FENCE METHODS FOR MIXED MODEL SELECTION

JIMING JIANG[†], J. SUNIL RAO[‡], ZHONGHUA GU[†] AND THUAN NGUYEN[†]
University of California, Davis[†] and Case Western Reserve University[‡]

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, but these do not naturally extend to non-standard situations such as mixed effects models, where simple definition of the sample size is not meaningful. This paper introduces a new class of strategies, known as fence methods, for mixed model selection, which includes linear and generalized linear mixed models. 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. We describe a variety of fence methods, based on the same principle but applied to different situations, including clustered and non-clustered data, linear or generalized linear mixed models, and Gaussian or non-Gaussian random effects. We show the broad applicability and study the performance of fence methods by giving a number of examples, each supported by simulation results or applied data analysis. In addition, we propose two variations of the basic fence method, one utilizes a stepwise procedure to handle situations of many predictors; the other introduces an adaptive approach of choosing a tuning constant involved in the fence method. We give sufficient conditions for consistency of fence and its variations, a desirable property for a good model selection procedure.

Key Words. Clustered Data, Consistency, Generalized Linear Mixed Models, Mixed Model Selection, Non-clustered Data.

1 Introduction

Many model search strategies involve trading off model fit with model complexity in a penalized goodness of fit measure. Such procedures usually amount to minimizing 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 (see Remark in section 2.2). The main difference between procedures 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, connecting the relative Kullback-Liebler discrepancy and the empirical log-likelihood function yields the Akaike's information criterion (AIC; Akaike 1973, 1974) where $\lambda_n = 2$. The idea has allowed major practical and theoretical advances in model selection and related fields (e.g., de Leeuw 1992). A number of similar criteria have since been proposed, for instance, the Bayesian information criterion (BIC; Schwarz 1978) in which $\lambda_n = \log(n)$; a criterion due to Hannan and Quinn (HQ; Hannan and Quinn 1979) in which $\lambda_n = c\log\{\log(n)\}$ and c is a constant > 2; and the generalized information criterion (GIC; Nishii 1984, Shibata 1984) in which λ_n assumes other values.

Although these criteria are widely used, difficulties are often encountered, especially in some non-conventional situations. A broad class of such non-conventional cases are mixed effects models, including linear and generalized linear mixed models. For example, consider the following linear mixed model, $y_{ij} = x_{ij}'\beta + u_i + v_j + e_{ij}$, $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$, 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., Hartley and Rao 1967, 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. 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), where $\lambda_n = \log(n)$ (it does not make sense to let $n = m_1 \cdot m_2$). In fact, in cases of correlated observations, such as the example here, 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, and that they have means zero and variances σ_u^2, σ_v^2 and σ_e^2, respectively. 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 under the assumed model.

Even in conventional cases, there are still some practical issues regarding the use of these model selection criteria. For example, the BIC is known to have the tendency of overly penalizing bigger models. In other words, the penalizer, $\log(n)$, may be a little too much in some cases (see, for example, section 4 below). 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 $c > 0$. Here consistency means that, as $n \to \infty$, the probability that the procedure selects the optimal model (i.e., a true model with minimal dimension; see below) goes to one. However, practically, the choice of c does matter. 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 almost the same.

In a way, model selection and estimation are viewed as two components of a process called model identification. 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 GIC's suitable for linear mixed model selection and proved consistency of their procedures. The authors also studied finite sample performance of their procedures by simulations. 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 $y_{ij} = x_{ij}'\beta + u_i + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, 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 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. See further discussion in the sequel. 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.

The rest of the paper is organized as follows. In section 2 we describe in detail a new procedure for mixed model selection, called *fence* method. A variation of the procedure known as F-B fence is also proposed. In section 3 we consider estimation of a standard deviation, which plays an important role in the fence method, and show how to utilize the fence in various situations involving clustered and non-clustered data. In sections 4 and 5 we give a number of examples, each supported by results of simulations or real data analyses, to illustrate the application of fence in various situations. The examples include linear mixed models and GLMMs with clustered and non-clustered data. In section 6 we propose an adaptive method of choosing a tuning constant involved in the fence procedure. In section 7 we address the issue of consistency of different fence methods. Some further discussion and concluding remarks are made in section 8. The proofs are given in section 9.

2 The fence method

The essential part of this procedure is a quantity $Q_M = Q_M(y, \theta_M)$, 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 this paper, we use the terms "true model" and "correct model" interchangeably. Below are some examples of Q_M.

1. Maximum likelihood (ML) model selection. If the model specifies the full distribution of y up to the parameter vector θ_M, an example of Q_M is the negative of the log-likelihood under M, i. e., $Q_M = -\log\{f_M(y|\theta_M)\}$, where $f_M(\cdot|\theta_M)$ is the joint pdf of y with respect to a measure ν under M, given that θ_M is the true parameter vector. To see that $\mathrm{E}(Q_M)$ is minimized when M is a true model and θ_M the true parameter vector under M, let $f(y)$ denote the true pdf of y. We have

$$\begin{aligned} -\mathrm{E}(Q_M) &= \int \log\{f_M(y|\theta_M)\} f(y)\nu(dy) \\ &= \int \log\{f(y)\} f(y)\nu(dy) + \int \log\left\{\frac{f_M(y|\theta_M)}{f(y)}\right\} f(y)\nu(dy) \\ &\leq \int \log\{f(y)\} f(y)\nu(dy) + \log\left\{\int \frac{f_M(y|\theta_M)}{f(y)} f(y)\nu(dy)\right\} \\ &= \int \log\{f(y)\} f(y)\nu(dy), \end{aligned} \tag{2}$$

using the concave-function inequality. The lone term on the right side of (2) is equal to $-\mathrm{E}(Q_M)$ when M is a true model and θ_M the true parameter vector.

2. Mean and variance/covariance (MVC) model selection. If the model is only specified by the mean and covariance matrix of y, it is called a mean and variance/covariance model, or MVC model. In this case, we may consider $Q_M = |(T' V_M^{-1} T)^{-1} T' V_M^{-1} (y - \mu_M)|^2$, where μ_M and V_M are the mean vector and covariance matrix under M, and T is a given $n \times s$ matrix of full rank

$s \le n$. To see that $\mathrm{E}(Q_M)$ is minimized when $\mu_M = \mu$, $V_M = V$, where μ and V denote the true mean vector and covariance matrix, note that

$$\begin{aligned} \mathrm{E}(Q_M) &= \mathrm{tr}\{(T'V_M^{-1}T)^{-1}T'V_M^{-1}VV_M^{-1}T(T'V_M^{-1}T)^{-1}\} \\ &\quad +|(T'V_M^{-1}T)^{-1}T'V_M^{-1}(\mu_M - \mu)|^2. \end{aligned} \tag{3}$$

The first term is the trace of the covariance matrix of the weighted least squares (WLS) estimator of β with the weight matrix $W = V_M^{-1}$ in the linear regression $y = T\beta + \epsilon$, where $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = V$. Since the covariance matrix of the WLS estimator is minimized when $W = V^{-1}$, i. e., $V_M = V$, the first term on the right side of (3) is minimized when $V_M = V$. On the other hand, the second term is zero when $\mu_M = \mu$.

3. Extended GLMM selection. Jiang and Zhang (2001) proposed an extension of GLMM, in which only the conditional mean of the response given the random effects is parametrically specified. It is assumed that, given a vector α of random effects, the responses $y_1, \ldots, y_n$ are conditionally independent such that $\mathrm{E}(y_i|\alpha) = h(x_i'\beta + z_i'\alpha)$, $1 \le i \le n$, where $h(\cdot)$ is a known function, β is a vector of unknown fixed effects, and x_i, z_i are known vectors. Furthermore, it is assumed that $\alpha \sim N(0, \Sigma)$, where the covariance matrix Σ depends on a vector ψ of variance components. Let β_M and ψ_M denote β and ψ under M, and $g_{M,i}(\beta_M, \psi_M) = \mathrm{E}\{h_M(x_i'\beta_M + z_i'\Sigma_M^{1/2}\xi)\}$, where h_M is the function h under M, Σ_M is the covariance matrix under M evaluated at ψ_M, and the expectation is taken with respect to $\xi \sim N(0, I_m)$ (which does not depend on M). Here m is the dimension of α and I_m the m-dimensional identity matrix. We consider the following

$$Q_M = \sum_{i=1}^{n}\{y_i - g_{M,i}(\beta_M, \psi_M)\}^2. \tag{4}$$

It is easy to see that the Q_M given above satisfies the basic requirement, i.e., $\mathrm{E}(Q_M)$ is minimized when M is a true model and $\theta_M = (\beta_M', \psi_M')'$ is the true parameter vector under M. In fact,

(4) corresponds to the Q_M in MVC model selection just discussed with $T = I$, the identity matrix. Note that, since V is not parametrically specified under the assumed model, it needs not get involved in Q_M. Therefore, (4) is a natural choice for Q_M in this case.

2.1 Building the fence

Given a specific Q_M, let $\hat{Q}_M = Q_M(y, \hat{\theta}_M)$, where $\hat{\theta}_M$ is the minimizer of Q_M over $\theta_M \in \Theta_M$, the parameter space under M, that is, $\hat{Q}_M = \inf_{\theta_M \in \Theta_M} Q_M(\theta_M, y)$. A model is called *optimal* if it is a true model with the smallest dimension. Here the dimension of a model M, $|M|$, is understood as the dimension of θ_M. However, it will be seen later that the method developed here is, in fact, flexible in this regard. Notice carefully that the optimal model would be selected by minimizing Q_M if one knew the true value of θ_M. However, Q_M is something we do not have the luxury of knowing and thus must base our selection on $\hat{Q}_M$. The initial thought was to consider something similar to (1), that is, a criterion function of the form

$$\hat{Q}_M + \lambda_n |M|. \tag{5}$$

However, we encountered the same problem as described earlier for a procedure based on (1). Although we know that, under regularity conditions, as long as $\lambda_n/n \to 0$ and $\lambda_n/\sqrt{n} \to \infty$, the procedure based on (5) is consistent, this only gives the order of λ_n. In other words, there is a constant involved, which in case of moderate sample size could make a bigger difference than n itself (see our earlier discussion regarding BIC and HQ). It took us some time to figure out an alternate solution. We arrived at the following thought.

Let $\tilde{M} \in \mathcal{M}$ be such that $\hat{Q}_{\tilde{M}} = \min_{M \in \mathcal{M}} \hat{Q}_M$, where $\mathcal{M}$ represents the set of candidate models. We assume that $\mathcal{M}$ contains a true model. Note that in many cases, $\tilde{M}$ can be determined without

any calculation. For example, if $\mathcal{M}$ contains a full model, say $M_{\rm f}$, that is, a model such that all other models in $\mathcal{M}$ are submodels of $M_{\rm f}$, then, clearly, $\tilde{M} = M_{\rm f}$ and, since $\mathcal{M}$ contains a true model, $M_{\rm f}$ is also a true model. In general, $\mathcal{M}$ may not contain a full model, but the following lemma shows that, at least in large sample, $\tilde{M}$ is expected to be a correct model.

Lemma 1. Under the assumptions A1 - A5 in section 7, we have with probability tending to one that $\tilde{M}$ is a true model.

The proof of Lemma 1 follows directly from that of Theorem 1 in the sequel.

However, the main question is, "Are there other correct models in $\mathcal{M}$ with smaller dimension than $\tilde{M}$?" To answer this question, we need to know what the difference $\hat{Q}_M - \hat{Q}_{\tilde{M}}$ is likely to be when M is a true model, and how the difference might be different when M is an incorrect model. Suppose that M^* is a correct model. As it turns out (see arguments in the next section), if M is also a correct model, an appropriate measure of the difference $\hat{Q}_M - \hat{Q}_{M^*}$ is its standard deviation, denoted by σ_{M,M^*}. On the other hand, if M is an incorrect model, the difference $\hat{Q}_M - \hat{Q}_{M^*}$ is expected to be much larger. This leads to the following procedure. For simplicity, let us first consider the case that $\tilde{M}$ is unique.

1. Find $\tilde{M}$ such that $\hat{Q}_{\tilde{M}} = \min_{M\in\mathcal{M}} \hat{Q}_M$. (See the remark following the definition of $\tilde{M}$.)

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

$$\hat{Q}_M \quad \leq \quad \hat{Q}_{\tilde{M}} + \hat{\sigma}_{M,\tilde{M}}. \tag{6}$$

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 quantity $\hat{Q}_{\tilde{M}} + \hat{\sigma}_{M,\tilde{M}}$ serves as a "fence" to confine the true models (with dimensions

smaller than $|\tilde{M}|$) and exclude the incorrect ones. For such a reason, the procedure is called *fence*. Note that the fence depends on M, i.e., for different M the fence is different.

2.2 The fence algorithm

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

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 (6) holds$\}$; if $\mathcal{M}_1 \neq \emptyset$, stop (no need for any more computation!). 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 (6) 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.

In case that $\tilde{M}$ is not unique, all one has to do is to redefine $\tilde{\mathcal{M}}$ in step 3 of fence as $\tilde{\mathcal{M}} = \{M \in \mathcal{M} : |M| = |\tilde{M}|, \hat{Q}_M = \hat{Q}_{\tilde{M}}\} \cup \tilde{\mathcal{M}}_-$.

Remark: The notion of model simplicity (or complexity) deserves further attention. Most generally, we refer to the *effective degrees of freedom* used in fitting a particular model. Ye (1998) uses the term generalized degrees of freedom (GDF) defined as the sum over data cased of the average sensitivity of changes in the fit of the estimated model mean to a small change in the

response, and thus measures the flexibility of a particular model or modeling procedure. Since this definition can literally apply to any type of model, GDF might not have a closed form expression but can be computed by simulation. Hodges and Sargent (1998) presented an effective degrees of freedom developed for hierarchical and other richly parameterized models, which for the case of linear mixed models and conditional on the random effect variances coincides with Ye's GDF.

2.3 Extension and variation

An extension of fence that takes into account the issue of consistency is given by the same steps 1-3 above with (6) replaced by

$$\hat{Q}_M \quad \leq \quad \hat{Q}_{\tilde{M}} + c_n \hat{\sigma}_{M,\tilde{M}}, \tag{7}$$

where c_n is a sequence that $\to \infty$ slowly as $n \to \infty$. A similar effective algorithm can be outlined.

It might appear that, like λ_n, the choice of c_n is also subjective. However, there are some major differences. In BIC, for example, the criterion is to choose a single model that minimizes (1). In other words, one has to be "exactly right", therefore the constant λ_n is important. In contrast, in fence one only needs to separate a subset of models. In other words, one only needs to be "about right", therefore the constant c_n is less important. Furthermore, the influence of c_n is not to the same extent as λ_n. To put it in a different way, the choice of λ_n is a first-order problem, while that of c_n is a second-order one. For example, typically, $\hat{Q}_M$, is of the order n. Thus, the order of λ_n in (5) is somewhere between $\sqrt{n}$ and n (see the discussion below (5)). On the other hand, the order of c_n in (7) is, essentially, that of $(\hat{Q}_M - \hat{Q}_{\tilde{M}})/\hat{\sigma}_{M,\tilde{M}}$, which is $O(1)$ if M is correct. In other words, the new procedure is less sensitive with respect to c_n than the previous ones to λ_n, which is confirmed by our simulation studies (see section 4). Nevertheless, in a finite sample situation the choice of c_n

may still make a difference. The issue of how to choose c_n will be addressed in section 6.

