8	28.6	-163.1	-31.4	-362.8	-94.0
11	-26.5	20.2	-60.8	-21.8	-220.3	-39.4
12	-10.7	33.3	-2373.3	-486.1	-210.4	-76.5
13	-13.9	27.5	-504.8	-74.0	-94.5	-18.4
14	-10.3	32.9	-173.1	-35.3	-188.2	-64.1
15	-17.5	25.7	-1023.6	-329.5	-163.0	-58.1
16	-4.4	38.1	-154.6	-46.6	-211.9	-65.2
17	-12.1	28.4	-335.8	-48.9	-197.6	-72.8
18	-11.4	27.1	-171.2	-22.0	-148.4	-59.9
19	-14.2	27.7	-230.6	-69.6	-56.7	-17.5
20	-18.4	28.3	-1089.6	-308.1	-104.1	-43.7

4.2. SAE after GIC

Now let us revisit the example of Subsection 2.2, but this time with $A \neq 0$. The model selection is done by the same BIC procedure, which is a special case of the generalized information criteria (GIC), whose consistency for linear mixed model selection has been proved by Jiang and Rao (2003). We consider the second scenario, in which Model 2 is the true underlying model. Two settings were considered in Subsection 2.2: $a = 4$ and $a = 16$. Here, we take a middle ground by considering $a = 8$. The true value of A is either 1 or 2. We compare McJack with the bootstrap and a method called PR, which estimates the MSPE by the P-R MSPE estimator (Prasad and Rao 1990; see Subsection 4.1) under the selected model, in estimating the area-specific log-MSPEs.

An issue that was previously not considered is whether it makes a difference to use the same random seed in computing the Monte-Carlo MSPEs based on $\hat{\psi}$ and all of the $\hat{\psi}_{-j}, 1 \leq j \leq m$. If the same random seed is used, the Monte-Carlo log-MSPE, $b(\psi)$, is a function of ψ only. In other words, the only difference between $b(\hat{\psi})$ and $b(\hat{\psi}_{-j}), 1 \leq j \leq m$ is due to the difference

between the M-estimators, $\hat{\psi}, \hat{\psi}_{-j}, 1 \le j \le m$. On the other hand, if different random seeds are used, for example, if one lets the computer choose the seed, the difference between $b(\hat{\psi})$ and $b(\hat{\psi}_{-j}), 1 \le j \le m$ is not only due to the difference between the M-estimators, but also due to the random seeds. In this particular simulation study, we also investigate this problem. Throughout this simulation, the Monte-Carlo sample size for computing the Monte-Carlo MSPE is chosen as $K = 1000$.

Figure 4 presents two boxplots of the %RBs of PR, bootstrap (BT), and McJack (MJ) for the case $A = 1$. To the left is the plots with the seed fixed for $b(\hat{\psi})$ and all of the $b(\hat{\psi}_{-j})$'s (but the seeds are different between different simulation runs). To the right is the plots with the seeds randomly selected by the computer [so the seeds for $b(\hat{\psi})$ and $b(\hat{\psi}_{-j})$'s are all different]. The results are based on $N_{\text{sim}} = 1000$ simulation runs. It is seen that the overall patterns are very similar. The difference in the scales are likely due to the fact that the covariate, $x_{2,i}$, were generated differently; but, once generated, they were fixed throughout the simulations (i.e., the same $x_{2,i}$'s were used for all of the simulation runs). Thus, we conclude that, at least for this simulation study, the Monte-Carlo sample size, K, is large enough that there is no essential difference between the fixed and random seeds. One set of the detailed results, for the fixed seed case, are presented in Table 3, where the Covariate column are the $x_{2,i}$'s for this simulation, and log(MSPE) are the true (simulated) log-MSPEs. It is seen that the PR method always over-estimate the true log-MSPE, sometimes substantially. One explanation is that the P-R MSPE estimator depends not only on the model for the covariates but also on the estimator of A, under the model. Because, typically, a selected model provides the best fit to the data, as a result, the estimated A, under the selected model, is smaller. A similar finding was reported in Datta *et al.* (2011), who noted that, if more effective covariates are added to the (Fay-Herriot) model, the estimated A is smaller; as a result, the variance A becomes insignificant in hypothesis testing. Thus, estimating A under a fixed model may over-estimate the MSPE. Also note that, even if the P-R MSPE estimator is computed under the selected model, it is derived under a fixed model, which happens to be the one selected. In other words, although this fixed model is the selected model this time around with the current data, it may not be the selected model when a new set of data is generated.

More specifically, for the first 10 small areas the over-estimation by the P-R method is of much higher order than for the last 10 small areas. Again, there are some explanations. First of all, the PR method computes the MSPE estimator based on the selected model, and the same model is used for all of

