Sankhyā : The Indian Journal of Statistics
2003, Volume 65, Part 1, pp 23-42

CONSISTENT PROCEDURES FOR MIXED LINEAR MODEL SELECTION

By JIMING JIANG
University of California, Davis, USA
and
J. SUNIL RAO
Case Western Reserve University, Cleveland, USA

SUMMARY. We consider the problem of selecting the fixed and random effects in a mixed linear model. Two kinds of selection problems are considered. The first is to select the fixed covariates from a set of candidate predictors when the random effects are not subject to selection; the second is to select both the fixed covariates and the random effect factors. Our selection criteria are similar to the generalized information criterion (GIC), but we show that a naive GIC does not work for the second kind of selection problem. Asymptotic theory is developed in which we give sufficient conditions for consistency of the selection criteria proposed. Finite sample performance of the selection procedures are investigated by simulation studies.

1. Introduction

Model selection and estimation are two components of a process called model identification. The former determines the form of the model, leaving only some undetermined coefficients or parameters. The latter finds estimators of the unknown parameters. A pioneering work on model selection criteria is Akaike's information criterion (AIC, Akaike (1972)). One of the earlier applications of AIC and other procedures such as the Bayesian information criterion (BIC) is the determination of the orders of an autoregressive moving-average (ARMA) time series model (e.g., Choi (1992)). Similar methods have also been applied to regression model selection (e.g., Rao and Wu (1989), Bickel and Zhang (1992), Shao (1993), and Zheng and Loh (1995)). Most of these model selection procedures are (asymptotically) equivalent to the generalized information criterion (GIC, e.g., Nishii (1984), Shibata (1984)).

Paper received November 2001; revised March 2002.
AMS (2000) *subject classification.* 62J05.
Keywords and phrases. Consistency, mixed effects models, model selection.

The purpose of this paper is to develop model selection procedures for mixed linear models. These models have found broad applications in various fields. There is extensive literature on parameter estimation in mixed linear models (e.g., Searle, Casella, and McCulloch (1992)), so one component of the model identification has been well-studied. However, so far the other component has been neglected, that is, the mixed model selection problem. This paper attempts to fill in an important gap in this field.

We consider the following mixed linear model:

$$y = X\beta + Z\alpha + \epsilon\,, \tag{1.1}$$

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.

In Section 2 we first consider the model selection problem when the random effect factors are not subject to selection, i.e., selecting the fixed covariates with fixed random factors. In this case, a procedure similar to GIC is proposed, and consistency of the selection criterion is established. In Section 3 we consider the more general problem when both the fixed covariates and the random effect factors are subject to selection. Due to special properties of the mixed linear models, a naive GIC will not work in this case (see reason given therein). Therefore, we propose a method which divides the random effect factors into several groups, and applies different model selection procedures for different groups simultaneously. Again, sufficient conditions are given, under which the combined procedure is consistent. In Section 4 we consider two simulated examples, which correspond to problems discussed in the previous two sections, respectively. As will be seen, our simulation results show strong agreement with the theory established. Some discussion and remarks are given in Section 5. All the technical proofs are given in the Appendix.

2. Selection with Fixed Random Factors

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:

$$y = X\beta + \zeta\,, \tag{2.1}$$

where ζ is a vector of correlated errors, and everything else is as in (1.1). 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 $\mathcal{L}(M)$: $P_M = M(M'M)^- M'$; 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:

$$C_N(k) = |y - X(k)\hat{\beta}(k)|^2 + \lambda_N |k| = |P_{X(k)}^\perp y|^2 + \lambda_N |k| \,, \tag{2.2}$$

$k \in \mathcal{A}$, 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.,

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

(see the second paragraph in Section 5 for discussion on the choice of estimation methods); 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: $\omega_N = \min_{1 \le j \le q} |P_{X(\{j\}^c)}^\perp X_j|^2$, $\nu_N = \max_{1 \le j \le q} |X_j|^2$, and $\rho_N = \lambda_{\max}(ZGZ') + \lambda_{\max}(R)$, where $\lambda_{\max}$ means largest eigenvalue. Let $\hat{k}$ be the minimizer of (2.2) 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

$$P(\hat{k} \neq k_0) \;\longrightarrow\; 0 \,. \tag{2.3}$$

THEOREM 1. *Suppose that $\nu_N > 0$ for large N,*

$$\rho_N/\nu_N \;\longrightarrow\; 0 \,, \quad \text{while} \quad \liminf(\omega_N/\nu_N) \;>\; 0 \,. \tag{2.4}$$

Then, (2.3) holds for any λ_N such that

$$\lambda_N/\nu_N \;\longrightarrow\; 0 \quad \text{and} \quad \rho_N/\lambda_N \;\longrightarrow\; 0 \,. \tag{2.5}$$

NOTE 1. If (2.4) holds, then there always exists λ_N that satisfies (2.5). For example, take $\lambda_N = \sqrt{\rho_N \nu_N}$.

NOTE 2. Typically, we have $\nu_N \sim N$. To see what the order of ρ_N may turn out to be, consider a special but important case in mixed linear models: Suppose that $Z = (Z_1, \ldots, Z_s)$, where each Z_r is a standard design matrix in the sense that it consists only of 0's and 1's, there is exactly one 1 in each row, and at least one 1 in each column. Likewise, $\alpha = (\alpha_1', \ldots, \alpha_s')'$ such that $Z\alpha = Z_1\alpha_1 + \cdots + Z_s\alpha_s$, where α_r is a m_r-dimensional vector of uncorrelated random effects with mean 0 and variance σ_r^2. Furthermore, ϵ is a vector of uncorrelated errors with mean 0 and variance σ_0^2. Finally, $\alpha_1, \ldots, \alpha_s, \epsilon$ are uncorrelated. Let n_{rk} be the number of 1's in the kth column of Z_r. Note that n_{rk} is the number of appearance of the kth component of α_r, and $Z_r'Z_r = \text{diag}(n_{rk}, 1 \le k \le m_r)$. Thus, $\lambda_{\max}(ZGZ') \le \sum_{r=1}^s \sigma_r^2 \lambda_{\max}(Z_r Z_r') = \sum_{r=1}^s \sigma_r^2 \max_{1\le k\le m_r} n_{rk}$. Also, we have $\lambda_{\max}(R) = \sigma_0^2$. It follows that $\rho_N = O\left(\max_{1\le r\le s} \max_{1\le k\le m_r} n_{rk}\right)$. Therefore, (2.5) is satisfied provided that $\lambda_N/N \to 0$ and $\max_{1\le r\le s} \max_{1\le k\le m_r} n_{rk}/\lambda_N \to 0$.