As mentioned, fence has the computational advantage that it starts with 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. For example, in quantitative trait loci (QTL) mapping, variance components arising from the trait genes, polygenic and environmental effects are often used to model the covariance structure of the phenotypes given the identity by descent (IBD) sharing matrix (e.g., Almsay and Blangero 1998). Such a model is usually complex due to the large number of putative trait loci. To make the fence procedure computationally more attractive to large and complex models, we propose the following variation of fence for situations of complex models with many predictors.

To be more specific, we focus on the extended GLMMs introduced earlier in this section. Let $X = (x_i')_{1\le i\le n}$ and $Z = (z_i')_{1\le i\le n}$. We assume that there is a collection of covariate vectors $X_1, \ldots, X_K$, from which the columns of X are to be selected. Furthermore, we assume that there is a collection of matrices $Z_1, \ldots, Z_L$ such that $Z\alpha = \sum_{s\in S} Z_s\alpha_s$, where $S \subset \{1, \ldots, L\}$, and each α_s is a vector of i.i.d. random effects with mean 0 and variance σ_s^2. The subset S is subject to selection. The parameters under an extended GLMM are the fixed effects and variances of the random effects. Note that in this case the full model corresponding to $X\beta + Z\alpha = \sum_{k=1}^{K} X_k\beta_k + \sum_{l=1}^{L} Z_l\alpha_l$ is among the candidate models. Thus, 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. We call such a variation of fence the forward-backward (F-B) fence.

The theoretical properties of fence and F-B fence will be explored in section 7, where consistency of both procedures will be established.

3 Estimation of σ_{M,M^*}

An important step of the fence method is the calculation of $\hat{\sigma}_{M,\tilde{M}}$. Although for consistency (see section 7) it is not required that $\hat{\sigma}_{M,M^*}$ be a consistent estimator of σ_{M,M^*}, as long as the former has the correct order, in practice, it is desirable to use a consistent estimator whenever possible. This is because, even if $\hat{\sigma}_{M,M^*}$ has the correct order, there is always a constant involved, which may be difficult to choose. A smaller constant is apparently to the benefit of larger models and thus results in overfitting; on the other hand, a larger constant would be in favor of smaller models, and hence prompts underfitting. Therefore, to balance the two sides, the best way would be to use a consistent estimator of σ_{M,M^*}, so that one can be less worried about the constant. Here consistency is in the sense that $\hat{\sigma}_{M,M^*} = \sigma_{M,M^*} + o(\sigma_{M,M^*})$ or, equivalently, $\hat{\sigma}_{M,M^*}/\sigma_{M,M^*} \to 1$, in a suitable sense (e. g., in probability). We first consider the case of clustered data.

3.1 Clustered observations

Clustered data arise naturally in many fields, including analysis of longitudinal data (e. g., Diggle *et al.* 1994) and small area estimation (e. g., Rao 2003). Let $y_i = (y_{ij})_{1\le j\le k_i}$ represent the vector of observations in the ith cluster, and $y = (y_i)_{1\le i\le m}$. We assume that $y_1, \ldots, y_m$ are independent. Examples of linear mixed models and GLMMs with clustered data are given in sections 4 and 5.

Furthermore, we assume that Q_M is *additive* in the sense that

$$Q_M \;=\; \sum_{i=1}^{m} Q_{M,i}, \tag{8}$$

where $Q_{M,i} = Q_{M,i}(y_i, \theta_M)$. We consider some examples.

Example 1. For ML model selection (see section 2), since $f_M(y|\theta_M) = \prod_{i=1}^m f_{M,i}(y_i|\theta_M)$ when the data is clustered, where $f_{M,i}(\cdot|\theta_M)$ is the joint pdf of y_i under M and θ_M, we have $Q_M = -\sum_{i=1}^m \log\{f_{M,i}(y_i|\theta_M)\}$. Thus, (8) holds with $Q_{M,i} = -\log\{f_{M,i}(y_i|\theta_M)\}$.

Example 2. Consider MVC model selection (see section 2). Let $T = \mathrm{diag}(T_1, \ldots . T_m)$, where T_i is $k_i \times s_i$ and $1 \le s_i \le k_i$, we have $Q_M = \sum_{i=1}^m |(T_i' V_{M,i}^{-1} T_i)^{-1} T_i' V_{M,i}^{-1}(y_i - \mu_{M,i})|^2$, where $\mu_{M,i}$ and $V_{M,i}$ are the mean vector and covariance matrix of y_i under M and θ_M. Thus, (8) holds with $Q_{M,i} = |(T_i' V_{M,i}^{-1} T_i)^{-1} T_i' V_{M,i}^{-1}(y_i - \mu_{M,i})|^2$.

Example 3. Note that the Q_M defined for extended GLMM selection (see section 2) always satisfies (8), even if the data is not clustered.

Denote, with a little abuse of the notation, the minimizer of $\mathrm{E}(Q_M)$ over $\theta_M \in \Theta_M$ by θ_M. Let M^* denote a correct model. We give approximations to $\mathrm{E}(\hat{Q}_M - \hat{Q}_{M^*})^2$ in two different situations.

Lemma 2. Suppose that the following regularity conditions are satisfied: i) $\mathrm{E}(\partial Q_M/\partial\theta_M) = 0$, and $\mathrm{tr}\{\mathrm{Var}(\partial Q_{M,i}/\partial\theta_M)\} \le c$ for some constant c; ii) there is a constant B_M such that $Q_M(\tilde{\theta}_M) > Q_M(\theta_M)$, if $|\tilde{\theta}_M| > B_M$; iii) there are constants $c_j > 0$, $j = 1, 2, 3$ such that $\mathrm{E}(|\hat{\theta}_M - \theta_M|^8) \le$

$c_1 m^{-4}$, $\mathrm{E}(|\partial Q_M/\partial\theta_M|^4) \le c_2 m^2$, and $\mathrm{E}(\sup_{|\tilde{\theta}_M|\le B_M} \|\partial^2 \tilde{Q}_M/\partial\theta_M\partial\theta_M'\|^4) \le c_3 m^4$; iv) there are constants $a, b > 0$ such that $am \le \mathrm{var}(Q_M - Q_{M^*}) \le bm$, if $M \ne M^*$; v) for any incorrect model M, we have $\mathrm{E}(Q_M - Q_{M^*}) = O(m)$. Then, we have $\mathrm{E}(\hat{Q}_M - \hat{Q}_{M^*}) = O(1)$, $\mathrm{var}(\hat{Q}_M - \hat{Q}_{M^*}) = \mathrm{var}(Q_M - Q_{M^*})\{1 + o(1)\} = O(m)$, if M is correct; and $\mathrm{E}(\hat{Q}_M - \hat{Q}_{M^*})^2 = \mathrm{var}(Q_M - Q_{M^*}) + O(m^2) = O(m^2)$, if M is incorrect.

The proof is given in subsection 9.1. Note that i) is satisfied if $\mathrm{E}(Q_M)$ can be differentiated inside the expectation, that is, $\partial \mathrm{E}(Q_M)/\partial\theta_M = \mathrm{E}(\partial Q_M/\partial\theta_M)$. Also note that ii) implies that $|\hat{\theta}_M| \le B_M$. Since a measure of the difference $\hat{Q}_M - \hat{Q}_{M^*}$ is $\sqrt{\mathrm{E}(\hat{Q}_M - \hat{Q}_{M^*})^2}$, Lemma 2 suggests a difference between a true model and an incorrect one: If M is a true model, $\hat{Q}_M - \hat{Q}_{M^*}$ may be measured by $\sigma_{M,M^*} = \mathrm{sd}(\hat{Q}_M - \hat{Q}_{M^*}) \approx \mathrm{sd}(Q_M - Q_{M^*})$; otherwise, $\hat{Q}_M - \hat{Q}_{M^*}$ is expected to be much larger since $\mathrm{sd}(Q_M - Q_{M^*}) = O(\sqrt{m})$.

It is not difficult to obtain an estimator of σ_{M,M^*}. By (8) and independence, it is easy to show that $\mathrm{var}(Q_M - Q_{M^*}) = \mathrm{E}[\sum_{i=1}^m (Q_{M,i} - Q_{M^*,i})^2 - \sum_{i=1}^m \{\mathrm{E}(Q_{M,i}) - \mathrm{E}(Q_{M^*,i})\}^2]$. Thus, an estimator of σ^2_{M,M^*} is the *observed variance* given by

$$\hat{\sigma}^2_{M,M^*} = \sum_{i=1}^m (\hat{Q}_{M,i} - \hat{Q}_{M^*,i})^2 - \sum_{i=1}^m \{\hat{\mathrm{E}}(Q_{M,i}) - \hat{\mathrm{E}}(Q_{M^*,i})\}^2, \tag{9}$$

where $\hat{Q}_{M,i} = Q_{M,i}(y_i, \hat{\theta}_M)$, $\hat{Q}_{M^*,i} = Q_{M^*,i}(y_i, \hat{\theta}_{M^*})$, $\hat{\mathrm{E}}(Q_{M,i}) = \mathrm{E}_{M^*,\hat{\theta}_{M^*}}\{Q_{M,i}(y_i, \hat{\theta}_M)\}$, and $\hat{\mathrm{E}}(Q_{M^*,i}) = \mathrm{E}_{M^*,\hat{\theta}_{M^*}}\{Q_{M^*,i}(y_i, \hat{\theta}_{M^*})\}$, in which the expectations are with respect to y_i under model M^* and evaluated at $\hat{\theta}_{M^*}$.

It should be pointed out that (9) only gives an estimator of σ^2_{M,M^*} in the most general situation. In some special cases there may be better ways of estimating σ^2_{M,M^*} that give more accurate results. See Example 4 in the sequel.

3.2 Non-clustered observations

We now consider the situations where the observations cannot be divided into independent clusters. Such data arise, for example, in linear mixed models and GLMMs with crossed random effects. We consider three such cases: Gaussian mixed models, non-Gaussian linear mixed models and extended GLMMs. Some examples are given in sections 4 and 5.

1. Gaussian mixed models. A Gaussian model is characterized by its mean vector μ_M and covariance matrix V_M, hence Gaussian model selection is all about selecting μ_M and V_M. A Gaussian mixed model can be expressed as $y = X\beta + Z\alpha + \epsilon$, where X is a matrix of known covariates, β is a vector of unknown fixed effects, Z is a known matrix, α is a vector of random effects, and ϵ is a vector of errors. It is assumed that α and ϵ are jointly normally distributed with $\mathrm{Var}(\alpha) = G$, $\mathrm{Var}(\epsilon) = R$ and $\mathrm{cov}(\alpha, \epsilon) = 0$, where G and R are the covariance matrices under the assumed model. It is clear that Gaussian mixed model is a special case of Gaussian model with $\mu_M = X_M\beta_M$ and $V_M = R_M + Z_M G_M Z_M'$, where X_M, β_M, Z_M, G_M and R_M are the corresponding matrices or vector under model M. Nevertheless, the result of this subsection applies to Gaussian models in general. Both ML and MVC methods (see section 2) apply to this case.

Lemma 3. For ML model selection, we have

$$\mathrm{var}(Q_M - Q_{M^*}) = \frac{1}{2}\mathrm{tr}\{(V_M^{-1}V_{M^*} - I)^2\} + (\mu_M - \mu_{M^*})'V_M^{-1}V_{M^*}V_M^{-1}(\mu_M - \mu_{M^*}).$$

For MVC model selection, we have

$$\begin{aligned}\mathrm{var}(Q_M - Q_{M^*}) &= 2\left(\mathrm{tr}\left[\{(T'V_M^{-1}T)^{-2}T'V_M^{-1}V_{M^*}V_M^{-1}T\}^2\right] - \mathrm{tr}\left\{(T'V_{M^*}^{-1}T)^{-2}\right\}\right) \\ &\quad +4(\mu_M - \mu_{M^*})'C_M V_{M^*} C_M(\mu_M - \mu_{M^*}),\end{aligned}$$

where $C_M = V_M^{-1}T(T'V_M^{-1}T)^{-2}T'V_M^{-1}$.

The proof follows directly from the covariance properties of multivariate normal distribution (e.g., Searle 1971, section 2.5). $\hat{\sigma}^2_{M,M^*}$ is then obtained by replacing μ_M, V_M, μ_{M^*} and V_{M^*} by $\hat{\mu}_M$, $\hat{V}_M$, $\hat{\mu}_{M^*}$ and $\hat{V}_{M^*}$, respectively., where $\hat{\mu}_M$ is μ_M with θ_M replaced by $\hat{\theta}_M$, etc.

2. Non-Gaussian linear mixed models. Consider a non-Gaussian linear mixed model (e. g., Jiang 1996). Since normality is not assumed, Lemma 3 is not valid. The main difference is that, unlike the Gaussian case, under a non-Gaussian linear mixed model, the expressions for $\mathrm{var}(Q_M - Q_{M^*})$ may involve higher (3rd and 4th) moments of the random effects and errors, which are not part of θ_M. As a result, estimators of these higher moments are not directly available. However, we can use a method known as partially observed information developed by Jiang (2005) to obtain an estimate of $\mathrm{var}(Q_M - Q_{M^*})$, hence $\hat{\sigma}^2_{M,M^*}$. The detail is omitted.

3. Extended GLMMs. Consider the Q_M introduced by (4). Write $\xi_{M,i} = g^2_{M,i}(\beta_M, \psi_M) - 2y_i g_{M,i}(\beta_M, \psi_M)$, $\xi_{M^*,i} = \xi_{M,i}$ with M replaced by M^*, and $d_i = \xi_{M,i} - \xi_{M^*,i}$. Also, let $\delta_i = g_{M,i}(\beta_M, \psi_M) - g_{M^*,i}(\beta_{M^*}, \psi_{M^*})$. Here θ_M represents the vector that minimizes $\mathrm{E}(Q_M)$ over Θ_M and θ_{M^*} the true parameter vector under M^*, a true model.

Lemma 4. Suppose that the following conditions are satisfied: i) $\mathrm{E}(y_i^2)$, $1 \le i \le n$ are bounded; and there is a sequence $a_n > 0$ such that $\mathrm{E}(|\hat{\theta}_M - \theta_M|^8) = O(a_n^{-4})$, $M \in \mathcal{M}$; ii) Condition ii) of Lemma 2; iii) $\xi_{M,i}$ is continuously differentiable with respect to θ_M, $1 \le i \le n$, and the following holds: $\mathrm{E}(\sup_{|\tilde{\theta}_M| \le B_M} \|\partial^2 \xi_{M,i}/\partial\theta_M \partial\theta'_M|_{\tilde{\theta}_M}\|^4) = O(1)$, $M \in \mathcal{M}$; iv) there is a constant $c > 0$ such that $\sum_{z_i'\Sigma z_j \ne 0} \delta_i \delta_j \mathrm{cov}(y_i, y_j) \ge c|S|$, where $S = \{(i,j) : z_i'\Sigma z_j \ne 0\}$, Σ is the true covariance matrix of α and $|A|$ the cardinality of A; Σ_{M^*} is positive definite and $z_i \ne 0$ for

any i; and $n^2/a_n^2|S| \to 0$, as $n \to \infty$. Then, we have $\sigma^2_{M,M^*} = \text{var}(\sum_{i=1}^n d_i)\{1+o(1)\}$ and

$$\text{var}\left(\sum_{i=1}^n d_i\right) = 4\left\{\text{E}\left(\sum_{i=1}^n \delta_i^2 y_i^2\right) + \sum_{i\neq j}\delta_i\delta_j g_{M^*,i,j}(\beta_{M^*},\psi_{M^*})1_{(z_i'\Sigma_{M^*}z_j\neq 0)} - \sum_{i,j}\delta_i\delta_j g_{M^*,i}(\beta_{M^*},\psi_{M^*})g_{M^*,j}(\beta_{M^*},\psi_{M^*})1_{(z_i'\Sigma_{M^*}z_j\neq 0)}\right\}, \quad (10)$$

where $g_{M^*,i,j}(\beta_{M^*},\psi_{M^*}) = \text{E}\{h_{M^*}(x_i'\beta_{M^*} + z_i'\Sigma_{M^*}^{1/2}\xi)h_{M^*}(x_j'\beta_{M^*} + z_j'\Sigma_{M^*}^{1/2}\xi)\}$, $\xi \sim N(0, I_m)$.

