Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 311–326
http://www.ssca.org.in/journal

Mixed Model Selection with Applications to Small Area Estimation

J. Sunil Rao[1] **and J. N. K. Rao**[2]
[1]*Division of Biostatistics, University of Minnesota, Twin Cities, USA*
[2]*Department of Mathematics and Statistics, Carleton University, Ottawa, Canada*

Received: 30 July 2024; Revised: 28 August 2024; Accepted: 31 August 2024

Abstract

Mixed models have widespread appeal in many areas of statistical modeling including small area estimation. Here we review a variety of different approaches for linear mixed model selection eventually arriving at the specific problem of selecting variables in small area models ranging from parametric and non-parametric area and unit level models to subarea small area models.

Key words: Subareas

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Many model search strategies involve trading off model fit with model complexity in a penalized goodness of fit measure. Asymptotic properties for these types of procedures in settings like linear regression and ARMA time series have been studied. Yet, these strategies do not generalize naturally to more complex models, such as those for modeling correlated data or those that involve adaptive estimation. In these cases, penalties and model complexity may not be naturally defined.

Since the introduction of Akaike's information criterion (AIC, Akaike 1973, 1974), a number of similar criteria have been proposed, including the Bayesian information criterion (BIC; Schwarz 1978), a criterion due to Hannan and Quinn (HQ; 1979), and the generalized information criterion (GIC; Nishii 1984, Shibata 1984). These procedures essentially amount to minimize a criterion function, which may be expressed as

$$\hat{D}_M + \lambda_n |M|, \tag{1}$$

where M represents a candidate model, $\hat{D}_M$ is a measure of lack of fit by M, and $|M|$ denotes the dimension of M, usually in terms of the number of estimated parameters under M. The difference is made by λ_n, where n is the sample size. This is called a "penalizer", although some authors refer $\lambda_n|M|$ as the penalizer. For example, $\lambda_n = 2$ for AIC; $\lambda_n = \log(n)$ for

Corresponding Author: J. Sunil Rao
Email: js-rao@umn.edu

BIC; and $\lambda_n = c\log\{\log(n)\}$ for HQ, where c is a constant greater than 2 (Bozdogan 1987, pp. 359).

1.1. Contributions of C. R. Rao

It should be no surprise that Professor C.R. Rao has made important contributions to model selection (in addition to the many other fundamental results he has given the field). Two specific examples are found in Rao and Wu (1989) and Bai *et al.* (1999). In both cases, the problem under study was that of linear model seletion. Specifically, the authors considered the (possibly overfit) linear model,

$$y = X\beta + \epsilon, \tag{2}$$

where y is a n-vector of response observations, X is a known design matrix, β is a p-vector of unknown regression parameters, and ϵ is a random error n-vector. Certain components of β may or may not be zero. There are thus 2^p total submodels, one of which is assumed to be the true model generating the responses.

Rao and Wu (1989) developed a criterion in the family of (1) with a flexible penalty function and proved strong consistency of model selection (that is, finding the true model). Their method allowed a wider range of penalty functions thus leading to improved small sample performance by adaptively choosing the best penalty function from the collection of candidate ones. Specifically, they entertained $\lambda_n = \alpha n^\gamma$ where $\gamma < 1$. They called a combination (α, γ) of interest if all of the models in a collection of new perturbed models built off the fitted full model are correctly selected. There may in fact be more than one combination of (α, γ) that share this property and thus Rao and Wu (1989) suggested that additional work is warranted in choosing among them. In Bai *et al.* (1999), they derived a particular choice of λ_n based on observed data, which makes it random. They then proved that the consistency property can still hold.

2. Mixed model selection

Consider the following mixed linear model:

$$y \;=\; X\beta + Z\alpha + \epsilon\,, \tag{3}$$

where $y = (y_i)_{1\le i\le N}$ is a vector of observations; $\beta = (\beta_j)_{1\le j\le p}$ is a vector of unknown regression coefficients (the fixed effects); $\alpha = (\alpha_j)_{1\le j\le m}$ is a vector of unobservable random variables (the random effects); $\epsilon = (\epsilon_i)_{1\le i\le N}$ is a vector of errors; and X, Z are known matrices. We assume that $E(\alpha) = 0$, $\mathrm{Var}(\alpha) = G$; $E(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = R$, where G and R may involve some unknown parameters such as variance components; and α and ϵ are uncorrelated.

2.1. Random factors not subject to selection

In this section, we consider the model selection problem when the random part of the model, *i.e.*, $Z\alpha$, is not subject to selection. Let $\zeta = Z\alpha + \epsilon$. Then, the problem is closely related to a regression model selection problem with correlated errors. Consider the

following general linear model:

$$y \;=\; X\beta + \zeta\ , \tag{4}$$

where ζ is a vector of correlated errors, and everything else is as in (3). We assume that there are a number of candidate vectors of covariates, $X_1, \ldots, X_q$, from which the columns of X are to be selected. Let $K = \{1, \ldots, q\}$. Then, the set of all possible models can be expressed as $\mathcal{B} = \{k : k \subseteq K\}$, and there are 2^q possible models. Let $\mathcal{A}$ be a subset of $\mathcal{B}$ that is known to contain the true model, so the selection will be within $\mathcal{A}$. In an extreme case, $\mathcal{A}$ may be $\mathcal{B}$ itself. For any matrix M, let $\mathcal{L}(M)$ be the linear space spanned by the columns of M; P_M the projection onto $\mathcal{L}(M)$: $P_M = M(M^TM)^-M^T$; and $P_M^\perp$ the orthogonal projection: $P_M^\perp = I - P_M$. For any $k \in \mathcal{B}$, let $X(k)$ be the matrix whose columns are X_j, $j \in k$, if $k \neq \emptyset$; and $X(k) = 0$ if $k = \emptyset$. We consider the following criterion for model selection:

$$C_N(k) \;=\; |y - X(k)\hat{\beta}(k)|^2 + \lambda_N|k| \;=\; |P^\perp_{X(k)}y|^2 + \lambda_N|k|\ , \tag{5}$$