EXAMPLE 1. Consider the following simple mixed linear model:

$$y_{ij} \quad = \quad \beta_0 + \beta_1 x_{ij} + \alpha_i + \epsilon_{ij} \ , \tag{2.6}$$

$i = 1, \ldots, m$, $j = 1, \ldots, n$, where β_0, β_1 are unknown coefficients (the fixed effects). It is assumed that the random effects $\alpha_1, \ldots, \alpha_m$ are uncorrelated with mean 0 and variance σ^2. Furthermore, assume that the errors ϵ_{ij}'s have the following exchangeable correlation structure: Let $\epsilon_i = (\epsilon_{ij})_{1\le j\le n}$. Then, $\text{Cov}(\epsilon_i, \epsilon_{i'}) = 0$ if $i \ne i'$, and $\text{Var}(\epsilon_i) = \tau^2\{(1-\rho)I + \rho J\}$, where I is the identity matrix and J matrix of 1's. Finally, assume that the random effects are uncorrelated with the errors. Suppose that $m \to \infty$, and

$$0 < \liminf \left[\frac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x}_{\cdot\cdot})^2\right] \quad \le \quad \limsup \left[\frac{1}{mn}\sum_{i=1}^m \sum_{j=1}^n x_{ij}^2\right] < \infty, \tag{2.7}$$

where $\bar{x}_{\cdot\cdot} = (mn)^{-1}\sum_{i=1}^m \sum_{j=1}^n x_{ij}$. Then, it is easy to show that all the conditions of Theorem 1 are satisfied. In fact, in this case, $\rho_N \sim n$, while $\nu_N \sim \omega_N \sim mn$.

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 (2.1) as

$$X\beta \;=\; \sum_{j=1}^{q} \beta_j X_j \tag{2.8}$$

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

$$(|P^{\perp}_{X_{-j}} y|^2 - |P^{\perp}_{X} y|^2)/(|P^{\perp}_{X_{-j}} X_j|^2 \delta_N) \;>\; 1\,. \tag{2.9}$$

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. *Suppose that $\eta_N > 0$ for large N, and*

$$\rho_N/\eta_N \;\longrightarrow\; 0\,. \tag{2.10}$$

Then, (2.3) holds for any δ_N such that

$$\delta_N \;\longrightarrow\; 0 \quad \textit{and} \quad \rho_N/(\eta_N\delta_N) \;\longrightarrow\; 0\,. \tag{2.11}$$

EXAMPLE 1 (CONTINUED). It is easy to show that, under exactly the same conditions (i.e., $m \to \infty$ and (2.7) holds), $\eta_N \sim mn$. Recall that $\rho_N \sim n$. Thus, all the conditions of Theorem 2 are satisfied.

3. Selection with Random Factors

In this section, we assume that $Z\alpha$ in (1.1) can be expressed as

$$Z\alpha \;=\; \sum_{j=1}^{s} Z_j \alpha_j\,, \tag{3.1}$$

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 (1.1) 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 (3.1) 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

$$y = X\beta + \sum_{j \in l_0} Z_j \alpha_j + \epsilon \,, \tag{3.2}$$

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 $\sigma_j^2 = 0$, $j \in L \setminus l_0$.

Before we go on, we would like to point out some differences between selecting the fixed covariates X_j, as we did in the previous section, and selecting the random effect 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_j's are linearly independent, in a mixed linear model it is possible to have $j \neq j'$ but $\mathcal{L}(Z_j) \subset \mathcal{L}(Z_{j'})$. For example, see Example 2 below. Because of these features, the selection of random factors cannot be handled the same way as in Section 2.

It should be noted that a hypothesis testing approach to selecting random factors could be used but will no doubt rely on normality assumptions about the distributions of the random factors. We have not made any such assumptions here. In addition, such testing-based approaches are known to not be consistent. Thus, we are going to take the following approach.

3.1 *The basic idea.* First, note that in Section 2 we have a procedure to determine the fixed part of the model, which leads to a selection $\hat{k}$ that satisfies (2.3). 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 $\lambda_{\max}(ZGZ')$. However, under (3.1), $\lambda_{\max}(ZGZ') \leq \sum_{j=1}^s \sigma_j^2 \|Z_j\|^2$, where for any matrix M, $\|M\| = [\lambda_{\max}(M'M)]^{1/2}$. Thus, an upper bound for the order of $\lambda_{\max}(ZGZ')$ 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, whose proof is obvious, states that the combined procedure, which determines both the fixed and random parts of the model, is consistent.

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

We now describe how to obtain $\hat{l}$. First divide 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 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.

As pointed out earlier (see the discussion below (3.2)), the selection of random factors cannot be treated the same way as that of fixed factors, because the design matrices $Z_1, \ldots, Z_s$ are usually linearly dependent. Intuitively, a selection procedure will not work if there is linear dependence among the candidate design matrices, because of identifiability problems. (To consider a rather extreme example, suppose that Z_1 is a design matrix consist of 0's and 1's such that there is exactly one 1 in each row, and $Z_2 = 2Z_1$. Then, to have $Z_1\alpha_1$ in the model means that there is a term α_{1i}; while to have $Z_2\alpha_2 = 2Z_1\alpha_2$ in the model means that there is a corresponding term $2\alpha_{2i}$. However, it makes no difference in terms of a model, because both α_{1i} and α_{2i} are random effects with mean 0 and certain variances.) However, by grouping the random effect factors we have divided the Z_j's into several groups such that there is linear independence within each group. This is the motivation behind grouping. To illustrate such a procedure, and also to show that such a division of groups does exist in typical situations, consider the following example.

EXAMPLE 2. Consider the following random effects model:

$$y_{ijkl} = \mu + a_i + b_j + c_k + d_{ij} + f_{ik} + g_{jk} + h_{ijk} + e_{ijkl}\ , \qquad (3.3)$$

$i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$, $k = 1, \ldots, m_3$, $l = 1, \ldots, n$, where μ is an unknown mean; a, b, c are random main effects; d, f, g, h are (random) two-

and three-way interactions; and e is error. The model can be written as

$$y = X\mu + Z_1a + Z_2b + Z_3c + Z_4d + Z_5f + Z_6g + Z_7h + e\,,$$

where $X = 1_N$ with $N = m_1m_2m_3n$, $Z_1 = I_{m_1} \otimes 1_{m_2} \otimes 1_{m_3} \otimes 1_n, \ldots, Z_4 = I_{m_1} \otimes I_{m_2} \otimes 1_{m_3} \otimes 1_n, \ldots$, and $Z_7 = I_{m_1} \otimes I_{m_2} \otimes I_{m_3} \otimes 1_n$. Here I_r and 1_r represent the r-dimensional identity matrix and vector of 1's, and $\otimes$ means Kronecker product. It is easy to see that the Z_j's are not linearly independent. For example, $\mathcal{L}(Z_j) \subset \mathcal{L}(Z_4)$, $j = 1, 2$, and $\mathcal{L}(Z_j) \subset \mathcal{L}(Z_7)$, $j = 1, \ldots, 6$. Also, $\mathcal{L}(X) \subset \mathcal{L}(Z_j)$ for any j. Suppose that $m_j \to \infty$, $j = 1, 2, 3$, while n is bounded. Then, the first group consists of Z_7; the second group Z_4, Z_5, Z_6; and the third group Z_1, Z_2, Z_3. If n also $\to \infty$, the largest random factor does not exist. However, one still has these three groups. It is easy to see that the Z_j's within each group are linearly independent.

Suppose that the Z_j's are divided into h groups such that $L = L_1 \cup \cdots \cup L_h$. We give 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.

