Biostatistics (2012), **13**, 2, *pp.* 303–314
doi:10.1093/biostatistics/kxr046
Advance Access publication on December 17, 2011

Restricted fence method for covariate selection in longitudinal data analysis

THUAN NGUYEN*, JIMING JIANG

Department of Public Health and Preventive Medicine, Oregon Health and Science University, Portland, OR 97239, USA and Department of Statistics, University of California, Davis, CA 95616, USA
nguythua@ohsu.edu

SUMMARY

Fence method (Jiang *and others* 2008. Fence methods for mixed model selection. *Annals of Statistics* **36**, 1669–1692) 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 *and others* (2009. A simplified adaptive fence procedure, *Statistics & Probability Letters* **79**, 625–629) simplified the adaptive fence method of Jiang *and others* (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 the fence with that of the restricted maximum likelihood. Furthermore, we propose to use the wild bootstrap for choosing 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 other procedures of variable selection, including the information criteria and shrinkage methods, in simulation studies. The method is further illustrated by an example of real-data analysis.

Keywords: Covariate variable selection; Longitudinal data; Restricted fence method; Wild boostrapping.

1. INTRODUCTION

Recently, Jiang *and others* (2008) developed a new strategy for model selection, known as the "fence" methods. 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. Jiang *and others* (2009) developed a simplified adaptive fence (SAF) procedure to reduce the computational burden of the adaptive fence of Jiang *and others* (2008) (see Section 2.2 for more details). A summary of the fence methods is provided in the supplementary appendix available at *Biostatistics* online. On the other hand, even with the SAF, 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 measures of lack-of-fit for every bootstrap sample if, for example, the number of candidate variables is

*To whom correspondence should be addressed.

fairly large. Furthermore, as in Jiang *and others* (2008), the adaptive fence or SAF involve 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 propose a restricted fence procedure that combines the idea of the fence with that of the restricted maximum likelihood (REML). We show how to implement the restricted fence via a wild bootstrap procedure, whose validity is discussed. Finite sample performance of the restricted fence is studied as well as its comparison with the information criteria and shrinkage methods of variable selection in a number of simulation studies. The method is further illustrated using a real-data example. Further results and technical derivations are deferred to the supplementary appendix available at *Biostatistics* online.

2. RESTRICTED FENCE PROCEDURE

2.1 *Method*

The REML is well known in mixed model analysis (e.g. Jiang, 2007). The idea is to first apply a transformation to the data to get rid of the (nuisance) fixed effects. Maximum likelihood (ML) is then applied with the transformed data to estimate the variance components. The transformation is constructed so that there is no loss of information in estimating the variance components. Our idea is to combine the ideas of REML and SAF to come up with a strategy for variable selection. We focus on longitudinal studies in which the mean response is often of main interest. As a result, selection of the fixed covariates that are directly associated with the mean is of main interest. Quite often in such studies, the number of candidate covariates, or variables, is fairly large. Thus, as noted, direct application of the fence may encounter computational difficulties. To reduce the computational difficulty, we first apply a transformation to the data that is orthogonal to a (large) subset of the candidate variables to make them "disappear." The SAF is then applied to the remaining (small) subset of the candidate variables. 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.

Consider a linear mixed model that can be expressed as

$$y = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + e,$$

where $\mathbf{X}$ 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, $\mathbf{Z}$ is a known matrix, $\mathbf{u}$ is a vector of random effects, and $\mathbf{e}$ is a vector of errors. Write $\epsilon = \mathbf{Z}\mathbf{u} + \mathbf{e}$. Note that, by combining the $\mathbf{Z}\mathbf{u}$ with $\mathbf{e}$, 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) in 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 *and others*, 2002, chapter 8). Suppose that $\mathbf{X}$ can be expressed as $\mathbf{X} = [\mathbf{X}_1\mathbf{X}_2]$, where $\mathbf{X}_1 = (x_{ij})_{1\leqslant i\leqslant n, j\in S_1}$, and $\mathbf{X}_2 = (x_{ij})_{1\leqslant i\leqslant n, j\in S_2}$, S_1 is a subset of S, the index set of all the candidate variables, 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

$$\mathbf{y} = \mathbf{X}\beta + \epsilon = \mathbf{X}_1\beta^{(1)} + \mathbf{X}_2\beta^{(2)} + \epsilon,$$

where $\mathbf{y} = (y_i)_{1\leqslant i\leqslant n}$, $\beta = (\beta_j)_{j\in S}$, $\beta^{(1)} = (\beta_j)_{j\in S_1}$, and $\beta^{(2)} = (\beta_j)_{j\in S_2}$. Let $p_j = \text{rank}(X_j)$, $j = 1, 2$. Let $\mathbf{A}$ be a $n \times (n - p_2)$ matrix such that $\mathbf{A}'\mathbf{A} = I_{n-p_2}$ and $\mathbf{A}'\mathbf{X}_2 = 0$. It follows that