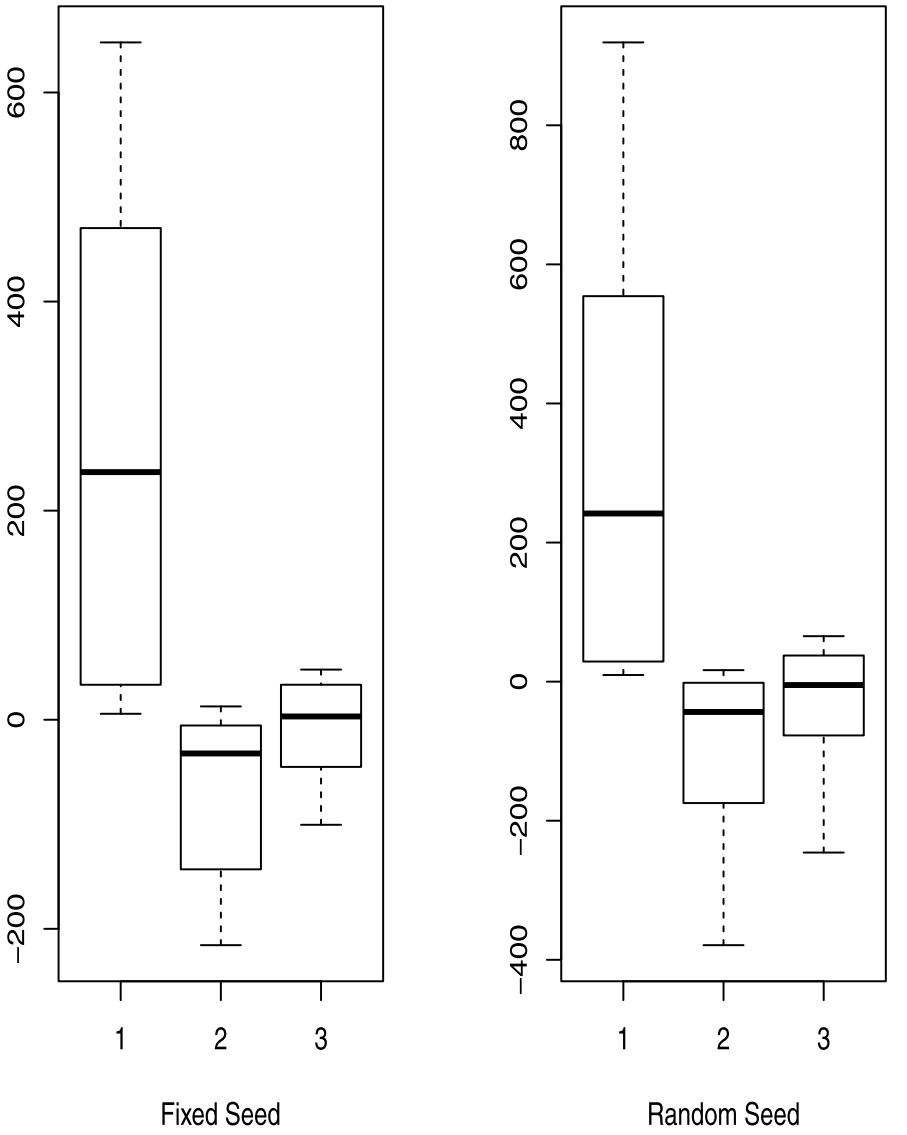

Figure 4: Boxplots of %RB. $A = 1$. In each plot, from left to right: 1–PR, 2–bootstrap, 3–McJack. Scales are different due to the values of the covariates that were generated.

the small areas. It is possible that the seleted model is more effective or, in some cases, less damaging, for some small areas than for the others. Recall that the sampling variance, D_i, is 1 for the first 10 small areas, and 8 for the

Table 3: Comparison of PR, BT, and MJ: $A = 1$, Fixed Seed

Area	Covariate	log(MSPE)	%RB PR	%RB BT	%RB MJ
1	-0.302	-0.281	446.6	-143.2	-44.3
2	-0.180	-0.272	459.1	-147.4	-47.5
3	0.072	-0.283	445.1	-113.8	-29.6
4	-0.763	-0.233	519.5	-176.0	-67.2
5	0.025	-0.297	428.8	-108.1	-24.9
6	0.449	-0.242	505.7	-92.4	-29.9
7	-0.643	-0.276	453.8	-142.9	-46.0
8	-1.657	-0.207	575.8	-47.9	-17.8
9	-0.109	-0.256	481.5	-155.4	-53.4
10	-1.025	-0.179	647.9	-215.6	-100.5
11	0.615	0.602	35.7	4.2	42.1
12	0.783	0.596	38.1	12.7	48.0
13	-0.617	0.567	41.8	-5.1	40.9
14	-0.488	0.554	44.9	-5.9	41.8
15	-0.531	0.612	31.2	-14.0	29.7
16	1.316	0.771	10.4	10.9	30.8
17	0.042	0.576	39.2	-9.8	36.2
18	-0.506	0.627	28.0	-16.6	25.6
19	-1.754	0.805	5.7	6.5	26.7
20	-1.008	0.684	19.2	-9.8	24.0

last 10 small areas. The leading term of the true MSPE is $AD_i/(A+D_i)$ for area i (e.g., Jiang 2010, p. 445), which is increasing with D_i; on the other hand, the part in the P-R MSPE estimator that is most model-dependent is the g_2 term, but this term is $O(m^{-1})$, compared to the order $O(1)$ of the leading term. Thus, in a way, the P-R MSPE estimator is less affected by the selected model for the last 10 small areas than for the first 10. Also, note that the true log-MSPE goes to the denominator when computing the %RB; when the denominator is smaller in absolute value, which is confirmed by the log(MSPE) column of Table 3, the corresponding %RB tends to be larger. As for the bootstrap MSPE estimator, it works very well for the last 10 small areas but seems to significantly under-estimate for the first 10 small areas. In comparison, McJack has a (much) more robust performance overall.

Figure 5 presents the comparison for the case $A = 2$. The pattern is similar, except that magnitude of the differences for the first 10 small areas is much amplified. On the other hand, the %RB performance for the last 10 small areas are better than the case $A = 1$ for all three methods. The detailed results are omitted.