3.2 *Group one.* First, let us consider the first group. Write $B = \mathcal{L}(X, Z_1, \ldots, Z_s)$, $B_{-j} = \mathcal{L}(X, Z_u,\ u \in L \setminus \{j\})$, $j \in L_1$; $r = N - \text{rank}(B)$, $r_j = \text{rank}(B) - \text{rank}(B_{-j})$; $R = |P_B^\perp y|^2$, $R_j = |(P_B - P_{B_{-j}})y|^2$. If M is a matrix, we define $\|M\|_2 = [\text{tr}(M'M)]^{1/2}$. For any $1 < \rho < 2$, let $\hat{l}_1$ be the set of indexes in L_1 such that

$$(r/R)(R_j/r_j) \;>\; 1 + r^{(\rho/2)-1} + r_j^{(\rho/2)-1}\,. \tag{3.4}$$

Let $l_{01} = \{j \in L_1 : \sigma_j^2 > 0\}$.

LEMMA 1. *Suppose that $r \to \infty$; and $r_j \to \infty$, $\liminf(\|P_{B_{-j}}^\perp Z_j\|_2^2/r_j) > 0$, $\|P_{B_{-j}}^\perp Z_j\|_2^2/r_j^2 \to 0$, and $\|Z_j'P_{B_{-j}}^\perp Z_j\|_2^2/r_j^2 \to 0$, $j \in L_1$. Then, $P(\hat{l}_1 = l_{01}) \to 1$.*

EXAMPLE 2 (CONTINUED). Suppose that $m_t \to \infty$, $t = 1, 2, 3$, while n is bounded but $n \geq 2$. Then, group one corresponds to a single index $j = 7$. Furthermore, it is easy to see that $r = m_1m_2m_3(n-1)$, $r_7 \geq d = m_1m_2m_3 - m_1m_2 - m_2m_3 - m_3m_1$, $nd \leq \|P_{B_{-7}}^\perp Z_7\|_2^2 \leq N = m_1m_2m_3n$, and $\|Z_7'P_{B_{-7}}^\perp Z_7\|_2^2 \leq nN$. It follows that all the conditions of Lemma 1 are satisfied.

3.3 *Group two.* We now consider the second group. Let $B_1 = \mathcal{L}(X, Z_u, u \in L \setminus L_1)$, $B_2 = \mathcal{L}(X, Z_u, u \in L \setminus L_1 \setminus L_2)$, $B_{1,-j} = \mathcal{L}(X, Z_u, u \in L \setminus L_1 \setminus \{j\})$, $j \in L_2$, and $B_1(l_2) = \mathcal{L}(X, Z_u, u \in (L \setminus L_1 \setminus L_2) \cup l_2)$, $l_2 \subseteq L_2$. Consider

$$C_{1,N}(l_2) \;=\; |P^{\perp}_{B_1(l_2)} y|^2 + \lambda_{1,N}|l_2| \,, \tag{3.5}$$

$l_2 \subseteq L_2$, where $\lambda_{1,N}$ is a positive number satisfying conditions specified below. Let $\hat{l}_2$ be the minimizer of $C_{1,N}$ over $l_2 \subseteq L_2$, and $l_{02} = \{j \in L_2 : \sigma_j^2 > 0\}$. Let $L_0 = \{0\}$, and $Z_0 = I$, the identity matrix; $\rho_{1,N} = \max_{j \in L_0 \cup L_1} \|P_{B_1} Z_j\|_2^2$, $\nu_{1,N} = \min_{j \in L_2} \|P^{\perp}_{B_{1,-j}} Z_j\|_2^2$, and $\gamma_{1,N} = \max_{j \in L_2} (\|P^{\perp}_{B_2} Z_j\|^2 / \|P^{\perp}_{B_{1,-j}} Z_j\|_2^2)$.

LEMMA 2. *Suppose that $\nu_{1,N} > 0$ for large N,*

$$\rho_{1,N}/\nu_{1,N} \;\longrightarrow\; 0 \,, \quad \text{and} \quad \gamma_{1,N} \;\longrightarrow\; 0 \,. \tag{3.6}$$

Then, $P(\hat{l}_2 \neq l_{02}) \to 0$ for any $\lambda_{1,N}$ such that

$$\lambda_{1,N}/\nu_{1,N} \;\longrightarrow\; 0 \quad \textit{and} \quad \rho_{1,N}/\lambda_{1,N} \;\longrightarrow\; 0 \,. \tag{3.7}$$

NOTE. As noted below Theorem 1, if (3.6) holds, then there always exists $\lambda_{1,N}$ that satisfies (3.7) (e.g., $\lambda_{1,N} = \sqrt{\rho_{1,N}\nu_{1,N}}$).

EXAMPLE 2 (CONTINUED). Group two corresponds to three indexes: $j = 4, 5, 6$. It is easy to see that $B_1 = \mathcal{L}(Z_4, Z_5, Z_6)$, $B_{1,-4} = \mathcal{L}(Z_5, Z_6)$, etc. Thus, $\|P_{B_1}\|_2^2 = \operatorname{rank}(B_1) \le f = m_1 m_2 + m_2 m_3 + m_3 m_1$, $\|P_{B_1} Z_7\|_2^2 \le nf$, and $\|P^{\perp}_{B_2} Z_4\|^2 \le \|Z_4\|^2 = m_3 n$, etc. Finally, it is easy to verify that $P_{Z_5} P_{Z_6} = P_{Z_6} P_{Z_5}$. Thus, by Lemma 3 in the Appendix, $P_{B_{1,-4}} = P_{(Z_5\ Z_6)} \le P_{Z_5} + P_{Z_6}$. It follows that $\operatorname{tr}(Z_4' P_{B_{1,-4}} Z_4) \le \operatorname{tr}(Z_4' P_{Z_5} Z_4) + \operatorname{tr}(Z_4' P_{Z_6} Z_4) = n(m_2 m_3 + m_3 m_1)$, and hence $\|P^{\perp}_{B_{1,-4}} Z_4\|_2^2 = \operatorname{tr}(Z_4' Z_4) - \operatorname{tr}(Z_4' P_{B_{1,-4}} Z_4) \ge n(m_1 m_2 m_3 - m_2 m_3 - m_3 m_1)$, etc. Thus, all the conditions of Lemma 2 are satisfied.

3.4 *General.* The above procedure can be extended to the remaining groups. In general, let $B_t = \mathcal{L}(X, Z_u, u \in L \setminus L_1 \setminus \cdots \setminus L_t)$, $1 \le t \le h$; $B_{t,-j} = \mathcal{L}(X, Z_u, u \in (L \setminus L_1 \setminus \cdots \setminus L_t \setminus \{j\})$, $j \in L_{t+1}$, and $B_t(l_{t+1}) = \mathcal{L}(X, Z_u, u \in (L \setminus L_1 \setminus \cdots \setminus L_{t+1}) \cup l_{t+1})$, $l_{t+1} \subseteq L_{t+1}$, $1 \le t \le h-1$. Define

$$C_{t,N}(l_{t+1}) \;=\; |P^{\perp}_{B_t(l_{t+1})} y|^2 + \lambda_{t,N}|l_{t+1}| \,, \quad l_{t+1} \subseteq L_{t+1} \,; \tag{3.8}$$

where $\lambda_{t,N}$ is a positive number satisfying certain conditions specified below. Let $\hat{l}_{t+1}$ be the minimizer of $C_{t,N}$ over $l_{t+1} \subseteq L_{t+1}$, and $l_{0\underline{t+1}} = \{j \in L_{t+1} :$

$\sigma_j^2 > 0\}$. By exactly the same proof as that of Lemma 2, one can prove the consistency of $\hat{l}_{t+1}$, $2 \le t \le h-1$. Let $\rho_{t,N} = \max_{j \in L_0 \cup \cdots \cup L_t} \|P_{B_t} Z_j\|_2^2$, $\nu_{t,N} = \min_{j \in L_{t+1}} \|P^{\perp}_{B_{t,-j}} Z_j\|_2^2$, $\gamma_{t,N} = \max_{j \in L_{t+1}} (\|P^{\perp}_{B_{t+1}} Z_j\|^2 / \|P^{\perp}_{B_{t,-j}} Z_j\|_2^2)$. Then, we have the following theorem for determining the combination of $l_{01}, \ldots, l_{0h}$.