$k \in \mathcal{A}$, where $|k|$ represents the cardinality of k; $\hat{\beta}(k)$ is the ordinary least squares (OLS) estimator of $\beta(k)$ for the model $y = X(k)\beta(k) + \zeta$, *i.e.*,

$$\hat{\beta}(k) \;=\; [X(k)^TX(k)]^-X(k)^Ty$$

and λ_N is a positive number satisfying certain conditions specified below. Note that $P_{X(k)}$ is understood as 0 if $k = \emptyset$. Denote the true model by k_0. If $k_0 \neq \emptyset$, we denote the corresponding X and β by X and $\beta = (\beta_j)_{1\le j\le p}$ ($p = |k_0|$), and assume that $\beta_j \neq 0$, $1 \le j \le p$. This is, of course, reasonable because otherwise the model can be further simplified. If $k_0 = \emptyset$, X, β, and p are understood as 0. For $1 \le j \le q$, Let $\{j\}^c$ represent the set $K \setminus \{j\}$. We define the following sequences: $\omega_N = \min_{1\le j\le q} |P^\perp_{X(\{j\}^c)}X_j|^2$, $\nu_N = \max_{1\le j\le q} |X_j|^2$, and $\rho_N = \lambda_{\max}(ZGZ^T) + \lambda_{\max}(R)$, where $\lambda_{\max}$ means largest eigenvalue. Let $\hat{k}$ be the minimizer of (5) over $k \in \mathcal{A}$, which will be our selection of the model. The following theorem gives sufficient conditions under which the selection is consistent in the sense that

$$P(\hat{k} \neq k_0) \;\longrightarrow\; 0\ . \tag{6}$$

Theorem 1. (Jiang and Rao 2003) Suppose that $\nu_N > 0$ for large N,

$$\rho_N/\nu_N \;\longrightarrow\; 0\ , \quad \text{while} \quad \liminf(\omega_N/\nu_N) \;>\; 0\ . \tag{7}$$

Then, (4) holds for any λ_N such that

$$\lambda_N/\nu_N \;\longrightarrow\; 0 \;\text{ and }\; \rho_N/\lambda_N \;\longrightarrow\; 0\ . \tag{8}$$

The above procedure requires selecting $\hat{k}$ from all subset of $\mathcal{A}$. Note that $\mathcal{A}$ may contain as many as 2^q subsets. When q is relatively large, alternative procedures have been proposed, in the (fixed effects) linear model context, which require less computation [*e.g.*, Zheng and Loh (1995)]. In the following, we consider an approach which is similar, in spirit, to Rao and Wu (1989). First, note that one can always express $X\beta$ in (4) as

$$X\beta \;=\; \sum_{j=1}^{q} \beta_j X_j \tag{9}$$

with the understanding that some of the coefficients β_j may be zero. It follows that $k_0 = \{1 \le j \le q : \beta_j \neq 0\}$. Let $X_{-j} = (X_u)_{1 \le u \le q, u \neq j}$, $1 \le j \le q$, $\eta_N = \min_{1 \le j \le q} |P^{\perp}_{X_{-j}} X_j|^2$, and δ_N be a sequence of positive numbers satisfying conditions specified below. Let $\hat{k}$ be the subset of K such that

$$(|P^{\perp}_{X_{-j}} y|^2 - |P^{\perp}_{X} y|^2)/(|P^{\perp}_{X_{-j}} X_j|^2 \delta_N) \;\; > \;\; 1 \, . \tag{10}$$

The following theorem states that, under suitable conditions, $\hat{k}$ is a consistent selection. Recall that ρ_N is defined above Theorem 1.

Theorem 2. (Jiang and Rao 2003) Suppose that $\eta_N > 0$ for large N, and

$$\rho_N/\eta_N \;\; \longrightarrow \;\; 0 \, . \tag{11}$$

Then, (6) holds for any δ_N such that

$$\delta_N \;\; \longrightarrow \;\; 0 \;\; \text{and} \;\; \rho_N/(\eta_N \delta_N) \;\; \longrightarrow \;\; 0 \, . \tag{12}$$

2.2. Selection of random factors

We now assume that $Z\alpha$ in (3) can be expressed as

$$Z\alpha \;\; = \;\; \sum_{j=1}^{s} Z_j \alpha_j \, , \tag{13}$$

where $Z_1, \ldots, Z_s$ are known matrices; each α_j is a vector of independent random effects with mean 0 and variance σ_j^2, which is unknown, $1 \le j \le s$. Furthermore, we assume that ϵ in (3) is a vector of independent errors with mean 0 and variance $\tau^2 > 0$, and $\alpha_1, \ldots, \alpha_s, \epsilon$ are independent. Such assumptions are customary in the mixed model context (*e.g.*, Searle, Casella, and McCulloch (1992), pp 233-234), therefore (13) represents a fairly general class of mixed linear models. If $\sigma_j^2 > 0$, we say that α_j is in the model; otherwise, it is not. Therefore, the selection of random factors is equivalent to simultaneously determining which of the variance components $\sigma_1^2, \ldots, \sigma_s^2$ are positive, and which of them are zero. The true model can be expressed as

$$y \;\; = \;\; X\beta + \sum_{j \in l_0} Z_j \alpha_j + \epsilon \, , \tag{14}$$

where $X = (X_j)_{j \in k_0}$ and $k_0 \subseteq K$ (see section 2); $l_0 \subseteq L = \{1, \ldots, s\}$ such that $\sigma_j^2 > 0$, $j \in l_0$, and $\sigma_j^2 = 0$, $j \in L \setminus l_0$.

There are some important differences between selecting the fixed covariates X_j and selecting the random factors. One difference is that, in selecting the random factors, we are going to determine whether the vector α_j, not a given component of α_j, should be in the model. In other words, the components of α_j are all "in" or all "out". Another difference is that, unlike selecting the fixed covariates, where it is reasonable to assume that the X_js are linearly independent, in a mixed linear model it is possible to have $j \neq j^T$ but $\mathcal{L}(Z_j) \subset \mathcal{L}(Z_{j^T})$.

First, note that in section 2.1 we discussed a procedure to determine the fixed part of the model, which leads to a selection $\hat{k}$ that satisfies (6). Note that the only place that the determination of $\hat{k}$ might use knowledge about Z, and hence about l_0, is through λ_N, which depends on the order of $\lambda_{\max}(ZGZ^T)$. However, under (13), $\lambda_{\max}(ZGZ^T) \leq \sum_{j=1}^{s} \sigma_j^2 \|Z_j\|^2$, where for any matrix M, $\|M\| = [\lambda_{\max}(M^T M)]^{1/2}$. Thus, an upper bound for the order of $\lambda_{\max}(ZGZ^T)$ is $\max_{1\leq j\leq s} \|Z_j\|^2$, which does not depend on l_0. Therefore, $\hat{k}$ could be determined without knowing l_0. In any case, we may write $\hat{k} = \hat{k}(l_0)$, be it dependent on l_0 or not. Now, suppose that a selection for the random part of the model, *i.e.*, a determination of l_0, is $\hat{l}$. We then define $\hat{k} = \hat{k}(\hat{l})$. The following theorem shows that the combined procedure is consistent.