$\mathbf{AA}' = \mathbf{P}_{\mathbf{X}_2^\perp} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}_2}$, where $\mathbf{P}_{\mathbf{X}_2} = \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$. Then, we have $\mathbf{z} = \mathbf{A}'\mathbf{y} = \tilde{\mathbf{X}}_1\beta_1 + \eta$, where $\tilde{\mathbf{X}}_1 = \mathbf{A}'\mathbf{X}_1$, and $\eta = \mathbf{A}'\epsilon$. Here, we assume, for simplicity, that $\mathbf{X}_2$ is of full rank p_2 (otherwise, $(\mathbf{X}_2'\mathbf{X}_2)^{-1}$ will be replaced by the generalized inverse).

Note that, by applying the transformation A' to the data, the matrix $\mathbf{X}_2$, which is typically of much higher dimension, has disappeared from the model. Also note that $\mathrm{E}(\eta) = 0$. Thus, one can apply the SAF method based on the transformed data $\mathbf{z}$ to the subset of candidate variables corresponding to $\mathbf{X}_1$ which is usually of much lower dimension. Also note that, although the matrix $\mathbf{A}$ is introduced here, its explicit form is not needed for the application of the fence method. For example, if $Q_M = \mathrm{RSS}$, the residual sum of squares, then it can be shown that the Q_M based on $\mathbf{z}$ is given by

$$\hat{Q}_M = \mathbf{y}'\mathbf{P}_{\mathbf{X}_2^\perp \ominus \mathbf{X}_1}\mathbf{y} \tag{2.1}$$

with $\mathbf{P}_{\mathbf{X}_2^\perp \ominus \mathbf{X}_1} = \mathbf{P}_{\mathbf{X}_2^\perp} - \mathbf{P}_{\mathbf{X}_2^\perp}\mathbf{X}_1(\mathbf{X}_1'\mathbf{P}_{\mathbf{X}_2^\perp}\mathbf{X}_1)^{-1}\mathbf{X}_1'P_{\mathbf{X}_2^\perp}$ (see supplementary appendix available at *Biostatistics* online). Furthermore, for applying the SAF (Jiang *and others*, 2009), one can bootstrap under the full model restricted to S_1 without having to know or estimate β_2. In fact, let

$$\hat{\beta}_1 = (\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1)^{-1}\tilde{\mathbf{X}}_1'\mathbf{z} = (\mathbf{X}_1'\mathbf{P}_{\mathbf{X}_2^\perp}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{P}_{\mathbf{X}_2^\perp}\mathbf{y}. \tag{2.2}$$

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

$$\hat{Q}_M^* = (\mathbf{X}_1\hat{\beta}_1 + \epsilon^*)'\mathbf{P}_{\mathbf{X}_2^\perp \ominus \mathbf{X}_1}(\mathbf{X}_1\hat{\beta}_1 + \epsilon^*), \tag{2.3}$$

where ϵ^* is the vector of bootstrapped errors ϵ (see Section 3.2 for more detail).

The point is that S_1 can be any subset of the candidate variables. Thus, by dividing S into a number of subsets and applying the above method to every subset, a number of variables are selected from each subset (or no variable is selected from the subset). Finally, the SAF is applied to all the variables that are picked up from the subsets to select the final set of covariate variables.

2.2 *Algorithm*

A numerical algorithm for the restricted fence procedure is given below:

1. For the candidate variables x_j, $1 \leqslant j \leqslant J$, determine a division $S = \{1, \ldots, J\} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_q$, where $\mathcal{S}_r$, $1 \leqslant r \leqslant q$ are subsets of S (not necessarily disjoint).
2. Let $S_1 = \mathcal{S}_1$ and $S_2 = S \setminus S_1$. Apply the SAF using the measure of lack-of-fit (2.1) to select the variables among x_j, $j \in S_1$. The SAF consists of the following steps:
   - 2.1. Estimating the parameters under the restricted full model ($= \{x_j, j \in S_1\}$).
   - 2.2. Bootstrapping under the restricted full model; for each bootstrapped sample, select the optimal model using the fence [see (A.2) of the supplementary appendix available at *Biostatistics* online] for each c among a grid $0 < c_1 < \cdots < c_K$.
   - 2.3. For each c among the grid, compute the frequency, over the bootstrap samples, that each candidate model is selected as the optimal model; compute the maximum frequency, denoted by p^*. Note that p^* depends on c.
   - 2.4. Find a peak in the middle of the plot of p^* against c; let c^* be the c corresponding to the peak; use the fence with $c = c^*$ to select the final optimal model.
3. Apply the same procedure as Step 2 to $\mathcal{S}_2, \ldots, \mathcal{S}_q$.
4. Apply another SAF to the subset of variables selected in Step 2 and Step 3 (combined; considered as the new candidate variables) to select the final variables.