THEOREM 4. *Suppose that the conditions of Lemma 1 are satisfied; and*

$$\rho_{t,N}/\nu_{t,N} \longrightarrow 0\,, \quad \gamma_{t,N} \longrightarrow 0\,, \quad 1 \le t \le h-1\,. \tag{3.9}$$

Then, for any $\lambda_{t,N}$ such that

$$\lambda_{t,N}/\nu_{t,N} \longrightarrow 0 \quad \textit{and} \quad \rho_{t,N}/\lambda_{t,N} \longrightarrow 0\,, \quad 1 \le t \le h-1\,, \tag{3.10}$$

we have $P(\hat{l}_1 = l_{01}, \ldots, \hat{l}_h = l_{0h}) \to 1$.

NOTE. Unlike $\hat{k}$ in Section 3.1 (see discussion above Theorem 3), here $\hat{l}_t$ does not depend on $\hat{l}_{t'}$, $t' < t$. In fact, $\hat{l}_1, \ldots, \hat{l}_h$ can be obtained simultaneously. Let $\hat{l} = \bigcup_{t=1}^{h} \hat{l}_t$. Then, $\hat{l}$ satisfies the requirement of Theorem 3.

EXAMPLE 2 (CONTINUED). Consider $t = h = 3$, which is the final group. It is easy to see that $B_2 = \mathcal{L}(Z_1, Z_2, Z_3)$, $B_{2,-1} = \mathcal{L}(Z_2, Z_3)$, etc.; $\|P_{B_2}\|_2^2 = \text{rank}(B_2) \le g = m_1 + m_2 + m_3$, and $\|P_{B_2} Z_7\|_2^2 \le ng$. Furthermore, by Lemma 3 in the Appendix, it is easy to show that $\|P_{B_2} Z_4\|_2^2 \le (m_1 + m_2 + 1) m_3 n$; $\text{tr}(Z_1' P_{B_{2,-1}} Z_1) \le 2 m_2 m_3 n$, and hence $\|P^{\perp}_{B_{2,-1}} Z_1\|_2^2 \ge (m_1 - 2) m_2 m_3 n$, etc. Finally, $\|P^{\perp}_{B_3} Z_1\|^2 \le \|Z_1\|^2 = m_2 m_3 n$, etc. Thus, all the conditions of Theorem 4 are satisfied.

4. Two Simulated Examples

In this section, we consider two simulated examples. The first is regarding selection with fixed random factors discussed in Section 2; the second is a case of selection with random factors discussed in Section 3.

4.1 *A simulation regarding Section 2.* We consider a model similar to Example 1 except that more than one fixed covariates may be involved, i.e., $\beta_0 + \beta_1 x_{ij}$ is replaced by $x_{ij}' \beta$, where x_{ij} is a vector of covariates and β a vector of unknown regression coefficients. We examine by simulation the probability of correct selection and also the overfitting ($a1$) and underfitting ($a2$) probabilities respectively of various GIC's for some given model parameters and sample sizes. Five GIC's with different choices of λ are considered:

(1) $\lambda = 2$, which corresponds to the C_p method; (2) $\lambda = \log n$. The latter choice satisfies the conditions of Theorem 1 for the case of a single random effect factor in the true underlying model, which includes the current case. A total of 500 realizations of each simulation were run.

In the simulation the number of fixed factors was $q = 5$ with $\mathcal{A}$ being all subsets of $\{1, ..., 5\}$. The first column of X is $\mathbf{1}$ and the other four columns of X are generated randomly from $N(0, 1)$ distributions but are fixed throughout the simulation. Three β are considered: $(2, 0, 0, 4, 0)'$, $(2, 0, 0, 4, 8)'$ and $(2, 9, 0, 4, 8)'$, which correspond to $k_0 = \{1, 4\}$, $\{1, 4, 5\}$ and $\{1, 2, 4, 5\}$, respectively.

We consider the case where the correlated errors have varying degrees of exchangeable structure as described in Example 1, where four values of ρ were considered: $0, 0.2, 0.5, 0.8$. Variance components σ and τ were both taken to be equal to 1. We take the number of clusters (m) to be 50 and 100 and the number or repeats on a cluster to be fixed at $n = 5$. Table 1 presents the results.

Table 1. Selection probabilities under simulated Example 1

Model	ρ	% correct		$a1$		$a2$	
		$\lambda_N = 2$	$\log(N)$	2	$\log(N)$	2	$\log(N)$
$M1(m = 50)$	0	59	94	41	6	0	0
	.2	64	95	36	5	0	0
	.5	59	90	40	9	1	1
	.8	52	93	47	5	1	2
$M1(m = 100)$	0	64	97	36	3	0	0
	.2	57	94	43	6	0	0
	.5	58	96	42	3	0	1
	.8	61	96	39	4	0	0
$M2(m = 50)$	0	76	97	24	3	0	0
	.2	76	97	24	3	0	0
	.5	73	96	27	4	0	0
	.8	68	94	31	4	1	2
$M2(m = 100)$	0	76	99	24	1	0	0
	.2	70	97	30	3	0	0
	.5	70	98	30	2	0	0
	.8	72	98	28	2	0	0
$M3(m = 50)$	0	90	99	10	1	0	0
	.2	87	98	13	2	0	0
	.5	84	98	16	2	0	0
	.8	78	95	21	3	1	2
$M3(m = 100)$	0	87	99	13	1	0	0
	.2	87	99	13	1	0	0
	.5	80	99	20	1	0	0
	.8	78	96	21	3	1	1

TABLE 2. SELECTION PROBABILITIES UNDER SIMULATED EXAMPLE 2

Model	% correct all		
	$\lambda_{t,N} = 2$	$\log(N)$	$N/\log(N)$
Model 1	27	96	98
Model 2	0	4	100

4.2 *A simulation regarding Section 3.* We now consider a model similar to Example 2 but with (possibly) more fixed covariates and fewer random effect factors, namely,

$$y_{ijk} = x'_{ijk}\beta + a_i + b_j + c_{ij} + e_{ijk} , \tag{4.1}$$

$i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$, $k = 1, \ldots, n$, where, as before, x_{ijk} are elements of a vector of fixed covariates generated once from a standard normal distribution with the first column being all 1, β a vector of unknown regression coefficients. Furthermore, a_i, b_j, and c_{ij} are random effects; and e_{ijk} is an error.

