Restricted fence method for covariate selection in longitudinal data analysis

THUAN NGUYEN[1], JIMING JIANG

Oregon Health and Science University and University of California, Davis

Department of Public Health and Preventive Medicine, Portland, OR 97239

nguythua@ohsu.edu

SUMMARY

Fence method (Jiang *et al.* 2008) is a recently proposed strategy for model selection. It was motivated by the limitation of the traditional information criteria in selecting parsimonious models in some nonconventional situations, such as mixed model selection. Jiang, Nguyen & Rao (2009) simplified the adaptive fence method of Jiang *et al.* (2008) to make it more suitable and convenient to use in a wide variety of problems. Still, the current modification encounters computational difficulties when applied to high dimensional and complex problems. To address this concern, we proposed a restricted fence procedure that combines the idea of fence with that of the restricted maximum likelihood (REML). Furthermore, we propose a robust bootstrap procedure to choose adaptively the tuning parameter used in the restricted fence. We focus on problems of longitudinal studies, and demonstrate the performance of the new procedure and its comparison with the information criteria in a simulation study. The method is further illustrated by a real-data analysis.

Key Words. Longitudinal data, Model selection, Restricted fence method, Robust boostrapping.

[1]To whom correspondence should be addressed.

1. INTRODUCTION

Recently, Jiang *et al.*(2008) developed a new strategy for model selection, known as the *fence methods*. The authors noted a number of limitations of the traditional model selection strategies when applied to mixed model situations. For example, the BIC procedure (Schwarz 1978) relies on the effective sample size which is often unclear in mixed effects models. The fence method avoids such limitations, and therefore are suitable to mixed model selection problems. The basic idea is to build a statistical fence, or barrier, to carefully isolate a subgroup of what are known as the correct models. Once the fence is constructed, the optimal model is selected from those within the fence according to a criterion which can incorporate quantities of practical interest. More specifically, the fence is constructed via the following inequality

$$\hat{Q}_M - \hat{Q}_{\tilde{M}} \;\leq\; c\hat{\sigma}_{M,\tilde{M}}, \tag{1}$$

where $Q_M = Q_M(y, \theta_M)$ be a measure of lack-of-fit, y represents the vector of observations, M indicates a candidate model, and θ_M denotes the vector of parameters under M. Here by lack-of-fit we mean that Q_M satisfies the basic requirement that $\mathrm{E}(Q_M)$ is minimized when M is a true model, and θ_M the true parameter vector under M. Furthermore, $\hat{Q}_M = \inf_{\theta_M \in \Theta_M} Q_M$, where Θ_M is the parameter space under M, and $\tilde{M}$ is a model that minimizes $\hat{Q}_M$ among $M \in \mathcal{M}$, the set of candidate models. Finally, $\hat{\sigma}_{M,\tilde{M}}$ is an estimate of the standard deviation of $\hat{Q}_M - \hat{Q}_{\tilde{M}}$. The tuning constant c on the right side of (1) can be chosen as fixed number (e.g., $c = 1$) or adaptively.

The calculation of $\hat{Q}_M$ is often straightforward. For example, in many cases Q_M can be chosen as the negative log-likelihood or residual sum of squares. On the other hand, the computation of $\hat{\sigma}_{M,\tilde{M}}$ can be quite challenging. Even if an expression can be obtained for $\hat{\sigma}_{M,\tilde{M}}$, its accuracy as an estimate of the standard deviation cannot be guaranteed in a finite sample situation. For such a reason, this step of the fence method has complicated its appli-

cability to many areas. Jiang, Nguyen & Rao (2009) developed a simplified adaptive fence procedure that avoids such difficulties. In the simplified procedure, the fence inequality (1) is replaced by

$$\hat{Q}_M - \hat{Q}_{\tilde{M}} \quad \leq \quad c, \tag{2}$$

In this simplified procedure, $\hat{\sigma}_{M,\tilde{M}}$ is merged into the tuning constant c, which is then chosen adaptively. See Jiang *et al.* (2008) for more details. As in the latter paper, the simplified adaptive fence is shown to be consistent under some regularity conditions, and have outstanding finite sample performance (see Jiang, Nguyen & Rao 2009 for more details).

On the other hand, even with the simplified procedure, one may still encounter computational difficulties when applying the fence to high dimensional and complex problems. The main difficulty rests in the evaluation of a large number of $\hat{Q}_M$'s, if, for example, the number of candidate variables is fairly large (if there are k candidate variables, a total of 2^k different $\hat{Q}_M$'s would have to be evaluated). Furthermore, as in Jiang *et al.* (2008), the adaptive choice of the tuning constant c in (2) involves bootstrapping under the full model. Such a procedure may not be robust, and can be time-consuming, if the full model is complex. To address these concerns, we first proposed a restricted fence procedure that combines the idea of the fence with that of the restricted maximum likelihood, or REML. We then show how to implement the restricted fence via a robust bootstrap procedure. The validity of the robust bootstrap is discussed,and a comparison is made with the generalized estimating equations (GEE). Finite sample performance of the restricted fence is studied, as well as its comparison with the traditional information criteria, in asimulation study. The method is further illustrated using a real-data example. Although the focus of the current paper is linear mixed models used in longitudinal data analysis, the basic idea applies to linear mixed models in general.

2. RESTRICTED FENCE PROCEDURE

The idea of restricted fence may be viewed as a combination of REML and the simplified adaptive fence procedure discussed in the previous section. Our focus is longitudinal studies, in which mean response is often of main interest. As a result, the selection of fixed covariates that are directly associated with the mean is of main interest. Quite often in longitudinal studies, the number of candidate covariates, or variables, is fairly large. Thus, as mentioned in the previous section, direct application of the fence may encounter computational difficulties. The computational difficulty may be greatly reduced by using the restricted fence, as described below. First, we apply a transformation to the data that is orthogonal to a (large) subset of candidate variables to make them "disappear". The simplified adaptive fence procedure is then applied to the remaining (small) subset of variables. It should be noted that, in practice, the candidate variables are not grouped arbitrarily in order to come up with the subsets. Instead, the candidate variables are usually grouped by similarities or possible associations. For example, in the real data example discussed in Subsection 4.3, the variables are grouped based on biological interest. The term "restricted" is used because the first step of the proposed procedure involves the same transformation of the data as in REML (e.g., Jiang 2007, p. 13); however, there is no estimation of the variance components as the REML does.