Theorem 3. (Jiang and Rao 2003) Suppose that $P(\hat{l} \neq l_0) \to 0$ and $P(\hat{k}(l_0) \neq k_0) \to 0$. Then, $P(\hat{k} = k_0 \text{ and } \hat{l} = l_0) \to 1$.

How does one actually obtain $\hat{l}$? Jiang and Rao (2003) divided the vectors $\alpha_1, \ldots, \alpha_s$, or, equivalently, the matrices $Z_1, \ldots, Z_s$ into several groups. The first group is called the "largest random factors". Roughly speaking, those are Z_j, $j \in L_1 \subseteq L$ such that $\text{rank}(Z_j)$ is of the same order as N, the sample size. We can assume that $\mathcal{L}(X, Z_u, u \in L \setminus \{j\}) \neq \mathcal{L}(X, Z_u, u \in L)$, $j \in L_1$, where $\mathcal{L}(M_1, \ldots, M_t)$ represents the linear space spanned by the columns of the matrices $M_1, \ldots, M_t$. Such an assumption is reasonable because Z_j is supposed to be "largest", and hence should have contribution to the linear space. The second group consists of Z_j, $j \in L_2 \subseteq L$ such that $\mathcal{L}(X, Z_u, u \in L \setminus L_1 \setminus \{j\}) \neq \mathcal{L}(X, Z_u, u \in L \setminus L_1)$, $j \in L_2$. The ranks of the matrices in this group are of lower order of N. Similarly, the third group consists of Z_j, $j \in L_3 \subseteq L$ such that $\mathcal{L}(X, Z_u, u \in L \setminus L_1 \setminus L_2 \setminus \{j\}) \neq \mathcal{L}(X, Z_u, u \in L \setminus L_1 \setminus L_2)$, and so on. Note that if the first group, *i.e.*, the largest random factors, does not exist, the second group becomes the first, and other groups also move on. Jiang and Rao (2003) gave a procedure that determines the indexes $j \in L_1$ for which $\sigma_j^2 > 0$; then a procedure that determines the indexes $j \in L_2$ for which $\sigma_j^2 > 0$; and so on.

3. Fence methods

Although criteria like (1) are broadly used, difficulties are often encountered, especially in some non-conventional situations. For example, consider the following linear mixed model written at the unit level,

$$y_{ij} = x_{ij}^T \beta + u_i + v_j + e_{ij}, i = 1, \ldots, m_1, j = 1, \ldots, m_2, \tag{15}$$

where x_{ij} is a vector of known covariates, β is a vector of unknown regression coefficients (the fixed effects), u_i, v_j are random effects, and e_{ij} is an additional error term. It is assumed that u_i's, v_j's and e_{ij}'s are independent, and that, for the moment, $u_i \sim N(0, \sigma_u^2)$, $v_j \sim N(0, \sigma_v^2)$, $e_{ij} \sim N(0, \sigma_e^2)$. It is well-known (*e.g.*, Harville 1977, Miller 1977) that, in this case, the effective sample size for estimating σ_u^2 and σ_v^2 is not the total sample size $m_1 \cdot m_2$, but m_1 and m_2, respectively, for σ_u^2 and σ_v^2. Now suppose that one wishes to select the fixed covariates, which are components of x_{ij}, under the assumed model structure, using BIC. Then, it is not clear what should be in place of n in (1). In fact, in cases of correlated observations, such as the example above, the definition of "sample size" is often unclear.

Furthermore, suppose that normality is not assumed in the above linear mixed model. In fact, the only distributional assumptions are that the random effects and errors are independent, have zero mean and constant variances. Now, suppose that one, again, wishes to select the fixed covariates using AIC, BIC, or HQ. It is not clear how to do this because the likelihood is unknown.

Even in the conventional case, there are still practical issues regarding these criteria. For example, BIC is known to have the tendency of overly penalizing. In such a case, one may wish to replace the penalizer by $c\log(n)$, where c is a constant less than one. Question is: What c? Asymptotically, the choice of c does not make a difference in terms of consistency so long as $c > 0$. However, practically, it does. For example, comparing BIC with HQ, the penalizer of the latter is lighter in its order ($\log\{\log(n)\}$ vs $\log(n)$), but there is a constant c involved in HQ. If $n = 100$, we have $\log(n) = 4.6$ and $\log\{\log(n)\} = 1.5$, hence, if the constant c in HQ is chosen as 3, BIC and HQ are the same.

Finally, the definition of $|M|$ in (1) can also cause difficulties. In some circumstances like ordinary linear regression, this is simply the number of parameters in M, but in other situations where nonlinear, adaptive models are fitted, this can be substantially more (*e.g.*, Hastie and Tibshirani 1990, Friedman 1991, Ye 1998).

While there is extensive literature on parameter estimation in linear and generalized linear mixed models, the other component, that is, mixed model selection, has received much less attention. Only recently have some results emerge in the area of linear mixed model selection. Datta and Lahiri (2001) discussed a model selection method based on computation of the frequentist's Bayes factor in choosing between a fixed effects model and a random effects model. They focused on the following one-way balanced random effects model for the sake of simplicity: $y_{ij} = \mu + u_i + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, k$, where the u_i's and e_{ij}'s are normally distributed with mean zero and variances σ_u^2 and σ_e^2, respectively. As noted by the authors, the choice between a fixed effects model and a random effects one in this case is equivalent to testing the following one-sided hypothesis H_0: $\sigma_u^2 = 0$ vs H_1: $\sigma_u^2 > 0$. In fact, hypothesis testing may be regarded as a special case of model selection, but not all model selection problems can be formulated as hypothesis testing (see further discussion in subsection 8.1). Jiang and Rao (2003) developed various generalized information criteria (GICs) suitable for linear mixed model selection. Meza and Lahiri (2005) demonstrated the limitations of Mallows' C_p statistic in selecting the fixed covariates in a nested error regression model which is a special case of the linear mixed models. The nested error regression model is defined as $y_{ij} = x_{ij}^T\beta + u_i + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where y_{ij} is the observation, x_{ij} is a vector of fixed covariates, β is a vector of unknown regression coefficients, and u_i's and e_{ij}'s are the same as in the model above considered by Datta and Lahiri (2001). Simulation studies carried out by Meza and Lahiri (2005) showed that the C_p method without modification does not work well in the current mixed model setting when the variance σ_u^2 is large; on the other hand, a modified C_p criterion developed by these latter authors by adjusting the intra-cluster correlations performs similarly as the C_p in regression settings. Another related paper is that of Vaida and Blanchard (2005) who proposed a conditional AIC where the penalty term in this CAIC is related to the effective degrees of freedom for a linear mixed model proposed by Hodges and Sargent (2001) which reflects an intermediate level of model complexity between a full fixed effects model and a corresponding mixed model conditional on the random effects variances.