Two scenarios are examined. The first model has $\beta = (2, 0, 0, 4, 0)'$ and only random factor $Z1$ is included in the model. The second model, has the same parameter β as the first model but now has random factors $Z1$ and $Z3$ in the model. For the first model, we set $\sigma_1 = 1$, $\tau = 1.5$, $m_1 = 20$ and $n = 3$. For the second model, we set $\sigma_1 = 1$, $\sigma_3 = 1.5$, $\tau = 1.5$, $m_1 = 20$, $m_2 = 20$, and $n = 3$. All random effects and errors are assumed independent, and the random effects are assumed independent of the errors.

Once again, percent correct selection is examined with a further caveat. Now the results are broken down further to look at percent correct selection of both the fixed and random parts and percent correct selection of only the fixed parts as well as overfitting $a1$ and underfitting $a2$ probabilities. A total of 100 realizations of each simulation were run. We now examine three different possible values of the penalty parameter. We again look at $\lambda_{t,N} = 2, \log(N), N/\log(N)$. Notice that in this particular example, $\lambda = \log(N)$ will only work for the first model since it contains only one random effect factor. For the second model, the conditions of Theorems 1,3 and 4 are satisfied however for $\lambda = N/\log(N)$. Table 2 gives the probabilities of correct selection of both the random and fixed components of the underlying model, and Table 3 gives the various empirical error probabilities including incorrect selection of random factors but correct selection of fixed ones, overfitting and underfitting probabilities. It is clear that the new method performs admirably in both model settings where the probability of correct selection remains very high even when the number of random factors present in the model increases past one. The BIC however does not share this property (see

TABLE 3. ERROR PROBABILITIES UNDER SIMULATED EXAMPLE 2

Model	% correct fixed only, $a1$, $a2$		
	$\lambda_{t,N} = 2$	$\log(N)$	$N/\log(N)$
Model 1	39,34,0	2,2,0	2,0,0
Model 2	68,31,0	93,3,0	0,3,0

discussion below). Using a penalty factor of $\lambda_N = 2$ clearly is suboptimal and tends to produce overfit models as expected.

5. Discussion

In the previous section, we exam the behaviour of BIC, which corresponds to $\lambda_N = \log(N)$. It is seen that BIC performs well in the case of single random effect factor, but not so in the case of multiple random effect factors. This is, of course, exactly as predicted by our theory. An intuitive interpretation is the following: BIC is associated with the information associated with a given model structure. In the classic case the information is affected by a single factor — the sample size. In a mixed model with single random effect factor and assuming the number of observations corresponding to each random effect is fixed, the information is, again, affected by a single factor, which is the number of clusters m, so it is essentially the same as the classic case. However, in mixed linear models with multiple random effect factors there may be more than one factor that may increase to infinity (e.g., m_1, m_2, ...). This feature of mixed linear models has long been recognized (e.g., Hartley and Rao (1967), Miller (1977), Jiang (1996)). As a consequence, the information may be affected by more than one factors (see Jiang (1996)). Therefore, the classic definition of BIC is inappropriate for mixed linear models with multiple random effect factors. In this paper, we have provided new conditions for consistent model selection in this more complex class of models that take into account the factors that affect model structure information.

Although the standard methods of fitting a mixed linear model are restricted maximum likelihood (REML) or maximum likelihood (ML), these methods are known to produce efficient estimators only when the random effects and errors are normally distributed. Note that in this paper we have not made such an assumption. In fact, throughout the paper, the random effects and errors are not even assumed to be i.i.d. In such a case, one cannot assume that the REML or ML based methods will be naturally more efficient in model selection than the least squares (LS) based methods, although comparison of these methods would be an interesting topic for future study. On the other hand, the LS based methods are computationally much

simpler, which is important for model selection problems because one may have to compute for a large number of candidate models.

In this paper we do not purport to provide an optimal way of choosing the best penalty parameter for a finite data set. But what we do provide are the necessary and sufficient conditions for consistency of selection that are expressed in terms of the penalty parameter. Note that adaptive strategies for picking the penalty parameter is an unsolved problem even in fixed effects linear models.

Acknowledgements. The research of J. Jiang is supported, in part, by U.S. National Science Foundation Grant SES-9978101; the two authors contribute equally to this work.

Appendix

PROOF OF THEOREM 1. First, we note that for any $k \in \mathcal{A}$,

$$\begin{aligned} & C_N(k) - C_N(k_0) \\ &\quad = \; y'[P_X - P_{X(k)}]y + (|k| - p)\lambda_N \\ &\quad = \; \beta' X' P^{\perp}_{X(k)} X\beta + 2\beta' X' P^{\perp}_{X(k)}\zeta + \zeta'[P_X - P_{X(k)}]\zeta + (|k| - p)\lambda_N \\ &\quad \geq \; \beta' X' P^{\perp}_{X(k)} X\beta + 2\beta' X' P^{\perp}_{X(k)}\zeta - \zeta' P_{X(k)}\zeta + (|k| - p)\lambda_N \; . \qquad \text{(A.1)} \end{aligned}$$

Note that $X'[P_X - P_{X(k)}]X = X' P^{\perp}_{X(k)} X$; and (A.1) holds even if k or k_0 is $\emptyset$.

Suppose that $k \in \mathcal{A}$ but k does not $\supseteq k_0$. Then, $k_0 \neq \emptyset$, and $k \neq K$. Also, there is a column of X, say X_j, such that $j \notin k$. It follows that $k \subseteq \{j\}^c$, and hence $P_{X(k)} \leq P_{X(\{j\}^c)}$. Thus, by (A.1), we have

$$\begin{aligned} C_N(k) - C_N(k_0) &\geq \beta' X' P^{\perp}_{X(\{j\}^c)} X\beta + 2\beta' X' \cdots \\ &= \beta_j^2 |P^{\perp}_{X(\{j\}^c)} X_j|^2 + 2\beta' X' \cdots \\ &\geq \beta_j^2 \omega_N + 2\beta' X' P^{\perp}_{X(k)}\zeta - \zeta' P_{X(k)}\zeta + (|k| - p)\lambda_N \; . \qquad \text{(A.2)} \end{aligned}$$