Consider a linear mixed model that can be expressed as

$$y = X\beta + Zu + e,$$

where X is a matrix of covariates whose columns are to be selected from a (large) set of candidates, β is the corresponding regression coefficients or fixed effects, Z is a known matrix, u is a vector of random effects, and e is a vector of errors. Write $\epsilon = Zu + e$. Note that, by combining the Zu with e, the random effects have "disappeared". However, typically, in longitudinal studies the main interest is the mean response. Although the random effects are used to model the correlations in the observations, there is little interest

in inference about the random effects themselves. This is different from some other areas such as small area estimation (e.g., Rao 2003), which estimation (or prediction) of random effects (or mixed effects) is of main interest. Therefore, we focus on the marginal model, which is standard for the generalized estimating equation (GEE) approach (e.g., Diggle *et al.* 2002, ch. 8). Suppose that X can be expressed as $X = [X_1 \ X_2]$, where $X_1 = (x_{ij})_{1\le i\le n, j\in S_1}$, and $X_2 = (x_{ij})_{1\le i\le n, j\in S_2}$, S_1 is a subset of S, the index set of all the candidates, and $S_2 = S \setminus S_1$. Here S_1 corresponds to the smaller subset and S_2 the larger one. Then the model can be expressed as

$$y = X\beta + \epsilon = X_1\beta^{(1)} + X_2\beta^{(2)} + \epsilon,$$

where $y = (y_i)_{1\le i\le n}$, $\beta = (\beta_j)_{j\in S}$, $\beta^{(1)} = (\beta_j)_{j\in S_1}$, $\beta^{(2)} = (\beta_j)_{j\in S_2}$, and $\epsilon = (\epsilon_i)_{1\le i\le n} \sim N(0, \sigma^2 I_n)$. Let $p_j = \mathrm{rank}(X_j)$, $j = 1, 2$. Let A be a $n \times (n - p_2)$ matrix such that

$$A'A = I_{n-p_2}, \quad A'X_2 = 0.$$

It follows that $AA' = P_{X_2^\perp} = I_n - P_{X_2}$, where $P_{X_2} = X_2(X_2'X_2)^{-1}X_2'$. Then, we have

$$z = A'y = \tilde{X}_1\beta_1 + \eta,$$

where $\tilde{X}_1 = A'X_1$, and $\eta = A'\epsilon$.

Note that, by applying the transformation A' to the data, the matrix X_2, which is typically of much higher dimension, has disappeared from the model. Thus, one can apply the simplified adaptive fence method to the subset of candidates corresponding to X_1 which is usually in much lower dimension. Also note that, although the matrix A is introduced here, its explicit form is not needed for the application of fence method. For example, if Q_M is chosen as RSS, the residual sum of squares, then it can be shown that

$$\hat{Q}_M \ = \ y'P_{X_2^\perp \ominus X_1}y \tag{3}$$

with $P_{X_2^\perp \ominus X_1} = P_{X_2^\perp} - P_{X_2^\perp} X_1 (X_1' P_{X_2^\perp} X_1)^{-1} X_1' P_{X_2^\perp}$ (see Appendix). Furthermore, for the adaptive fence procedure, one can bootstrap under the full model restricted to S_1 without having to know or estimate β_2. In fact, let

$$\hat{\beta}_1 = (\tilde{X}_1' \tilde{X}_1)^{-1} \tilde{X}_1' z = (X_1' P_{X_2^\perp} X_1)^{-1} X_1' P_{X_2^\perp} y. \tag{4}$$

Then, the bootstrap version of $\hat{Q}_M$ is given by

$$\hat{Q}_M^* \;=\; (X_1 \hat{\beta}_1 + \epsilon^*)' P_{X_2^\perp \ominus X_1} (X_1 \hat{\beta}_1 + \epsilon^*), \tag{5}$$

where ϵ^* is the vector of bootstrap errors ϵ (see Appendix for details of the derivation).

We then apply the simplified adaptive fence procedure (Jiang, Nguyen and Rao 2009) to select the covariates within the subset S_1. The same procedure is then applied to a new subset S_1, and so on. Typically, if the total number of covariates selected from each subset put together is still fairly large, another simplified adaptive fence may be applied to the selected covariates to come up with the final set of selected covariates. See Subsection 4.3 for an illustration.

3. LONGITUDINAL STUDIES

3.1 CHARACTERISTICS

There have been remarkable developments in the analysis of longitudinal data over the past twenty-five years (e.g., Diggle *et al.* 2002). The subject has been widely studied in clinical trials, medicine, epidemiology, psychology, education and economics, among other fields. The literature in these fields has contributed to the rapid development of statistical methods for longitudinal data analysis. Longitudinal data provides an opportunity to study the progresses of characteristics of interest over time. In fact, an assessment of the within-subject changes in the response over time can only be achieved with a longitudinal study design.

In comparison to longitudinal studies, the response in a cross-sectional study is measured at a single occasion, therefore, only estimates of the subject-population characteristics can be obtained. In other words, a cross-sectional study design is not capable of capturing information on the response changes during a course of time.

The objectives of longitudinal data analysis are threefolds. These include: 1) to capture the inter-correlation within the subjects; 2) to separate between *cohort* and *age* effects; and 3) to borrow strength across subjects. Such analyses would allow researchers to distinguish changes over time within individuals (aging effect) from differences among subjects at their baseline levels (cohort effect), and make efficient use of the available information (i.e., borrowing strength across subjects, which are often assumed independent). Furthermore, valid inferences can be made more robust to model assumptions, such as in the GEE approach (see below for further discussion).

