Hindawi Publishing Corporation
Advances in Statistics
Volume 2014, Article ID 830821, 14 pages
http://dx.doi.org/10.1155/2014/830821

Review Article

The Fence Methods

Jiming Jiang

University of California, Davis, CA 95618, USA

Correspondence should be addressed to Jiming Jiang; jiang@wald.ucdavis.edu

Received 3 February 2014; Revised 20 June 2014; Accepted 3 July 2014; Published 24 July 2014

Academic Editor: Lynn Kuo

This paper provides an overview of a recently developed class of strategies for model selection, known as the fence methods. It also offers directions of future research as well as challenging problems.

1. Introduction

On the morning of March 16, 1971, Hirotugu Akaike, as he was taking a seat on a commuter train, came out with the idea of a connection between the relative Kullback-Liebler discrepancy and the empirical log-likelihood function, a procedure that was later named Akaike's information criterion, or AIC (Akaike [1, 2]; see Bozdogan [3] for the historical note). The idea has allowed major advances in model selection and related fields. See, for example, de Leeuw [4]. A number of similar criteria have since been proposed, including the Bayesian information criterion (BIC; Schwarz [5]), a criterion due to Hannan and Quinn (HQ; [6]), and the generalized information criterion (GIC; Nishii [7], Shibata [8]). All of the information criteria can be expressed as

$$\mathrm{GIC}(M) = \widehat{Q}_M + \lambda_n |M|, \tag{1}$$

where $\widehat{Q}_M$ is a measure of lack-of-fit by the model, M; $|M|$ is the dimension of M, defined as the number of free parameters under M; and λ_n is a penalty for complexity of the model, which may depend on the effective sample size, n.

Although the information criteria are broadly used, difficulties are often encountered, especially in some nonconventional situations. We discuss a number of such cases below.

(1) The Effective Sample Size. In many cases, the effective sample size, n, is not the same as the number of data points. This often happens when the data are correlated. Take a look at two extreme cases. In the first case, the observations are independent; therefore, the effective sample size should be the same as the number of observations. In the second case, the data are so much correlated that all of the data points are identical. In this case, the effective sample size is 1, regardless of the number of data points. A practical situation may be somewhere between these two extreme cases, such as cases of mixed effects models (e.g., Jiang [9]), which makes the effective sample size difficult to determine.

(2) The Dimension of a Model. The dimension of a model, $|M|$, can also cause difficulties. In some cases, such as the ordinary linear regression, this is simply the number of parameters under M, but in other situations, where nonlinear, adaptive models are fitted, this can be substantially different. Ye [10] developed the concept of generalized degrees of freedom (gdf) to track model complexity. For example, in the case of multivariate adaptive regression splines (Friedman [11]), k nonlinear terms can have an effect of approximately $3k$ degrees of freedom. While a general algorithm in its essence, the gdf approach requires significant computations. It is not at all clear how a plug-in of gdf for $|M|$ in (1) affects the selection performance of the criterion.

(3) Unknown Distribution. In many cases, the distribution of the data is not fully specified (up to a number of unknown parameters); as a result, the likelihood function is not available. For example, suppose that normality is not assumed under a linear mixed model (LMM; e.g., Jiang [9]). Then, the likelihood function is typically not available. Now suppose that one wishes to select the fixed covariates using AIC, BIC, or HQ. It is not clear how to do this because the first term on the right side of (1) is not available. Of course, one could still blindly use those criteria, pretending that the data are

normal, but the criteria are no longer what they mean to be. For example, Akaike's bias approximation that led to the AIC (Akaike [1]) is no longer valid.

(4) Finite-Sample Performance and the Effect of a Constant. Even in the conventional situation, there are still practical issues regarding the use of these criteria. For example, the BIC is known to have the tendency of overly penalizing "bigger" models. In other words, the penalizer, $\lambda_n = \log n$, may be a little too much in some cases. In such a case, one may wish to replace the penalizer with $c\log(n)$, 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 of model selection, so long as $c > 0$. However, practically, it does. As another example, comparing BIC with HQ, the penalizer in HQ is lighter in its order, that is, $\log n$ for BIC and $c\log\log n$ for HQ, where $c > 2$ is a constant. However, if $n = 100$, we have $\log n \approx 4.6$ and $\log\log n \approx 1.5$; hence, if c is chosen as 3, BIC and HQ are almost the same.