The proof is given in subsection 9.2. $\hat{\sigma}^2_{M,M^*}$ is then obtained as a *partially observed variance*:

$$\hat{\sigma}^2_{M,M^*} = 4\left\{\sum_{i=1}^n \hat{\delta}_i^2 y_i^2 + \sum_{i\neq j}\hat{\delta}_i\hat{\delta}_j g_{M^*,i,j}(\hat{\beta}_{M^*},\hat{\psi}_{M^*})1_{(z_i'\hat{\Sigma}_{M^*}z_j\neq 0)} - \sum_{i\neq j}\hat{\delta}_i\hat{\delta}_j g_{M^*,i}(\hat{\beta}_{M^*},\hat{\psi}_{M^*})g_{M^*,j}(\hat{\beta}_{M^*},\hat{\psi}_{M^*})1_{(z_i'\hat{\Sigma}_{M^*}z_j\neq 0)}\right\}, \quad (11)$$

where $\hat{\delta}_i$ is δ_i with β_M, ψ_M, β_{M^*} and ψ_{M^*} replaced by $\hat{\beta}_M$, $\hat{\psi}_M$, $\hat{\beta}_{M^*}$ and $\hat{\psi}_{M^*}$, respectively, and $\hat{\Sigma}_{M^*}$ is Σ_{M^*} with ψ_{M^*} replaced by $\hat{\psi}_{M^*}$.

4 Simulations

In this section, we study the performance of the fence methods through a number of simulated examples. These examples include linear mixed models and GLMMs, and are classified as clustered data and non-clustered data. Subsections 4.1, 4.2 are examples of clustered data, while subsection 4.3 is an example of non-clustered data.

4.1 Linear mixed models (clustered data)

We consider selection in the following linear mixed model (see Jiang and Rao 2003),

$$y_{ij} = x_{ij}'\beta + \alpha_i + \epsilon_{ij}\ , \quad (12)$$

$i = 1, \ldots, m$, $j = 1, \ldots, K$, where x_{ij} is a vector of covariates and β a vector of unknown regression 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 $\epsilon_i = (\epsilon_{ij})_{1 \le j \le K}$. Then, $\mathrm{Cov}(\epsilon_i, \epsilon_{i'}) = 0$ if $i \neq i'$, and $\mathrm{Var}(\epsilon_i) = \tau^2\{(1-\rho)I + \rho J\}$, where I is the identity matrix and J matrix of 1's. Finally, the random effects are uncorrelated with the errors.

We examine by simulation the probability of correct selection and also the overfitting and underfitting probabilities of various GIC's developed in Jiang and Rao (2003), which are similar to (1) for this problem. Two GIC's with different choices of λ_n are considered: (1) $\lambda_n = 2$, which corresponds to the C_p method; (2) $\lambda_n = \log n$ where $n = mK$ which corresponds to the BIC method. The latter choice satisfies the conditions of Theorem 1 in Jiang and Rao (2003) for consistent model selection for the case of a single random effect factor in the true underlying model with bounded cluster size, which includes the current case. A total of 100 realizations of each simulation were run. The first column of X is **1** and the other four columns of X are generated randomly from $N(0,1)$ distributions but are fixed throughout the simulation. Three β's are considered: $(2,0,0,4,0)$, $(2,9,0,4,8)$ and $(1,2,3,2,3)$.

We consider the case where the errors have varying degrees of exchangeable structure. Four values of ρ were considered: $0, 0.2, 0.5, 0.8$. The random effects and errors were simulated from Normal distributions with σ and τ both taken to be equal to 1. We set the number of clusters (m) to be 100 and the number of observations within a cluster to be $K = 5$. The ML fence method is applied for this simulation with $c_n = 1.1$ for all situations.

Summary: The results are presented in Table 1. The fence method has robust selection performance in most situations considered. In cases where the true model was relatively small in

dimension, the fence method suffers some from overfitting. The overfitting proneness in these few situations is less than that found when using C_p but more than that found when using BIC. Selection performance in the second situation with a larger true model with high signal is solid for the fence method. However, in the last situation with the optimal model being the full model with all weak covariates, both BIC and C_p tend to underfit. The fence method still shines having excellent performance with comparatively little or no underfitting empirically observed (note that overfitting is not possible in this situation since the true model is the model with the full complement of fixed effects). The effect of increasing correlation in the errors (i. e., clustering) is to act as a means of reducing effective sample size for selection. The end result is that as the correlation between observations within a cluster increases, selection performance for all methods degrades somewhat.

4.2 Generalized linear mixed models (clustered data)

Consider the following simulated example of GLMM selection with three candidate models.

Model I: Given the random effects $\alpha_1, \ldots, \alpha_m$, binary responses y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, k$ are conditionally independent such that, $\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_i + \alpha_i$, where $p_{ij} = \text{P}(y_{ij} = 1|\alpha)$; β_0, β_1 are fixed parameters; $x_i = 0$, $1 \leq i \leq [m/2]$ and $x_i = 1$, $[m/2] + 1 \leq i \leq m$ ($[x]$ is the integer part of x). The random effects are independent and distributed as $N(0, \sigma^2)$.

Model II: Same as Model I except that $\beta_1 = 0$.

Model III: Same as Model I except that $\beta_0 = \beta_1 = 0$.

We first study consistency of the MVC and ML model selection procedures in the situation where the data is generated from one of the candidate models. In other words, a true model belongs to the class of candidate models. Throughout the simulation, T was chosen as a block-diagonal matrix (see Example 2) with $T_i = T_1$, $1 \leq i \leq m$, where T_1 is a $k \times l$ matrix with $l = [k/2]$, whose

Table 1: **Simulation Results: Linear Mixed Model Selection.** *Reported are probabilities of correct selection(underfitting, overfitting) as percentages estimated empirically from 100 realizations of the simulation. C_p and BIC results for models 1 and 2 were taken from Jiang and Rao (2003).*

True Model	ρ	C_p	BIC	Fence (ML)
$\beta' = (2, 0, 0, 4, 0)$	0	64(0,36)	97(0,3)	94(0,6)
	0.2	57(0,43)	94(0,6)	91(0,9)
	0.5	58(0,42)	96(1,3)	86(0,14)
	0.8	61(0,39)	96(0,4)	72(0,28)
$\beta' = (2, 9, 0, 4, 8)$	0	87(0,13)	99(0,1)	100(0,0)
	0.2	87(0,13)	99(0,1)	100(0,0)
	0.5	80(0,20)	99(0,1)	99(0,1)
	0.8	78 (1,21)	96(1,3)	94(0,6)
$\beta' = (1, 2, 3, 2, 3)$	0	85(15,0)	81(19,0)	100(0,0)
	0.2	79(21,0)	73(27,0)	100(0,0)
	0.5	74(26,0)	64(36,0)	97(3,0)
	0.8	44(56,0)	26(74,0)	94(6,0)

Table 2: **Simulation Results: Consistency.** *The columns for MVC and ML are probabilities of correct selection, reported as percentages estimated empirically from 100 realizations of the simulation. The numbers in parentheses are the percentages of selection of the other two models in order of increasing index of the model.*

True Model	m	k	l	β_0	β_1	σ	c_n	MVC	ML
I	100	4	2	-.5	1	1	1	82(5,13)	94(3,3)
I	200	4	2	-.5	1	1	1.1	97(1,2)	99(0,1)
II	100	4	2	-.5	NA	1	1	87(4,9)	88(5,7)
II	200	4	2	-.5	NA	1	1.1	93(4,3)	98(2,0)
III	100	4	2	NA	NA	1	1	87(3,10)	91(2,7)
III	200	4	2	NA	NA	1	1.1	96(0,4)	91(1,8)

entries are generated from a Uniform$[0, 1]$ distribution, and then fixed. The simulation results are summarized in Table 2, with each result based on 100 simulations.

We next study robustness of the MVC and ML fence procedures in the case where no true model (with respect to ML) is among the candidate models. We consider such a case, in which the binary responses y_{ij} are generated as follows. Suppose that $(X_1, \ldots, X_k)$ has a multivariate normal distribution such that $\mathrm{E}(X_j) = \mu$, $\mathrm{var}(X_j) = 1$, $1 \le j \le k$ and $\mathrm{cor}(X_s, X_t) = \rho$, $1 \le s \ne t \le k$. Then, let $Y_j = 1_{(X_j>0)}$, $1 \le j \le k$. Denote the joint distribution of $(Y_1, \ldots, Y_k)$ by NB(μ, ρ) (here NB refers to "Normal-Bernoulli"). We then generate the data such that $y_1, \ldots, y_m$ are independent, and the distribution of $y_i = (y_{ij})_{1 \le j \le k}$ follows one of the following models.

Model A: $y_i \sim \mathrm{NB}(\mu_1, \rho_1)$, $i = 1, \ldots, [m/2]$, and $y_i \sim \mathrm{NB}(\mu_2, \rho_2)$, $i = [m/2] + 1, \ldots, m$, where μ_j, ρ_j, $j = 1, 2$ are chosen to match the means, variances and covariances under Model I.

Note that one can do so because the means, variances and covariances under Model I depend only on three parameters, while there are four parameters under Model A.

Model B: $y_i \sim \mathrm{NB}(\mu, \rho)$, $i = 1, \ldots, m$, where μ and ρ are chosen to match the mean, variance and covariance under Model II. Note that, under Model II, the mean, variance and covariance depend on two parameters.

Model C: Same as Model B except that μ and ρ are chosen to match the mean, variance and covariance under Model III. Note that, under Model III, the mean is equal to $1/2$, the variance is $1/4$, while the covariance depends on a single parameter σ.

If the data is generated from Model A, Model I is a correct model with respect to MVC; similarly, if the data is generated from Model B, both Model I and II are correct with respect to MVC; and, if the data is generated from Model C, Models I - III are all correct in the sense of MVC. However, no model (I, II or III) is correct from an ML standpoint. The simulation results are summarized in Table 3, in which β_0^*, β_1^* and σ^* correspond to the parameters under the models in Table 2 with the matching mean(s), variance(s) and covariance(s). Again, each result is based on 100 simulations.

Summary: It is seen in Table 2 and Table 3 that the numbers increase as m increases (and c_n slowly increases), a good indication of consistency. In Table 2, with the exception of one case (III/200), ML outperforms MVC, which is not surprising. What is a bit of surprise is that ML also seems quite robust in the situation where the true model is not among the candidate models (therefore the objective is to select a candidate model that is closest to the reality). In fact, Table 3 shows that even in the latter case, ML still outperforms MVC (again with the exception of one case - III/200). However, one has to keep in mind that there are many ways of model misspecification, and here we only considered one of them (which misspecifies a NB as a GLMM). Furthermore, MVC has computational advantage over ML, which is important in cases such as GLMM selection. Note

Table 3: **Simulation Results: Robustness.** *The columns for MVC and ML are probabilities of correct selection, reported as percentages estimated empirically from 100 realizations of the simulation. The numbers in parentheses are the percentages of selection of the other two models in order of increasing index of the model.* β_0^*, β_1^* *and* σ^* *are the matching parameters.*

True Model	m	k	l	β_0^*	β_1^*	σ^*	c_n	MVC	ML
A	100	4	2	-.5	1	1	1	83(7,10)	91(5,4)
A	200	4	2	-.5	1	1	1.1	97(2,1)	99(0,1)
B	100	4	2	-.5	NA	1	1	80(3,17)	91(4,5)
B	200	4	2	-.5	NA	1	1.1	95(3,2)	97(3,0)
C	100	4	2	NA	NA	1	1	83(8,9)	86(4,10)
C	200	4	2	NA	NA	1	1.1	91(1,8)	90(1,9)

that the computational burden usually increases with the sample size; on the other hand, the larger sample performance of MVC ($m = 200$) is quite close to that of ML.

A compromise would be to use MVC in cases of large sample, and ML in cases of small or moderate sample. Alternatively, one may use MVC for an initial round of model selection to narrow down the number of candidate models, and ML for a final round of model selection. For example, one may use MVC for steps 1 and 2 of fence (see section 2) to identify the subclass $\tilde{\mathfrak{M}}$, and then apply ML (with steps 1 - 3) within $\tilde{\mathfrak{M}}$ to identify the optimal model.

4.3 Gaussian mixed model selection (non-clustered data)

We consider the problem of selecting a Gaussian linear mixed model for non-clustered observations. There are three candidate models. These are:

Model I. $y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + v_j + e_{ij}$, $i = 1, \ldots, a$, $j = 1, \ldots, b$, where β_0 and β_1 are unknown coefficients, u_i, v_j are random effects, and e_{ij} is an error. It is assumed that u_i's, v_j's and e_{ij}'s are independent with $u_i \sim N(0, \sigma_1^2)$, $v_j \sim N(0, \sigma_2^2)$ and $e_{ij} \sim N(0, \sigma_0^2)$.

Model II. $y_{ij} = \beta_0 + u_i + v_j + e_{ij}$, where everything is the same as in Model I.

Model III. $y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + e_{ij}$, where everything is the same as in Model I.

In the simulation, the x_{ij}'s are generated from a Poisson(1) distribution and, once generated, fixed throughout the simulation.

We consider the fence ML model selection (see section 2), which seems to be the natural choice in this case. We consider fence without c_n (or $c_n = 1$). Four sample size configurations are considered: (i) $a = b = 10$; (ii) $a = 10, b = 20$; (iii) $a = 20, b = 10$; and (iv) $a = b = 20$. Note that the effective sample sizes here are a and b, not the product ab, so these correspond to situations of relatively small sample size. For each sample size configuration, three cases are considered. In the first case, the data is generated under Model I with the following true parameters: $\beta_0 = 0.5$, $\beta_1 = 0.2$, $\sigma_j^2 = 1.0$, $j = 0, 1, 2$. In this case, Model I is the only true model and therefore the optimal model. In the second case, the data is generated under Model II with the following true parameters: $\beta_0 = 0.5$, $\sigma_j^2 = 1.0$, $j = 0, 1, 2$. In this case, Model I and Model II are both true models with Model II being the optimal model. In the third case, the data is generated under Model III with the following true parameters: $\beta_0 = 0.5$, $\beta_1 = 0.2$, $\sigma_j^2 = 1.0$, $j = 0, 2$. In this case, Model I and Model III are both true models with Model III being the optimal model.