3.2 INFERENCE ABOUT PARAMETERS OF MAIN INTEREST

In most longitudinal studies, the main interests are associated with the so-called mean response. For example, how does the mean response relate to some of the covariates, such as age, sex, body mass index and blood pressure? how does the treatment (e.g., drug) affect the mean response? and how does the mean response change over time? As mentioned earlier, in longitudinal studies, responses collected from the same individual over time are expected to be correlated. As a result, classical statistical analysis such as linear regression may not be appropriate for longitudinal data. For example, by ignoring the response correlation the standard error calculation may be inaccurate, which leads to falsely rejected or accepted null hypotheses; or confidence intervals with incorrect coverage probability or unnecessarily wide. In fact, this has been the main reason that mixed effects models are widely used in longitudinal data analysis. The simplest model assumes that there is

a random effect corresponding to each individual. The responses from the same individual are therefore correlated for sharing the same random effects. On the other hand, as mentioned earlier, the responses from different individuals are assumed to be independent. More generally, these models may be expressed as

$$y_i = X_i\beta + Z_i\alpha_i + \epsilon_i, \quad i = 1, \ldots, n \tag{6}$$

where n represents the number of individuals (subjects) involved in the study; y_i represents the vector of responses from the ith individual collected over time; and X_i is the matrix of covariates corresponding to the same individual. Note that some of these covariates may be time dependent (such as blood pressure and blood serum measures); others may not be time dependent (such as sex and age, if the duration of the study is relatively short). Furthermore, β is a vector of unknown regression coefficients which are often related to the question of main interest; Z_i is known as a design matrix, α_i is a vector of random effects associated with the ith individual, and ϵ_i is a vector of additional errors. It is assumed that $y_1, \ldots, y_n$ are independent, but the components of y_i are correlated due to the structure of this model. It is also assumed that the random effects and errors have mean zero. Therefore, the mean response is represented by $X_i\beta$. Typically, the mean response consists of two parts. The first part is a linear function of covariates, which is the same as what one has in linear regression; the second part is time-dependent which may involve a function of time and some time-dependent covariates.

Note that the model (6) is a special case of the linear mixed model. If normality is assumed (which means that the random effects and errors are normally distributed), the model can be fitted by maximum likelihood or REML (e.g., Jiang 2007).

The mixed model approach requires considerable modeling of the covariance structure of the data, and hence may suffer from model misspecification. An alternative approach is GEE, which is more robust to misspecification of the covariance structure of the data. This

approach was first proposed by Liang & Zeger (1986). It is assumed that $\mathrm{E}(y_i) = X_i\beta$ and $\mathrm{Var}(y_i) = V_i$, an unknown covariance matrix. Note that the random effects are not explicitly involved in the model. In other words, the model is a *marginal* one. If the V_i's were known, the best linear unbiased estimator (BLUE) of β would be given by

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i' V_i^{-1} X_i\right)^{-1} \sum_{i=1}^{n} X_i' V_i^{-1} y_i. \tag{7}$$

However, the V_i's are usually unknown in real life. Therefore, we replace V_i^{-1} in (7) by a known (symmetric) matrix, say, W_i, called the *working covariance matrix*. For example, $W_i = I$, the identity matrix of a suitable dimension. This replacement does not affect the consistency of the estimator (Liang & Zeger 1986), but may reduce the efficiency of the estimator. Nevertheless, the covariance matrix of the estimator (7) can be estimated by the following "sandwich estimator":

$$\left(\sum_{i=1}^{n} X_i' W_i X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i' W_i \hat{V}_i W_i X_i\right) \left(\sum_{i=1}^{n} X_i' W_i X_i\right)^{-1},$$

where $\hat{V}_i = (y_i - X_i\hat{\beta})(y_i - X_i\hat{\beta})'$. Moreover, by using an iterative procedure developed by Jiang *et al.* (2007) it is possible to obtain an estimator of β that is asymptotically as efficient as the estimator (7) as if the V_i's were known.

4. RESTRICTED FENCE FOR LONGITUDINAL DATA

While there is an extensive literature on modeling the correlation structures, parameter estimation, and inference about the mean response (e.g., Jones 1993, Hand & Crowder 1995, Diggle *et al.* 2002, Jiang 2007), longitudinal model selection has received much less attention. In particular, there is a lack of theoretical development regarding model selection criteria due to the nonconventional features of longitudinal data (see Jiang *et al.* 2008). Although a practioners may employ a number of heuristic selection criteria, such

as AIC (Akaike, 1973), BIC (Schwarz, 1978), HQ (Hannan & Quinn, 1979), and CAIC (or consistent AIC; see Bozdogan, 1987), the theoretical bases for these methods have not been justified in the longitudinal setting. In fact, our simulation results (see section 4.1) showed that some of these methods may perform poorly in selecting parsimonious models for longitudinal studies.

On the other hand, fence method, which was designed for non-conventional problems, applies naturally to longitudinal model selection problems. The main goal of this section is to study performance of the restricted fence method, introduced in Section 2, in selection of important covariate variables among a considerably large number of candidates. Such a problem is motivated practical problems of longitudinal data analysis, which often involves many (potential) covariate variables.

The measure of lack-of-fit, Q_M, for the restricted fence is chosen as the RSS, as in Section 2. The measure is computationally easy to operate, which is very important for high dimensional model selection problems. Note that an explicit expression of $\hat{Q}_M$ is given by (3). Furthermore, in our simulation study, restricted fence based on RSS performs very well as compared to other methods (see the following subsections for further discussion).