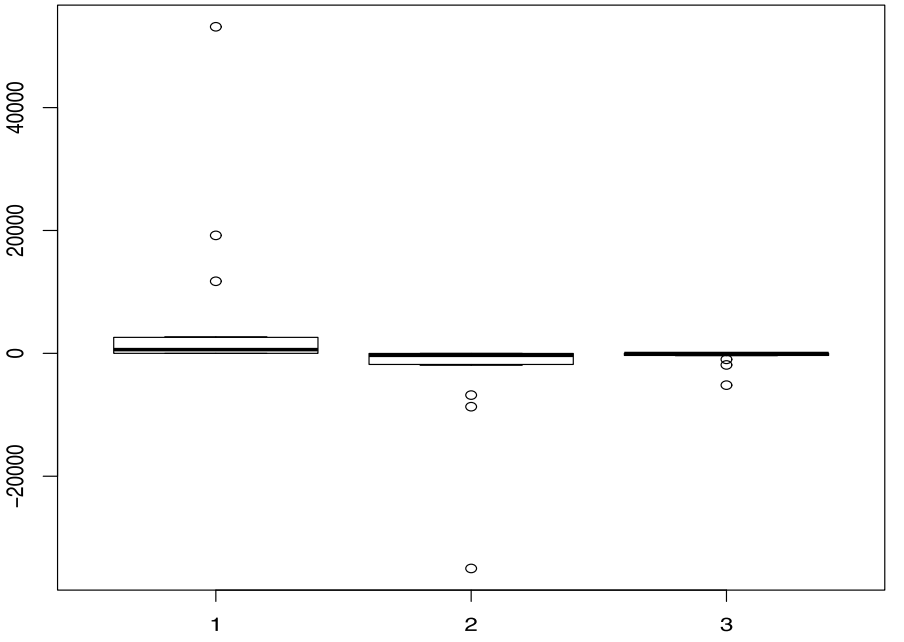

Figure 5: Boxplots of %RB. $A = 2$ (Fixed Seed). From left to right: 1–PR, 2–bootstrap, 3–McJack.

5. A real data example

Morris and Christiansen (1995) presented a data set involving 23 hospitals (out of a total of 219 hospitals) that had at least 50 kidney transplants during a 27 month period (see Table 4). The y_i's are graft failure rates for kidney transplant operations, that is, y_i = number of graft failures/n_i, where n_i is the number of kidney transplants at hospital i during the period of interest. The variance for the graft failure rate, D_i, is approximated by $(0.2)(0.8)/n_i$, where 0.2 is the observed failure rate for all of the hospitals. Thus, D_i is assumed known. In addition, a severity index, s_i, is available for each hospital, which is the average fraction of females, blacks, children and extremely ill kidney recipients at hospital i. Ganesh (2009) proposed a Fay-Herriot model for the graft failure rates, which is (2) with $x_i'\beta = \beta_0 + \beta_1 s_i$. Jiang *et al.* (2010) suggests that, in a way, the optimal model for this data is a cubic model, that is, (2) with $x_i'\beta = \beta_0 + \beta_1 s_i + \beta_2 s_i^2 + \beta_3 s_i^3$, which is also used in Datta *et al.* (2011).

We analyze the data under the latter model for the mean function but with selection of the random effect factor using the strategy of Datta *et*

Table 4: The Hospital Data, Estimates, and Measures of Uncertainty

Area	y_i	s_i	$\sqrt{D_i}$	$\hat{\theta}_i$	DHM	BT	MJ	$\tilde{\theta}_i$	MJ
1	.302	.112	.055	.221	.015	.029	.038	.238	.034
2	.140	.206	.053	.186	.013	.027	.019	.178	.019
3	.203	.104	.052	.214	.014	.029	.038	.215	.036
4	.333	.168	.052	.215	.011	.028	.044	.240	.040
5	.347	.337	.047	.349	.047	.047	.047	.349	.047
6	.216	.169	.046	.215	.011	.026	.030	.218	.024
7	.156	.211	.046	.183	.015	.027	.026	.176	.021
8	.143	.195	.046	.195	.011	.026	.032	.184	.034
9	.220	.221	.044	.177	.018	.029	.040	.186	.040
10	.205	.077	.044	.168	.015	.029	.048	.177	.049
11	.209	.195	.042	.195	.011	.026	.030	.199	.027
12	.266	.185	.041	.203	.010	.026	.029	.221	.026
13	.240	.202	.041	.189	.012	.026	.030	.203	.030
14	.262	.108	.036	.218	.014	.026	.021	.235	.018
15	.144	.204	.036	.188	.013	.025	.028	.174	.026
16	.116	.072	.035	.155	.017	.028	.038	.141	.042
17	.201	.142	.033	.228	.015	.025	.025	.221	.025
18	.212	.136	.032	.229	.015	.025	.025	.226	.025
19	.189	.172	.031	.213	.010	.023	.017	.205	.019
20	.212	.202	.029	.189	.012	.024	.038	.199	.034
21	.166	.087	.029	.189	.013	.024	.036	.180	.030
22	.173	.177	.027	.209	.010	.023	.032	.194	.034
23	.165	.072	.025	.155	.017	.022	.022	.159	.019

al. (2011), that is, by testing for the presence of the random effects, v_i. At $\alpha = 0.05$ level of significance, the test statistic (see Subsection 4.1) $T = 24.3$, while the critical value of χ^2_{19} is 30.1. Thus, the null hypothesis that $A = 0$ is not rejected. As a result, $\hat{\theta}_i = x_i'\hat{\beta}$ is used as the estimate of θ_i, according to Datta *et al.* (2011). However, the main issue is how to assess the uncertainty. We apply the three different methods investigated in Subsection 4.1 to this data, and obtain the square roots of the estimated MSPEs, denoted by DHM, BT, and MJ, respectively. Here the MSPE estimates are obtained by taking the exponentials of the corresponding log-MSPE estimates. The Monte-Carlo sample size for BT and MJ is $K = 4000$. The results are presented in Table 4. It is seen that the measures of uncertainty by DHM are always smaller than those by BT and MJ. This is not surprising because DHM does not take into account the potential variation in model selection. As for the comparison between BT and MJ, the latter measures are larger in most cases.

As another comparison, we also computed the standard EBLUPs (i.e., without testing the presence of the random effects) and their corresponding McJack estimates of $\sqrt{\text{MSPE}}$. The results are presented in the last two columns of Table 4, where $\tilde{\theta}_i$ represents the EBLUPs and MJ the corresponding estimated $\sqrt{\text{MSPE}}$s. Note that the same data was also analyzed by Datta *et al.* (2011), who stated that, because the estimated MSPEs for DHM are much smaller than those for EBLUP, the DHM method is "significantly more accurate". The results of our analysis show that this is not necessarily the case when additional variation in the model selection (by testing) is taken into account, and estimated correctly: Out of the 23 small areas, only 5 have smaller estimated $\sqrt{\text{MSPE}}$ for DHM as compared to EBLUP when comparing the MJs for both (column 8 vs column 10).