It should be pointed out that all these studies are limited to linear mixed models, while model selection in generalized linear mixed models (GLMMs) has never been seriously addressed in the literature in a general way (there are some fully Bayesian approaches for special cases like logistic mixed effects models - see Kinney and Dunson (2007) for example). In fact, our earlier simulation results suggested that in the case of GLMM selection, a procedure like GIC is much more sensitive to the choice of λ_n than in linear mixed model selection. It is these concerns, such as the above, that motivated the development of a new principle for model selection that is potentially less subjective, and applicable to both linear mixed models and GLMMs.

Jiang, Rao *et al.* (2008) proposed a new procedure for model selection, called the fence methods. An essential part of this procedure is a measure of lack-of-fit, denoted by $Q_M = Q_M(y, \theta_M)$, where M indicates the candidate model, y is an $n \times 1$ vector of observations, θ_M represents the vector of parameters under M, such that $\mathrm{E}(Q_M)$ is minimized when M is a true model and θ_M the true parameter vector under M. Here by true model we mean that M is a correct model but not necessarily the most efficient one. In the sequel we use the terms "true model" and "correct model" interchangeably. One example of Q_M is the negative log-likelihood function under a parametric model. Another example is the residual sum of squares (RSS) under a parametric or semiparametric model. For more examples, see Jiang, Rao *et al.* (2008).

The idea involves a procedure to isolate a subgroup of what are known as correct models (of which the optimal model is a member). This is accomplished by constructing a statistical *fence*, or barrier, to carefully eliminate incorrect models. Once the fence is constructed, the optimal model is selected from amongst those within the fence according to a criterion which can be made flexible and incorporate scientific or economical concerns. The fence is built by checking the following inequality for every candidate model M,

$$\hat{Q}_M - \hat{Q}_{\tilde{M}} \;\leq\; c_n \hat{\sigma}_{M,\tilde{M}}, \tag{16}$$

where $\hat{Q}_M = \inf_{\theta_M \in \Theta_M} Q_M(\theta_M, y)$, $\hat{Q}_{\tilde{M}} = \min_{M \in \mathcal{M}} \hat{Q}_M$, and $\mathcal{M}$ represents the set of candidate models. Here $\hat{\sigma}_{M,\tilde{M}}$ is an estimate of the standard deviation of the left side of (16). Finally, c_n is a tuning constant chosen below.

The motivation of (16) is to exam the closeness of $\hat{Q}_M$ to its lower bound - when the measure of lack-of-fit is close enough to the minimum the model is considered correct. The reason for the appearance of $\hat{\sigma}_{M,\tilde{M}}$ on the right side is that, when M is correct, this is an appropriate measure of the left side. Still, the constant c_n plays an important role for the finite sample performance of fence. Therefore, Jiang, Rao *et al.* (2008) proposed the following method to choose c_n adaptively.

1. Fence procedure with fixed c_n.

1. ind $\tilde{M}$ such that $\hat{Q}_{\tilde{M}} = \min_{M \in \mathcal{M}} \hat{Q}_M$.
2. For each $M \in \mathcal{M}$ such that $|M| < |\tilde{M}|$, compute $\hat{\sigma}_{M,\tilde{M}}$, an estimator of $\sigma_{M,\tilde{M}}$. Then, M belongs to $\tilde{\mathcal{M}}_-$, the set of "true" models with $|M| < |\tilde{M}|$ if (2) holds.
3. Let $\tilde{\mathcal{M}} = \{\tilde{M}\} \cup \tilde{\mathcal{M}}_-$, $m_0 = \min_{M \in \tilde{\mathcal{M}}} |M|$, and $\mathcal{M}_0 = \{M \in \tilde{\mathcal{M}} : |M| = m_0\}$. Let M_0 be the model in $\mathcal{M}_0$ such that $\hat{Q}_{M_0} = \min_{M \in \mathcal{M}_0} \hat{Q}_M$. M_0 is the selected model.

The following outlines an effective algorithm for fence. Let $d_1 < d_2 < \cdots < d_L$ be all the different dimensions of the models $M \in \mathcal{M}$.

The fence algorithm: i) Find $\tilde{M}$. ii) Compute $\hat{\sigma}_{M,\tilde{M}}$ for all $M \in \mathcal{M}$ such that $|M| = d_1$; let $\mathcal{M}_1 = \{M \in \mathcal{M} : |M| = d_1$ and (16) holds$\}$; if $\mathcal{M}_1 \neq \emptyset$, stop. Let M_0 be the model in $\mathcal{M}_1$ such that $\hat{Q}_{M_0} = \min_{M \in \mathcal{M}_1} \hat{Q}_M$; M_0 is the selected model. iii) If $\mathcal{M}_1 = \emptyset$, compute $\hat{\sigma}_{M,\tilde{M}}$ for all $M \in \mathcal{M}$ such that $|M| = d_2$; let $\mathcal{M}_2 = \{M \in \mathcal{M} : |M| = d_2$ and (16) holds$\}$; if $\mathcal{M}_2 \neq \emptyset$, stop. Let M_0 be the model in $\mathcal{M}_2$ such that $\hat{Q}_{M_0} = \min_{M \in \mathcal{M}_2} \hat{Q}_M$; M_0 is the selected model. iv) Continue until the program stops (it will at some point).