4.1. ROBUST BOOTSTRAPPING

As mentioned, an important step of the adaptive fence is bootstrapping. This is relatively straightforward in the illustrative example in Section 2, which was used to introduce the restricted fence method. In fact, in this case, all we need to do is to (i) obtain an estimate of the variance of ϵ_i under the full model, say, $\hat{\sigma}^2_{\rm f}$; and (ii) bootstrap the components of ϵ^* independently from the $N(0, \hat{\sigma}^2_{\rm f})$. However, under the mixed linear model (6), the situation is more complicated.

Ideally, the bootstrapping should be done under the full model of (6). To do so, one needs to (a) estimate the parameters, which include the fixed effects β and all the variance

components associated with the distributions of α_i and ϵ_i; (b) draw samples $\alpha_i^*, \epsilon_i^*, i = 1, \ldots, n$ from the assumed distributions of α_i and ϵ_i, respectively, treating the estimated variance components as the true parameters; and (c) use $y_i^* = X_{\mathrm{f},i}\hat{\beta}_{\mathrm{f}} + Z_i\alpha_i^* + \epsilon_i^*$, $i = 1, \ldots, n$ to generate the bootstrap samples, where $X_{\mathrm{f},i}$ is the covariate vector under the full model, and $\hat{\beta}_{\mathrm{f}}$ the estimator of β under the full model. We call such a procedure linear mixed model bootstrapping.

However, there are practical reasons that bootstrapping under the full linear mixed model as above may not be robust. For example, the standard procedures of fitting the linear mixed model (6), which are maximum likelihood (ML) and restricted maximum likelihood (REML), involve numerically solving nonlinear maximization problems or equations. Although these procedures are available in standard software packages, such as SAS, S-plus and R, non-convergence, false convergence, and convergence to local maximums often occur in practice. In such cases, the variance components under the full linear mixed model may be poorly estimated, which results in poor bootstrap approximations of the distributions of the random effects and errors, as in step (b) above. It is observed, for example, in some of our simulation studies (not shown), in which we found that the restricted fence performs significantly better using the robust bootstrapping method, described below, than using the linear mixed model boostrapping.

In the robust bootstrapping procedure, we first estimate the fixed effects β_1 under the restricted full model that includes all of the variables in the S_1 subset (see Section 2). This is naturally done by minimizing the Q_M that we are using for the restricted fence, which is the RSS. The estimator is given by (4) with $X_1 = X_{\mathrm{f},1}$ and $X_2 = X_{\mathrm{f},2}$, and is denoted by $\hat{\beta}_{\mathrm{f},1}$, where $X_{\mathrm{f},j}$ is the full X_j, $j = 1, 2$. Thus, $X_{\mathrm{f},1}$ corresponds to the restricted full model, which has much less covariates than the full model for X (see Section 2). Also note that $\hat{\beta}_{\mathrm{f},1}$ is a GEE estimator based on $z = A'y$ with the working covariance matrix chosen as the identity matrix (see Subsection 3.2), where A is defined in Section

2 with $X_j = X_{\mathrm{f},j}, j = 1, 2$. Here, for simplicity, we have assumed that $X'_{\mathrm{f},1} P_{X_{\mathrm{f},2}^\perp} X_{\mathrm{f},1}$ is non-singular; otherwise, the generalized inverse should be used. This procedure has the computational advantage in that the numbers of fixed effects that need to be estimated is much smaller than that under the full model of X. For example, in our simulation study, the full model of X has 30 fixed effects, while the full model of X_1 has somewhere between 6 to 8 fixed effects. Next, we write model (6) as

$$y \;=\; X\beta + \zeta, \tag{8}$$

where $y = (y_i)_{1 \le i \le m}$, $X = (X_i)_{1 \le i \le m}$, and ζ represents the rest of the model involving the random effects and errors. We then assume a *working distribution* for the error vector ζ such that, under the working distribution, the components of ζ are independent and distributed as $N(0, \sigma^2)$, where σ^2 is an unknown variance. This unknown variance is then estimated by the standard unbiased estimator,

$$\hat{\sigma}^2 = \frac{\hat{Q}_{M_{\mathrm{f}}}}{n - p_{\mathrm{f}}} = \frac{y' P_{X_{\mathrm{f},2}^\perp \ominus X_{\mathrm{f},1}} y}{n - p_{\mathrm{f}}}, \tag{9}$$

where $\hat{Q}_M$ is given by (3), and $p_{\mathrm{f}} = p_{\mathrm{f},1} + p_{\mathrm{f},2}$ with $p_{\mathrm{f},j} = \mathrm{rank}(X_{\mathrm{f},j}), j = 1, 2$ (see Appendix). Given $\hat{\sigma}^2$, we generate ϵ^* by $\epsilon^* = \hat{\sigma}\xi$, where the components of ξ are generated independently from the $N(0, 1)$ distribution, and then use (5) to compute $\hat{Q}^*_M$, the bootstrap version of $\hat{Q}_M$, for the simplified adaptive fence for selecting the covariates for X_1.

The rationale behind the robust bootstrap may be compared to that of the GEE (see Subsection 3.2). In GEE, the means of the responses are correctly specified but the covariance matrices may be misspecified. Nevertheless, the GEE estimator is consistent (Liang & Zeger 1986), even thought it may not be efficient. In the robust bootstrap procedure, the bootstrapped $\hat{Q}_M$, that is, (5), depends on $X_{\mathrm{f},1}\hat{\beta}_{\mathrm{f},1} + \epsilon^*$. The first term is correctly specified. This is because the LS estimator of $\beta_{\mathrm{f},1}$, which is a special GEE estimator (see above), is consistent. On the other hand, the covariance matrix of ϵ^* may be misspecified, but this

does not affect the consistency property of the model selection. Note that only selection of the fixed covariates are considered here. By a very similar argument as that in Jiang *et al.* (2008) (or Jiang, Nguyen & Rao 2009), the consistency property of the restricted fence with robust bootstrapping can be rigorously established. For the most part, the consistency of fence rests on a single requirement, that is, the values of $\hat{Q}_M$ are well-separated between correct and incorrect models. It can be shown that the $\hat{Q}^*_M$ given by (5) has the latter property. Note that here correct and incorrect models are in the sense of (correct or incorrect) specification of $\mathrm{E}(y)$ as linear functions of the covariates. The technical conditions and proof are omitted. On the other hand, we would like to mention some empirical results in this regard.