Summary: For each combination of sample size configuration and case, 100 simulations were run. Table 4 reports the percentages of simulations (out of the 100) in which fence has selected the optimal model. For comparison purposes, the method of Jiang and Rao (2003) (Case 2) was also run for each setting. Their method is based on minimizing an information criterion which

Table 4: **Gaussian Model Selection.** *Reported are probabilities of correct selection as percentages estimated empirically from 100 realizations of the simulation. Table entries correspond to the fence method with* $c_n = 1$*, and the method of Jiang and Rao (2003) Case 2 using* $\lambda_{t,n} = 2, log(n)$ *and* $n/log(n)$ *respectively in parentheses.*

Optimal Model	$a = b = 10$	$a = 10, b = 20$	$a = 20, b = 10$	$a = b = 20$
Model I	34 (35, 14, 0)	92 (67, 42, 0)	85 (64, 34, 0)	97 (87, 71, 0)
Model II	97 (31, 25, 0)	80 (56, 52, 0)	79 (56, 63, 0)	96 (63, 82, 0)
Model III	92 (38, 27, 0)	98 (74, 53, 0)	98 (46, 38, 0)	99 (71, 81, 0)

trades off a goodness-of-fit measure with a (penalized) model complexity term. Consistency of selection was proved by imposing specific requirements on the penalty term. In this simulation, three different penalty terms were entertained: $\lambda_{t,n} = 2, log(n), n/log(n)$. The last two of these satisfy the conditions for consistency. Note that only the empirical percentages of correct selection of both random and fixed effects is presented in Table 4. Clearly, there are many types of potential selection errors that can be made. These will be discussed model by model in turn. Models I and II represent situations where the true model includes the full complement of random effects but Model II includes only the intercept fixed effect term. What is evident with the Jiang and Rao (2003) method using $\lambda_{t,n} = 2$, is that selection performance for Model I tends to be uniformly better than that for Model II across all settings of a and b. This is because overfitting of random effects is not an issue and underfitting of random effects structure vanishes quite quickly in a or b - even at these smallish sample sizes. However with Model II, selection performance degrades across all settings of a and b. This can actually be attributed to overfitting in the fixed effects part of the model. Contrast this to the performance using $\lambda_{t,n} = n/log(n)$. Here the misses can be attributed

to underfitting primarily in the random effects and to a lesser extent in the fixed effects. Clearly, meeting consistency requirements has not translated into good finite sample performance. Model III represents a situation with the full complement of fixed effects but only one of the random effects related to a. Here again performance of the Jiang and Rao (2003) methods is not much improved. With $\lambda_{t,n} = 2$, the sensitivity to overfitting (in the random effects) starts to become a little apparent. With the other choices, underfitting in the fixed effects is still an issue due to the small signal to noise ratio in this simulation. A synopsis of these three Models under these four settings leads one to conclude that the choice of the penalty term makes a difference on selection performance, a point we made earlier in section 1, and how this plays out really depends on the underlying true model.

What is lovely about running these comparisons is that it helps to illuminate the robustness of the fence method. It is seen that, despite the relatively small sample size, the low signal to noise ratio, and the variety of potential selection errors, the performance of fence is quite good in all but one case. The exception occurs when $a = b = 10$ and data is generated from Model I. A closer look at this case reveals that all the misses went to Model II, which has the same random effect factors but no covariates (i.e., $\beta_1 = 0$). Some possible explanations are: (1) weak signal/noise ratio (note that the true $\beta_1 = 0.2$, while all three variance components are equal to 1.0); (2) small sample size. In this case, $\sigma_{M,\tilde{M}}$ is estimated using the Gaussian formula derived in subsection 3.2.1. Since all the variance components are involved in this formula, they have to be estimated. As mentioned, the effective sample size for estimating σ_1^2 is $a = 10$, and that for estimating σ_2^2 is $b = 10$. With such small sample sizes, these estimators are not expected to be accurate.

5 Real data analyses

In this section, we give a number of examples, each supported by results of real data analysis, to illustrate the application of fence to various problems of mixed model selection. As in the previous section, the examples are classified as clustered data (subsections 5.1 and 5.2) and non-clustered data (subsection 5.3).

5.1 Analysis of Gc genotype data

Human group-specific component (Gc) is the plasma transport protein for Vitamin D. Polymorphic electrophoretic variants of Gc are found in all human populations. Daiger *et al.* (1984) presented data involving a series of monozygotic (MZ) and dizygotic (DZ) twins of known Gc genotypes to determine the heritability of quantitative variation in Gc. These included 31 MZ twin pairs, 13 DZ twin pairs, and 45 unrelated controls. For each individual, the concentration of Gc was available along with additional information about the sex, age and Gc genotype of the individual. The genotypes are distinguishable at the Gc structural locus, classified as 1-1, 1-2 and 2-2.

Lange (2002) considered three statistical models for the Gc genotype data. Let y_{ij} represent the Gc concentration measured for the jth person who is one of the ith identical twin pair, $i = 1, \ldots, 31$, $j = 1, 2$. Furthermore, let y_{ij} represent the Gc concentration measured for the jth person who is one of the $(i - 31)$th fraternal twin pairs, $i = 32, \ldots, 44$, $j = 1, 2$. Finally, Let y_i represent the Gc concentration for the $(i - 44)$th person among the unrelated controls, $i = 45, \ldots, 89$. Then, the first model, Model I, can be expressed as $y_{ij} = \mu_{1-1} 1_{(g_{ij}=1-1)} + \mu_{1-2} 1_{(g_{ij}=1-2)} + \mu_{2-2} 1_{(g_{ij}=2-2)} + \mu_{\text{male}} 1_{(s_{ij}=\text{male})} + \mu_{\text{age}} a_{ij} + \epsilon_{ij}$, $i = 1, \ldots, 44$, $j = 1, 2$, where g_{ij}, s_{ij} and a_{ij} represent the genotype, sex and age of the jth person in the i twin pair (identical or fraternal), and ϵ_{ij} is an

error which will be further specified later. If we let x_{ij} denote the vector whose components are $1_{(g_{ij}=1-1)}$, $1_{(g_{ij}=1-2)}$, $1_{(g_{ij}=2-2)}$, $1_{(s_{ij}=\text{male})}$ and a_{ij}, and β denote the vector whose components are μ_{1-1}, μ_{1-2}, μ_{2-2}, μ_{male} and μ_{age}, then the model can be expressed as $y_{ij} = x'_{ij}\beta + \epsilon_{ij}, \quad i = 1, \ldots, 44, j = 1, 2$. Similarly, we have $y_i = \mu_{1-1}1_{(g_i=1-1)} + \mu_{1-2}1_{(g_i=1-2)} + \mu_{2-2}1_{(g_i=2-2)} + \mu_{\text{male}}1_{(s_i=\text{male})} + \mu_{\text{age}}a_i + \epsilon_i$, $i = 45, \ldots, 89$, where g_i, s_i and a_i are the genotype, sex and age of the $(i-44)$th person in the unrelated control group, and ϵ_i is an error which will be further specified. Let x_i denote the vector whose components are $1_{(g_i=1-1)}$, $1_{(g_i=1-2)}$, $1_{(g_i=2-2)}$, $1_{(s_i=\text{male})}$ and a_i, and β be the same as above, then we have $y_i = x'_i\beta + \epsilon_i, \quad i = 45, \ldots, 89$.

We now specify the distributions for the errors. Let $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2})'$, $i = 1, \ldots, 44$. We assume that ϵ_i, $i = 1, \ldots, 89$ are independent. Furthermore, we assume that ϵ_i is bivariate normal with means zero, variance σ^2_{tot} and correlation coefficient ρ_{ident}, $i = 1, \ldots, 31$, where σ^2_{tot} is the unknown total variance, and ρ_{ident} the unknown correlation coefficient between identical twins. Similarly, we assume that ϵ_i is bivariate normal with means zero, variance σ^2_{tot} and correlation coefficient ρ_{frat}, $i = 32, \ldots, 44$, where ρ_{frat} is the unknown correlation coefficient between fraternal twins. Finally, we assume that $\epsilon_i \sim N(0, \sigma^2_{\text{tot}})$, $i = 45, \ldots, 89$.

The second model, Model II, is the same as Model I except under the constraint $\rho_{\text{frat}} = \rho_{\text{ident}}/2$. The third model, Model III, is the same as Model I except under the constraints $\mu_{1-1} = \mu_{1-2} = \mu_{2-2}$. It is clear that all three models are Gaussian models. We apply the fence method to this dataset to select an optimal model from the candidate models. More specifically, we consider ML model selection (see section 2) with $c_n = 1$. Note that, since Model II and III are submodels of Model I (in other words, Model I is the full model), we may take $\tilde{M}$ as Model I. The analysis resulted in the following values for $\hat{Q}_M$: $\hat{Q}_{\text{I}} = 337.777$, $\hat{Q}_{\text{II}} = 338.320$ and $\hat{Q}_{\text{III}} = 352.471$. Furthermore, we obtained $\hat{\sigma}_{\text{II,I}} = 1.367$ and $\hat{\sigma}_{\text{III,I}} = 4.899$. Thus, Model II is in the fence while Model III is out. In

conclusion, the analysis has selected Model II as the optimal model. This result is consistent with the finding of Lange (2002), who indicated that a "likelihood ratio test shows that there is virtually no evidence against the assumption $\rho_{\text{frat}} = \rho_{\text{ident}}/2$."

5.2 Prenatal care for pregnancy

This real-data example is an application of the F-B fence procedure (see section 2). Rodriguez and Goldman (2001) considered a dataset from a survey conducted in Guatemala regarding the use of modern prenatal care for pregnancies where some form of care was used (Pebley *et al.* 1996). While Rodriguez and Goldman focused on assessing the performance of the approximation method they developed in fitting a three-level variance component logistic model, we consider applying the fence method in selection of the fixed covariates in the variance component logistic model. The models are described as follows.

Suppose that given the random effects at community levels u_i, $1 \leq i \leq m$ and random effects at family levels v_{ij}, $1 \leq i \leq m$, $1 \leq j \leq n_i$, binary responses y_{ijk}, $1 \leq i \leq m$, $1 \leq j \leq n_i$, $1 \leq k \leq n_{ij}$ are conditionally independent with $\pi_{ijk} = \mathrm{E}(y_{ijk}|u,v) = \mathrm{P}(y_{ijk} = 1|u,v)$. Furthermore, suppose that the random effects are independent with $u_i \sim N(0,\sigma^2)$ and $v_{ij} \sim N(0,\tau^2)$. The following models for the conditional means are considered such that under model M, $\mathrm{logit}(\pi_{ijk}) = X'_{M,ijk}\beta_M + u_i + v_{ij}$, where $X_{M,ijk}$ is a subvector of the full set of fixed covariates and β_M the corresponding vector of regression coefficients.

Let $\psi = (\sigma^2, \tau^2)'$. The vector of parameters under model M is $\theta_M = (\beta'_M, \psi')'$. Define

$$Q_M = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{ij}}\{y_{ijk} - g_{M,ijk}(\theta_M)\}^2, \tag{13}$$

where $g_{M,ijk}(\theta_M) = \mathrm{E}\{h(X'_{M,ijk}\beta_M + u_i + v_{ij})\}$ and $h(x) = e^x/(1+e^x)$. Using the method devel-

oped in subsection 3.2.3, an estimate of σ^2_{M,M^*} can be obtained (detail omitted). The expectations involved in Q_M are evaluated by numerical integration. Since the number of covariates considered is quite large, to keep the computational time manageable we apply the F-B fence procedure introduced in section 2 (with $c_n = 1$).

The data analysis has selected the following variables (in the order that they were selected in the forward procedure): Proportion indigenous (1981), Modern toilet in household, Husband's education secondary or better, Husband's education primary, Television watched daily, Distance to nearest clinic, Mother's education primary, Television not watched daily, Mother's education secondary or better, Indigenous (no Spanish), Indigenous (Spanish), Mother age, Husband agriculture employee, Husband agriculture self-employee, Child age, Birth order 4-6, and Husband's education missing. There are some interesting differences between the fixed effects discovered by the fence versus those found by standard maximum likelihood analysis using a 5% significance level as reported in Rodriguez and Goldman (2001). First, Husband's education overall (primary or higher relative to the reference group of no education for the husband) was found to be an important predictor whereas Rodriguez and Goldman found that only Husband's secondary education was important. Our more uniform finding is also in line with the finding for Mother's education. The implication is that education of some kind is important for both the mother and husband to have. A similar kind of finding was observed for variables corresponding to husband's profession. We found that regardless of what type of agricultural employment the husband had, it was an important predictor overall. Rodriguez and Goldman report that only non-self employed agricultural jobs for the husband mattered. The fence method also uniquely found that watching television (daily or not) was an important predictor. This can be intuitively justified since it provides a medium for women to learn more about modern prenatal health care methods and thus make it more likely for them to choose to use such methods.

Other findings were in line with those of Rodriguez and Goldman.

5.3 Modeling the salamander-mating experiments (non-clustered data)

Finally, we consider the well-known salamander-mating data originally reported by McCullagh and Nelder (1989, section 14.5). The data was collected from mating experiments involving two populations of salamanders, Rough Butt (RB) and White Side (WS). These populations, which are geographically isolated from one another, are found in the southern Appalachian mountains of the eastern United States. The question whether the geographic isolation had created barriers to the animals' interbreeding was thus of great interest to biologists studying speciation.

The data was studied extensively by fitting GLMMs (e.g., Breslow and Clayton 1993, Drum and McCullagh 1993, Lin and Breslow 1996, Jiang 1998 and Booth and Hobert 1999). However, in most studies it has been assumed that a different group of animals (20 for each sex) are used in each mating experiment, although, in reality, the same group of animals were repeatedly used in two of the three experiments. The GLMMs used in these studies assumed that no further correlation among the data exists given the random effects. However, the responses in this case should be considered longitudinal, because repeated measures were collected from the same subjects (once in the summer and once in the fall). Therefore, serial correlation may still exist among the repeated responses given the random effects. Alternatively, one could pool the responses from the two experiments involving the same group of animals, as suggested by McCullagh and Nelder (1989, section 4.1), so let $y_{ij\cdot} = y_{ij1} + y_{ij2}$, where y_{ij1} and y_{ij2} represent the responses from the summer and first fall experiments, respectively, that involved the same (ith) female and (jth) male. This avoids the issue of conditional independence, but brings in a new problem: The pooled response $y_{ij\cdot}$ may not be *binomial* given the random effects.

In general, pooling the responses from the repeated measures over time will maintain conditional independence, but may destroy the (conditional) exponential family, another key assumption of GLMM. To address such concerns, Jiang and Zhang (2001) proposed an extended version of GLMM, in which the (conditional) exponential family assumption is dropped. The authors considered two models for the conditional means, with *logit* or *probit* links, respectively, which correspond to models I and III below, and fitted both models to the data. Following the latter approach, we pool the data from the two experiments involving the same group of salamanders, so let y_{ij1} be the observed proportion of successful matings between the ith female and jth male in the two experiments. Let y_{ij2} be the indicator of successful mating between the ith female and jth male in the last experiment involving a new set of animals.

We assume that given the random effects, $u_{k,i}$, $v_{k,j}$, $k = 1, 2$, $i, j = 1, \ldots, 20$, which are independent and normally distributed with mean 0 and variances σ^2 and τ^2, respectively, the responses y_{ijk}, $(i, j) \in P$, $k = 1, 2$ are conditionally independent, where P represents the set of pairs (i, j) determined by the design, which is partially crossed; u and v represent the female and male, respectively; $1, \ldots, 10$ correspond to RB, and $11, \ldots, 20$ to WS. Furthermore, we consider the following models for the conditional means.

Model I: $\mathrm{E}(y_{ijk}|u, v) = h_1(\beta_0 + \beta_1 \mathrm{WS_f} + \beta_2 \mathrm{WS_m} + \beta_3 \mathrm{WS_f} \times \mathrm{WS_m} + u_{k,i} + v_{k,j})$, $(i, j) \in P$, $k = 1, 2$, where $h_1(x) = e^x/(1 + e^x)$; $\mathrm{WS_f}$ is an indicator for WS female (1 for WS and 0 for RB), $\mathrm{WS_m}$ is an indicator for WS male and $\mathrm{WS_f} \times \mathrm{WS_m}$ represents the interaction.