Finally, there is one area, #5, for which all of the uncertainty measures give essentially the same results, 0.047 (although, to the fourth digit, the DHM measure is still smaller than its BT and MJ counterparts). This case corresponds to the "outlier" for this data, according to Jiang, Nguyen and Rao (2011). As noted by the latter authors (also see Jiang *et al.* 2010), without this case, a quadratic, instead of cubic, mean function would fit the data well. However, there is an over-fitting problem for this particular area, that is, the outlier causes the cubic fit to be "perfect" for this area. This means that the fitted cubic function goes through exactly the data point; as a result, the direct estimate, y_5, is equal to the regression estimate, $x_5'\hat{\beta}$. It follows that there is no difference between the EBLUP and the direct and synthetic estimates, regardless of the value of D_5 and how one estimates A. Thus, in this case, every method essentially reduces to the direct estimate, $y_5 = 0.347$, and its measure of uncertainty, $\sqrt{D_5} = 0.047$.

Another real-data example on estimation of median income of four-person families is also considered. The details are deferred to Supplementary Material (http://intlpress.com/site/pub/pages/journals/items/amsa/content/vols/0003/0002/s001).

Concluding remark. We have shown that the impact of model selection in accuracy measures may be complicated. If the accuracy measure only focuses on the variance, model selection is likely to add additional variation to the measure. This is shown, for example, in Subsection 2.2, where the EBLUP is an unbiased estimator, hence the MSPE reduces to the variance. On the other hand, if the accuracy measure is the MSPE, the overall impact of model selection depends on the relative contributions of the bias and variance as in the identity MSPE = (prediction bias)^2 + prediction variance. As further discussed in Supplementary Material, model selection helps to reduce the bias but this may be at the cost of adding more variation. Because,

in practice, it is difficult to predict in which way, and how much, the overall impact is, the best strategy is to obtain an accurate MSPE estimator. We have shown that the latter can be done via McJack.

6. Proofs

6.1. Proof of Theorem 1

Throughout this proof, ψ denotes the true parameter vector. Let $\widetilde{b(\psi)}$ denote (10) with $\tilde{b}(\cdot)$ replaced by $b(\cdot)$. Also, c denotes a positive, generic constant, whose value may be different at different places. By Theorem 5.2 of Jiang *et al.* (2002), we have

$$\mathrm{E}_y\{\widetilde{b(\psi)} - b(\psi)\} = o(m^{-1-\gamma}), \tag{20}$$

where $\gamma = [(d-2)/(2d+1)] \wedge \nu > 0$, and E_y denotes expectation with respect to y. Because the left side of (20) does not depend on ξ, the equation also holds with E_y replaced by E.

Let E_ξ and P_ξ denote expectation and probability with respect to ξ. Consider

$$\begin{aligned} & \widehat{b(\psi)} - \widetilde{b(\psi)} \\ &= \hat{b}(\hat{\psi}) - b(\hat{\psi}) - \frac{m-1}{m}\sum_{j=1}^{m}\{\hat{b}(\hat{\psi}_{-j}) - b(\hat{\psi}_{-j}) + b(\hat{\psi}) - \hat{b}(\hat{\psi})\}. \end{aligned} \tag{21}$$

Let $\tilde{\psi}$ be a fixed parameter vector such that $|\tilde{\psi} - \psi| \le w$. Then, we have

$$\begin{aligned} \hat{b}(\tilde{\psi}) - b(\tilde{\psi}) &= \{\hat{b}(\tilde{\psi}) - b(\tilde{\psi})\}1_{(c_1/2 \le \tilde{s}(\tilde{\psi}) \le 2c_2)} \\ &\quad + \{\hat{b}(\tilde{\psi}) - b(\tilde{\psi})\}1_{(\tilde{s}(\tilde{\psi}) < c_1/2)} + \{\hat{b}(\tilde{\psi}) - b(\tilde{\psi})\}1_{(\tilde{s}(\tilde{\psi}) > 2c_2)} \\ &= I_1 + I_2 + I_3. \end{aligned} \tag{22}$$

First note that, by *A8*, we have $\mathrm{P}_\xi\{\tilde{s}(\tilde{\psi}) < c_1/2\} \le \mathrm{P}_\xi\{|\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})| > c_1/2\} \le (c_1/2)^{-q/2}\mathrm{E}_\xi\{|\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})|^{q/2}\}$. Next, write $u_k = \{\hat{\theta}_i^{(k)} - \theta_i^{(k)}\}^2$ and note that $\mathrm{E}_\xi(u_1) = s(\tilde{\psi})$. By Marcinkiewicz-Zygmund inequality (e.g., Jiang 2010, p. 150), we have

$$\mathrm{E}_\xi\{|\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})|^{q/2}\} = \frac{1}{K^{q/2}}\mathrm{E}_\xi\left[\left|\sum_{i=1}^{K}\{u_k - \mathrm{E}_\xi(u_1)\}\right|^{q/2}\right]$$

$$\begin{aligned} &\leq \frac{c}{K^{q/2}} \mathrm{E}_\xi \left[\sum_{k=1}^{K} \{u_k - \mathrm{E}_\xi(u_1)\}^2 \right]^{q/4} \\ &\leq \frac{c}{K^{q/4}} \times \frac{1}{K} \sum_{k=1}^{K} \mathrm{E}_\xi[|u_k - \mathrm{E}_\xi(u_1)|^{q/2}] \\ &\leq \frac{c}{K^{q/4}}, \end{aligned}$$

using Jensen's inequality for the second-to-last step, and *A8* for the last step. It follows, by *A8* and the definition of $\hat{b}(\cdot), b(\cdot)$ that

$$|\mathrm{E}_\xi(I_2)| \leq cm^\rho K^{-q/4}. \tag{23}$$

By essentially the same argument, we also have

$$|\mathrm{E}_\xi(I_3)| \leq cm^\rho K^{-q/4}. \tag{24}$$

Now suppose that $c_1/2 \leq \tilde{s}(\tilde{\psi}) \leq 2c_2$. We also know that $c_1 \leq s(\tilde{\psi}) \leq c_2$ by *A8*. Thus, for sufficiently large m, we have $\hat{b}(\tilde{\psi}) = \tilde{b}(\tilde{\psi})$. By Taylor series expansion, we have