In some of our simulation studies (not shown) we compared the performance of the restricted fence using the robust boostrap with that using the linear mixed model boostrap. It turns out that the restricted fence with the robust bootstrapping significantly perform better than the one with the linear mixed model bootstrapping. We have also considered robust bootstrapping under the unrestricted full model based on (9). In this case, the parameters β under the full model of X is estimated; then ζ^* is sampled from $N(0, \tilde{\sigma}^2 I_n)$, where $\tilde{\sigma}^2 = |y - X_{\mathrm{f}}\hat{\beta}_{\mathrm{f}}|^2/(n-p)$, where X_{f} is the full model for X and $\hat{\beta}_{\mathrm{f}}$ the corresponding LS estimator. Simulation results for the restricted fence under this kind of robust bootstrapping (which requires more computation; see above) are almost no difference from those shown in Tables 1 & 2.

4.2 SIMULATION STUDY

We begin with a simulation study. Consider the following model for the sake of simplicity

$$y_{ij} = x'_{ij}\beta + \upsilon_i + \epsilon_{ij}, \quad i = 1, \ldots, n; j = 1, \ldots, T, \tag{10}$$

where i represents the i^{th} subject, and j the j^{th} time point; $\upsilon_i \sim N(0, \sigma_v^2)$, $\epsilon_{ij} \sim N(0, \sigma_e^2)$, and υ_i's and ϵ_{ij}'s are independent.

The data were simulated to mimic a real dataset regarding a bone turnover study collected over three time points. The outcome of interest was a measure of a marker of bone formation, measured over time with respect to dietary groups. The data include several covariates. Some of them are independent to time (e.g., gender, dietary treatment). Most of them are time-dependent variables. There are 30 variables as the total.

We consider three cases: $n = 50,\ n = 100,\ \text{and}\ n = 150$, where n is the number of subjects. In each case, similar to the study design, we assign half number of the subjects to each dietary group. The continuous covariates are generated under the normal distribution with the mean and standard deviation equal to those obtained from the real dataset with respect to time and group categories, disregarding the missing values.

The true model used for the simulation includes the 5 variables. We use two indicator variables for the three time points. Therefore, the true regression coefficients are $\beta = (1, 1, 1, 1, .05, .25, .001)'$, corresponding to the intercept, the unit change in the response from time 1 to time 2, and from time 1 to time 3, and the other three continuous measures. These coefficients are set to be similar to those obtained from the real data under the full model. The variances of the subject-specific random effects and random errors, σ_v^2 and σ_e^2 are set to be 1, which is close to their estimates from the real data. The number of bootstrap samples for the (restricted) simplified adaptive fence is 100. A total of 100 simulations are run under each sample size.

For the restricted fence, we divide all potential predictors into four groups according to biological considerations, and the number of variables in each group ranges from 6 to 8. We then apply the simplified adaptive fence procedure to each group based on the transformed data (see Section 2). The results are reported in Tables 1 & 2. A same-data comparison is made with four of the traditional information criteria, AIC, BIC , HQ, and

CAIC. This means that, all five methods, the restricted fence, AIC, BIC , HQ, and CAIC, are applied to each simulated data set, and the results of selected variables are obtained. The information criteria are based on the following criterion functions, respectively, where $|M|$ is the dimension of the model M defined as the number of parameters, including the fixed effects and variance components:

$$\begin{aligned}
\mathrm{AIC}(M) &= \log(\mathrm{RSS}) + |M|\frac{2}{n}, \\
\mathrm{BIC}(M) &= \log(\mathrm{RSS}) + |M|\frac{\log(n)}{n}, \\
\mathrm{HQ}(M) &= \log(\mathrm{RSS}) + 3|M|\frac{\log\{\log(n)\}}{n}, \\
\mathrm{CAIC}(M) &= \log(\mathrm{RSS}) + |M|\frac{\log(n)+1}{n}.
\end{aligned}$$

Due to the high dimensionality and complex data structure, the Forward and Backward procedures are incorporated with the four above information criteria procedures. More specifically, we first run the Forward selection procedure, and stop when we get 50% of all variables. The Forward selection is then followed by the Backward elimination procedure. We then apply AIC, BIC, HQ and CAIC to the sequence of models generated by the Forward/Backward procedure (F/B), and choose the model with minimum AIC, BIC, HQ, or CAIC, depending on which criterion we are examining, as our optimal model. Note that the same F/B procedure is used, for example, Browman and Speed (2002) in their δ BIC procedure. In this simulation, since there are 30 potential variables, the Forward procedure stops when 15 variables are collected, which is then followed by the Backward elimination.

Table 1 & 2 show the performance of the restricted fence comparing with those of the (F/B) BIC, CAIC, HQ and AIC procedures. The first three are known to be consistent model selection procedures, at least for conventional problems. The results show that the BIC, CAIC, HQ and AIC procedures tend to overfit in all cases, although the BIC procedure appears to perform better than the other three. In the case of small sample size ($n =$

50), the restricted fence seems to be reluctant in choosing a higher dimensional model, resulting underfitting. The same feature has been observed in Jiang *et al.* (2008); and Jiang, Nguyen & Rao (2009). The AIC procedure appears to perform the worst amongst all procedures. It is well-known that the AIC procedure is inconsistent in the situation that a finite dimensional true model exists, and is among the candidate models. This appears to be the case in our simulation study. Overall, the restricted fence seems to outperform, significantly, all the other procedures, both in terms of the (empirical) probability of correct selection and in terms of the (empirical) means and standard deviations of the numbers of correctly and incorrectly selected variables. Some plots of p^* vs. c are shown in Figures 1 & 2. These plots are similar to those in Jiang, Nguyen & Rao (2009) for the simplified adaptive fence. The plots are obtained from the third simulation, which is randomly chosen.