In fact, there have been a number of modifications of the BIC aiming at improving the finite-sample performance. For example, Broman and Speed [12] proposed a δ-BIC method by relacing the $\lambda_n = \log n$ in BIC by $\delta\log n$, where δ is a constant carefully chosen to optimize the finite-sample performance. However, the choice of δ relies on extensive Monte-Carlo simulations, is case-by-case, and, in particular, depends on the sample size. Therefore, it is not easy to generalize the δ-BIC method.

(5) Criterion of Optimality. Strictly speaking, model selection is hardly a purely statistical problem—it is usually associated with a problem of practical interest. Therefore, it seems a bit unnatural to let the criterion of optimality in model selection be determined by purely statistical considerations, such as the likelihood and K-L information. Other considerations, such as scientific and economic concerns, need to be taken into account. For example, what if the optimal model selected by the AIC is not to the best interest of a practitioner, say, an economist? In the latter case, can the economist change one of the selected variables, and do so "legitimately"? Furthermore, the dimension of a model, $|M|$, is used to balance the model complexity through (1). However, the minimum-dimension criterion, also known as *parsimony*, is not always as important. For example, the criterion of optimality may be quite different if prediction is of main interest.

These concerns, such as the above, led to the development of a new class of strategies for model selection, known as the *fence* methods, first introduced by Jiang et al. [13]. Also see Jiang et al. [14]. The idea consists of a procedure to isolate a subgroup of what are known as correct models (those within the fence) via the inequality

$$Q(M) - Q\left(\widetilde{M}\right) \le c, \tag{2}$$

where $Q(M)$ is the measure of lack-of-fit for model M, $\widetilde{M}$ is a "baseline model" that has the minimum Q, and c is a cut-off. The optimal model is then selected from the models within the fence according to a criterion of optimality that can be flexible; in particular, the criterion can incorporate the problem of practical interest. Furthermore, the choice of the measure of lack-of-fit, Q, is also flexible and can incorporate problem of interest.

To see how the fence helps to resolve the difficulties of the information criteria, note that the (effective) sample size is not used in the fence procedure, although the cut-off c, when chosen adaptively (see Section 2), may implicitly depend on the effective sample size. Depending on the criterion of optimality for selecting the optimal model within the fence, the dimension of the model may be involved, but the criterion does allow flexibility. Also, the measure of lack-of-fit, Q, does not have to be the negative log-likelihood, as in the information criteria. For example, the residual sum of squares (RSS) is often used as Q, which does not require complete specification of the distribution. Furthermore, a date-driven approach is introduced in Section 2 for choosing the cut-off or tuning constant, c, that optimizes the finite-sample performance. Finally, the criterion of optimality for selecting the model within the fence can incorporate practical interests.

It should be noted that, as far as consistency is concerned, which is, by far, the most important theoretical property for a model selection procedure, the basic underlying assumptions for the fence are the same as those for the traditional model selection approaches, such as the information criteria and cross-validation (CV; e.g., Shao [15]). For the most part, it is assumed that the space of candidate models is finite, which contains a true model and that the sample size goes to infinity while the model space remains the same.

There is a simple numerical procedure, known as the *fence algorithm*, which applies when model simplicity is used as the criterion to select the model within the fence. Given the cut-off c in (2), the algorithm may be described as follows: check the candidate models, from the simplest to the most complex. Once one has discovered a model that falls within the fence and checked all the other models of the same simplicity (for membership within the fence), one stops. One immediate implication of the fence algorithm is that one does not need to evaluate all the candidate models in order to identify the optimal one. This leads to potentially computational savings.

To highlight the main contributions of the current paper, we have provided a short description of the fence idea. Several variations of the fence will be discussed, including the adaptive fence, restricted fence, invisible fence, and similar ideas developed in other fields. We conclude the paper with some challenging problems and future directions.

2. Adaptive Fence

Finite-sample performance of the fence depends heavily on the choice of the cut-off, or tuning parameter, c in (2). In a way, this is similar to one of the difficulties with the information criteria noted in the previous section. Jiang et al. [13] came up with an idea, known as *adaptive fence* (AF), to let the data "speak" on how to choose this cut-off. Let $\mathscr{M}$ denote the set of candidate models. To be more specific, assume that the minimum-dimension criterion is used in selecting the models within the fence. Furthermore, assume that there is a correct model in $\mathscr{M}$ as well as a full model, M_{f},