$$\begin{aligned} \hat{b}(\tilde{\psi}) - b(\tilde{\psi}) &= \tilde{b}(\tilde{\psi}) - b(\tilde{\psi}) \\ &= \log\{\tilde{s}(\tilde{\psi})\} - \log\{s(\tilde{\psi})\} \\ &= \frac{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})}{s(\tilde{\psi})} - \frac{\{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})\}^2}{2s(\tilde{\psi})^2} + \frac{\{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})\}^3}{3\eta^3}, \end{aligned} \tag{25}$$

where η lies between $s(\tilde{\psi})$ and $\tilde{s}(\tilde{\psi})$; hence, we have $\eta \geq c_1/2$. It follows that

$$\begin{aligned} &\left| \mathrm{E}_\xi \left[\frac{\{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})\}^3}{3\eta^3} 1_{(c_1/2 \leq \tilde{s}(\tilde{\psi}) \leq 2c_2)} \right] \right| \\ &\leq \frac{8}{3c_1^3} \mathrm{E}_\xi\{|\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})|^3\} \\ &\leq cK^{-3/2}, \end{aligned} \tag{26}$$

using an earlier inequality. Similarly, we have

$$\left| \mathrm{E}_\xi \left[\frac{\{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})\}^2}{2s(\tilde{\psi})^2} 1_{(c_1/2 \leq \tilde{s}(\tilde{\psi}) \leq 2c_2)} \right] \right| \leq cK^{-1}. \tag{27}$$

Furthermore, note that $\mathrm{E}_\xi\{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})\} = 0$, thus, we have

$$\begin{aligned}
& \left| \mathrm{E}_\xi \left[\frac{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})}{s(\tilde{\psi})} 1_{(c_1/2 \le \tilde{s}(\tilde{\psi}) \le 2c_2)} \right] \right| \\
= & \frac{1}{s(\tilde{\psi})} \left| \mathrm{E}_\xi [\{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})\} 1_{(\tilde{s}(\tilde{\psi}) < c_1/2 \text{ or } \tilde{s}(\tilde{\psi}) > 2c_2)}] \right| \\
\le & \frac{\mathrm{E}_\xi [\{\tilde{s}(\tilde{\psi}) + s(\tilde{\psi})\} 1_{(\tilde{s}(\tilde{\psi}) < c_1/2)}]}{s(\tilde{\psi})} \\
& + \frac{\mathrm{E}_\xi [\{\tilde{s}(\tilde{\psi}) + s(\tilde{\psi})\} 1_{(\tilde{s}(\tilde{\psi}) > 2c_2)}]}{s(\tilde{\psi})}.
\end{aligned}$$

By Hölder and Jensen's inequalities, *A8* and an earlier result, we have

$$\begin{aligned}
& \mathrm{E}_\xi [\{\tilde{s}(\tilde{\psi}) + s(\tilde{\psi})\} 1_{(\tilde{s}(\tilde{\psi}) < c_1/2)}] \\
\le & [\mathrm{E}_\xi \{\tilde{s}(\tilde{\psi}) + s(\tilde{\psi})\}^{q/2}]^{2/q} [\mathrm{P}_\xi \{\tilde{s}(\tilde{\psi}) < c_1/2\}]^{1-2/q} \\
\le & c \left\{ \frac{1}{K} \sum_{k=1}^{K} \mathrm{E}_\xi (u_k^{q/2}) + c_2^{q/2} \right\}^{2/q} K^{-(q/4)(1-2/q)} \\
\le & c K^{-(q-2)/4}.
\end{aligned}$$

Similarly, we have $\mathrm{E}_\xi [\{\tilde{s}(\tilde{\psi}) + s(\tilde{\psi})\} 1_{(\tilde{s}(\tilde{\psi}) > 2c_2)}] \le cK^{-(q-2)/4}$. It follows that

$$\left| \mathrm{E}_\xi \left[\frac{\tilde{s}(\tilde{\psi}) - s(\tilde{\psi})}{s(\tilde{\psi})} 1_{(c_1/2 \le \tilde{s}(\tilde{\psi}) \le 2c_2)} \right] \right| \le cK^{-(q-2)/4}. \tag{28}$$

Combining (24)–(28), and the fact that $(q-2)/4 \ge 1$ by *A8*, we conclude that

$$|\mathrm{E}_\xi(I_1)| \le cK^{-1}. \tag{29}$$

Thus, combining (22)–(24), and (29), we have

$$|\mathrm{E}_\xi \{\hat{b}(\tilde{\psi}) - b(\tilde{\psi})\}| \le c \left[m^\rho K^{-q/4} + K^{-1} \right], \text{ if } |\tilde{\psi} - \psi| \le w, \tag{30}$$

where c does not depend on $\tilde{\psi}$.

Now, for any $0 \le j \le m$, we have

$$\mathrm{E}\{\hat{b}(\hat{\psi}_{-j}) - b(\hat{\psi}_{-j})\} = \mathrm{E}_y [\mathrm{E}_\xi \{\hat{b}(\hat{\psi}_{-j}) - b(\hat{\psi}_{-j}) | \hat{\psi}_{-j}\}] = \mathrm{E}_y \{\Delta(\hat{\psi}_{-j})\},$$

where $\Delta(\tilde{\psi}) = \mathrm{E}_\xi \{\hat{b}(\hat{\psi}_{-j}) - b(\hat{\psi}_{-j}) | \hat{\psi}_{-j} = \tilde{\psi}\} = \mathrm{E}_\xi \{\hat{b}(\tilde{\psi}) - b(\tilde{\psi})\}$ by *A7*. Therefore, we have

$$\begin{aligned} & \mathrm{E}\{\hat{b}(\hat{\psi}_{-j}) - b(\hat{\psi}_{-j})\} \\ (31) \qquad = \; & \mathrm{E}_y\{\Delta(\hat{\psi}_{-j}) 1_{(|\hat{\psi}_{-j}-\psi| \le w)}\} + \mathrm{E}_y\{\Delta(\hat{\psi}_{-j}) 1_{(|\hat{\psi}_{-j}-\psi| > w)}\}. \end{aligned}$$