It is easy to show that $E\zeta' P_{X(k)}\zeta = \mathrm{tr}(P_{X(k)}[ZGZ' + R]) \leq \rho_N \mathrm{tr}(P_{X(k)}) \leq |k|\rho_N$. Thus, the third term of the right side of (A.2) is $\rho_N O_P(1)$, where, and hereafter, $O_P(1)$ represents a term that is bounded in probability. Similarly, we have

$$\begin{aligned} E(\beta' X' P^{\perp}_{X(k)}\zeta)^2 &= \beta' X' P^{\perp}_{X(k)}(ZGZ' + R) P^{\perp}_{X(k)} X\beta \\ &\leq \rho_N \lambda_{\max}(X' P^{\perp}_{X(k)} X)|\beta|^2 \;\leq\; p|\beta|^2 \rho_N \nu_N \; , \end{aligned}$$

because $\lambda_{\max}(X'P^{\perp}_{X(k)}X) \le \lambda_{\max}(P^{\perp}_{X(k)})\lambda_{\max}(X'X) \le \mathrm{tr}(X'X) \le p\nu_N$. Thus, the second term on the right side of (A.2) is $\sqrt{\rho_N\nu_N}O_P(1)$. Therefore, we have

$$\begin{aligned} & C_N(k) - C_N(k_0) \\ & \quad \ge \beta_j^2\omega_N + \sqrt{\rho_N\nu_N}O_P(1) + \rho_N O_P(1) + \lambda_N O(1) \\ & \quad = \nu_N\left[\beta_j^2\left(\frac{\omega_N}{\nu_N}\right) + O_P(1)\sqrt{\frac{\rho_N}{\nu_N}} + O_P(1)\left(\frac{\rho_N}{\nu_N}\right) + O(1)\left(\frac{\lambda_N}{\nu_N}\right)\right] . \quad \text{(A.3)} \end{aligned}$$

Because $\beta_j \ne 0$, by (2.4) and (2.5), the right side of (A.3) is > 0 with probability tending to one. It follows that $P(\hat{k} = k) \to 0$.

Now, assume that $k \in \mathcal{A}$, $k \supseteq k_0$, but $k \ne k_0$. Then, $|k| \ge p + 1$. Also, since $\mathcal{L}(X(k)) \supset \mathcal{L}(X)$, we have $P^{\perp}_{X(k)}X = 0$. Thus, by (A.1), we have

$$C_N(k) - C_N(k_0) \quad \ge \quad \lambda_N - \zeta' P_{X(k)}\zeta . \quad \text{(A.4)}$$

By an earlier argument, the second term on the right side of (A.4) is $\rho_N O_P(1)$. Therefore, we have

$$C_N(k) - C_N(k_0) \quad \ge \quad \lambda_N\left[1 + O_P(1)\left(\rho_N/\lambda_N\right)\right] . \quad \text{(A.5)}$$

By (2.5), the right side of (A.5) is > 0 with probability tending to one. It follows that, again, $P(\hat{k} = k) \to 0$.

Because there are only finitely many k's in $\mathcal{A}$, (2.3) follows. □

Proof of Theorem 2. First, suppose that $j \in k_0$. Then, since, by (2.1) and (2.8), $P^{\perp}_{X_{-j}}y = \beta_j P^{\perp}_{X_{-j}}X_j + P^{\perp}_{X_{-j}}\zeta$, $P^{\perp}_{X}y = P^{\perp}_{X}\zeta$, we have

$$\begin{aligned} |P^{\perp}_{X_{-j}}y|^2 - |P^{\perp}_{X}y|^2 & = \beta_j^2|P^{\perp}_{X_{-j}}X_j|^2 + 2\beta_j X_j' P^{\perp}_{X_{-j}}\zeta + \zeta' P_X\zeta - \zeta' P_{X_{-j}}\zeta \\ & \ge \beta_j^2|P^{\perp}_{X_{-j}}X_j|^2 + 2\beta_j X_j' P^{\perp}_{X_{-j}}\zeta - \zeta' P_{X_{-j}}\zeta . \quad \text{(A.6)} \end{aligned}$$

Since $E\zeta' P_{X_{-j}}\zeta = \mathrm{tr}[P_{X_{-j}}(R + ZGZ')] \le (q-1)\rho_N$, the third term on the right side of (A.6) can be expressed as $\rho_N O_P(1)$. Also, since $E|X_j' P^{\perp}_{X_{-j}}\zeta|^2 = \mathrm{tr}[P^{\perp}_{X_{-j}}X_j X_j' P^{\perp}_{X_{-j}}(R + ZGZ')] \le \rho_N|P^{\perp}_{X_{-j}}X_j|^2$, the second term on the right side of (A.6) can be expressed as $\sqrt{\rho_N}|P^{\perp}_{X_{-j}}X_j|O_P(1)$. It follows that

$$\begin{aligned} & \frac{|P^{\perp}_{X_{-j}}y|^2 - |P^{\perp}_{X}y|^2}{|P^{\perp}_{X_{-j}}X_j|^2\delta_N} \\ & \quad \ge \delta_N^{-1}\left[\beta_j^2 + \left(\frac{\rho_N}{|P^{\perp}_{X_{-j}}X_j|^2}\right)^{1/2}O_P(1) + \left(\frac{\rho_N}{|P^{\perp}_{X_{-j}}X_j|^2}\right)O_P(1)\right] \\ & \quad \ge \delta_N^{-1}\left[\beta_j^2 - \left(\frac{\rho_N}{\eta_N}\right)^{1/2}O_P(1) - \left(\frac{\rho_N}{\eta_N}\right)O_P(1)\right] , \end{aligned}$$

which exceeds 1 with probability tending to one. Thus, $P(j \in \hat{k}) \to 1$.

Now, suppose that $j \notin k_0$. Then, by the first equation in (A.6),

$$|P^{\perp}_{X_{-j}} y|^2 - |P^{\perp}_{X} y|^2 \;\; \leq \;\; \zeta' P_X \zeta \; . \tag{A.7}$$

Since $E\zeta' P_X \zeta = \mathrm{tr}[P_X(R+ZGZ')] \leq q\rho_N$, the right side of (A.7) is $\rho_N O_P(1)$. It follows that

$$\frac{|P^{\perp}_{X_{-j}} y|^2 - |P^{\perp}_{X} y|^2}{|P^{\perp}_{X_{-j}} X_j|^2 \delta_N} \;\; \leq \;\; \left(\frac{\rho_N}{|P^{\perp}_{X_{-j}} X_j|^2 \delta_N}\right) O_P(1) \;\; \leq \;\; \left(\frac{\rho_N}{\eta_N \delta_N}\right) O_P(1) \; ,$$

which goes to 0 with probability tending to one. Thus, $P(j \in \hat{k}) \to 0$. □

Proof of Lemma 1. Throughout the proof, c represents a constant that may have different values at different places.

Suppose that $\sigma_j^2 = 0$. We show that $P(j \in \hat{l}_1) \to 0$. Since

$$\left(\frac{r}{R}\right)\left(\frac{R_j}{r_j}\right) - 1 \;\; = \;\; \left(\frac{r}{R}\right)\left[\left(\frac{R_j}{r_j} - \tau^2\right) - \left(\frac{R}{r} - \tau^2\right)\right] \; , \tag{A.8}$$

$j \in \hat{l}_1$ implies that either $|(R/r) - \tau^2| > r^{\rho/2-1}(R/r)$ or $|(R_j/r_j) - \tau^2| > r_j^{\rho/2-1}(R/r)$. We have $P(|(R/r) - \tau^2| > r^{\rho/2-1}(R/r)) \leq P(R/r < \tau^2/2) + P(|(R/r) - \tau^2| > (\tau^2/2) r^{\rho/2-1}) = I_1 + I_2$; $P(|(R_j/r_j) - \tau^2| > r_j^{\rho/2-1}(R/r)) \leq P(R/r < \tau^2/2) + P(|(R_j/r_j) - \tau^2| > (\tau^2/2) r_j^{\rho/2-1}) = I_1 + I_3$. By Lemma 2.1 of Jiang (1997), we have $I_1 \leq P(|\epsilon' P^{\perp}_B \epsilon - E\epsilon P^{\perp}_B \epsilon| > \tau^2 r/2) \leq c r^{-2} \|P^{\perp}_B\|_2^2 = c r^{-1}$. Similarly, $I_2 \leq c r^{1-\rho}$. Also, since $\sigma_j^2 = 0$, we have $(P_B - P_{B_{-j}})y = (P_B - P_{B_{-j}})\epsilon$. Thus, again, by Lemma 2.1 of Jiang (1997), $I_3 = P(|\epsilon'(P_B - P_{B_{-j}})\epsilon - E\cdots| > \tau^2 r_j^{\rho/2}/2) \leq c r_j^{-\rho} \|P_B - P_{B_{-j}}\|_2^2 \;\; = \;\; c r_j^{1-\rho}$. It follows that $I_t \to 0$, $t = 1, 2, 3$.