4.3 A REAL DATA EXAMPLE

A clinical trial, Soy Isoflavones for Reducing Bone Loss (SIRBL), was conducted at multi-centers (Iowa State University, and University of California at Davis - UCD). Only part of the data collected at UCD will be analyzed here. The data includes 56 healthly post-menopausal women (45 - 65 years of age) as part of a randomized, double-blind, and placebo-controlled study. The data were collected over three time points - baseline, after 6 and 12 months. One problem of interest is to model the Cytokines (IL1BBLLA, TNFA-BLLA, IL6BLLA) - inflammatory markers - over time on gene expression for IFNb and cFos, along with other variables, which include Dietary treatment (Soy isoflavones): A (8mg/d); B (120 mg/d); C (placebo), and covariates such as Age, Weight, Height, body mass index (BMI), sitting height (HtSitCm), WaistCir. In addition, Bone Mineral Content - femoral neck bone mineral content (FNBMC); lumbar spine total bone mineral content (LSTBMC); whole body total bone mineral content (WBTBMC); hip measures including hip total bone mineral content (HTBMC), TrocBMC, TibTrBMC; Bone Mineral Density

- femoral neck bone mineral density (FNBMD); lumbar spine total bone mineral density (LSTBMD); whole body total bone mineral density (WBTBMD); hip measures including total bone mineral density (HTBMD), TrocBMD; TibTrBMD; Bone Mineral Area - femoral neck area (FNArea); lumbar spine total area(LSTArea); whole body total area (WBTArea); hip total area (HTArea), and others such as lumbar spine TScore (LSTScore), lumbar spine ZScore (LSZScore), whole body TScore (WBTScore), whole body ZScore (WBZScore), hip TScore (HipTScore), hip ZScore (HipZScore). We are interested in finding a subset of relevant variables/covariates that contribute to the variation of Cytokines.

Here we only report the results of data analysis for IL1BBLLA. The covariate variables are grouped into 4 groups according to biological interest, and the restricted fence method is applied in very much the same way as in our simulation study (see Subsection 4.2). The result of variable selection are compared with other procedures. Table 3 show the selected optimal models by different procedures.

The main objective of the study was to examine whether Soy Isoflavones treatment affects the bone metabolism. This treatment effect is selected by the restricted fence and AIC, but not by BIC, CAIC and HQ, in modeling IL1BBLLA (Table 3). The Weight variable was thought to be relevant, and is picked up by AIC and HQ, but not by other procedures; however, the BMI variable, which is a function of weight and height, is picked up by the restricted fence procedure. As also seen in the same table, the BMC and BMD for lumbar spine and hip measures are picked up by the restricted fence, but not by any other procedure. Apparently in this analysis, BIC, CAIC and HQ have over-penalized; as a result, their optimal models do not pick up relevant important covariates (e.g., BMD, BMC). As for AIC, it is able to pick up femoral neck area (FNArea) and lumbar spine total area (LSTArea), which are related to bone areal size (i.e., prefix-Area) and considered relevant covariates. However, after consulting with an expert scientist in this field, we are confirmed that BMD and BMC are more important variables than Area measures in this

case. Therefore, the results of the Fence data analysis are more clinically relevant.

ACKNOWLEDGEMENTS

The research was supported, in part, by NSF Grants DMS-02-03676 and DMS-04-02824. The authors are grateful to Dr. Marta Van Loan for kindly providing two data sets from her research laboratory at the USDA Western Human Nutrition Research Center, and for consultation regarding interpretations of the results of our data analysis.

A. APPENDIX

A.1. DERIVATIONS OF (3) and (5)

Consider $y = X\beta + \epsilon$, then $z = A'y = A'X_1\beta_1 + A'X_2\beta_2 + A'\epsilon$, where $X = [X_1 \;\; X_2]$, $A_{n\times(n-p_2)}$ is such that $A'X_2 = 0$, $\mathrm{rank}(A) = n - p_2$ and $A'A = I_{n-p_2}$. Hence, $AA' = A(A'A)^{-1}A' = P_A = P_{X_2^\perp}$. Therefore $z = \tilde{X}_1\beta_1 + \eta$. It follows that

$$\begin{aligned} Q_M &= |z - \tilde{X}_{M,1}\beta_{M,1}|^2 \\ &= |A'y - A'X_{M,1}\beta_{M,1}|^2 \\ &= |A'(y - X_{M,1}\beta_{M,1})|^2 \\ &= (y - X_{M,1}\beta_{M,1})'AA'(y - X_{M,1}\beta_{M,1}) \\ &= (y - X_{M,1}\beta_{M,1})'P_{X_2^\perp}(y - X_{M,1}\beta_{M,1}), \end{aligned}$$

where $P_{X_2^\perp} = I - P_{X_2}$ and $P_{X_2} = X_{M,2}(X'_{M,2}X_{M,2})^{-1}X'_{M,2}$. Thus, we have

$$\begin{aligned} \hat{Q}_M &= |z - \tilde{X}_{M,1}\hat{\beta}_{M,1}|^2, \\ \hat{\beta}_{M,1} &= (\tilde{X}'_{M,1}\tilde{X}_{M,1})^{-1}\tilde{X}'_{M,1}z \\ &= (X'_{M,1}AA'X_{M,1})^{-1}X'_{M,1}AA'y \\ &= (X'_{M,1}P_{X_2^\perp}X_{M,1})^{-1}X'_{M,1}P_{X_2^\perp}y. \end{aligned}$$