so that every model in $\mathscr{M}$ is a submodel of $M_{\rm f}$. It follows that $\widetilde{M} = M_{\rm f}$ in (2). First note that, ideally, one wishes to select c that maximizes the probability of choosing the optimal model, here defined as a correct model that has the minimum dimension among all of the correct models. This means that one wishes to choose c that maximizes

$$P = \mathrm{P}\left(M_c = M_{\rm opt}\right), \tag{3}$$

where $M_{\rm opt}$ represents the optimal model and M_c is the model selected by the fence (2) with the given c. However, two things are unknown in (3): (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 $\mathscr{M}$ imply that $M_{\rm f}$ is a correct model. Therefore, it is possible to bootstrap under $M_{\rm f}$. For example, one may estimate the parameters under $M_{\rm f}$ and then use a model-based (or parametric) bootstrap to draw samples under $M_{\rm f}$. This allows us to approximate the probability P on the right side of (3).

To solve problem (ii), we use the idea of maximum likelihood. Namely, let $p^*(M) = \mathrm{P}^*(M_c = M)$, where $M \in \mathscr{M}$ and P^* denote the empirical probability obtained by the bootstrapping. In other words, $p^*(M)$ is the sample proportion of times out of the total number of bootstrap samples that model M is selected by the fence with the given c. Let $p^* = \max_{M\in\mathscr{M}} p^*(M)$. Note that p^* depends on c. The idea is to choose c that maximizes p^*. It should be kept in mind that the maximization is not without restriction. To see this, let M_* denote a model in $\mathscr{M}$ that has the minimum dimension. Note that if $c = 0$, then $p^* = 1$ because the procedure always chooses $M_{\rm f}$. Similarly, $p^* = 1$ for very large c, if M_* is unique (because, when c is large enough, every $M \in \mathscr{M}$ is in the fence; hence, the procedure always chooses M_*). Therefore, what one looks for is "a peak in the middle" of the plot of p^* against c. See Figure 1(a) for an illustration, where $c = c_n$. The highest peak in the middle of the plot (corresponding to approximately $c = 9$) gives the optimal choice. Here is another look at the AF. 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, one is looking for the model (using the fence) that is most supported by the data or, in other words, one that has the highest posterior probability. The latter is estimated by bootstrapping. One then pulls off the c that maximizes the posterior probability.

There are also some technical issues regarding the situation when the optimal model is either $M_{\rm f}$ or M_*. Nevertheless, these issues are mainly of theoretical interest. For example, in most cases of variable selection there are a set of candidate variables and only some of them are important. This means that the optimal model is neither $M_{\rm f}$ nor M_*. We refer the (technical) details on how to handle these extreme cases to Jiang et al. [13].

Note. In the original paper of Jiang et al. [13], the fence inequality (2) was presented with the c on the right side replaced by $c\widehat{\sigma}_{M,\widetilde{M}}$, where $\widehat{\sigma}_{M,\widetilde{M}}$ is an estimated standard deviation of the left side. Although, in some special cases, such as when Q is the negative log-likelihood, $\widehat{\sigma}_{M,\widetilde{M}}$ is easy to obtain, the computation of $\widehat{\sigma}_{M,\widetilde{M}}$, in general, can be time-consuming. This is especially the case for the AF, which calls for repeated computation of the fence under the bootstrap samples. Jiang et al. [14] proposed to merge the factor $\widehat{\sigma}_{M,\widetilde{M}}$ with the tuning constant c, which leads to (2), and use the AF idea to choose the tuning constant adaptively. The latter authors called this modification *simplified adaptive fence* and showed that it enjoys similarly impressive finite-sample performance as the original AF (see below). Thus, throughout this paper, we refer to (2) as the AF.

3. Restricted Fence

The AF has been shown to have outstanding finite-sample performance (Jiang et al. [13, 14]). On the other hand, the method may encounter computational difficulties when applied to high-dimensional and complex problems. The main difficulty rests in the evaluation of a large number of not only the $Q(M)$'s but also their bootstrapped versions, if, for example, the number of candidate variables is large.