3. RESTRICTED FENCE FOR LONGITUDINAL DATA

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? One important feature of longitudinal data is that responses collected from the same individual over time are expected to be correlated. Ignoring such correlations may lead to incorrect standard error calculations, confidence intervals, and p values. In fact, this has been a main reason that mixed effects models are widely used in longitudinal data analysis. A linear mixed effects model may be expressed as

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\alpha_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where n is the number of individuals (subjects) involved in the study; $\mathbf{y}_i$ is the vector of responses from the ith individual collected over time; and $\mathbf{X}_i$ is the matrix of covariates corresponding to the ith individual. Furthermore, β is a vector of unknown regression coefficients related to the question of main interest; Z_i is a known 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 $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are independent, but the components of $\mathbf{y}_i$ are usually correlated due to the structure of the model. It is also assumed that the random effects and errors have mean zero. Under the normality assumption and a parametric model for the covariance structure of the data, the model may be fitted by ML or REML. However, this approach requires strong parametric modeling and hence may suffer from model misspecification. An alternative approach is the GEE method (Liang and Zeger, 1986; also see Diggle *and others*, 2002), which does not have to specify the covariance structure.

While there is an extensive literature on modeling the correlation structures, parameter estimation, and inference about the mean response (e.g. Diggle *and others*, 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 the longitudinal data (see Jiang *and others*, 2008). Although a practioners may employ a number of heuristic selection criteria, such as the AIC (Akaike, 1973), BIC (Schwarz, 1978), HQ (Hannan and 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 below) show that some of these methods may perform poorly in selecting parsimonious models for longitudinal studies. On the other hand, the fence methods (Jiang *and others*, 2008) were developed for dealing with nonconventional model selection problems. In particular, the restricted fence method introduced in Section 2 applies naturally to longitudinal variable selection problems. Such a problem is motivated by practical problems of longitudinal studies, which often involve many potential 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 (2.1). Furthermore, in our simulation study, restricted fence based on RSS performs very well as compared with other methods (see below for further discussion).

3.1 *Simulation study*

We consider the following linear mixed model: $y_{ij} = x'_{ij}\beta + v_i + \epsilon_{ij}$, $i = 1, \ldots, n$; $j = 1, \ldots, T$, where i represents the ith subject, and j the jth time point; $v_i \sim N(0, \sigma_v^2)$, $\epsilon_{ij} \sim N(0, \sigma_e^2)$, and v_is and ϵ_{ij}s are independent.

The data were simulated to mimic a real data set regarding a bone turnover study. The data for the study were collected over 3 time points (within 12 months study period). The participants were women

(18–40 years of age) in 2 dietary groups, vegan, or omnivore. The outcome of interest was a marker of bone formation (Osteocalcin), measured over time with respect to dietary groups. The covariates including 30 variables. See supplementary appendix available at *Biostatistics* online for a list of these variables.

As measurements were collected at 3 time points, we have $T = 3$. We consider 3 cases: $n = 50$, $n = 100$, and $n = 150$, where n is the number of subjects. In the real data set, there were 48 participants, 24 of them are vegan and the others are omnivore. Thus, we set up the simulations in a similar way so that half of the subjects are in each dietary group. The continuous covariates are generated under the normal distribution with the mean and standard deviation (SD) equal to those obtained from the real data set with respect to time and group categories, disregarding the missing values.

The true model used for the simulation includes the following variables: time, dietary group, height, *N*-telopeptide, and crude calcium balance. The true regression coefficients are $\beta = (1, 1, 1, 1, 0.05, 0.25, 0.001)'$, corresponding to the intercept, 2 time-point indicators, the dietary indicator, and the 3 continuous variables. These coefficients are set to be similar to those obtained from the real data under the full model except for the coefficient of crude calcium balance. The latter variable is known to be associated with the bone metabolism (Anderson and Garner, 1995). Yet, it is not a significant variable according to the real-data analysis in that its coefficient under the full model was quite small. Thus, in the simulation, we bring this variable into the true model (due to its practical interest) by increasing the value of its coefficient to 0.001. 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 fence is 100. A total of 100 simulations are run under each sample size.