Model II: Same as Model I except dropping the interaction term.

Model III: Same as Model I with h_1 replaced by h_2, where $h_2(x) = \Phi(x)$, the cdf of $N(0, 1)$.

Model IV: Same as Model III except dropping the interaction term.

The models are special cases of the extended GLMMs introduced in section 2 (also see sub-

section 3.2.3). We apply the fence method therein (with $c_n = 1$) to this case. The analysis has yielded the following values of $\hat{Q}_M$ for $M =$ I, II, III and IV: $39.5292, 44.3782, 39.5292, 41.6190$, hence $\tilde{M} =$ I or III. If $\tilde{M} =$ I, then $\hat{\sigma}_{M,\tilde{M}} = 1.7748$ for $M =$ II and $\hat{\sigma}_{M,\tilde{M}} = 1.1525$ for $M =$ IV. Therefore, neither $M =$ II nor $M =$ IV fall within the fence. If $\tilde{M} =$ III, then $\hat{\sigma}_{M,\tilde{M}} = 1.68$ for $M =$ II and $\hat{\sigma}_{M,\tilde{M}} = 1.3795$ for $M =$ IV. Thus, once again, neither $M =$ II nor $M =$ IV are inside the fence. In conclusion, the fence method has selected both Model I and Model III (either one) as the optimal model. Interestingly, these are exactly the ones fitted by Jiang and Zhang (2001) using a different method, although the authors had not considered it a model selection problem. The eliminations of Model II and Model IV are consistent with many of the previous studies (e.g., Karim and Zeger 1992, Breslow and Clayton 1993, Lin and Breslow 1996), which have found the interaction term significant, although the majority of these studies have focused on logit models.

6 Adaptive fence procedure

In this section we address the issue of choosing the tuning constant c_n involved in (7). According to Theorem 1 in the sequel, for consistency of the fence one needs $c_n \to \infty$ at a certain rate, but there are many c_n's that satisfy this requirement. Also note that although for the consistency it is not required that $\hat{\sigma}_{M,M^*}$ be a consistent estimator of σ_{M,M^*} as long as it has the right order (see the first paragraph of section 3), there is always a constant involved which may make a difference in a finite sample situation. Therefore, the focus here is finite sample performance.

We now introduce the idea of an adaptive 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 $M_{\rm f} \in \mathcal{M}$, hence $\tilde{M} = M_{\rm f}$ in (7); and that every model in $\mathcal{M} \backslash \{M_{\rm f}\}$ is a submodel of a model in $\mathcal{M}$ with one

less parameter than $M_{\rm f}$. 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

$$P \quad = \quad \mathrm{P}(M_0 = M_{\rm opt}), \tag{14}$$

where $M_{\rm opt}$ 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 (14): (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 distribution P on the right side of (14).

To solve problem (ii), we use the idea of maximum likelihood. Namely, let $p^*(M) = \mathrm{P}^*(M_0 = M)$, where $M \in \mathcal{M}$ and 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.

Here is another look at the method. Typically, the optimal model is the model from which the data is generated, then this model should be the most likely given the data. Thus, given c_n, one is looking for the model (using the fence procedure) that is most supported by the data or, in other

words, one that has the highest (posterior) probability. The latter is estimated by a bootstrapping procedure. Note that although the bootstrap samples are generated under the full model, they are almost the same as those generated under the optimal model. This is because the estimates corresponding to the zero parameters are expected to be close to zero, provided that the parameter estimators under the full model are consistent. One then pulls off the c_n that maximizes the (posterior) probability and this is the optimal choice, denoted by c_n^*.

The procedure does not work, however, if there is no peak in the middle. Typically, this happens when the optimal model is one of the extreme cases - either $M_{\rm f}$ or M_*. To handle such cases we run screen tests for the extreme cases before searching for the peak in the middle. The first is called full model test. The idea is the following. Define $\mathcal{M}_{\rm f-1}$ as the set of all models with one less parameter than $M_{\rm f}$ (see above). Suppose that when $M_{\rm f}$ is the optimal model, we have $\mathrm{E}(\hat{Q}_M - \hat{Q}_{M_{\rm f}}) \sim a_n, \forall M \in \mathcal{M}_{\rm f-1}$. Here $u_n \sim v_n$ means that both u_n/v_n and v_n/u_n are bounded. On the other hand, if $M_{\rm f}$ is not the optimal model, there is $M \in \mathcal{M}_{\rm f-1}$ which is a true model, hence $\mathrm{E}(\hat{Q}_M - \hat{Q}_{M_{\rm f}}) = O(b_n)$, where $b_n = o(a_n)$. It follows that $\min_{M\in\mathcal{M}_{\rm f-1}} \mathrm{E}(\hat{Q}_M - \hat{Q}_{M_{\rm f}}) = O(b_n)$. Therefore, we consider

$$q_n \quad = \quad \frac{\{\min_{M\in\mathcal{M}_{\rm f-1}} \mathrm{E}(\hat{Q}_M - \hat{Q}_{M_{\rm f}})\}^2}{a_n b_n}. \tag{15}$$

In practice, q_n is replaced by its bootstrap estimate, q_n^*, obtained as above. If $q_n^* < 1$, the full model test passes; otherwise, the full model test fails, in which case we assign $c_n^* = 0$. In case the full model test passes, we follow with a minimum model test. For simplicity, we assume that there is a unique $M_* \in \mathcal{M}$ that has the minimum dimension. Note that this is not a serious restriction because in most cases one can add a (trivial) model to $\mathcal{M}$, if necessary, which then becomes the unique M_*. Suppose that $\mathrm{E}(\hat{Q}_{M_*} - \hat{Q}_{M_{\rm f}}) = O(g_n)$ if M_* is incorrect; and the order becomes $O(h_n)$ if M_* is

correct (hence optimal), where $h_n = o(g_n)$. We then consider

$$r_n \quad = \quad \frac{\{\mathrm{E}(\hat{Q}_{M_*} - \hat{Q}_{M_\mathrm{f}})\}^2}{g_n h_n}. \tag{16}$$

Let r_n^* be the bootstrap version of r_n. If $r_n^* > 1$ the minimum model test passes; otherwise, the minimum model test fails, in which case we assign c_n^* as the upper bound of a sequence of values considered (see below). In case both tests pass, we start searching for the peak in the middle. Quite often there are more than one c_n's at which p^* reaches the peak. Let c_n^* be the median of those c_n's.

The last thing one needs to determine is at which values of c_n to evaluate p^*. Theoretically, the range of c_n is $[0, \infty)$, but practically one needs an upper bound. This can be determined as follows. Note that any c_n greater than or equal to $B = (\hat{Q}_{M_*} - \hat{Q}_{M_\mathrm{f}})/\hat{\sigma}_{M_*,M_\mathrm{f}}$ makes no difference to the fence procedure. This is because then (7) is satisfied by M_*, hence $M_0 = M_*$ (recall that M_* is unique by our simplicity assumption). Therefore, we choose the upper bound of c_n as the smallest integer $\geq B$, i.e., $B^* = [B] + 1$. We then divide the interval $[0, B^*]$ by subintervals of equal length, and consider the end points, for example, $c_n = 0.5(k-1)$, $k = 1, 2, \ldots, 2B^* + 1$.

To demonstrate the method, we consider a special class of simple mixed models that are of strong practical interest in small area estimation (e.g., Rao 2003).

Example 4. (Fay-Herriot model) the Fay-Herriot model is widely used in small area estimation. It was first proposed to estimate the per-capita income of small places with population less than 1000 (Fay and Herriot 1979). The model can expressed as $y_i = x_i'\beta + v_i + e_i$, $i = 1, \ldots, m$, where x_i is a vector of known covariates, β is a vector of unknown regression coefficients, v_i's are area-specific random effects and e_i's represent sampling errors. It is assumed that v_i, e_i are independent with $v_i \sim N(0, A)$ and $e_i \sim N(0, D_i)$. The variance A is unknown, but the sampling variances D_i's are assumed known.

Let $X = (x_i')_{1\le i\le m}$, so that the model can be expressed as $y = X\beta + v + e$, where $y = (y_i)_{1\le i\le m}$, $v = (v_i)_{1\le i\le m}$ and $e = (e_i)_{1\le i\le m}$. The first column of X is assumed to be 1_m which corresponds to the intercept. The rest of the columns of X are to be selected from a set of candidate covariate vectors $X_2, \ldots, X_K$, which include the true covariate vectors. First note that by applying the following transformation we can simplify the problem to the case $D_i = 1$. Let $D = 1 + \max_{1\le i\le m} D_i$. Draw independent samples $u_1, \ldots, u_m$ independent with the v_i's and e_i's such that $u_i \sim N(0, D-D_i)$, $1 \le i \le m$. Then, let $\tilde{y}_i = (y_i+u_i)/\sqrt{D}$, $\tilde{x}_i = x_i/\sqrt{D}$, $\tilde{v}_i = v_i/\sqrt{D}$ and $\tilde{e}_i = (e_i + u_i)/\sqrt{D}$. Consider $\tilde{y}_i$'s as the new observations. Then, we have $\tilde{y}_i = \tilde{x}_i'\beta + \tilde{v}_i + \tilde{e}_i$, $i = 1, \ldots, m$, where $\tilde{v}_i$, $\tilde{e}_i$, $i = 1, \ldots, m$ are independent with $\tilde{v}_i \sim N(0, \tilde{A})$, $\tilde{A} = A/D$ and $\tilde{e}_i \sim N(0, 1)$. Thus, without loss of generality, we let $D_i = 1$, $1 \le i \le m$.

Consider the fence ML model selection (see section 2). It is easy to show that, in this case, $\hat{Q}_M = (m/2)\{1+\log(2\pi)+\log(|P_{X^\perp} y|^2/m)\}$, where $P_{X^\perp} = I_m - P_X$ and $P_X = X(X'X)^{-1}X'$. Here we assume for simplicity that X is of full rank. It follows that

$$\hat{Q}_M - \hat{Q}_{M_{\mathrm{f}}} \;=\; \frac{m}{2}\log\left(\frac{|P_{X^\perp} y|^2}{|P_{X_{\mathrm{f}}^\perp} y|^2}\right).$$

Furthermore, it can be shown that, when M is a true model, we have

$$\hat{Q}_M - \hat{Q}_{M_{\mathrm{f}}} \;=\; \frac{m}{2}\log\left(1 + \frac{K-p}{m-K-1}F\right),$$

where $p+1$ is the number of columns of X, and $F \sim F_{K-p,m-K-1}$. Therefore, $\sigma_{M,M_{\mathrm{f}}}$ is completely known given $|M|$ and can be evaluated accurately (e. g., by numerical integration).

We carry out a simulation study to evaluate the performance of the adaptive method. We consider a (relatively) small sample situation with $m = 30$. With $K = 5$, $X_2, \ldots, X_5$ were generated from the $N(0, 1)$ distribution, and then fixed throughout the simulation. The candidate models include all possible models with at least an intercept (thus there are $2^4 = 16$ candidate models). We

Table 5: **Fence methods with different c_n's in the Fay-Herriot model**

Optimal Model	1	2	3	4	5
Adaptive c_n	100	100	100	99	100
$c_n = \log\log(n)$	52	63	70	83	100
$c_n = \log(n)$	96	98	99	96	100
$c_n = \sqrt{n}$	100	100	100	100	100
$c_n = n/\log(n)$	100	91	95	90	100
$c_n = n/\log\log(n)$	100	0	0	0	6

consider five cases in which the data y is generated from the model $y = \sum_{j=1}^{5} \beta_j X_j + v + e$, where $\beta' = (\beta_1, \ldots, \beta_5) = (1,0,0,0,0)$, $(1,2,0,0,0)$, $(1,2,3,0,0)$, $(1,2,3,2,0)$ and $(1,2,3,2,3)$, denoted by Model 1, 2, 3, 4, 5, respectively. The true value of A is 1 in all cases. The number of bootstrap samples for the evaluation of the p^*'s is set at 100.

In addition to the adaptive method, we consider five different (non-adaptive) c_n's ($n = m$ in this case), which satisfy the consistency requirements given in Theorem 1 in the sequel (note that these requirements reduce to $c_n \to \infty$ and $c_n/n \to 0$ in this case). These are $c_n = \log\log(n)$, $\log(n)$, $\sqrt{n}$, $n/\log(n)$ and $n/\log\log(n)$. Reported in Table 5 are percentage of times, out of 100 simulations, that the optimal model was selected by each method.

It seems that performance of the fence with $c_n = \log(n)$, $\sqrt{n}$ or $n/\log(n)$ is fairly close to that of the adaptive fence. In any particular situation, one might get lucky to find a good c_n value by chance, but one may not be lucky all the time. For example, we have observed that in the case of a mixed logistic model (e. g., subsection 4.2) $c_n = \log(n)$ may not work as well as $c_n = 1$ in a finite sample situation even though the latter does not satisfy the consistency requirements. Furthermore,

as mentioned in section 1, for more complicated mixed models the definition of the sample size may not simply be the total number of observations (or the number of clusters). In such cases something like $\log(n)$ or $n/\log(n)$ may not make sense. See subsections 4.1 and 4.3 for our simulation results. In the next section we show that the adaptive fence procedure is indeed consistent, as expected.

The top figure of Figure 1 shows a plot of p^* against c_n in the adaptive procedure based on the first simulated dataset generated under Model 4. To show an overall picture, the plot was extended beyond the upper bound B^* in the adaptive procedure, which was 24 in this case. A smoothed version is also plotted. The plot shows two peaks in the middle, which is not unusual. In practice, when there are multiple peaks in the middle, one should pick the highest one. This is supported by our theoretical result, namely, Theorem 3 in the sequel, which shows that c_n^* is an approximate global maximum of p^*. On the other hand, this strategy does not always work in a finite sample situation. For example, the strategy is responsible for the only failure of the adaptive c_n out of a total of 500 simulations (100 under each model; see Table 5). A closer examination shows that, in this case, there were two peaks in the middle; unfortunately, the higher peak led one to the wrong choice - Model 3 instead of Model 4 (the lower peak led to the right choice). The bottom figure shows parallel boxplots of the c_n^*'s obtained from the simulations under the five models.

Remark: It turns out that requiring the existence of a full model or other known true model from which to draw bootstrap samples is not much of a practical problem, because in essence the adaptive fence can be done in two steps. In the first step, one could use the fence with a fixed c_n (e. g., $c_n = 1$) to select a true model (which may not be optimal). Then in the second step, one applies the adaptive fence procedure with bootstrap samples drawn under the true model selected in the first step. Note that in the first step, one does not need c_n to increase in order to select (with probability tending to one) a true model. In fact, we applied this very procedure to the same simulated datasets

as above and found the exact same result - that we found the optimal model 499 out of 500 times and the time we missed, was the very same time we missed above.

7 Consistency of fence, F-B fence and adaptive fence

We assume that the following A1 - A4 hold for each $M \in \mathcal{M}$, where, as before, θ_M represents a parameter vector at which $\mathrm{E}(Q_M)$ attains its minimum, and $\partial Q_M/\partial\theta_M$, etc. represent derivatives evaluated at θ_M. Similarly, $\partial\tilde{Q}_M/\partial\theta_M$, etc. represent derivatives evaluated at $\tilde{\theta}_M$.

A1. Q_M is three-times continuously differentiable with respect to θ_M; and

$$\mathrm{E}\left(\frac{\partial Q_M}{\partial\theta_M}\right) \;=\; 0. \tag{17}$$

A2. Condition ii) of Lemma 2.

A3. The equation $\partial Q_M/\partial\theta_M = 0$ has an unique solution.