Nguyen and Jiang [16] proposed a method to overcome the computational difficulty, called *restricted fence* (RF). The idea was motivated by the restricted maximum likelihood (or residual maximum likelihood, REML), a well-known method in linear mixed model analysis (e.g., Jiang [9]). Consider, for example, a problem of variable selection in linear regression. Let $x_1, \ldots, x_p$ denote the column vectors corresponding to all of the candidate covariates, or predictors. Let X_1 be the data matrix corresponding to a (small) subset of the predictors, say, x_j, $j \in S \subset \{1, \ldots, p\}$, and X_2 the data matrix corresponding to the rest of the predictors. Then, the full model can be expressed as $y = X\beta + \epsilon = X_1\beta^{(1)} + X_2\beta^{(2)} + \epsilon$, where β is the vector of regression coefficients, $\beta^{(1)}$, $\beta^{(2)}$ are subvectors of β corresponding to X_1, X_2, respectively, and ϵ is the vector of regression errors. We then apply a REML-type transformation to the data. Let $n = \dim(y)$ and $p_2 = \mathrm{rank}(X_2)$. Let A be a $n \times (n - p_2)$ matrix satisfying

$$A'A = I_{n-p_2}, \qquad A'X_2 = 0. \tag{4}$$

Then, the transformed data can be expressed as $z = A'y = A'X_1\beta^{(1)} + A'\epsilon = \widetilde{X}_1\beta^{(1)} + \eta$, where the columns of $\widetilde{X}_1$ are $A'x_j$, $j \in S$, and $\eta = A'\epsilon$. Furthermore, if $\epsilon \sim N(0, \sigma^2 I_n)$, where σ^2 is an unknown variance, and I_n is the n-dimensional identity matrix, then $\eta \sim N(0, \sigma^2 I_{n-p_2})$. It is seen that, by applying the transformation, one has reduced the model from $y = X\beta + \epsilon$ to $z = \widetilde{X}_1\beta^{(1)} + \eta$, and $\widetilde{X}_1$ has (much) lower dimension than X. In other words, a larger subset of the variables, X_2, has "disappeared." The AF procedure, introduced in the previous section, is then applied to the model for z to select a subset of variables from $A'x_j$, $j \in S$ or, equivalently, x_j, $j \in S$. Nguyen and Jiang [16] showed that the RF does not depend on the choice of A so long as (4) is satisfied.

In general, to apply the RF, we first divide the predictors into a number of subsets. Let $\{1, \ldots, p\} = \bigcup_{k=1}^{r} S_k$ where the subsets S_k, $1 \le k \le r$ do not have to be disjoint. We then apply the procedure described above to each subset S_k, $1 \le$