For the restricted fence, we divide all potential predictors into 4 groups according to biological considerations. The variable groups are Group A-1: 1–8, Group A-2: 9–15, Group A-3: 16–22, and Group A-4: 23–30, where the variable numbers correspond to those listed in Section A.3 of the supplementary appendix available at *Biostatistics* online. The true variables correspond to the numbers 1, 2, 3, 4, 9, 10, and 27 on the list. We then apply the SAF procedure to each group based on the transformed data (see Section 2). The results are reported in Table 1. Same data comparisons are made with 4 of the traditional information criteria, AIC, BIC, HQ, and CAIC. Due to the high dimensional and complex data structure, the forward/backward (F/B) procedure of Broman and Speed (2002) is applied, where the forward selection stops when 50% of the candidate variables are selected, which is then followed by the backward elimination. We then apply AIC, BIC, HQ, and CAIC to the sequence of models generated by the F/B procedure and choose the model with minimum AIC, BIC, HQ, or CAIC, respectively, as the optimal model.

In addition, there have been several shrinkage variable selection methods, following the Lasso (Tibshirani, 1996). We make the same data comparisons of our method with 2 of the most popular shrinkage methods, the adaptive Lasso (Zou, 2006) and the smoothly clipped absolute deviation method (SCAD; Fan and Li, 2001; also see Fan and Lv, 2008). It has been shown (Zou, 2006) that the Lasso is not consistent for model selection while the adaptive Lasso is. Therefore, our comparison focuses on the latter. It should also be pointed out that there have been recent work on simultaneous selection of fixed and random effects in linear mixed effects models using the shrinkage methods (Bondell *and others*, 2010; Ibrahim *and others*, 2011). However, because our focus is selection of the fixed covariates only, it seems more fair to compare with the shrinkage methods that focus on the fixed covariates, namely the adaptive Lasso and SCAD, even though the latter use regression-based measures of lack-of-fit. Interestingly, all the methods being compared, including our method (see the last paragraph before Section 3.1), AIC, BIC, HQ, and CAIC, use regression-based measures of lack-of-fit.

Table 1 summarizes the performance of the restricted fence comparing with those of the (F/B) BIC, CAIC, HQ, and AIC procedures as well as the adaptive Lasso and SCAD. For the adaptive Lasso, we use the function adalasso() from the parcor package in R, with the regularization parameter chosen by the cross-validation (which is the default method). For SCAD, we use the function GLMvanISISscad() in the SIS package in R, with the regularization parameter chosen by the BIC method. In Table 1, RF, AIC, BIC,

Table 1. *Empirical results. See Section 3.1 for notation*

n	Summary	RF	BIC	CAIC	HQ	AIC	ALASSO	SCAD
50	TP	53	37	33	28	0	15	0
	UF	26	46	39	39	14	32	29
	OF	21	17	28	33	86	53	71
	MC	6.73	6.54	6.61	6.61	6.86	6.66	6.63
	SD	(0.46)	(0.50)	(0.49)	(0.49)	(0.34)	(0.52)	(0.80)
	MIC	0.34	0.39	0.68	0.77	4.28	1.84	1.30
	SD	(0.60)	(0.69)	(0.88)	(0.88)	(1.93)	(1.98)	(0.48)
100	TP	85	65	54	47	0	30	0
	UF	5	7	4	3	1	2	44
	OF	10	28	42	50	99	68	56
	MC	6.94	6.92	6.96	6.97	6.99	6.98	6.28
	SD	(0.27)	(0.30)	(0.19)	(0.17)	(0.10)	(0.14)	(1.05)
	MIC	0.12	0.33	0.52	0.72	4.70	1.76	6.72
	SD	(0.38)	(0.53)	(0.65)	(0.87)	(2.03)	(2.05)	(0.58)
150	TP	96	82	69	56	0	37	0
	UF	3	0	0	0	2	1	86
	OF	1	18	31	44	98	62	14
	MC	6.97	7.00	7.00	7.00	6.98	6.99	5.39
	SD	(0.17)	(0.00)	(0.00)	(0.00)	(0.14)	(0.10)	(1.43)
	MIC	0.01	0.22	0.38	0.58	4.14	1.45	10.58
	SD	(0.10)	(0.50)	(0.63)	(0.76)	(2.01)	(1.85)	(0.93)

RF, restricted fence procedure; BIC, the F/B BIC procedure; CAIC, the F/B CAIC procedure; HQ, the F/B HQ procedure; AIC, the F/B AIC procedure; ALASSO, the adaptive Lasso procedure; SCAD, the SCAD procedure.

HQ, CAIC, and ALASSO stand for the restricted fence, the F/B AIC, BIC, HQ, CAIC, and adaptive Lasso, respectively. Note that the true model includes 7 variables. Here, true positive (TP) means identifying exactly the true variables; underfitting (UF) means that at least one true variable is missing in the selected model (which may also include extraneous variables); overfitting (OF) means that the selected model includes all the true variables plus at least one extraneous variable; TP, UF, and OF are in percentages of empirical probabilities; and MC (MIC) is the empirical mean number of correctly (incorrectly) selected variables. The corresponding empirical SD are in the parentheses. 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) mean and SD of the number of incorrectly selected variables. Some plots of p^* vs. c are shown in Figures 1 and 2 as illustrations.