Therefore, we have

$$\begin{aligned}
\hat{Q}_M &= |A'y - A'X_{M,1}(X'_{M,1}P_{X_2^\perp}X_{M,1})^{-1}X'_{M,1}P_{X_2^\perp}y|^2 \\
&= |A'\{I - X_{M,1}(X'_{M,1}P_{X_2^\perp}X_{M,1})^{-1}X'_{M,1}P_{X_2^\perp}\}y|^2 \\
&= |A'(I - P_M)y|^2 \\
&= y'(I - P'_M)AA'(I - P_M)y \\
&= y'(I - P'_M)P_{X_2^\perp}(I - P_M)y \\
&= y'\{P_{X_2^\perp} - P'_M P_{X_2^\perp} - P_{X_2^\perp}P_M + P'_M P_{X_2^\perp}P_M\}y \\
&= y'\{P_{X_2^\perp} - P_{X_2^\perp}X_{M,1}(X'_{M,1}P_{X_2^\perp}X_{M,1})^{-1}X'_{M,1}P_{X_2^\perp}\}y,
\end{aligned}$$

where $P_M = X_{M,1}(X'_{M,1}P_{X_2^\perp}X_{M,1})^{-1}X'_{M,i1}P_{X_2^\perp}$. Note that P_M is idempotent. Let

$$P_{X_2^\perp \ominus X_1} = P_{X_2^\perp} - P_{X_2^\perp}X_{M,1}(X'_{M,1}P_{X_2^\perp}X_{M,1})^{-1}X'_{M,1}P_{X_2^\perp}.$$

Then, we have the expression

$$\hat{Q}_M = y'P_{X_2^\perp \ominus X_1}y.$$

Thus, under a bootstrap sample, we have $y^* = X\hat{\beta} + \epsilon^*$ and

$$\begin{aligned}
\hat{Q}^*_M &= (X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \epsilon^*)'P_{X_2^\perp \ominus X_1}(X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \epsilon^*) \\
&= (X_1\hat{\beta}_1 + \epsilon^*)'P_{X_2^\perp \ominus X_1}(X_1\hat{\beta}_1 + \epsilon^*),
\end{aligned}$$

because $P_{X_2^\perp \ominus X_1}X_2 = 0$.

A.2. UNBIASEDNESS OF THE ESTIMATOR (9) UNDER THE WORKING DISTRIBUTION

Write $X_{\mathrm{f},j} = X_j, p_{\mathrm{f},j} = p_j, j = 1, 2$ and $\hat{\beta}_1 = \hat{\beta}_{\mathrm{f},1}$ for notation simplicity. Then, we have $\mathrm{E}(\hat{Q}_M) = \mathrm{E}(|z - \tilde{X}_1\hat{\beta}_1|^2) = \mathrm{E}(|P_{\tilde{X}_1^\perp}z|^2) = \mathrm{E}(z'P_{\tilde{X}_1^\perp}z) = \mathrm{E}(\eta'P_{\tilde{X}_1^\perp}\eta)$, where $\eta = A'\zeta$, A

satisfies the conditions in Section 2, and ζ is defined by (8). Under the working distribution, we have $\eta \sim N(0, \sigma^2 I_n)$. It follows that $\eta \sim N(0, \sigma^2 I_{n-p_2})$. Therefore, continuing with the derivation, we have $\mathrm{E}(\hat{Q}_M) = \mathrm{E}\{\mathrm{tr}(P_{\tilde{X}_1^\perp}\eta\eta')\} = \mathrm{tr}\{P_{\tilde{X}_1^\perp}\mathrm{E}(\eta\eta')\} = \sigma^2\mathrm{tr}(P_{\tilde{X}_1^\perp}) = \sigma^2\{n - p_2 - \mathrm{tr}(P_{\tilde{X}_1})\} = \sigma^2(n - p_2 - p_1)$.

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] Anderson, J. J. B. and Garner, S. C., eds (1995), Calcium and Phosphorus in Health and Disease, *CRC Press* Boca Raton, FL.

[3] Bozdogan, H. (1987), Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions, *Psychomatrika* 52, 345-370.

[4] Broman, K. W. and Speed, T. P. (2002), A model selection approach for the identification of quantitative trait loci in experiemental crosses, *J. Roy. Statist. Soc. B*, 64, 641-656.

[5] Diggle, P. J., Heagerty, P. J., Liang, K. Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed., Oxford Univ. Press.

[6] Hand, D. and Crowder, M. (1995), *Practical Longitudinal Data Analysis* , Chapman and Hall, New York.

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

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

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

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

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

[12] Jiang, J., Luan, Y. and Wang, Y. G. (2007), Iterative estimating equations: Linear convergence and asymptotic properties, *Ann. Statist.*, 35, 2233-2260.

[13] Jiang, J., Rao, J. S., Gu, Z. and Nguyen, T. (2008), Fence methods for mixed model selection, *Ann. Statist.*, 36, 1669-1692.

[14] Jiang, J., Nguyen, T. and Rao, J. S. (2009), A simplified adaptive fence procedure, *Statist. Probab. Letters*, 79, 625-629.

[15] Jones, R.H., (1993) *Longitudinal Data with Serial Correlation A State-Space Approach*, Chapman and Hall, New York.

[16] Laird, N. M. & Ware, J. M. (1982), Random effects models for longitudinal data, *Biometrics* 38, 963–974.

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

[18] Nguyen, T. (2008). New Procedures of Fence Methods and Their Applications, *Ph. D. Dissertation.*, Dept. of Statist., Univ. of Calif., Davis, CA.

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

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

[21] Sen, A. & Srivastava, M. (1990), *Regression Analysis*, Springer, New York.

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

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

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

[25] Shibata, R. (1986), Essays in Time Series and Allied Processes, *Journal of Applied Probability*, 23, 127-141.