By (30), the first term on the right side of (31) is bounded in absolute value by $c[m^{\rho}K^{-q/4} + K^{-1}]$. As for the second term, by the definition of $\hat{b}(\cdot), b(\cdot)$, and *A5*, it is bounded in absolute value by $cm^{\rho-d}$. Thus, in conclusion, we have

$$\begin{aligned} & |\mathrm{E}\{\hat{b}(\hat{\psi}_{-j}) - b(\hat{\psi}_{-j})\}| \\ (32) \qquad \le \; & c\left[m^{\rho}K^{-q/4} + K^{-1} + m^{\rho-d}\right], \;\; 0 \le j \le m. \end{aligned}$$

Combining (21), (32), we have

$$\begin{aligned} |\mathrm{E}\{\widehat{b(\psi)} - \widetilde{b(\psi)}\}| \; & \le \; c\left(m^{1+\rho}K^{-q/4} + \frac{m}{K} + m^{1+\rho-d}\right) \\ (33) \qquad & = \; o(m^{-1}), \end{aligned}$$

by *A9* and the conditions on d, q.

The result then follows by (20) (with E_y replaced by E) and (33).

6.2. Proof of Theorem 2

First, by (i)–(iv) of Jiang *et al.* (2002, p. 1803), it is easy to see that assumptions *A1–A6* are satisfied. Assumption *A7* is satisfied by the statement above *A1*. Thus, all we need is to verify assumption *A8*. Once again, in the arguments below, c denotes a positive constant whose value may be different at different places.

Suppose that the data is generated under the parameter vector $\tilde{\psi}$. Let $\tilde{\theta}_i$ denote the BP of θ_i. Then, we have $s(\tilde{\psi}) = \mathrm{MSPE}_{\tilde{\psi}}(\hat{\theta}_i) = \mathrm{MSPE}_{\tilde{\psi}}(\tilde{\theta}_i) + \mathrm{E}_{\tilde{\psi}}\{(\hat{\theta}_i - \tilde{\theta}_i)^2\} \ge \mathrm{MSPE}_{\tilde{\psi}}(\tilde{\theta}_i) = \tilde{A}D_i/(\tilde{A} + D_i)$. Thus, if $0 < A/2 \le \tilde{A} \le 2A$, where $\tilde{A}$ is the A component of $\tilde{\psi}$, and A is the true A, $s(\tilde{\psi})$ is clearly bounded away from 0.

On the other hand, we have $s(\tilde{\psi}) = \mathrm{E}_{\tilde{\psi}}(\hat{\theta}_i^2) - 2\mathrm{E}_{\tilde{\psi}}(\hat{\theta}_i\theta_i) + \mathrm{E}_{\tilde{\psi}}(\theta_i^2)$. By (3), we have $\mathrm{E}(\theta_i^2) \le 2(|x_{\mathrm{f},i}|^2|\tilde{\beta}_{\mathrm{f}}|^2 + \tilde{A}^2) \le c$, if, say $|\tilde{\beta}_{\mathrm{f}} - \beta_{\mathrm{f}}| \le 1$ and $\tilde{A} \le 2A$. Also, by (6), and Jensen's inequality, we have

$$(34) \qquad \hat{\theta}_i^2 \le \frac{\hat{A}_{\mathrm{f}}}{\hat{A}_{\mathrm{f}} + D_i} y_i^2 + \frac{D_i}{\hat{A}_{\mathrm{f}} + D_i} |x_{\mathrm{f},i}\hat{\beta}_{\mathrm{f}}|^2 \le y_i^2 + |x_{\mathrm{f},i}|^2|\hat{\beta}_{\mathrm{f}}|^2,$$

and, by (2), $\mathrm{E}_{\tilde{\psi}}(y_i^2) = \{\mathrm{E}_{\tilde{\psi}}(y_i)\}^2 + \mathrm{var}_{\tilde{\psi}}(y_i) = (x_{\mathrm{f},i}\tilde{\beta}_\mathrm{f})^2 + \tilde{A} + D_i \le |x_{\mathrm{f},i}|^2|\tilde{\beta}_\mathrm{f}|^2 + \tilde{A} + D_i \le c$. Define $P_\mathrm{f} = I_m - D^{-1/2}X_\mathrm{f}(X_\mathrm{f}'D^{-1}X_\mathrm{f})^{-1}X_\mathrm{f}'D^{-1/2}$. By Lemma 1 of Jiang (2000), with $\hat{V} = \hat{A}I_m + D$, $D = \mathrm{diag}(D_i, 1 \le i \le m)$, $X_D = D^{-1/2}X_\mathrm{f}$, $Z = D^{-1/2}$, $\Gamma = \hat{A}I_m$, and $\zeta = y - X_\mathrm{f}\tilde{\beta}_\mathrm{f}$, we have

$$\begin{aligned}
\hat{\beta}_\mathrm{f} &= (X_\mathrm{f}'\hat{V}^{-1}X_\mathrm{f})^{-1}X_\mathrm{f}'\hat{V}^{-1}y \\
&= \tilde{\beta}_\mathrm{f} + (X_\mathrm{f}'\hat{V}^{-1}X_\mathrm{f})^{-1}X_\mathrm{f}'\hat{V}^{-1}\zeta \\
&= \tilde{\beta}_\mathrm{f} + \{X_D'(I_m + Z\Gamma Z')^{-1}X_D\}^{-1}X_D'(I_m + Z\Gamma Z')^{-1}D^{-1/2}\zeta \\
&= \tilde{\beta}_\mathrm{f} + (X_D'X_D)^{-1}X_D'\{I_m - \hat{A}D^{-1}P_\mathrm{f}(I_m + \hat{A}P_\mathrm{f}D^{-1}P_\mathrm{f})^{-1}\}D^{-1/2}\zeta \\
&= \tilde{\beta}_\mathrm{f} + (X_\mathrm{f}'D^{-1}X_\mathrm{f})^{-1}X_\mathrm{f}'D^{-1}\zeta \\
&\quad -(X_\mathrm{f}'D^{-1}X_\mathrm{f})^{-1}X_\mathrm{f}'D^{-1}\hat{A}D^{-1/2}P_\mathrm{f}(I_m + \hat{A}P_\mathrm{f}D^{-1}P_\mathrm{f})^{-1}D^{-1/2}\zeta \\
&= \tilde{\beta}_\mathrm{f} + I_1 - I_2. \qquad (35)
\end{aligned}$$