3.2 *Wild bootstrapping*

As mentioned, a key step of the restricted fence is bootstrapping. This is relatively straightforward if the random effects are not present, that is, if the components of ϵ in Section 2.1 are i.i.d. In fact, in this case, all one needs to do is to (i) obtain an estimate of the variance of ϵ_i under the full model, say, $\hat{\sigma}_{\mathrm{f}}^2$; and (ii) bootstrap the components of ϵ^* independently from the $N(0, \hat{\sigma}_{\mathrm{f}}^2)$ distribution. However, under the mixed linear model (3.1), the situation is more complicated.

Ideally, the bootstrapping should be done under the full model of (3.1). 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

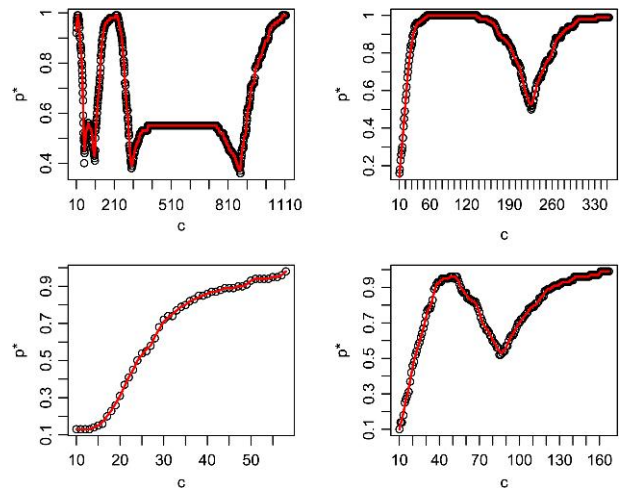

Fig. 1. Plots from the group stage of the restricted fence in one simulation. The four p^* vs. c plots correspond to the 4 bins of variables. Five variables are picked up from the upper left plot; 1 variable is selected from the upper right plot; no variable is selected from the lower left plot; and one more variable is picked from the lower right plot.

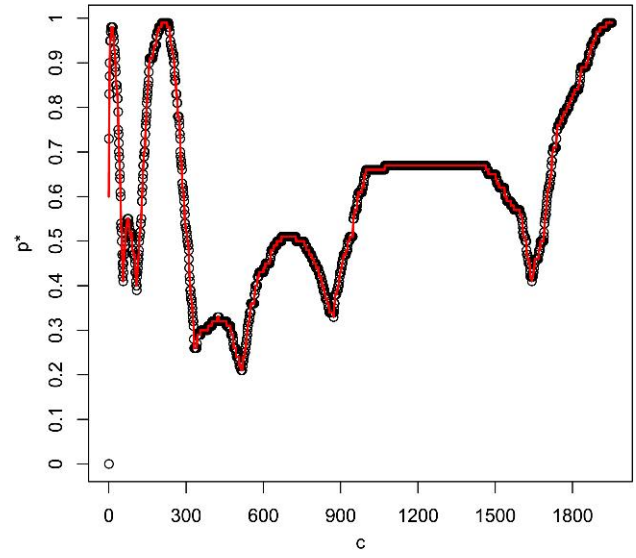

Fig. 2. Plot of p^* vs. c from the final stage of the restricted fence for the same simulation (as the one in Fig. 1). Seven variables are selected from this plot.

of α_i and ϵ_i, respectively, treating the estimated variance components as the true parameters; and (c) use $\mathbf{y}_i^* = \mathbf{X}_{\mathrm{f},i}\hat{\beta}_{\mathrm{f}} + \mathbf{Z}_i\alpha_i^* + \epsilon_i^*$, $i = 1, \ldots, n$, to generate the bootstrap samples, where $\mathbf{X}_{\mathrm{f},i}$ is the covariate matrix 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 (3.1), which are ML and 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, nonconvergence, 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, as in step (b) above. This is confirmed, for example, in our simulation studies in which we found that the restricted fence performs significantly better using the wild bootstrapping method, described below, than using the linear mixed model boostrapping.

The "wild bootstrap" was proposed by Liu (1988) following a suggestion of Wu (1986). Also see Beran (1986). Suppose that we are interested in estimating the mean function of the data, which are independent but not identically distributed. Suppose that we apply the classical bootstrap based on i.i.d. samples from the empirical population to estimate the sampling distribution of the estimator. Can the result still be asymptotically correct? Liu (1988) showed that the answer is yes even though the classical bootstrap does not seem intuitively appropriate here for the simple reason that the original data are not i.i.d. More specifically, it was shown that this wild bootstrap not only captures the first-order limit but also retains the second-order asymptotic properties. This suggests that the wild bootstrap is robust, at least to some extent, against distributional misspecifications.