In short, the algorithm may be described as follows: Check the candidate models, from the simplest to the most complex; once one has discovered a model that falls within the fence and checked all the other models of the same simplicity (for membership within the fence), one stops. One apparent advantage of the fence algorithm is that one needs not search the entire space of candidate models in order to find the optimal model. Here the optimality is defined in terms of minimal dimension, *i.e.*, $|M|$. However, as mentioned, the criterion of optimality is flexible.

2. Forward-backward (F-B) fence procedure. The fence algorithm searches from the simplest models and therefore may not need to search the entire model space in order to determine the optimal model. On the other hand, such a procedure may still involve a lot of evaluations when the model space is large. To make the fence procedure computationally more attractive to large and complex models, the following variation of fence was proposed for situations of complex models with many predictors.

To be more specific, we let $\tilde{M}$ be the full model. The idea is to use a forward-backward procedure to generate a sequence of candidate models, among which the optimal model is selected using the fence method. We begin with a forward procedure. Let M_1 be the model that minimizes $\hat{Q}_M$ among all models with a single parameter; if M_1 is within the fence, stop the forward procedure; otherwise, let M_2 be the model that minimizes $\hat{Q}_M$ among all models that add one more parameter to M_1; if M_2 is within the fence, stop the forward procedure; and so on. The forward procedure stops when the first model is discovered within the fence. The procedure is then followed by a backward elimination. Let M_k be the final model of the forward procedure. If no submodel of M_k with one less parameter is within the fence, M_k will be our selection; otherwise, M_k is replaced by M_{k+1} which is a submodel of M_k with one less parameter and is within the fence, and so on. This approach is called the forward-backward (F-B) fence.

3. Adaptive fence procedure. Recall that $\mathcal{M}$ denotes the set of candidate models, which includes a true model. To be more specific, we assume that there is a full model $M_{\mathrm{f}} \in \mathcal{M}$, hence $\tilde{M} = M_{\mathrm{f}}$ in (16); and that every model in $\mathcal{M} \setminus \{M_{\mathrm{f}}\}$ is a submodel of a model in $\mathcal{M}$ with one less parameter than M_{f}. Let M_* denote a model with minimum dimension among $M \in \mathcal{M}$. First note that, ideally, one wishes to select c_n that maximizes the probability of choosing the optimal model. Here for simplicity the optimal model is defined as a true model that has the minimum dimension among all true models. This means that one wishes to choose c_n that maximizes

$$P \;=\; \mathrm{P}(M_0 = M_{\mathrm{opt}}), \tag{17}$$

where M_{opt} represents the optimal model, and $M_0 = M_0(c_n)$ is the model selected by the

fence procedure with the given c_n. However, two things are unknown in (17): (i) under what distribution should the probability P be computed; and (ii) what is $M_{\rm opt}$?

To solve problem (i), note that the assumptions above on $\mathcal{M}$ imply that $M_{\rm f}$ is a true model. Therefore, it is possible to bootstrap under $M_{\rm f}$. For example, one may estimate the parameters under $M_{\rm f}$, then use a model-based bootstrap to draw samples under $M_{\rm f}$. This allows us to approximate the probability P on the right side of (17).

To solve problem (ii), we use the idea of maximum likelihood. Namely, let $p^*(M) = {\rm P}^*(M_0 = M)$, where $M \in \mathcal{M}$ and $\rm P^*$ denotes the empirical probability obtained by bootstrapping. Let $p^* = \max_{M\in\mathcal{M}} p^*(M)$. Note that p^* depends on c_n. The idea is to choose c_n that maximizes p^*. It should be kept in mind that the maximization is not without restriction. To see this, note that if $c_n = 0$ then $p^* = 1$ (because when $c_n = 0$ the procedure always chooses $M_{\rm f}$). Similarly, $p^* = 1$ for very large c_n, if M_* is unique (because when c_n is large enough the procedure always chooses M_*). Therefore, what one looks for is "the peak in the middle" of the plot of p^* against c_n. This procedure is also studied in detail in Jiang *et al.* (2008).

Jiang, Rao *et al.* (2008) established consistency of fence, F-B fence and adaptive fence methods under mild regularity conditions. Here consistency is in the sense that with probability tending to one as the sample size increases the procedure will select the optimal model.

3.1. Fence method for high dimensions and subtractive measures of fit

Computation in high dimensions (p large typically), can be a challenge. If m is large, as is typically the case, this could result in a large number of $\hat{Q}(M)$'s that have to be evaluated. Jiang *et al.* (2011) introduced the idea of a subtractive measure in their work on fence methods for gene set analysis (what they called the invisible fence). Let $1, \ldots, m$ denote the candidate elements. A measure $\hat{Q}$ is said to be *subtractive* if it can be expressed as

$$\hat{Q}(M) \;=\; s - \sum_{i\in M} s_i, \tag{18}$$

where s_i, $i = 1, \ldots, m$ are some nonnegative quantities computed from the data, M is a subset of $1, \ldots, m$, and s is some quantity computed from the data that does not depend on M. Typically we have $s = \sum_{i=1}^m s_i$, but the definition does not impose such a restriction. Also the nonnegativity constraint on the s_i's is only to ensure that $\hat{Q}(M)$ behaves like a measure of lack-of-fit, that is, larger model has smaller $\hat{Q}(M)$.

For a subtractive measure, the models that minimize $\hat{Q}(M)$ at different dimensions are found almost immediately. Let $r_1, r_2, \ldots, r_m$ be the ranking of the candidate elements in terms of decreasing s_i. Then, the model that minimizes $\hat{Q}(M)$ at dimension one is r_1; the model that minimizes $\hat{Q}(M)$ at dimension two is $\{r_1, r_2\}$; the model that minimizes $\hat{Q}(M)$ at dimension three is $\{r_1, r_2, r_3\}$, and so on. This is what Jiang *et al.* (2011) called a *fast algorithm* for implementing the fence approach.

3.2. Other approaches to mixed model selection

Muller *et al.* (2013) wrote a survey paper on linear mixed model selection and discussed some other methods not discussed above. These include the marginal AIC (Vaida and Blanchard 2005), the bootstrap biased-correct mAIC of Shang and Cavanaugh (2008), Srivastava and Kubokawa (2010), conditional AIC (Vaida and Blanchard 2005), the modified Schwarz approach of Pauler (1998), minimum description length (MDL) approaches, shrinkage methods and Bayesian methods. Interested readers are directed to that survey paper for more details.