A4. There is a sequence of positive numbers $a_n \to \infty$ and $0 \le \gamma < 1$ such that

$\partial Q_M/\partial\theta_M - \mathrm{E}(\partial Q_M/\partial\theta_M) = O_{\mathrm{P}}(a_n^{\gamma})$,

$\partial^2 Q_M/\partial\theta_M\partial\theta_M' - \mathrm{E}(\partial^2 Q_M/\partial\theta_M\partial\theta_M') = O_{\mathrm{P}}(a_n^{\gamma})$,

$\liminf a_n^{-1}\lambda_{\min}\{\mathrm{E}(\partial^2 Q_M/\partial\theta_M\partial\theta_M')\} > 0$,

$\limsup a_n^{-1}\lambda_{\max}\{\mathrm{E}(\partial^2 Q_M/\partial\theta_M\partial\theta_M')\} < \infty$, and there is $\delta_M > 0$ such that

$\sup_{|\tilde{\theta}_M-\theta_M|\le\delta_M} |\partial^3\tilde{Q}_M/\partial\theta_{M,j}\partial\theta_{M,k}\partial\theta_{M,l}| = O_{\mathrm{P}}(a_n)$, $1 \le j,k,l \le p_M$, where $p_M = \dim(\theta_M)$.

In addition, we assume the following. Recall that c_n is the constant in (7).

A5. $c_n \to \infty$; for any true model M^* and incorrect model M, we have $\mathrm{E}(Q_M) > \mathrm{E}(Q_{M^*})$, $\liminf(\sigma_{M,M^*}/a_n^{2\gamma-1}) > 0$ and $c_n\sigma_{M,M^*}/\{\mathrm{E}(Q_M) - \mathrm{E}(Q_{M^*})\} \to 0$.

A6. $\hat{\sigma}_{M,M^*} > 0$ and $\hat{\sigma}_{M,M^*} = \sigma_{M,M^*}O_{\mathrm{P}}(1)$ if M^* is true and M incorrect; and $\sigma_{M,M^*} \vee a_n^{2\gamma-1} = \hat{\sigma}_{M,M^*}O_{\mathrm{P}}(1)$ if both M and M^* are true.

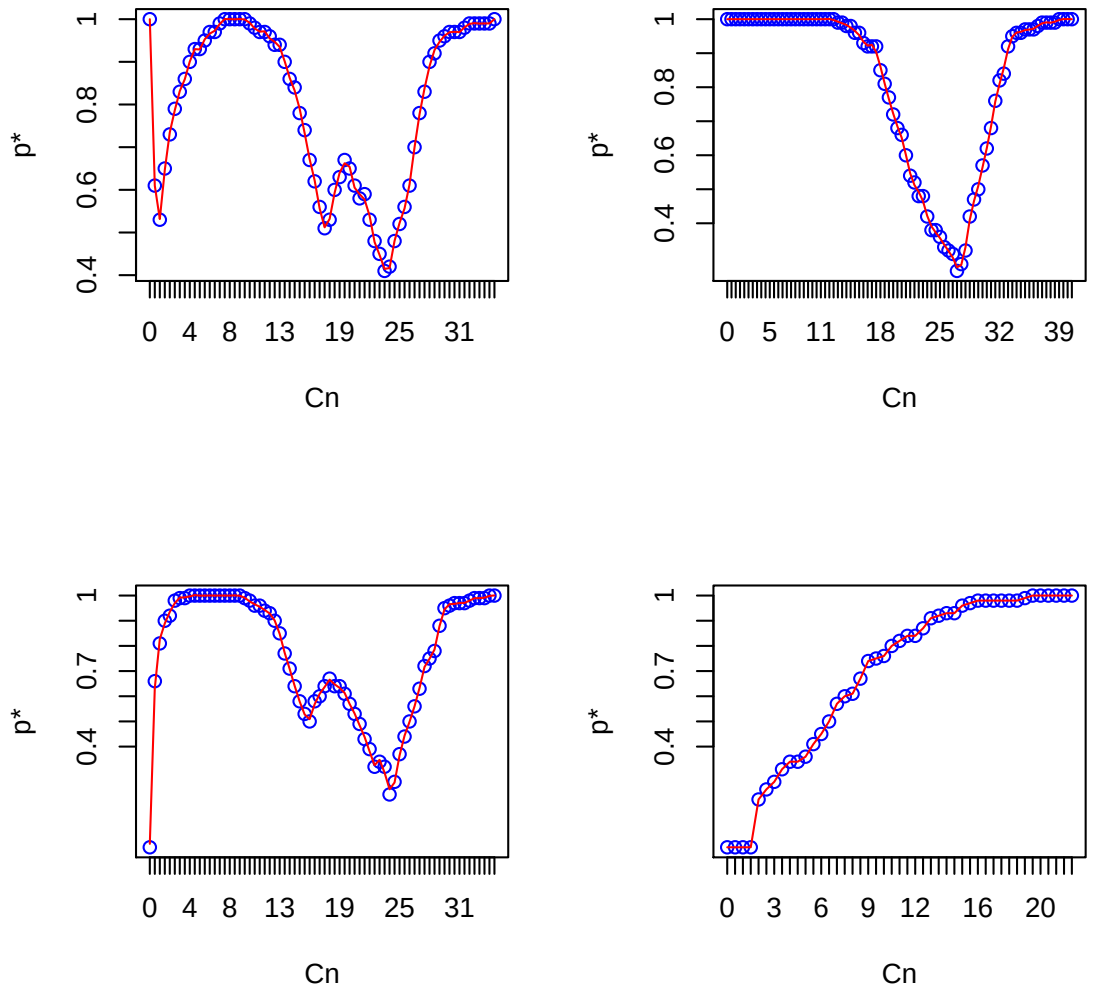

Figure 1: *Top figure: Plot of p^* versus c_n for the adaptive procedure for choosing c_n based on the first simulated dataset generated under Model 4. Notice the peak in the middle of the range of c_n from which the optimal value, c_n^*, is determined. Bottom figure: Parallel boxplots of the c_n^*'s based on the repeated simulations under the five different models under consideration.*

Note. See the remark following Lemma 2 regarding (17) and A2. To illustrate A4 and A5, consider the case of clustered responses (see subsection 3.1). Then, under regularity conditions, A4 holds with $a_n = m$ and $\gamma = 1/2$. Furthermore, we have $\sigma_{M,M^*} = O(\sqrt{m})$ and $\mathrm{E}(Q_M) - \mathrm{E}(Q_{M^*}) = O(m)$, provided that M^* is true, M is incorrect and some regularity conditions hold. Thus, A5 holds with $\gamma = 1/2$ and c_n being any sequence satisfying $c_n \to \infty$ and $c_n/\sqrt{m} \to 0$. Finally, A6 does not require that $\hat{\sigma}_{M,M^*}$ be a consistent estimator of σ_{M,M^*} - only that it has the same order as σ_{M,M^*}. However, see the discussion at the beginning of the previous section.

Lemma 5. Under A1 - A4, we have $\hat{\theta}_M - \theta_M = O_{\mathrm{P}}(a_n^{\gamma-1})$ and $\hat{Q}_M - Q_M = O_{\mathrm{P}}(a_n^{2\gamma-1})$.

Let M_0 be the model selected by fence using (7). The following theorem establishes consistency of the fence procedure.

Theorem 1. Under assumptions A1 - A6, we have with probability tending to one that M_0 is a true model with minimum dimension.

The proofs of Lemma 5 and Theorem 1 are given in subsections 9.3 and 9.4, respectively.

The next theorem establishes consistency of the F-B fence proposed in section 2. Note that the method is introduced in the case of extended GLMMs (also see subsection 5.2). Let $M_0^\dagger$ be the final model of the F-B fence procedure using (7).

Theorem 2. Under assumptions A1 - A6, we have with probability tending to one that $M_0^\dagger$ is a true model and no proper submodel of $M_0^\dagger$ is a true model.

Note that the consistency of the F-B fence is in the sense that (w. p. $\to 1$) $M_0^\dagger$ is a true model which cannot be further reduced or simplified. The proof is given in subsection 9.5.

Finally, we give sufficient conditions for the consistency of the adaptive fence procedure introduced in the previous section. For simplicity, assume that M_{opt} is unique. Consider the ratios $r_M = (\hat{Q}_M - \hat{Q}_{M_{\mathrm{f}}})/\hat{\sigma}_{M,M_{\mathrm{f}}}$, $M \in \mathcal{M}$. Let $\mathcal{M}_{\mathrm{w}\leq}$ denote the subset of incorrect models with dimen-

sion $\leq |M_{\rm opt}|$. Write $r_{\rm opt} = r_{M_{\rm opt}}$ and $r_{{\rm w}\leq} = \min_{M\in\mathcal{M}_{{\rm w}\leq}} r_M$. Denote the cumulative distribution functions of $r_{\rm opt}$ and $r_{{\rm w}\leq}$ by $F_{\rm opt}$ and $F_{{\rm w}\leq}$, respectively. Let $M_0(x)$ be the model selected by the fence procedure using (7) with $c_n = x$, and $P(x) = {\rm P}(M_0(x) = M_{\rm opt})$. Let $P^*(x)$ be the bootstrap version of $P(x)$. Denote the bootstrap sample size by n^*. Recall the definitions of a_n, b_n, q_n, q_n^* in (15), g_n, h_n, r_n, r_n^* in (16), and B^* above Example 4. We make the following assumptions.

A7. (Asymptotic distributional separation) if $M_{\rm opt} \notin \{M_{\rm f}, M_*\}$, then for any $\epsilon > 0$, there is $0 < \delta \leq 0.1$, $x_{n,1} < x_{n,2} < x_{n,3}$, and $N \geq 1$ such that when $n \geq N$ the following hold: $F_{\rm opt}(x_{n,1}) > 1-\epsilon$, $F_{{\rm w}\leq}(x_{n,3}) \leq \epsilon$, $P(x_{n,2}) > 1-\delta$, $1-4\delta < P(x_{n,j}) \leq 1-3\delta$, $j = 1, 3$; if $M_{\rm opt} = M_{\rm f}$, we have ${\rm P}(\min_{M\in\mathcal{M}, M\neq M_{\rm f}} \hat{Q}_M > \hat{Q}_{M_{\rm f}}) \to 1$ as $n \to \infty$.

A8. (Good bootstrap approximation) if $M_{\rm opt} \notin \{M_{\rm f}, M_*\}$, then for any δ, $\eta > 0$, there are $N \geq 1$, $N^* = N^*(n)$ such that, when $n \geq N$ and $n^* \geq N^*$, we have ${\rm P}(\sup_{x>0} |P^*(x) - P(x)| < \delta) > 1-\eta$; if $M_{\rm opt} = M_{\rm f}$, we have $q_n/q_n^* = O_{\rm P}(1)$; if $M_{\rm opt} = M_*$, we have $q_n^*/q_n = O_{\rm P}(1)$ and $r_n^*/r_n = O_{\rm P}(1)$.

For the most part, assumption A7 says that there is an asymptotic separation between the optimal model and the incorrect ones that matter in that the peak of $P(x)$ is distant from the area where $r_{{\rm w}\leq}$ concentrates. This is reasonable because, typically, $r_{\rm opt}$ is of lower order than $r_{{\rm w}\leq}$. Therefore, one can find an interval, $(x_{n,1}, x_{n,3})$, such that (7) is almost always satisfied by $M = M_{\rm opt}$ when $c_n \in (x_{n,1}, x_{n,3})$. On the other hand, $(x_{n,1}, x_{n,3})$ is distant from the area where $r_{{\rm w}\leq}$ concentrates, so that $r_{\rm opt} \leq c_n$, $r_{{\rm w}\leq} > c_n$ with high probability, if $c_n \in (x_{n,1}, x_{n,3})$. Thus, $P(x)$ is expected to peak in $(x_{n,1}, x_{n,3})$ while $F_{{\rm w}\leq}(x)$ stays low in the region.

Recall that p^* in the adaptive procedure is a function of c_n, i.e., $p^* = p^*(c_n)$. The following theorem establishes consistency of the adaptive fence. The proof is given in subsection 9.6.

Theorem 3. Under assumptions A7 and A8 the following hold.

(i) If $M_{\rm opt} \notin \{M_{\rm f}, M_*\}$, then with probability tending to one there is $c_n^* \in (0, \infty)$ which is at least a local maximum and approximate global maximum of p^* in the sense that for any $\delta, \eta > 0$, there is $N \geq 1$ and $N^* = N^*(n)$ such that $\mathrm{P}(p^*(c_n^*) \geq 1 - \delta) \geq 1 - \eta$, if $n \geq N$ and $n^* \geq N^*$.

(ii) In general, define c_n^* as

$$\begin{cases} 0, & \text{if } q_n^* > 1; \\ B^*, & \text{if } q_n^* \leq 1, r_n^* < 1; \\ \text{the } c_n^* \text{ in (i)}, & \text{if } q_n^* \leq 1, r_n^* \geq 1 \text{ and such a } c_n^* \text{ exists}; \\ 1, & \text{otherwise.} \end{cases}$$

Let M_0^* be the model selected by the fence procedure using (7) with $\tilde{M} = M_{\rm f}$ and c_n replaced by c_n^*. Then M_0^* is consistent in the sense that for any $\eta > 0$ there is $N \geq 1$ and and $N^* = N^*(n)$ such that $\mathrm{P}(M_0^* = M_{\rm opt}) \geq 1 - \eta$, if $n \geq N$ and $n^* \geq N^*$.

8 Further discussion and concluding remarks

8.1 A note on hypothesis testing

It is tempting to think of the fence method as similar to hypothesis testing for choosing between models. However, there are some clear and important differences. The fence method is sufficiently more general in nature. In many situations, models must be compared which are not related to one another by parameter restrictions (e. g., non-nested). There may be better ways to capture model complexity in these cases. In such situations, log-likelihood ratios (if a likelihood is available) of pairs of estimated models do not have a chi-square asymptotic distribution. As a result, pulling out appropriate critical values for testing can be quite complex often requiring much more restrictive assumptions about the underlying nature of the models being compared (Findley and Wei, 1989).

Even in the nested model situation, asymptotic null distribution approximations can be poor (e.g., in case of correlated responses or non-normality), or if a likelihood does not exist but some other goodness of fit measure is used, working out critical values for testing can prove problematic.

In addition, fence methods work when the true model does not exist or is not within the set of candidate or approximating models (see subsection 4.2). When such a class is misspecified, hypothesis testing procedures may lead to the simultaneous acceptance or rejection of multiple non-nested models. The former might be a consequence of lack of data, while the latter be indicative of the testing procedure being misspecified altogether (Gourierous and Monfort, 1995).

8.2 Concluding remarks

Fence is different from procedures like AIC, BIC in that there is no criterion function that is minimized. In other words, instead of trying to find an "optimal" model that minimizes a criterion function, fence proposes to carry out the optimization by two steps. The first step is to identify the set of true models (the ones that are in the fence) or, in case a true model does not exist, the models that best approximate the real-life problem. Note that although in this paper we have assumed the existence of a true model, the method can be easily extended to the situation where a true model does not exist, or is understood as the one that provides the best approximation (see subsection 4.2). On the other hand, the second step of fence, which identifies the model with minimal dimension within the fence, is quite flexible. For example, the dimension of a model may not be defined as the number of estimated parameters (e.g., Hastie and Tibshirani 1990, Ye 1998); or it may be replaced by some other considerations, such as economical concerns. In fact, practically speaking, optimality in model selection usually goes beyond statistics. Keeping this in mind, it appears that the fence procedure is easier to incorporate with other scientific or economical criteria than minimizing a

single criterion function determined before the scientific or economic problem.