In our bootstrapping procedure, we first estimate the fixed effects β_1 under the restricted full model. This is naturally done by minimizing the Q_M that is the RSS. The estimator is given by (2.2) with $\mathbf{X}_1 = \mathbf{X}_{\mathrm{f},1}$ and $\mathbf{X}_2 = \mathbf{X}_{\mathrm{f},2}$ and is denoted by $\hat{\beta}_{\mathrm{f},1}$, where $\mathbf{X}_{\mathrm{f},j}$ is the full $\mathbf{X}_j$, $j = 1, 2$. Thus, $\mathbf{X}_{\mathrm{f},1}$ corresponds to the restricted full model, which has much fewer covariates than the full model. Here, for simplicity, we assume that $\mathbf{X}'_{\mathrm{f},1}\mathbf{P}_{\mathbf{X}^{\perp}_{\mathrm{f},2}}\mathbf{X}_{\mathrm{f},1}$ is nonsingular. For example, in our simulation study, the full model of X has 30 fixed covariates, while the full model of $\mathbf{X}_1$ has either 7 or 8 fixed covariates. Next, we write (3.1) as $\mathbf{y} = \mathbf{X}\beta + \zeta$, where $\mathbf{y} = (y_i)_{1\leqslant i\leqslant m}$, $\mathbf{X} = (\mathbf{X}_i)_{1\leqslant i\leqslant m}$, 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 that is estimated by the standard unbiased estimator, $\hat{\sigma}^2 = \hat{Q}_{M_{\mathrm{f}}}/(n - p_{\mathrm{f}}) = y'\mathbf{P}_{X^{\perp}_{\mathrm{f},2}\ominus\mathbf{X}_{\mathrm{f},1}}\mathbf{y}/(n - p_{\mathrm{f}})$, where $\hat{Q}_M$ is given by (2.1), 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 supplementary appendix available at *Biostatistics* online). 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. We then use (2.3) to compute $\hat{Q}^*_M$, the bootstrap version of $\hat{Q}_M$ for the SAF in selecting the covariates for X_1.

In a way, our case is similar to what Liu considered. The underlying model is a linear mixed model, but we are bootstrapping under a regression model that has the same mean function asymptotically. In other words, the bootstrap draws samples that have the correct mean vector but potentially incorrect covariance matrix. Thus, we refer our bootstrap procedure also as wild bootstrap following Liu (1988). By using similar arguments as the latter, it can be shown that our wild bootstrap captures the first-order limit, which is what matters for the consistency of model selection.

There is also a similarity between our wild bootstrap procedure and the GEE (Liang and Zeger, 1986). In GEE, the means of the responses are correctly specified but the covariance matrices may be misspecified. Nevertheless, the GEE estimator is consistent, even thought it may not be efficient. In our wild bootstrap procedure, the bootstrapped $\hat{Q}_M$, that is, (2.3), 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, 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 *and others* (2008) (or Jiang *and others*, 2009), the consistency property of the restricted fence using the wild bootstrap 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 (2.1) has the latter property. The technical conditions and proof are omitted (see supplementary appendix available at *Biostatistics* online

for some empirical results). The results further support our conclusion that when the parameter estimation is unreliable due to computational instability, the simple and much more stable wild bootstrap may have an advantage.

3.3 *A real-data example*

A clinical trial, Soy Isoflavones for Reducing Bone Loss (SIRBL), was conducted at multicenters (Iowa State University and University of California at Davis-UCD). Only part of the data collected at UCD will be analyzed here. The data include 56 healthy postmenopausal women (45–65 years of age) as part of a randomized, double-blind, and placebo-controlled study. The data were collected over 3 time points—baseline, after 6 and 12 months. One problem of interest is to model the Cytokines (IL1BBLLA, TNFA-BLLA, and IL6BLLA)—inflammatory markers—over time on gene expression for IFNb and cFos along with other variables listed in Section A.3 of the supplementary appendix available at *Biostatistics* online.

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. More specifically, one of the authors worked closely with an expert scientist in the field, Dr Marta Van Loan of the USDA Western Human Nutrition Research Center located at UCD to determine what variables should be grouped together and finally came up with the grouping (see Section A.3 of the supplementary appendix available at *Biostatistics* online for details). The restricted fence method is then applied in very much the same way as in Section 3.1. The results are compared with other procedures, reported in Table 2.