Note that $\mathrm{E}_{\tilde{\psi}}(\zeta\zeta') = \tilde{A}I_m + D \le cD$ under the assumptions. Thus, we have

$$\begin{aligned}
\mathrm{E}_{\tilde{\psi}}(|I_1|^2) &= \mathrm{E}_{\tilde{\psi}}\left[\mathrm{tr}\{(X_\mathrm{f}'D^{-1}X_\mathrm{f})^{-1}X_\mathrm{f}'D^{-1}\zeta\zeta'D^{-1}X_\mathrm{f}(X_\mathrm{f}'D^{-1}X_\mathrm{f})^{-1}\}\right] \\
&= \mathrm{tr}\left\{(X_\mathrm{f}'D^{-1}X_\mathrm{f})^{-1}X_\mathrm{f}'D^{-1}\mathrm{E}_{\tilde{\psi}}(\zeta\zeta')D^{-1}X_\mathrm{f}(X_\mathrm{f}'D^{-1}X_\mathrm{f})^{-1}\right\} \\
&\le c\mathrm{tr}\{(X_\mathrm{f}'D^{-1}X_\mathrm{f})^{-1}\} \\
&\le \frac{c}{m\lambda_{\min}(m^{-1}X_\mathrm{f}'X_\mathrm{f})}. \qquad (36)
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
|I_2| &\le \|(X_\mathrm{f}'D^{-1}X_\mathrm{f})^{-1}X_\mathrm{f}'D^{-1}\| \cdot \|\hat{A}D^{-1/2}P_\mathrm{f}(I_m + \hat{A}P_\mathrm{f}D^{-1}P_\mathrm{f})^{-1} \\
&\quad \times |D^{-1/2}\zeta|.
\end{aligned}$$

By a similar argument as above, we have

$$\|(X_\mathrm{f}'D^{-1}X_\mathrm{f})^{-1}X_\mathrm{f}'D^{-1}\|^2 \le \frac{c}{m\lambda_{\min}(m^{-1}X_\mathrm{f}'X_\mathrm{f})}.$$

Next, let $\lambda_1 \ge \cdots \ge \lambda_m \ge 0$ be the eigenvalues $P_\mathrm{f}D^{-1}P_\mathrm{f}$. Then, we have

$$\|\hat{A}D^{-1/2}P_\mathrm{f}(I_m + \hat{A}P_\mathrm{f}D^{-1}P_\mathrm{f})^{-1}\|^2 = \max_{1\le i\le m}\frac{\hat{A}^2\lambda_i}{(1+\hat{A}\lambda_i)^2}.$$

If $\hat{A}\lambda_i = 0$, then $\hat{A}^2\lambda_i/(1+\hat{A}\lambda_i)^2 = 0$; and $\hat{A}^2\lambda_i/(1+\hat{A}\lambda_i)^2 \le \hat{A}^2\lambda_i/\hat{A}^2\lambda_i^2 = 1/\lambda_i$ otherwise. It follows that

$$\max_{1\le i\le m} \frac{\hat{A}^2\lambda_i}{(1+\hat{A}\lambda_i)^2} \quad \le \quad \frac{1}{\lambda_r},$$

where $r = \text{rank}(P_\text{f} D^{-1} P_\text{f})$. Because $P_\text{f} D^{-1} P_\text{f} u = 0$ if and only if $P_\text{f} u = 0$, we have $r = \text{rank}(P_\text{f})$, and P_f is a projection matrix, whose eigenvalues are 0 or 1. Also, because

$$P_\text{f} D^{-1} P_\text{f} \ge \frac{P_\text{f}^2}{\max_{1\le i\le m} D_i} = \frac{P_\text{f}}{\max_{1\le i\le m} D_i},$$

by a well-known eigenvalue inequality (e.g., DasGupta 2008, p. 669), we have

$$\lambda_r \ge \lambda_r \left(\frac{P_\text{f}}{\max_{1\le i\le m} D_i} \right) = \frac{\lambda_r(P_\text{f})}{\max_{1\le i\le m} D_i} = \frac{1}{\max_{1\le i\le m} D_i}.$$

Thus, in conclusion, we have

$$\|\hat{A} D^{-1/2} P_\text{f} (I_m + \hat{A} P_\text{f} D^{-1} P_\text{f})^{-1}\|^2 \le \max_{1\le i\le m} D_i.$$

Finally, it is easy to show that $\text{E}_{\tilde{\psi}}(|D^{-1/2}\zeta|^2) \le cm$. Thus, combining the results, we have

$$\text{E}_{\tilde{\psi}}(|I_2|^2) \quad \le \quad \frac{c}{\lambda_{\min}(m^{-1} X_\text{f}' X_\text{f})}. \tag{37}$$

The upper bound for $s(\tilde{\psi})$ follows from (34)–(37).

The last part of *A8* follows from the above arguments by noting that $\theta_i^{(k)} = x_{\text{f},i}' \tilde{\beta}_\text{f} + \tilde{A}^{1/2} \xi_i^{(k)}$, $y_i = \theta_i^{(k)} + \sqrt{D_i}\eta_i^{(k)}$, and $\xi_i^{(k)}, \eta_i^{(k)}$ are $N(0,1)$ random variables.