4. Mixed model selection and small area estimation

Small area estimation (SAE) has received increasing attention in recent literature. Here the term small area typically refers to a subpopulation or domain for which reliable statistics of interest cannot be produced 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.* In absence of adequate direct samples from the small areas, methods have been developed in order to "borrow strength". See Rao and Molina (2015) for a comprehensive account of various methods used in SAE. Statistical models, especially mixed effects models, have played important roles in SAE. See Jiang and Lahiri (2006) for an overview of mixed effects models in SAE.

While there is extensive literature on inference about small areas using mixed effects models, including estimation of small area means which is a problem of mixed model prediction, estimation of the mean squared error (MSE) of the empirical best linear unbiased predictor (EBLUP; see Rao 2003), and prediction intervals (*e.g.*, Chatterjee, Lahiri, and Li 2007), model selection in SAE has received much less attention. However, the importance of model selection in SAE has been noted by prominent researchers in this field (*e.g.*, Battese, Harter, and Fuller 1988, Ghosh and Rao 1994). Datta and Lahiri (2001) discussed a model selection method based on computation of the frequentist's Bayes factor in choosing between a fixed effects model and a random effects model. They focused on the following one-way balanced random effects model for the sake of simplicity: $y_{ij} = \mu + u_i + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, k$, where the u_i's and e_{ij}'s are normally distributed with mean zero and variances σ_u^2 and σ_e^2, respectively. As noted by the authors, the choice between a fixed effects model and a random effects one in this case is equivalent to testing the following one-sided hypothesis H_0: $\sigma_u^2 = 0$ vs H_1: $\sigma_u^2 > 0$. Note that, however, not all model selection problems can be formulated as hypothesis testing. Fabrizi and Lahiri (2004) developed a robust model selection method in the context of complex surveys. Meza and Lahiri (2005) demonstrated the limitations of Mallows' C_p statistic in selecting the fixed covariates in a nested error regression model (Battese, Harter, and Fuller 1988), defined as $y_{ij} = x_{ij}^T\beta + u_i + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where y_{ij} is the observation, x_{ij} is a vector of fixed covariates, β is a vector of unknown regression coefficients, and u_i's and e_{ij}'s are the same as in the model above considered by Datta and Lahiri (2001). Simulation studies carried out by Meza and Lahiri (2005) showed that the C_p method without modification does not work well in the current mixed model setting when the variance σ_u^2 is large; on the other hand, a modified C_p criterion developed by these latter authors by adjusting the intra-cluster correlations performs similarly as the C_p in regression settings. It should be pointed out that all these studies are

limited to linear mixed models, while model selection in SAE in a generalized linear mixed model (GLMM) setting has never been seriously addressed.

4.1. Fence methods for SAE model selection

One of the advantages of fence methods is that the criterion of optimality for selecting the models within the fence is flexible. In SAE the problem of main interest is the estimation, or prediction, of the small area means. For simplicity, consider the case of linear mixed models. Then, the small area mean is typically estimated by the best linear unbiased predictor, or BLUP. Because an important measure of the accuracy of BLUP is its MSE, it makes sense to take the latter into account. Therefore, we consider the following criterion for selecting models within the fence when linear mixed models are under consideration. Suppose that one is interested in a small-area specific mixed effect (*e.g.*, the small area mean), θ_i, which is a linear combination of fixed and random effects. Let $\tilde{\theta}_i$ be the BLUP of θ_i. Let $\theta = (\theta_i)_{1\le i\le m}$ and $\tilde{\theta} = (\tilde{\theta}_i)_{1\le i\le m}$. Then, $\mathrm{MSE}(\tilde{\theta}) = \mathrm{E}(|\tilde{\theta} - \theta|^2) = \sum_{i=1}^m \mathrm{E}(\tilde{\theta}_i - \theta_i)^2 = \sum_{i=1}^m \mathrm{MSE}(\tilde{\theta}_i)$. Furthermore, an explicit expression of $\mathrm{MSE}(\tilde{\theta}_i)$ can be obtained (*e.g.*, Rao 2003, pp. 137). Note that $\mathrm{MSE}(\tilde{\theta})$ typically depends on some unknown variance components. Let $\widehat{\mathrm{MSE}}(\tilde{\theta})$ be an estimator of $\mathrm{MSE}(\tilde{\theta})$, say, by replacing the variance components by their REML estimators (*e.g.*, Jiang 2007). A model within fence is selected if (i) it has the minimum dimension; and (ii) if there are more than one models chosen by (i), select the one that has the minimal $\widehat{\mathrm{MSE}}(\tilde{\theta})$.

An interesting example is that from Jiang *et al.* (2010) who considered model selection for non-parametric SAE models. Opsomer *et al.* (2008) proposed a spline-based nonparametric model for SAE. The idea is to approximate an unknown nonparametric small-area mean function by a penalized spline (P-spline). The authors then used a connection between P-splines and linear mixed models (Ruppert, Wand, and Carroll 2003) to formulate the approximating model as a linear mixed model, where the coefficients of the splines are treated as random effects. Consider, for simplicity, the case of univariate covariate. Then, a P-spline can be expressed

$$\tilde{f}(x) \;=\; \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \gamma_1 (x - \kappa_1)_+^p + \cdots + \gamma_q (x - \kappa_q)_+^p, \tag{19}$$

where p is the degree of the spline, q is the number of knots, κ_j, $1 \le j \le q$ are the knots, and $x_+ = x1_{(x>0)}$. Clearly, a P-spline is characterized by p, q, and also the location of the knots.

Jiang *et al.* (2010) developed a simplified version of the adaptive fence in order to choose p and q. First, since the optimal model is rarely either M_{f} or M_*, the minimal model (dimensionwise; *e.g.*, a model with only the intercept). Baseline adjustment and threshold checking are used to deal with these two cases (see Jiang *et al.* 2008). The baseline adjustment is done by generating an additional vector of covariates, say, X_{a}, so that it is unrelated to the data. Then, define the model M_{f}^* as M_{f} plus X_{a}, and replace M_{f} by M_{f}^*, but let $\mathcal{M}$ remain unchanged. This way one knows for sure that the new full model, M_{f}^*, is not an optimal model (because it is not a candidate model). The threshold checking inequality is given by $\hat{Q}_{M_*} - \hat{Q}_{M_{\mathrm{f}}^*} > d_*$, where d_* is the maximum of the left side of the threshold inequality computed under the bootstrap samples generated under M_*. In case the threshold inequality holds, we ignore the right tail of the plot of p^* against c_n that

eventually goes up and stays at one.