A good feature of the fence algorithm is that one needs not check all the models for membership within the fence (see the remark following the fence algorithm in section 2). Furthermore, if the candidate models include a full model, the first step of fence, i.e., the identification of $\tilde{M}$, does not require any computation (see the remark following the definition of $\tilde{M}$ in section 2). These features potentially save computational time, especially when the number of candidate models is large.

Finally, fence is conceptually simple. It takes knowledge about information theory and likelihood to understand the idea behind AIC, and Bayesian theory for BIC. But, apparently, everyone understands standard deviation. By the way, the name "fence" is also easily interpreted. In English, fence means a fence.

In this paper, we have demonstrated the robust performance of fence in various situations of linear or generalized linear mixed models as well as its broad applicability to problems in different fields, ranging from genetics, medical care to biology and surveys. In addition, we have introduced a stepwise fence procedure to handle situations of large number of predictors. Furthermore, we have proposed an adaptive procedure for choosing a tuning constant involved in the fence method. The adaptive procedure improves the finite sample performance of fence at a computational cost for bootstrapping. On the theoretical side, we have established consistency of the different fence procedures, with the proofs given in the next section.

9 Proofs

9.1 Proof of Lemma 2

By i), (8) and the fact that the clusters are independent, we have

$$\begin{aligned} \mathrm{E}\left|\frac{\partial Q_M}{\partial \theta_M}\right|^2 &= \mathrm{tr}\left\{\mathrm{Var}\left(\sum_{i=1}^m \frac{\partial Q_{M,i}}{\partial \theta_M}\right)\right\} \\ &= \sum_{i=1}^m \mathrm{tr}\left\{\mathrm{Var}\left(\frac{\partial Q_{M,i}}{\partial \theta_M}\right)\right\} = O(m). \end{aligned}$$

Thus, we have

$$\frac{\partial Q_M}{\partial \theta_M} = O_{\mathrm{P}}(\sqrt{m}). \tag{18}$$

By Taylor expansion, (18) and iii), we have $\hat{Q}_M = Q_M + (\partial Q_M/\partial\theta_M)'(\hat{\theta}_M - \theta_M) + (1/2)(\hat{\theta}_M - \theta_M)'\{\partial^2 Q_M/\partial\theta_M\partial\theta_M'|_{\tilde{\theta}_M}\}(\hat{\theta}_M - \theta_M) = Q_M + R_2$, where $\tilde{\theta}_M$ lies between θ_M and $\hat{\theta}_M$. Hereafter, R_2 represents a random variable whose second moment is bounded, but the definition of R_2 may change from place to place. Since the above holds for any $M \in \mathcal{M}$, we also have $\hat{Q}_{M^*} = Q_{M^*} + R_2$, hence $\hat{Q}_M - \hat{Q}_{M^*} = Q_M - Q_{M^*} + R_2$. Therefore, by iv), we have

$$\mathrm{E}(\hat{Q}_M - \hat{Q}_{M^*}) = \mathrm{E}(Q_M - Q_{M^*}) + O(1), \tag{19}$$

$$\mathrm{var}(\hat{Q}_M - \hat{Q}_{M^*}) = \mathrm{var}(Q_M - Q_{M^*}) + O(\sqrt{m}). \tag{20}$$

Thus, if M is correct, we have $\mathrm{E}(Q_M) = \mathrm{E}(Q_{M^*})$, hence $\mathrm{E}(\hat{Q}_M - \hat{Q}_{M^*}) = O(1)$, $\mathrm{var}(\hat{Q}_M - \hat{Q}_{M^*}) = \mathrm{var}(Q_M - Q_{M^*})\{1 + O(1/\sqrt{m})\} = O(m)$. On the other hand, if M is incorrect, we have, by v) and (19), (20), $\mathrm{E}(\hat{Q}_M - \hat{Q}_{M^*})^2 = \mathrm{var}(Q_M - Q_{M^*}) + O(m^2) = O(m^2)$.

9.2 Proof of Lemma 4

By Taylor expansion and conditions i) - iii), we have $\xi_{M,i} = \hat{\xi}_{M,i} + \partial\xi_{M,i}/\partial\theta_M'|_{\hat{\theta}_M}(\theta_M - \hat{\theta}_M) + (1/2)(\theta_M - \hat{\theta}_M)'(\partial^2\xi_{M,i}/\partial\theta_M\partial\theta_M'|_{\tilde{\theta}_{M,i}})(\theta_M - \hat{\theta}_M) = \hat{\xi}_{M,i} - (\partial\xi_{M,i}/\partial\theta_M'|_{\hat{\theta}_M})(\hat{\theta}_M - \theta_M) - R_{M,i}$,

where $\tilde{\theta}_{M,i}$ lies between θ_M and $\hat{\theta}_M$ and $\mathrm{E}(R_{M,i}^2) \leq ca_n^{-2}$ for some constant c. Furthermore, conditions ii) and iii) imply that $\partial\hat{Q}_M/\partial\theta_M' = \sum_{i=1}^n \partial\xi_{M,i}/\partial\theta_M'|_{\hat{\theta}_M} = 0$. Thus, we have $\hat{Q}_M = \sum_{i=1}^n y_i^2 + \sum_{i=1}^n \hat{\xi}_{M,i} = \sum_{i=1}^n y_i^2 + \sum_{i=1}^n \xi_{M,i} + (\sum_{i=1}^n \partial\xi_{M,i}/\partial\theta_M'|_{\hat{\theta}_M})(\hat{\theta}_M - \theta_M) - \sum_{i=1}^n R_{M,i} = \sum_{i=1}^n y_i^2 + \sum_{i=1}^n \xi_{M,i} - \sum_{i=1}^n R_{M,i}$. A similar expression is obtained for $\hat{Q}_{M^*}$. It follows that $\hat{Q}_M - \hat{Q}_{M^*} = \sum_{i=1}^n d_i + R$, where $R = \sum_{i=1}^n (R_{M^*,i} - R_{M,i})$. Thus, we have $\sigma_{M,M^*}^2 = \mathrm{var}(\sum_{i=1}^n d_i) + 2\mathrm{cov}(\sum_{i=1}^n d_i, R) + \mathrm{var}(R) = I_1 + 2I_2 + I_3$. It is easy to show that $I_3 \leq c_1 n^2 a_n^{-2}$ for some constant c_1. Furthermore, we have, by condition iv),

$$I_1 \;=\; 4 \sum_{z_i'\Sigma z_j \neq 0} \delta_i \delta_j \mathrm{cov}(y_i, y_j) \;\geq\; c_2|S| \tag{21}$$

for some constant $c_2 > 0$. It follows, again by condition iv), that $I_3 \leq o(1)I_1$ and, by Cauchy-Schwarz inequality, $I_2 \leq o(1)I_1$. It follows that $\sigma_{M,M^*}^2 = I_1\{1 + o(1)\}$.

We now derive (10) by using the first equation in (21). Note that if $z_i'\Sigma_{M^*}z_j = 0$, $z_i'\alpha$ and $z_j'\alpha$ are independent. Also, $z_i'\Sigma_{M^*}z_j = 0$ implies $i \neq j$, because otherwise one concludes $z_i = 0$, which contradicts condition iv). Thus, if $z_i'\Sigma_{M^*}z_j = 0$, we have $\mathrm{E}(y_iy_j) = \mathrm{E}\{\mathrm{E}(y_i|\alpha)\mathrm{E}(y_j|\alpha)\} = \mathrm{E}\{h_{M^*}(x_i'\beta_{M^*} + z_i'\alpha)h_{M^*}(x_j'\beta_{M^*} + z_j'\alpha)\} = \mathrm{E}(y_i)\mathrm{E}(y_j)$, hence $\mathrm{cov}(y_i, y_j) = 0$. On the other hand, if $z_i'\Sigma_{M^*}z_j \neq 0$ but $i \neq j$, it is easy to show that $\mathrm{cov}(y_i, y_j) = g_{M^*,i,j}(\beta_{M^*}, \psi_{M^*}) - g_{M^*,i}(\beta_{M^*}, \psi_{M^*})g_{M^*,j}(\beta_{M^*}, \psi_{M^*})$. Finally, note that $z_i'\Sigma_{M^*}z_i \neq 0$, and $\mathrm{var}(y_i) = \mathrm{E}(y_i^2) - g_{M^*,i}^2(\beta_{M^*}, \psi_{M^*})$. It is then easy to derive the expression (10).

9.3 Proof of Lemma 5

A2 and A3 imply that $\hat{\theta}_M$ is the unique solution to $\partial Q_M/\partial\theta_M = 0$, By Taylor expansion, we have, $\tilde{Q}_M - Q_M = (\partial Q_M/\partial\theta_M)'(\tilde{\theta}_M - \theta_M) + (1/2)(\tilde{\theta}_M - \theta_M)'(\partial^2 Q_M/\partial\theta_M\partial\theta_M')(\tilde{\theta}_M - \theta_M) + (1/6)\sum_{j,k,l}(\partial^3 Q_M^*/\partial\theta_{M,j}\partial\theta_{M,k}\partial\theta_{M,l})(\tilde{\theta}_{M,j} - \theta_{M,j})(\tilde{\theta}_{M,k} - \theta_{M,k})(\tilde{\theta}_{M,l} - \theta_{M,l}) = I_1 +$

$(1/2)I_2 + \frac{1}{6}I_3$ for any $\tilde{\theta}_M$, where $\partial^3 Q_M^*/\cdots$ represents the third derivatives evaluated at θ_M^*, which lies between θ_M and $\tilde{\theta}_M$. For any $\epsilon > 0$, by A1 and A4, there are $\delta > 0$ and $N_0 \geq 1$ such that $\lambda_{\min}\{\mathrm{E}(\partial^2 Q_M/\partial\theta_M\partial\theta_M')\} \geq \delta a_n$, $n \geq N_0$, and $L_1 > 0$ such that the probability is greater than $1-\epsilon$ that $|\partial Q_M/\partial\theta_M| \leq L_1 a_n^\gamma$, $\|\partial^2 Q_M/\partial\theta_M\partial\theta_M' - \mathrm{E}(\partial^2 Q_M/\partial\theta_M\partial\theta_M')\| \leq L_1 a_n^\gamma$, $\max_{j,k,l}\sup_{|\tilde{\theta}_M-\theta_M|\leq\delta_M}|\partial^3\tilde{Q}_M/\partial\theta_{M,j}\partial\theta_{M,k}\partial\theta_{M,l}| \leq L_1 a_n$. Now choose $L_2 > 0$ such that $\delta L_2 > 2L_1$. Let $\Theta_{M,L_2} = \{\tilde{\theta}_M : |\tilde{\theta}_M - \theta_M| \leq L_2 a_n^{\gamma-1}\}$, and $\bar{\Theta}_{M,L_2}$ be the boundary of Θ_{M,L_2}, i. e., $\bar{\Theta}_{M,L_2} = \{\tilde{\theta}_M : |\tilde{\theta}_M - \theta_M| = L_2 a_n^{\gamma-1}\}$. Then, choose $N_1 \geq 1$ such that $L_2 a_n^{\gamma-1} \leq \delta_M$, $n \geq N_1$. It follows that for $\tilde{\theta} \in \bar{\Theta}_{M,L_2}$, we have $|I_1| \leq L_1 L_2 a_n^{2\gamma-1}$, $I_2 \geq \delta L_2^2 a_n^{2\gamma-1} - L_1 L_2^2 a_n^{3\gamma-2}$, $|I_3| \leq L_1 a_n \left(\sum_j |\tilde{\theta}_{M,j} - \theta_{M,j}|\right)^3 \leq L_1 L_2^3 p_M^{3/2} a_n^{3\gamma-2}$, hence for all $\tilde{\theta} \in \bar{\Theta}_{M,L_2}$,

$$\tilde{Q}_M - Q_M \quad \geq \quad \frac{1}{2}L_2 a_n^{2\gamma-1}\left\{\delta L_2 - 2L_1 - L_1 L_2\left(1 + \frac{1}{3}L_2 p_M^{3/2}\right)a_n^{\gamma-1}\right\}. \tag{22}$$

If we choose $N_2 \geq 1$ such that, when $n \geq N_2$, the quantity inside $\{\cdots\}$ on the right side of (22) is positive, and let $N = N_0 \vee N_1 \vee N_2$, then we have, with probability greater than $1-\epsilon$, that $\tilde{Q}_M > Q_M, \forall\tilde{\theta} \in \bar{\Theta}_{M,L_2}$. It follows that $\mathrm{P}(|\hat{\theta}_M - \theta_M| < L_2 a_n^{\gamma-1}) \geq 1-\epsilon$, if $n \geq N$. This proves that $\hat{\theta}_M - \theta_M = O_{\mathrm{P}}(a_n^{\gamma-1})$.

By similar arguments, it can be shown that for any $\epsilon > 0$, there are constants L, L_1, L_2 and $N \geq 1$ such that, when $n \geq N$, $\hat{Q}_M - Q_M \leq L_1 L_2 a_n^{2\gamma-1} + \frac{1}{2}LL_2^2 a_n^{2\gamma-1} + (1/2)L_1 L_2^2 a_n^{3\gamma-2} + (1/6)L_1 L_2^3 p_M^{3/2} a_n^{3\gamma-2} \leq L_2\left\{L_1 + (1/2)(L+L_1)L_2 + \frac{1}{6}L_1 L_2^2 p_M^{3/2}\right\}a_n^{2\gamma-1}$ with probability $> 1-\epsilon$. This proves that $\hat{Q}_M - Q_M = O_{\mathrm{P}}(a_n^{2\gamma-1})$.

9.4 Proof of Theorem 1

For the most part, we show that, with probability tending to one (w. p. $\to 1$), all the true models (with $|M| < |\tilde{M}|$) are in the fence, and all the incorrect ones are out.

Let M be an incorrect model and M^* a true model. By Lemma 5 and A5, we have $\hat{Q}_M - \hat{Q}_{M^*} = Q_M - Q_{M^*} + \hat{Q}_M - Q_M - (\hat{Q}_{M^*} - Q_{M^*}) = Q_M - Q_{M^*} + O_{\mathrm{P}}(a_n^{2\gamma-1}) = \mathrm{E}(Q_M) - \mathrm{E}(Q_{M^*}) + \{Q_M - Q_{M^*} - \mathrm{E}(Q_M - Q_{M^*})\} + O_{\mathrm{P}}(a_n^{2\gamma-1}) = \mathrm{E}(Q_M) - \mathrm{E}(Q_{M^*}) + \sigma_{M,M^*} O_{\mathrm{P}}(1) = \{\mathrm{E}(Q_M) - \mathrm{E}(Q_{M^*})\}\{1 + o_{\mathrm{P}}(1)\}$. It follows that, w. p. $\to 1$, we have $\hat{Q}_M > \hat{Q}_{M^*}$. This implies that, w. p. $\to 1$, $\tilde{M}$ is a true model (because an incorrect model cannot be the minimizer).