Now, suppose that $\sigma_j^2 > 0$. We show that $P(j \in \hat{l}_1) \to 1$. Again, by (A.8), $(R_j/r_j) - \tau^2 > (2r^{\rho/2-1} + r_j^{\rho/2-1})(R/r)$ and $|(R/r) - \tau^2| \leq r^{\rho/2-1}(R/r)$ implies that $j \in \hat{l}_1$. By a previous result, $P(|R/r - \tau^2| \leq r^{\rho/2-1} R/r) \to 1$. Also, there is $\delta > 0$ such that for large N, $\|P^{\perp}_{B_{-j}} Z_j\|_2^2 \geq \delta r_j$. Thus, for large N,

$$\begin{aligned} & P\left[\frac{R_j}{r_j} - \tau^2 \leq (2r^{\rho/2-1} + r_j^{\rho/2-1})\left(\frac{R}{r}\right)\right] \\ & \qquad \leq \;\; P\left(\frac{R}{r} > \frac{3}{2}\tau^2\right) + P\left(\frac{R_j}{r_j} - \tau^2 < \frac{\delta}{2}\sigma_j^2\right) \;\; = \;\; J_1 + J_2 \; . \end{aligned}$$

Again, by an earlier argument, $J_1 \to 0$. Let $\xi = (\alpha_j' \ \epsilon')'$. Then, $R_j = \xi' A \xi$, where

$$A = \begin{pmatrix} Z_j' P_{B_{-j}}^{\perp} Z_j & Z_j' P_{B_{-j}}^{\perp} \\ P_{B_{-j}}^{\perp} Z_j & P_B - P_{B_{-j}} \end{pmatrix},$$

so that $ER_j = E\xi' A \xi = \tau^2 r_j + \sigma_j^2 \|P_{B_{-j}}^{\perp} Z_j\|_2^2$. Thus, once again, using Lemma 2.1 of Jiang (1997), we have

$$\begin{aligned} J_2 &\le P(|\xi' A \xi - E\xi' A \xi| > \delta \sigma_j^2 r_j/2) \\ &\le c r_j^{-2} \|A\|_2^2 = c \left[\frac{\|Z_j' P_{B_{-j}}^{\perp} Z_j\|_2^2}{r_j^2} + 2 \left(\frac{\|P_{B_{-j}}^{\perp} Z_j\|_2^2}{r_j^2} \right) + \frac{1}{r_j} \right]. \end{aligned}$$

Thus, $J_2 \to 0$, and this completes the proof. □

PROOF OF LEMMA 2. For notation simplicity, we write $\zeta_t = \sum_{u \in L_t} Z_u \alpha_u$, $t = 1, 2$; $B = B_1(l_2)$, $B_0 = B_1(l_{02})$, $P = P_B$, $P_0 = P_{B_0}$, and $\Delta = P_0 - P$. Then, since $y = X\beta + \sum_{u \in L \setminus L_1 \setminus L_2} Z_u \alpha_u + \zeta_2 + \zeta_1 + \epsilon$, we have $\Delta y = \Delta(\epsilon + \zeta_1 + \zeta_2)$. It follows that

$$\begin{aligned} C_{1,N}(l_2) - C_{1,N}(l_{02}) &= y' \Delta y + \lambda_{1,N}(|l_2| - |l_{02}|) \\ &= (\epsilon + \zeta_1 + \zeta_2)' \Delta (\epsilon + \zeta_1 + \zeta_2) + \lambda_{1,N}(|l_2| - |l_{02}|) \ . \quad \text{(A.9)} \end{aligned}$$

If l_2 does not $\supseteq l_{02}$, let $L_1 = \{j_{1,1}, \ldots, j_{1,a}\}$, $l_{02} = \{j_{2,1}, \ldots, j_{2,b}\}$,

$$W = (I \ Z_{j_{1,1}} \ \cdots \ Z_{j_{1,a}} \ Z_{j_{2,1}} \ \cdots \ Z_{j_{2,b}}),$$

$\xi = (\epsilon' \ \alpha_{j_{1,1}}' \ \cdots \ \alpha_{j_{1,a}}' \alpha_{j_{2,1}}' \ \cdots \ \alpha_{j_{2,b}}')'$, and $A = W' \Delta W$. Then, by (A.9),

$$\begin{aligned} C_{1,N}(l_2) - C_{1,N}(l_{02}) &= E\xi' A \xi + (\xi' A \xi - E\xi' A \xi) + \lambda_{1,N}(|l_2| - |l_{02}|) \\ &= I_1 + I_2 + I_3 \ . \quad \text{(A.10)} \end{aligned}$$

We have

$$\begin{aligned} I_1 &= \operatorname{tr}(\Delta W \operatorname{Var}(\xi) W') \\ &= \tau^2 [\operatorname{rank}(B_0) - \operatorname{rank}(B)] \\ &\quad + \sum_{u \in L_1} \sigma_u^2 (\|P_0 Z_u\|_2^2 - \|P Z_u\|_2^2) + \sum_{u \in l_{02}} \sigma_u^2 \|P^{\perp} Z_u\|_2^2 \\ &\ge \sum_{u \in l_{02} \setminus l_2} \sigma_u^2 \|P^{\perp} Z_u\|_2^2 - \tau^2 \operatorname{rank}(B) - \sum_{u \in L_1} \sigma_u^2 \|P Z_u\|_2^2 \\ &\ge c_1 \left(\sum_{u \in l_{02} \setminus l_2} \|P^{\perp} Z_u\|_2^2 \right) - c_2 \left(\sum_{u \in L_0 \cup L_1} \|P Z_u\|_2^2 \right), \quad \text{(A.11)} \end{aligned}$$

where c_1, c_2 are positive constants. Note that, for $u \in l_{02}$, we have $\Delta Z_u = P^{\perp} Z_u$, which is 0 if $u \in l_{01} \cap l_2$. Also, by Lemma 2.1 of Jiang (1997), it is easy to show that

$$I_2 \;=\; \|A\|_2 O_{L^2}(1)\,, \tag{A.12}$$

where $O_{L^2}(1)$ represents a term that is bounded in L^2. We now obtain a bound for $\|A\|_2$. Write $S = L_0 \cup L_1 \cup l_{02}$. Then, $A = (Z_u' \Delta Z_v)_{u,v \in S}$, and hence

$$\begin{aligned} \|A\|_2^2 &= \sum_{u,v \in S} \|Z_u' \Delta Z_v\|_2^2 \\ &= \sum_{u,v \in L_0 \cup L_1} \|Z_u' \Delta Z_v\|_2^2 + \sum_{u \in l_{02} \text{ or } v \in l_{02}} \|Z_u' P^{\perp} Z_v\|_2^2 \\ &= J_1 + J_2\,. \end{aligned} \tag{A.13}$$