References

Akaike, H. (1973), Information theory as an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki eds.), Akademiai Kiado, Budapest, 267–281. MR0483125

Berk, R., Brown, L., and Zhao, L. (2010), Statistical inference after model selection, *J. Quant. Criminol.* 26, 217–236.

Chen, S. and Lahiri, P. (2011), On the estimation of mean squared prediction error in small area estimation, *Calcutta Statist. Assoc. Bull.* 63, 249–252. MR2986705

Das, K., Jiang, J., and Rao, J. N. K. (2004), Mean squared error of empirical predictor, *Ann. Statist.* 32, 818–840. MR2060179

DasGupta, A. (2008), *Asymptotic Theory of Statistics and Probability*, Springer, New York. MR2664452

Datta, G. S., Hall, P., and Mandal, A. (2011), Model selection by testing for the presence of small-area effects, and applications to area-level data, *J. Amer. Statist. Assoc.* 106, 361–374. MR2816727

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. MR1885193

Efron, B. (1992), Jackknife-after-bootstrap standard errors and influence functions (with discussion), *J. Roy. Statist. Soc. Ser. B* 54, 83–127. MR1157715

Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.* 96, 1348–1360. MR1946581

Fay, R. E. and Herriot, R. A. (1979), Estimates of income for small places: an application of James-Stein procedures to census data, *J. Amer. Statist. Assoc.* 74, 269–277. MR0548019

Ganesh, N. (2009), Simultaneous credible intervals for small area estimation problems, *J. Multivariate Anal.* 100, 1610–1621. MR2535373

Gershunskaya, J. B. and Dorfman, A. H. (2013), Calibration and evaluation of generalized variance functions, *Proceedings of Joint Statistical Meetings, Survey Methods Section*, 2655–2669.

Hall, P. and Maiti, T. (2006), On parametric bootstrap methods for small area prediction, *J. Roy. Statist. Soc. Ser. B* 68, 221–238. MR2188983

Jiang. J. (2000), A matrix inequality and its statistical application, *Linear Algebra Appl.* 307, 131–144. MR1741921

Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York. MR2308058

Jiang, J. (2010), *Large Sample Techniques for Statistics*, Springer, New York. MR2675055

Jiang, J. (2014), The fence methods, in *Advances in Statistics*, Vol. 2014, 1–14, Hindawi Publishing Corp.

Jiang, J., Lahiri, P., and Wan, S.-M. (2002), A unified jackknife theory for empirical best prediction with M-estimation, *Ann. Statist.* 30, 1782–1810. MR1969450

Jiang, J., Nguyen, T., and Rao, J. S. (2010), Fence method for nonparametric small area estimation, *Survey Methodology* 36, 3–11.

Jiang, J., Nguyen, T., and Rao, J. S. (2011), Best predictive small area estimation, *J. Amer. Statist. Assoc.* 106, 732–745. MR2847987

Jiang, J. and Rao, J. S. (2003), Consistent procedures for mixed linear model selection, *Sankhya A* 65, 23–42. MR2016775

Lahiri, P. and Rao, J. N. K. (1995), Robust estimation of mean squared errors of small area estimators, *J. Amer. Statist. Assoc.* 90, 758–766. MR1340527

Lahiri, P. and Suntornchost, J. (2014), Variable selection for linear mixed models with applications to small area estimation, *Sankhya*, DOI 10.1007/s13571-015-0096-0. MR3434486

Leeb, H. (2009), Conditional predictive inference after model selection, *Ann. Statist.* 37, 2838–2876. MR2541449

Molina, I., Rao, J. N. K., and Datta, G. S. (2015), Samll area estimation under a Fay-Herriot model with preliminary testing for the presence of random area effects, *Survey Methodology*, in press.

Morris, C. N. and Christiansen, C. L. (1995), Hierarchical models for ranking and for identifying extremes with applications, *Bayes Statistics* 5, Oxford Univ. Press. MR1425411

Müller, S., Scealy, J. L., and Welsh, A. H. (2013), Model selection in linear mixed models, *Statist. Sci.* 28, 135–167. MR3112403

Pang, Z., Lin, B., and Jiang, J. (2016), Regularization parameter selections with divergent and NP-dimensionality via bootstrapping, *Austral. New Zealand J. Statist.* 58, 335–356.

Pfeffermann, D. (2013), New important developments in small area estimation, *Statist. Sci.* 28, 40–68. MR3075338

Prasad, N. G. N. and Rao, J. N. K. (1990), The estimation of mean squared errors of small area estimators, *J. Amer. Statist. Assoc.* 85, 163–171. MR1137362

Rao, J. N. K. and Molina, I. (2015), *Small Area Estimation*, 2nd ed., Wiley, New York. MR3380626

Rao, J. N. K. and Yu, M. (1994), Small area estimation by combining time series and cross-sectional data, *Canad. J. Statist.* 22, 511–528. MR1321472

Rao, C. R. and Wu, Y. (2001), On model selection, in *IMS Lecture Notes–Monograph Series* 38, 1–57. MR2000751

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Statist.* 6, 461–464. MR0468014

Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso, *J. Roy. Statist. Soc. Ser. B* 16, 385–395. MR1379242

Yoshimori, M. and Lahiri, P. (2014), A second-order efficient empirical Bayes confidence interval, *Ann. Statist.* 42, 1233–1261. MR3226156

Zimmerman, D., Pavlik, C., Ruggles, A., and Armstrong, M. P. (1999), An experimental comparison of ordinary and universal Kriging and inverse distance weighting, *Math. Geol.* 31, 375–390.

Jiming Jiang
Department of Statistics
University of California, Davis
Davis, CA 95616
USA
E-mail address: jimjiang@ucdavis.edu

P. Lahiri
Joint Program in Survey Methodology
1218 LeFrak Hall
7251 Preinkert Dr.
College Park, MD 20742
USA
E-mail address: plahiri@umd.edu

Thuan Nguyen
School of Public Health
840 SW Gaines St.
Portland, OR 97239
USA
E-mail address: nguythua@ohsu.edu

Received February 16, 2016