Jiang *et al.* (2010) also constructed a a (large sample) confidence lower bound, for example,

$$p^* - 1.96\sqrt{p^*(1-p^*)/B} \tag{20}$$

where B is the bootstrap sample size. When selecting c_n that maximize p^* we take (20) into account. More specifically, suppose that there are two peaks in the plot of p^* against c_n located at $c_{n,1}$ and $c_{n,2}$ such that $c_{n,1} < c_{n,2}$. Let p_1^* and p_2^* be the heights of the peaks corresponding to $c_{n,1}$ and $c_{n,2}$. As long as p_1^* is greater than the confidence lower bound at p_2^*, that is, (4) with $p^* = p_2^*$, we choose $c_{n,1}$ over $c_{n,2}$. Clearly, the selection is in favor of smaller c_n in order to be more conservative. (In other words, we are more concerned with underfit than overfit.)

Consistency of selection under mild regularity conditions was then proven in the following Theorem:

Theorem 4. (Jiang, Nguyen, and Rao 2010). Let M_0^* denote the model selected by the fence procedure with $c_n = c_n^*$. Also, let M_{opt} denote an optimal model defined as a true model with minimum dimension and minimum MSE($\tilde{\theta}$) among all the true models within the (same) minimum dimension. Under the regularity conditions given therein, there is c_n^* which is at least a local maximum and approximate global maximum of p^*, and the corresponding M_0^* is consistent in the sense that any $\delta, \eta > 0$, there are N, N^* such that

$$\mathrm{P}\{p^*(c_n^*) \geq 1-\delta\} \wedge \mathrm{P}(M_0^* = M_{\text{opt}}) \;\; \geq \;\; 1-\eta,$$

if $m \geq N$ and $B \geq N^*$.

4.2. Variable selection for area and subarea level SAE models

In this section, we focus on variable selection under area level models and subarea level SAE models which are extensively used in practice. A basic area level model, also called the Fay-Herriot model (FH; Fay and Herriot 1979), uses direct estimators $\hat{\theta}_i$ of area means $\theta_i (i = 1, \ldots, m)$ and associated area level covariates. Direct estimators are obtained from area-specific unit level data, taking survey design into account. Area level covariates are used to link the area means. This leads to a sampling model and a linking model given by $\hat{\theta}_i = \theta_i + e_i$ and $\theta_i = x_i^T \beta + v_i$ respectively, where e_i is the sampling error, β is the vector of model parameter, x_i is the $p \times 1$ vector of area level covariates and v_i is a random area effect. Further, e_i has mean 0 and known variance ψ_i, and the sampling errors are assumed to be independent. In practice, the sampling variances ψ_i are obtained by smoothing their direct estimators using generalized variance functions. The area effect v_i has mean 0 and variance σ_v^2, and the area effects are assumed to be independent. Combining the sampling model with the linking model leads to the FH model $\hat{\theta}_i = x_i^T \beta + v_i + e_i$ which is then used for variable selection. Note that the linking model alone cannot be used for variable selection because the area means θ_i are not known.

Because of the sampling errors in the FH model, standard methods for linear regression models, such as the AIC, BIC and Mallows' C_p used for variable selection, can lead to

biased variable selection when applied to the FH model. Han (2013) used a conditional AIC method for variable selection that accounts for the sampling errors in the FH model. This method is fairly complex, and practitioners might prefer simple modifications to standard methods that can account for sampling errors in the FH model. We give a brief description of a simple method of estimating the ideal variable selection criteria under the linking model that accounts for the sampling error (Lahiri and Suntornchost 2015). The resulting estimation error is shown to converge to 0 in probability as the number of areas increases, unlike the estimation error in the naïve criteria ignoring the sampling errors. The proposed method performed well in simulations unlike the naïve method that ignores the sampling errors in the FH model.

Let, $MSE_\theta = \frac{1}{m-p}\theta^T(I_m - P)\theta$ denote the ideal mean error sum of squares, where $\theta = (\theta_1, \ldots, \theta_m)^T$, I_m is the identity matrix of order m, and $P = X(X^TX)^{-1}X$ is the standard projection matrix based on the linking model. Then the estimator of MSE_θ is given by $mse_\theta = MSE_{\hat{\theta}} - \bar{\psi}_w$, where $MSE_{\hat{\theta}}$ is obtained by replacing θ by its direct estimator $\hat{\theta}$, and $\bar{\psi}_w = \frac{1}{m-p}\sum_{i=1}^m (1-h_{ii})\psi_i$ with $h_{ii} = x_i^T(X^TX)^{-1}x_i$. We simply replace MSE_θ by mse_θ in the ideal AIC, BIC and C_p which are functions of MSE_θ. For example, the resulting $AIC = mlog\{\frac{m-p}{m}mse_\theta\} + 2p$. In the case of small m, the estimator mse_θ could take a negative value and Lahiri and Suntornchost (2015) suggested a simple modification that leads to strictly positive estimator of MSE_θ.

Two-fold subarea models are also often used in practice to estimate subarea and area means. For example, Mohadjer et al. (2012) studied adult literacy for counties (subareas) sampled from states (areas) in the United States, using data from the 2003 U. S. National Assessment of Adult Literacy (NAAL). We have areas i and subareas j are sampled from area i. Direct estimators of subarea means $\theta_{ij}(j = 1, \ldots, n_i; i = 1, \ldots, m)$ and associated subarea level covariate vector are denoted as $\hat{\theta}_{ij}$ and x_{ij} respectively. A two-fold subarea model consists of a sampling model $\hat{\theta}_{ij} = \theta_{ij} + e_{ij}$ and a linking model $\theta_{ij} = x_{ij}^T\beta + b_{ij}$ respectively, where e_{ij} is the sampling error and $b_{ij} = v_i + u_{ij}$ is the sum of the random area effect v_i and subarea effect u_{ij}. The sampling errors e_{ij} are assumed to be independent with zero means and known variances. Further, the area effect is independent of the subarea effect, and the v_i and u_{ij} are independent and identically distributed with zero means and variances σ_v^2 and σ_u^2 respectively. Under the assumptions, the composite random effects b_{ij} are correlated for each area i with covariance matrix $\Sigma_i = \sigma_v^2 1_i 1_i^T + \sigma_u^2 I_i$ where 1_i is the unit vector of length n_i and I_i is the identity matrix of order n_i.