Furthermore, it is seen from this argument that, if M is incorrect, we have

$$\hat{Q}_M - \hat{Q}_{M^*} \quad = \quad c_n \hat{\sigma}_{M,M^*} \left[\frac{c_n \sigma_{M,M^*}}{\mathrm{E}(Q_M) - \mathrm{E}(Q_{M^*})} \left(\frac{\hat{\sigma}_{M,M^*}}{\sigma_{M,M^*}} \right) \{1 + o_{\mathrm{P}}(1)\}^{-1} \right]^{-1}. \qquad (23)$$

A5 and A6 imply that the quantity inside $[\cdots]$ in (23) is $o_{\mathrm{P}}(1)$. Therefore, w. p. $\to 1$, we have $\hat{Q}_M > \hat{Q}_{M^*} + c_n \hat{\sigma}_{M,M^*}$. It follows that $\mathrm{P}(|M| < |\tilde{M}|, M \in \tilde{\mathfrak{M}}_-) \le \mathrm{P}(\hat{Q}_M \le \hat{Q}_{\tilde{M}} + c_n \hat{\sigma}_{M,\tilde{M}}) \le \sum_{M^* \text{ is true}} \mathrm{P}(\hat{Q}_M \le \hat{Q}_{M^*} + c_n \hat{\sigma}_{M,M^*}, \tilde{M} = M^*) + \mathrm{P}(\tilde{M} \text{ is incorrect}) \le \sum_{M^* \text{ is true}} \mathrm{P}(\hat{Q}_M \le \hat{Q}_{M^*} + c_n \hat{\sigma}_{M,M^*}) + \mathrm{P}(\tilde{M} \text{ is incorrect}) \to 0$. If we let $E_1 = \cap_{M \text{ is incorrect}, |M| < |\tilde{M}|} \{M \notin \tilde{\mathfrak{M}}_-\}$, then $E_1^c = \cup_{M \text{ is incorrect}} \{|M| < |\tilde{M}|, M \in \tilde{\mathfrak{M}}_-\}$, hence $\mathrm{P}(E_1^c) \to 0$. This proves the "out" part.

On the other hand, if M and M^* are both true models, then, by the property of Q_M, we have $\mathrm{E}(Q_M) = \mathrm{E}(Q_{M^*})$. Therefore, by similar arguments and A6, we have $\hat{Q}_M - \hat{Q}_{M^*} = Q_M - Q_{M^*} + O_{\mathrm{P}}(a_n^{2\gamma-1}) = \hat{\sigma}_{M,M^*} O_{\mathrm{P}}(1)$. Since $c_n \to \infty$, we have, w. p. $\to 1$, $\hat{Q}_M \le \hat{Q}_{M^*} + c_n \hat{\sigma}_{M,M^*}$. It follows that $\mathrm{P}(|M| < |\tilde{M}|, M \notin \tilde{\mathfrak{M}}_-) \le \mathrm{P}(\hat{Q}_M > \hat{Q}_{\tilde{M}} + c_n \hat{\sigma}_{M,\tilde{M}}) \le \sum_{M^* \text{ is true}} \mathrm{P}(\hat{Q}_M > \hat{Q}_{M^*} + c_n \hat{\sigma}_{M,M^*}, \tilde{M} = M^*) + \mathrm{P}(\tilde{M} \text{ is incorrect}) \le \sum_{M^* \text{ is true}} \mathrm{P}(\hat{Q}_M > \hat{Q}_{M^*} + c_n \hat{\sigma}_{M,M^*}) + \mathrm{P}(\tilde{M} \text{ is incorrect}) \to 0$. If we let $E_2 = \cap_{M \text{ is true}, |M| < |\tilde{M}|} \{M \in \tilde{\mathfrak{M}}_-\}$, then $E_2^c = \cup_{M \text{ is true}} \{|M| < |\tilde{M}|, M \notin \tilde{\mathfrak{M}}_-\}$, hence $\mathrm{P}(E_2^c) \to 0$. This proves the "in" part.

Finally, note that $\{M_0 \text{ is optimal}\} \supset E_0 \cap E_1 \cap E_2$, where $E_0 = \{\tilde{M} \text{ is true}\}$.

9.5 Proof of Theorem 2

First note that, like the fence procedure, the F-B fence is guaranteed to stop at some point. This is because, otherwise, one keeps adding the parameters until one gets the full model, which automatically satisfies the fence inequality (note that in this case $\tilde{M}$ is chosen as the full model).

Next we show that, w. p. $\to 1$, $M_0^\dagger$ is a true model. Suppose that this is not the case. Then, there is an incorrect model, say, M, such that

$$\mathrm{P}(M_0^\dagger = M) \quad \geq \quad \delta, \tag{24}$$

where $\delta > 0$ is a constant. Since $\tilde{M}$ is a true model, we have by the proof of Theorem 1 that, w. p. $\to 1$, $\hat{Q}_M > \hat{Q}_{\tilde{M}} + c_n \hat{\sigma}_{M,\tilde{M}}$. On the other hand, $M_0^\dagger = M$ implies that $\hat{Q}_M \leq \hat{Q}_{\tilde{M}} + c_n \hat{\sigma}_{M,\tilde{M}}$ (because $M_0^\dagger$ has to satisfy the fence inequality). Thus, we have $\mathrm{P}(M_0^\dagger = M) \leq \mathrm{P}(\hat{Q}_M \leq \hat{Q}_{\tilde{M}} + c_n \hat{\sigma}_{M,\tilde{M}}) \to 0$, which contradicts (24).

We next show that, w. p. $\to 1$, no proper submodel of $M_0^\dagger$ is a true model. Suppose that this is not true. Then there is a true model M_1 and a constant $\delta > 0$ such that $\mathrm{P}(M_1 \subset M_0^\dagger) \geq \delta$. Hereafter the notation $M_1 \subseteq M_2$ ($M_1 \subset M_2$) means that M_1 is a (proper) submodel of M_2. Suppose that under $M_0^\dagger$, $X\beta + Z\alpha = \sum_{r \in R_0} X_r \beta_r + \sum_{s \in S_0} Z_s \alpha_s$, and, under M_1, the same expression holds with R_0, S_0 replaced by R_1, S_1, respectively. Define $R_{10} = R_1 \cup \{r_1, \ldots, r_{a-1}\}$, $S_{10} = S_0$, if $R_1 \subset R_0$, $S_1 \subseteq S_0$ and $R_0 \setminus R_1 = \{r_1, \ldots, r_a\}$; $R_{10} = R_0$, $S_{10} = S_1 \cup \{s_1, \ldots, s_{b-1}\}$, if $R_1 = R_0$, $S_1 \subset S_0$ and $S_0 \setminus S_1 = \{s_1, \ldots, s_b\}$; and $R_{10} = R_1$, $S_{10} = S_1$ otherwise. Let M_{10} be the model corresponding to R_{10} and S_{10}. Then, $M_1 \subset M_0^\dagger$ implies that $M_{10} \subset M_0^\dagger$ with one less parameter, hence we must have $\hat{Q}_{M_{10}} > \hat{Q}_{\tilde{M}} + c_n \hat{\sigma}_{M_{10},\tilde{M}}$ by the definition of $M_0^\dagger$. It follows that

$$\mathrm{P}(\hat{Q}_{M_{10}} > \hat{Q}_{\tilde{M}} + c_n \hat{\sigma}_{M_{10},\tilde{M}}) \quad \geq \quad \delta. \tag{25}$$

On the other hand, we have by the proof of Theorem 1 that for any true model M, w. p. $\to 1$,

$\hat{Q}_M \leq \hat{Q}_{\tilde{M}} + c_n\hat{\sigma}_{M,\tilde{M}}$. Since M_{10} is always a true model, it follows that $\mathrm{P}(\hat{Q}_{M_{10}} > \hat{Q}_{\tilde{M}} + c_n\hat{\sigma}_{M_{10},\tilde{M}}) \leq \sum_{M \text{ true}} \mathrm{P}(\hat{Q}_M > \hat{Q}_{\tilde{M}} + c_n\hat{\sigma}_{M,\tilde{M}}) \to 0$, which contradicts (25).

9.6 Proof of Theorem 3

(i) For any $\epsilon, \eta > 0$, let δ, $x_{n,j}$, $j = 1, 2, 3$, N and N^* be as in A7 and A8. Then, when $n \geq N$ and $n^* \geq N^*$, the following arguments hold with probability $> 1 - \eta$.

For $j = 1, 3$, we have $\mathrm{P}^*(x_{n,j}) > P(x_{n,j}) - \delta > 1 - 5\delta \geq 1/2$. It follows that $p^*(x_{n,j}) = \max_{M\in\mathfrak{M}} P^*(M_0(x_{n,j}) = M) = P^*(x_{n,j}) < P(x_{n,j}) + \delta \leq 1 - 2\delta$. Similarly, $p^*(x_{n,2}) = P^*(x_{n,2}) > P(x_{n,2}) - \delta > 1 - 2\delta$. Thus, there is $c_n^* \in (x_{n,1}, x_{n,3})$ which is the maximum of p^* over $[x_{n,1}, x_{n,3}]$. Furthermore, we have $p^*(c_n^*) \geq p^*(x_{n,2}) > 1 - 2\delta$.

(ii) If $M_{\mathrm{opt}} = M_{\mathrm{f}}$, then $q_n \sim a_n/b_n$, hence $q_n^{-1} = (b_n/a_n)O(1) = o(1)$. Also, by A8, for any $\eta > 0$, there is $L > 0$ such that $\mathrm{P}(q_n/q_n^* > L) < \eta$. Choose $N_1 \geq 1$ such that $q_n^{-1} < 1/L$ when $n \geq N_1$. Then, when $n \geq N_1$, we have, w. p. $> 1 - \eta$, $(q_n^*)^{-1} = q_n^{-1}(q_n/q_n^*) < 1$, hence $q_n^* > 1$, hence $c_n^* = 0$. On the other hand, by A7, there is $N_2 \geq 1$ such that $\mathrm{P}(\min_{M\in\mathfrak{M},M\neq M_{\mathrm{f}}} \hat{Q}_M > \hat{Q}_{M_{\mathrm{f}}}) > 1 - \eta$, if $n \geq N_2$. Let $N = N_1 \vee N_2$, then $\mathrm{P}(M_0^* = M_{\mathrm{f}}) > 1 - 2\eta$, if $n \geq N$.

If $M_{\mathrm{opt}} = M_*$, then, by similar arguments, it can be shown that $r_n^* = o_{\mathrm{P}}(1)$ and $q_n^* = o_{\mathrm{P}}(1)$. Thus, for any $\eta > 0$, there is $N \geq 1$ such that when $n \geq N$ we have, w. p. $> 1 - \eta$, $q_n^* \leq 1$ and $r_n^* < 1$, hence $c_n^* = B^*$, hence $M_0^* = M_*$.

If $M_{\mathrm{opt}} \notin \{M_{\mathrm{f}}, M_*\}$, note that

$$\{M_0^* = M_{\mathrm{opt}}\} \supset \{r_{\mathrm{opt}} \leq c_n^*, r_{\mathrm{w}\leq} > c_n^*\} \supset \{r_{\mathrm{opt}} \leq x_{n,1}, r_{\mathrm{w}\leq} > x_{n,3}\},$$

if $c_n^* \in (x_{n,1}, x_{n,3})$. Therefore, by (i), for any ϵ, $\eta > 0$, we have

$$\begin{aligned} \mathrm{P}(M_0^* = M_{\mathrm{opt}}) &\geq \mathrm{P}(M_0^* = M_{\mathrm{opt}}, c_n^* \in (x_{n,1}, x_{n,3})) \\ &\geq \mathrm{P}(r_{\mathrm{opt}} \leq x_{n,1}, r_{\mathrm{w}\leq} > x_{n,3}, c_n^* \in (x_{n,1}, x_{n,3})) \\ &\geq F_{\mathrm{opt}}(x_{n,1}) - F_{\mathrm{w}\leq}(x_{n,3}) - \mathrm{P}(c_n^* \notin (x_{n,1}, x_{n,3})) \\ &> 1 - 2\epsilon - \eta, \quad n \geq N, n^* \geq N^*. \end{aligned}$$

Acknowledgments

Jiming Jiang is partially supported by NSF grants DMS - 0203676 and DMS - 0402824. J. Sunil Rao is partially supported by NSF grants DMS - 0203724, DMS - 0405072 and NIH grant K25-CA89868.

References

[1] 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.

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

[3] Almasy, L. and Blangero, J. (1998), Multipoint quantitative-trait linkage analysis in general pedigrees, *Am. J. Hum. Genet.* 62, 1198-1211.

[4] Booth, J. G. and Hobert, J. P. (1999), Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, *J. Roy. Statist. Soc. B* 61, 265-285.

[5] Breslow, N. E. and Clayton, D. G. (1993), Approximate inference in generalized linear mixed models, *J. Amer. Statist. Assoc.* 88, 9-25.

[6] Daiger, S. P., Miller M. and Chakraborty (1984), Heritability of quantitative variation at the group-specific component (Gc) Locus, *Amer. J. Hum. Genet.* 36, 663-676.

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

[8] de Leeuw, J. (1992), Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle, in *Breakthroughs in Statistics* (S. Kotz and N. L. Johnson eds.), Springer, London, Vol. 1, 599-609.

[9] Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford Univ. Press.

[10] Drum, M. L. and McCullagh, P. (1993), REML estimation with exact covariance in the logistic mixed model, *Biometrics* 49, 677-689.

[11] Fay, R. E. and Herriot, R. A. (1979), Estimates of income for small places: An application of James-Stein procedure to census data, *J. Amer. Statist. Assoc.* 74, 269-277.

[12] Findley, D. F. and Wei, C. Z. (1989), Likelihood ratio procedures for comparing non-nested, possibly incorrect regressors, *Bureau of the Census Statistical Research Division Report Series: RR*-89/08.

[13] Gourieroux, C. and Monfort, A. (1995), *Statistics and Econometrics Models*, Cambridge University Press.

[14] Hannan, E. J. and Quinn, B. G. (1979), The determination of the order of an autoregression, *J. Roy. Statist. Soc. B* 41, 190-195.

[15] Hartley, H. O. and Rao, J. N. K. (1967), Maximum likelihood estimation for the mixed analysis of variance model, *Biometrika* 54, 93-108.

[16] Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall.

[17] Harville, D. A. (1977), maximum likelihood approaches to variance components estimation and related problems, *J. Amer. Statist. Assoc.* 72, 320-340.

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

[19] Jiang, J. (1996), REML estimation: asymptotic behavior and related topics, *Ann. Statist.* 24, 255-286.

[20] Jiang, J. (1998), Consistent estimators in generalized linear mixed models, *J. Amer. Statist. Assoc.* 93, 720-729.

[21] Jiang, J. (2005), Partially observed information and inference about non-Gaussian mixed linear models, *Ann. Statist.* 33, 2695-2731.

[22] Jiang, J. and Zhang, W. (2001), Robust estimation in generalized linear mixed models, *Biometrika* 88, 753-765.

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

[24] Karim, M. R. and Zeger, S. L. (1992), Generalized linear models with random effects: Salamander mating revisited, *Biometrics* 48, 631-644.

[25] Lange, K. (2002), *Mathematical and Statistical Methods for Genetic Analysis*, 2nd ed., Springer, New-York.

[26] Lin, X. and Breslow, N. E. (1996), Bias correction in generalized linear mixed models with multiple components of dispersion, *J. Amer. Statist. Assoc.* 91, 1007-1016.

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

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

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

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

[31] Pebley, A. R., Goldman, N. and Rodriguez, G. (1996), Prenatal and delivery care and childhood immunization in Guatamala; do family and community matter? *Demography* 33, 231-247.

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

[33] Rodriguez, G. and Goldman, N. (2001), Improved estimation procedure for multilevel models with binary responses: A case-study, *J. Roy. Statist. Soc. A* 164, 339-355.

[34] Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Statist.* 6, 461-464.

[35] Searle, S. R. (1971), *Linear Models*, Wiley, New York.

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

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

[38] Ye, J. (1998), On measuring and correcting the effects of data mining and model selection, *J. Amer. Statist. Assoc.* 93, 120-131.