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, AIC and SCAD, but not by the other methods. 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 and SCAD. As also seen in the same table, BMD for lumbar and spine measures (LSTBMD) is picked up by the restricted fence but not by any other procedure. Apparently, in this analysis, BIC, CAIC, HQ, and the adaptive Lasso have overpenalized; as a result, their optimal models do not pick up some relevant covariates, such as BMD and BMC (adaptive Lasso did not pick up any of the variables). 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. However, after consulting with the expert scientist in this field, we are confirmed that BMD and BMC are more important variables than area measures in this case. Thus, the results of the restricted fence data analysis

Table 2. *Empirical results under stronger signals. The true model includes the same* 7 *variables as in Table 1.* $n = 100$. *Notations are the same as in Table* 1. *The results are exactly the same under the* 2 *grouping strategies, A and B*

	RF	BIC	CAIC	HQ	AIC	ALASSO	SCAD
TP	100	70	57	49	0	62	0
UF	0	0	0	0	1	0	100
OF	0	30	43	51	99	38	0
MC	7.00	7.00	7.00	7.00	6.99	7.00	4.00
SD	(0.00)	(0.00)	(0.00)	(0.00)	(0.10)	(0.00)	(0.00)
MIC	0.00	0.33	0.52	0.71	4.70	0.92	6.24
SD	(0.00)	(0.53)	(0.65)	(0.88)	(2.03)	(1.60)	(0.69)

RF, restricted fence procedure; BIC, the F/B BIC procedure; CAIC, the F/B CAIC procedure; HQ, the F/B HQ procedure; AIC, the F/B AIC procedure; ALASSO, the adaptive Lasso procedure; SCAD, the SCAD procedure.

are more clinically relevant. Although SCAD has selected the most variables, it has missed the important variable LSTBMD. As for the total body area (WBodArea) that is uniquely picked up by SCAD, the variable is relatively less important, compared with the BMD and BMC, as noted. Our simulation study (see Tables 1 and) has suggested that SCAD has the tendency of missing important variables as well as selecting extraneous variables.

4. DISCUSSION

The restricted fence begins with the division of the candidate variables into several groups (see Step 1 of Section 2.2). We have indicated that, in practice, the grouping should be based on biological information (see Sections 3.1, 3.3). Nevertheless, there is concern on sensitivity of the variable selection result to the grouping. Suppose that the biological information is ignored in the grouping. Will the result be dramatically different? We carry out additional simulation studies to investigate this problem. Recall that, in the simulation study of Section 3.1, the candidate variables were divided into 4 Groups, A-1–A-4, according to biological considerations. Call this grouping Strategy A. In our additional simulation study, we ignore the biological consideration (which is something that we would not recommend in practice). Instead, we consider a different grouping strategy, called grouping Strategy B, by shuffling the variable numbers randomly (keeping 3 and 4 together, which are the time-point indicators—it does not make sense to separate them). We then divide the variables into 4 groups, with the same numbers of variables in the groups as Strategy A (i.e. 8, 7, 7, 8). The new groups are Group B-1: 5, 8, 16, 18, 22, 23, 24, 28; Group B-2: 1, 2, 9, 10, 19, 20, 21; Group B-3: 3, 4, 13, 14, 17, 29, 30; and Group B-4: 6, 7, 11, 12, 15, 25, 26, 27. We then run the simulations based on the new grouping. The results for the restricted fence corresponding to the part $n = 100$ in Table 1 are TP: 77; UF: 2; OF: 21; MC: 6.76 (0.49); MIC: 0.02 (0.14). (The corresponding results for the competing methods do not change, of course.) Comparing with Table 1, it is seen that grouping makes some difference, which suggests that information such as biological interest may help. On the other hand, even with the completely randomized grouping, the results have not changed dramatically; in particular, the restricted fence still outperforms the competing methods. This suggests that the restricted fence is somewhat robust with respect to the grouping.

The robustness of the restricted fence can be argued theoretically in large sample. Because of the consistency of the restricted fence (Jiang *and others*, 2008), in large sample, the procedure will select the correct variables (and nothing else) with high probability regardless of the grouping. Equivalently, one may argue in terms of "signal consistency" (Jiang *and others*, 2011), which is more appropriate in cases where the number of variables is comparable to the sample size. For the most part, signal consistency means that, as the signals (i.e. the absolute values of the true regression coefficients) increase (but with the sample size fixed), the probability of identifying the true variables (and nothing else) goes to one. As argued in Jiang *and others* (2011), it can be shown that the restricted fence is signal-consistent regardless of the grouping.

To verify the signal consistency of the restricted fence empirically, we consider again the 2 different grouping strategies. We run simulations with $n = 100$ and the following increased signals for the true variables: 1, 1, 1, 1, 0.5, 0.5, and 0.01 (in other words, the first 4 coefficients are unchanged, the fifth and seventh are 10 times as strong, and the sixth is twice as strong). The simulation results are presented in Table 3. In particular, the results for the restricted fence are exactly the same (which are perfect) under the 2 grouping strategies, A and B, indicating signal consistency of the restricted fence as aforementioned. Interestingly, the results also seem to suggest that the competing methods improve at a much slower rate as the signals increase, compared to the restricted fence.