Table 1: **Empirical Probabilities of Selection**. The true model includes seven variables and nothing else. *Fence - Restricted Fence procedure; BIC - the F/B BIC procedure, CAIC - the F/B consistent AIC procedure, HQ - the F/B Hannan & Quinn procedure, AIC - the F/B AIC procedure. Underfitting refers to the case that at least one true variable is missing in the selected model, but may include extraneous variable(s). Overfitting refers to the case that the selected model includes all the true variables plus at least one extraneous variable.*

Sample size	Empirical Probability	Fence	BIC	CAIC	HQ	AIC
n = 50	% detecting the true model	53	37	33	28	0
	% underfitting	26	46	39	39	14
	% overfitting	21	17	28	33	86
n = 100	% detecting the true model	85	65	54	47	0
	% underfitting	5	7	4	3	1
	% overfitting	10	28	42	50	99
n = 150	% detecting the true model	96	82	69	56	0
	% underfitting	3	0	0	0	2
	% overfitting	1	18	31	44	98

Table 2: **Empirical Means and standard deviations.** #C - number of correct variables in the selected model; #IC - number of incorrect variables in the selected model. *Fence - Restricted Fence procedure, BIC - the F/B BIC procedure, CAIC - the F/B consistent AIC procedure, HQ - the F/B Hannan & Quinn procedure, AIC - the F/B AIC procedure.*

Sample size	Mean (s.d.)	Fence	BIC	CAIC	HQ	AIC
n = 50	#C	6.73(.46)	6.54(.5)	6.61(.49)	6.61(.49)	6.86(.34)
	#IC	.34(.6)	.39(.69)	.68(.88)	.77(.88)	4.28(1.93)
n = 100	#C	6.94(.27)	6.92(.3)	6.96(.19)	6.97(.17)	6.99 (.1)
	#IC	.12(.38)	.33(.53)	.52(.65)	.72(.87)	4.7(2.03)
n = 150	#C	6.97(.17)	7(0)	7(0)	7(0)	6.98(.14)
	#IC	.01(.1)	.22(.5)	.38(.63)	.58(.76)	4.14(2.01)

Table 3: **Modeling IL1BBLLA.** *Fence - Restricted Fence procedure, BIC - the F/B BIC procedure, CAIC - the F/B consistent AIC procedure, HQ - the F/B Hannan & Quinn procedure, AIC - the F/B AIC procedure. Variable(s) highlighted with the blue color is/are selected variable(s) by the corresponding procedure.*

Fence	BIC	CAIC	HQ	AIC
Soy treatment	Soy treatment	Soy treatment	Soy treatment	Soy treatment
Weight	Weight	Weight	Weight	Weight
BMI	BMI	BMI	BMI	BMI
WaistCir	WaistCir	WaistCir	WaistCir	WaistCir
LSTBMC	LSTBMC	LSTBMC	LSTBMC	LSTBMC
LSTBMD	LSTBMD	LSTBMD	LSTBMD	LSTBMD
TibTrBMD	TibTrBMD	TibTrBMD	TibTrBMD	TibTrBMD
FNArea	FNArea	FNArea	FNArea	FNArea
LSTArea	LSTArea	LSTArea	LSTArea	LSTArea

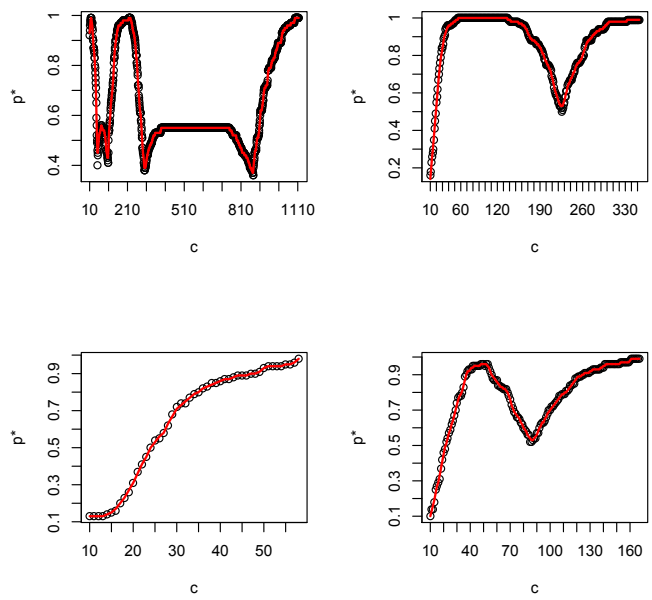

Figure 1: *Plots from step 1 of the restricted fence procedure in one simulation. The four* p^* $vs.$ c *plots correspond to to the four bins of variables. Five variables are picked up from the upper left plot; one variable is selected from the upper right plot; no variable selected from the lower left plot; and one more variable is picked from the lower right plot.*

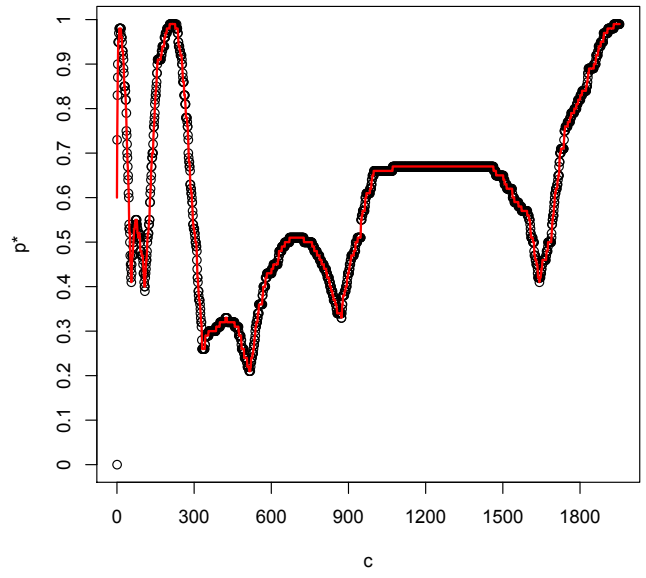

Figure 2: *Plot* p^* $vs.$ c *from step 2 of the restricted fence procedure of one simulation. Seven variables selected from this plot.*