We cannot treat the linking model for the two-fold case as a FH-type model on the subarea means because the composite random effects b_{ij} are correlated. It is necessary to transform the covariance matrix Σ_i to a diagonal covariance matrix with equal diagonal elements across areas i, and then apply the variable selection method to the transformed linking model to get the ideal error mean sum of squares. Cai *et al.* (2020) obtained a parameter-free transformation matrix A_i of order $(n_i - 1) \times n_i$ and full rank that makes the covariance matrix of $A_i b_i$ diagonal with equal diagonal elements across $i = 1, \ldots, m$, where $b_i = (b_{i1}, \ldots, b_{im})^T$ (Li and Lahiri (2019) used a similar transformation in the context of unit level models). The transformed linking model is then used to get the ideal mean square error sum of squares MSE_{θ^*} and its estimator mse_{θ^*} along the lines of the method used for the FH linking model. Note that the transformed vector $\theta_i^* = A_i\theta_i$ has length $n_i - 1$ unlike the vector

θ_i with elements $\theta_{ij}, j = 1, \ldots, n_i$, and as a result each area loses one degree of freedom after transformation. The variable selection criteria can then be computed using mse^*_θ, as in the case of the FH model. Cai *et al.* (2020) report simulation results showing that the proposed transformation method performs well in variable selection, unlike the naive method treating the linking model as a FH-type model ignoring the correlations, especially as σ_v^2 increases.

Three-fold models linking sub-subarea means to related covariates and random effects at the area, subarea and sub-subarea levels are also used in practice to estimate sub-subarea means as well as subarea means. For example, the Program for the International Assessment of Adult Competencies (PIAAC) in the United States used a three-fold model with census divisions as areas, states within a census division as subareas and counties within a state as sub-subareas. Cai and Rao (2022) extended the two-fold model variable selection method of Cai *et al.* (2020) to variable selection to variable selection under three-fold models.

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.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

Bai, Z., Rao, C. R., and Wu, Y. (1999). Model selection with data-oriented penalty. *Journal of Statistical Planning and Inference*, **77**, 103-117.

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

Bozdogan, H. (1994). Editor's general preface, in *Engineering and Scientific Applications*, Vol. 3 (H. Bozdogan ed.), *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, pages ix-xii. Kluwer Academic Publishers, Dordrecht, Netherlands.

Cai, S. and Rao, J. N. K. (2022). Selection of auxiliary variables for three-fold linking models in small area estimation: a simple and effective method. *Stats*, **5**, 128-138.

Cai, S., Rao, J. N. K., Dumitrescu. L., and Chatrchi, G. (2020). Effective transformation -based variable selection under two-fold subarea models in small area estimation. *Statistics in Transition*, **21**, 68 – 83.

Chatterjee, S., Lahiri, P., and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models. *Annals of Statistics*, **36**, 1221-1245.

Datta, G. S. and Lahiri, P. (2001). Discussions on a paper by Efron and Gous. *Model Selection*, IMS Lecture Notes/Monograph 38.

Fabrizi, E. and Lahiri, P. (2004). A new approximation to the Bayes information criterion in finite population sampling. Tech. Report, Department of Mathematics, University of Maryland.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269-277.

Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, **19**, 1-141.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, **9**, 55-93.

Han, B. (2013). Conditional Akaike information criterion in the Fay-Herriot model. *Statistical Methodology*, **11**, 53-67.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society B*, **41**, 190-195.

Harville, D. A. (1977). Maximum likelihood approaches to variance components estimation and related problems. *Journal of the American Statistical Association*, **72**, 320-340.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, New York.

Hodges, J. S. and Sargent, D. J. (2001), Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, **88**, 367-379.

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

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussion). *Test*, **15**, 1-96.

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

Jiang, J., Rao, J. S., Gu, Z., and Nguyen, T. (2008). Fence methods for mixed model selection. *Annals of Statistics*, **36**, 1669-1692.

Jiang, J., Nguyen, T., and Rao, J. S. (2010). A simplified adaptive fence procedure. *Statistics and Probability Letters*, **79**, 625-629.

Jiang, J., Rao, J. S., and Nguyen, T. (2011). Invisible fence methods and the identification of differentially expressed gene sets. *Statistics and Its Interface*, **4**, 403-415.

Kinney, S. K. and Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, **63**, 690-698.

Kubokawa, T. (2011). Conditional and unconditional methods for selecting variables in linear mixed models. *Journal of Multivariate Analysis*, **102**, 641–660.

Lahiri, P. and Suntornchost, J. (2015). Variable selection for linear mixed models with application to small area estimation. *Sankhya B*, **77**, 312-320.

Li, Y. and Lahiri, P. (2019). A simple adaptation of variable selection software for regression models to select variables in nested error regression models. *Sankhya B*, **81**, 302 – 317.

Meza, J. and Lahiri, P. (2005). A note on the Cp statistic under the nested error regression model. *Survey Methodology*, **31**, 105-109.

Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of analysis of variance. *Annals of Statistics*, **5**, 746-762.

Mohadjer, L., Rao, J. N. K., Liu, B., Krenzke, T., and Van De Kerckhove, W. (2012). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Journal of the Indian Society of Agricultural Statistics*, **66**, 55-63.

Muller, S., Scealy, J. S., and Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, **28**, 135-167.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, **12**, 758-765.

Opsomer, J. D., Breidt, F. J., Claeskens, G., Kauermann, G., and Ranalli, M. G. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society B*, **70**, 265-286.

Pauler, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika*, **85**, 13–27.

Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika*, **76**, 369-374.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley, New York.

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

Ruppert, R., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. John Wiley and Sons, New York.

Shang, J. and Cavanaugh, J. E. (2008). Bootstrap variants of the Akaike information criterion for mixed model selection. *Computational Statistics and Data Analysis*, **52**, 2004–2021.

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika*, **71**, 43-49.

Srivastava, M. S. and Kubokawa, T. (2010). Conditional information criteria for selecting variables in linear mixed models. *Journal of Multivariate Analysis*, **101**, 1970–1980.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed effects models. *Biometrika*, **92**, 351-370.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120-131.

Zheng, X., and Loh, W. Y. (1995). Consistent variable selection in linear models. *Journal of the American Statistical Association*, **90**, 151-156.