It should be noted that consistency or signal consistency are theoretical properties indicating what to expect in the "ideal" situations. In the practical and most likely less ideal situations, a careful design for the grouping could make a difference as is shown. In short, if there is knowledge about the candidate variables,

Table 3. *Modeling IL1BBLLA*

Variable	RF	BIC	CAIC	HQ	AIC	ALASSO	SCAD
Soy treatment	×				×		×
Weight				×	×		
BMI	×						×
WaistCir	×				×		×
HipBMD							×
LSTBMC	×						×
LSTBMD	×						
TibTrBMC							×
TibTrBMD	×	×	×	×	×		×
FNArea					×		
LSTArea					×		
WBodArea							×

RF, restricted fence procedure; BIC, the F/B BIC procedure; CAIC, the F/B CAIC procedure; HQ, the F/B HQ procedure; AIC, the F/B AIC procedure; ALASSO, the adaptive Lasso procedure; SCAD, the SCAD procedure. The × indicates variable selected. Variables not listed were not selected by any of the methods.

such as biological interests, the knowledge should be used in the grouping. This was illustrated in Section 3.3 with a data example. For the most part, we recommend that the (bio)statistician work closely with the expert scientist(s) in the field to determine what grouping strategy is reasonable. It is also important to take into account the relationships between the variables (see Section A.4 of the supplementary appendix available at *Biostatistics* online). Keep in mind that, by definition (see Step 1 of Section 2.2), the groups need not be disjoint. Finally, for computational efficiency, the group sizes should be kept relatively small, typically 5–10 variables in each group if possible.

See Section A.4 of the supplementary appendix available at Biostatistics online for further simulation results that show another aspect of robustness of the restricted fence to "bad" groupings.

Finally, there is interest in comparing the restricted fence with the adaptive fence (Jiang *and others*, 2008) or SAF (Jiang *and others*, 2009). Although the latter are not computationally feasibles for the simulation setting considered in Section 3.1, which has 30 candidate variables, we have provided limited simulation results for a (much) lower dimensional problem. The comparison is in terms of both the selection performance and the computational costs. See Section A.5 of the supplementary appendix available at Biostatistics online.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors are grateful to Dr Marta Van Loan for kindly providing 2 data sets from her research laboratory at the USDA Western Human Nutrition Research Center located at UC-Davis and for consultation and helpful discussions. The authors also wish to thank an Associate Editor and 2 referees for their insightful comments. *Conflict of Interest:* None declared.

Funding

The research was supported, in part, by National Science Foundation (DMS-02-03676 and DMS-04-02824).

REFERENCES

AKAIKE, H. (1973). Information theory as an extension of the maximum likelihood principle. In: Petrov, B. N. and Csaki, F. (editors), *Second International Symposium on Information Theory*. Budapest, Hungary: Akademiai Kiado, pp. 267–281.

ANDERSON, J. J. B. AND GARNER, S. C. (editors) (1995). *Calcium and Phosphorus in Health and Disease.* Boca Raton, FL: CRC Press.

BERAN, R. (1986). Discussion of "Jackknife, bootstrap and other resampling methods in regression analysis" by C. F. J. Wu. *Annals of Statistics* **14**, 1295–1298.

BONDELL, H. D., KRISHNA, A. AND GHOSH, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66**, 1069–1077.

BOZDOGAN, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychomatrika* **52**, 345–370.

BROMAN, K. W. AND SPEED, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Joural of Royal Statistical Society, Series B* **64**, 641–656.

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K. Y. AND ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd edition. Oxford: Oxford University Press.

FAN, J. AND LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association* **96**, 1348–1360.

FAN, J. AND LV, J. (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society Series B* **70**, 849–911.

HANNAN, E. J. AND QUINN, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society Series B* **41**, 190–195.

IBRAHIM, J. G., ZHU, H., CARCIA, R. I. AND GUO, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67**, 495–503.

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

JIANG, J., NGUYEN, T. AND RAO, J. S. (2009). A simplified adaptive fence procedure. *Statistics & Probability Letters* **79**, 625–629.

JIANG, J., NGUYEN, T. AND RAO, J. S. (2011). Invisible fence methods and the identification of differentially expressed gene sets. *Statistics and Its Interface* **4**, 403–415.

JIANG, J., RAO, J. S., GU, Z. AND NGUYEN, T. (2008). Fence methods for mixed model selection. *Annals of Statistics* **36**, 1669–1692.

LIANG, K. Y. AND ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

LIU, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Annals of Statistics* **16**, 1696–1708.

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

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

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B* **16**, 385–395.

WU, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics* **14**, 1261–1295.

ZOU, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

[*Received April 13, 2011; revised November 13, 2011; accepted for publication November 14, 2011*]