If $u, v \in L_0 \cup L_1$, then

$$\|Z_u' \Delta Z_v\|_2 \le \|Z_u' P_0 Z_v\|_2 + \|Z_u' P Z_v\|_2 \le \|P_0 Z_u\|_2 \|P_0 Z_v\|_2 + \|P Z_u\|_2 \|P Z_v\|_2.$$

It follows that

$$J_1 \;\le\; 2\left[\left(\sum_{u \in L_0 \cup L_1} \|P_0 Z_u\|_2^2\right)^2 + \left(\sum_{u \in L_0 \cup L_1} \|P Z_u\|_2^2\right)^2\right]\,. \tag{A.14}$$

If $u \in l_{02}$, then $\|Z_u' P^{\perp} Z_v\|_2 \le \|P^{\perp} Z_v\| \cdot \|P^{\perp} Z_u\|_2$. Thus,

$$J_2 \;\le\; 2 \sum_{u \in l_{02}} \|Z_u' P^{\perp} Z_v\|_2^2 \;\le\; 2\left(\sum_{u \in S} \|P^{\perp} Z_u\|^2\right)\left(\sum_{u \in l_{02}} \|P^{\perp} Z_u\|_2^2\right)\,. \tag{A.15}$$

Combining (A.13) — (A.15), we have

$$\begin{aligned} \|A\|_2 \;\le\; \sqrt{2}\Bigg[&\sum_{u \in L_0 \cup L_1} \|P_0 Z_u\|_2^2 + \sum_{u \in L_0 \cup L_1} \|P Z_u\|_2^2 + \Bigg(\sum_{u \in L_0 \cup L_1} \|P^{\perp} Z_u\|^2 \\ &+ \sum_{u \in l_{02} \setminus l_2} \|P^{\perp} Z_u\|^2\Bigg)^{1/2} \left(\sum_{u \in l_{02} \setminus l_2} \|P^{\perp} Z_u\|_2^2\right)^{1/2}\Bigg]\,. \end{aligned} \tag{A.16}$$

Note that, if $u \notin l_2$, then $B = B_1(l_2) \subseteq B_{1,-u}$, hence $P \le P_{B_{1,-u}}$, and therefore $\|P^{\perp} Z_u\|_2 \ge \|P^{\perp}_{B_{1,-u}} Z_u\|_2$. Also, we have $B_2 \subseteq B \subseteq B_1$, hence $\|P Z_u\| \le$

$\|PZ_u\|_2 \le \|P_{B_1}Z_u\|_2$, $\|P_0Z_u\|_2 \le \|P_{B_1}Z_u\|_2$, and $\|P^\perp Z_u\| \le \|P^\perp_{B_2}Z_u\|$. By these inequalities, and (A.8), (A.11), (A.12), and (A.16), it is easy to see that

$$C_{1,N}(l_2) - C_{1,N}(l_{02}) \;\ge\; \left(\sum_{u\in l_{02}\setminus l_2} \|P^\perp_{B_{1,-u}}Z_u\|_2^2\right)[c + o_P(1)]\,,$$

where c is a positive constant, and $o_P(1)$ a term that $\to 0$ in probability. It follows that $P(\hat{l}_2 = l_2) \to 0$.

If $l_2 \supseteq l_{02}$ but $\neq l_{02}$, then $|l_2| \ge |l_{02}| + 1$. Furthermore, we have $B_0 \subseteq B$, hence $\Delta\zeta 2 = 0$. Thus, by (A.9),

$$\begin{aligned} C_{1,N}(l_2) - C_{1,N}(l_{02}) &\ge (\epsilon+\zeta_1)'\Delta(\epsilon+\zeta_1) + \lambda_{1,N} \\ &\ge \lambda_{1,N} - (\epsilon+\zeta_1)'P(\epsilon+\zeta_1)\,. \end{aligned} \tag{A.17}$$

It is easy to show that $E(\epsilon+\zeta_1)'P(\epsilon+\zeta_1) \le c\sum_{u\in L_0\cup L_1}\|P_{B_1}Z_u\|_2^2$ for some constant c. Thus, the second term on the right side of (A.17) is $o_P(\lambda_{1,N})$, i.e., it converges to 0 in probability if divided by $\lambda_{1,N}$. It follows that $P(\hat{l}_2 = l_2) \to 0$. Since there are only finite many choices for l_2, this completes the proof. □

LEMMA 3. *Let A and B be matrices with the same number of rows such that $P_AP_B = P_BP_A$. Then, $P_{(A\ B)} \le P_A + P_B$.*

PROOF. First, let $x \in \mathcal{L}[(A\ B)]$. Then, $x = a + b$, where $a \in \mathcal{L}(A)$, $b \in \mathcal{L}(B)$. It is easy to show that $x'P_Ax + x'P_Bx - x'x = |P_Ab + P_Ba|^2 \ge 0$. Next, for any vector x, $x = x_1 + x_2$, where $x_1 \in \mathcal{L}[(A\ B)]$, $x_2 \perp \mathcal{L}[(A\ B)]$. It follows that $x'P_{(A\ B)}x = x_1'x_1 \le x_1'P_Ax_1 + x_1'P_Bx_1 = x'(P_A + P_B)x$. □

References

AKAIKE H. (1972). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int Symp Information Theory, Supp to Problems of Control and Information Theory* B.N. Petron and F. Csak, eds., Akademiai kaido, Budapest, 267-281.

BICKEL, P.J. and ZHANG, P. (1992). Variable selection in nonparametric regression with categorical covariates. *J. Amer. Statist. Assoc.* **87**, 90-97.

CHOI, B.S. (1992). *ARMA Model Identification*, Springer-Verlag, New York.

HARTLEY, H.O. and RAO, J.N.K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* **54**, 93-108.

JIANG, J. (1996). REML estimation: Asymptotic behavior and related topics, *Ann. Statist.* **24**, 255-286.

Jiang, J. (1997). Wald consistency and the method of sieves in REML estimation, *Ann. Statist.* **25**, 1781-1803.

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

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

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

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

Shao, J. (1993). Linear model selection by cross-validation, *J. Amer. Statist. Assoc.* **88**, 486-494.

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

Zheng, X. and Loh, W.Y. (1995). Consistent variable selection in linear models. *J. Amer. Statist. Assoc.* **90**, 151-156.

Jiming Jiang
Department of Statistics
University of California
One Shields Avenue
Davis, CA 95616, USA
E-mail: jiang@wald.ucdavis.edu

J. Sunil Rao
Department of Biostatistics and Epidemiology
Case Western Reserve University
10900 Euclid Ave
Cleveland, OH 44106, USA
E-mail: sunil@hal.cwru.edu