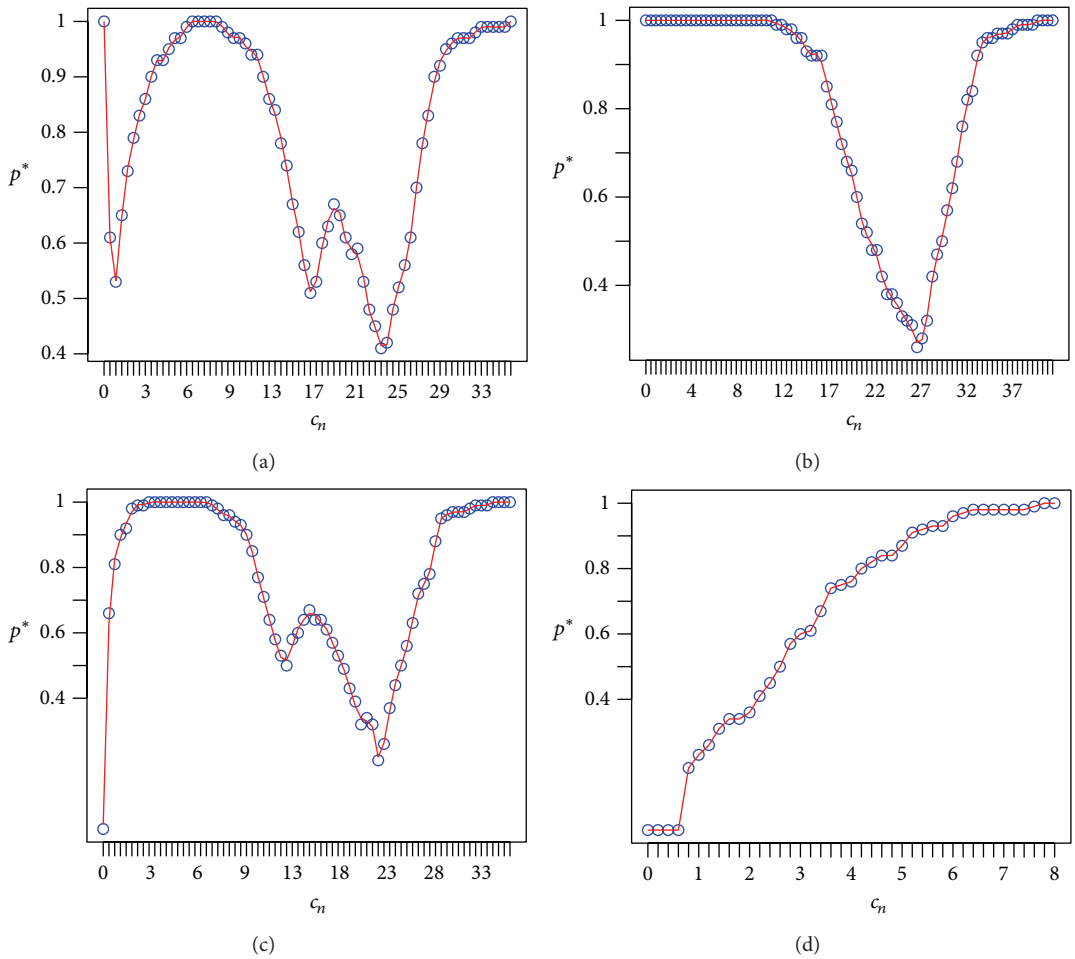

FIGURE 1: A few plots of p^* against $c_n = c$.

$k \leq r$. Some variables are selected from each subset, or no variable is selected from the subset. The final set of variables are selected by combining all the variables selected from the subsets; alternatively, another round of AF may be applied to the variables selected from the subsets to select the final set of variables. We illustrate with an example.

Example 1. Broman and Speed [12] considered a model selection problem in Quantitative Trait Loci (QTL) mapping in a backcross experiment with 9 chromosomes, each with 11 markers. This led to a (conditional) linear regression model with 99 candidate variables, one corresponding to each marker, namely, $y_i = \beta_0 + \sum_{j \in M} \beta_j x_{ij} + \epsilon_i$, $i = 1, \ldots, n$, where y_i is the phenotype of the ith subject, $x_{ij} = 1$ or $x_{ij} = 0$ according to whether the ith subject has the homozygote or heterozygote genotype at the jth marker, and M is a subset to be selected. The errors, ϵ_i, are assumed to be independent $N(0, \sigma^2)$. Nguyen et al. [17] applied RF to this problem, where each subset S_k corresponds to a half-chromosome.

It should be noted that, in order to apply the RF, one must have $p_2 < n$; in other words, there is a dimensionality restriction. For example, in Example 1, the total number of markers is 99, plus an intercept. So, if $p_1 = 6$, one has $p_2 = 99 + 1 - 6 = 94$; thus, the sample size, n, has to be larger than 94. In the simulation studies of Nguyen et al. [17], the authors have considered $n = 250$ or larger. In case the dimension of the model is larger than what the RF can handle, one may first run a dimension reduction using a shrinkage method, such as the Lasso (see Section 7) before applying the RF.

In Nguyen and Jiang [16], the authors offered some tips on how the grouping, or partitioning, of the variables should be done. Namely, the statistician is recommended to work closely with the expert in the subject field (e.g., a scientist or a medical doctor) to get suggestions. The same strategy should also be used in case different partitions lead to different results. See Section 10.3 for further discussion.

4. Invisible Fence

Another variation of the fence that is intended for high-dimensional selection problems is *invisible fence* (IF; Jiang et al. [18]). A critical assumption in Jinag et al. [13] is that there exists a correct model among the candidate models. Although the assumption is necessary in establishing consistency of the fence, it limits the scope of applications because, in practice, a correct model simply may not exist, or exist but not among the candidate models. We can extend the fence by dropping this critical assumption.

Note that the measure of lack-of-fit, Q in (2), typically has the expression $Q(M) = \inf_{\theta_M \in \Theta_M} Q(M, y, \theta_M)$ for some measure Q. A vector $\theta_M^* \in \Theta_M$ is called an optimal parameter vector under M with respect to Q if it minimizes $\mathrm{E}\{Q(M, Y, \theta_M)\}$; that is,

$$\mathrm{E}\left\{Q\left(M, Y, \theta_M^*\right)\right\} = \inf_{\theta_M \in \Theta_M} \mathrm{E}\left\{Q\left(M, Y, \theta_M\right)\right\} \equiv Q\left(M\right), \quad (5)$$

where the expectation is with respect to the true distribution of Y (which may be unknown, but not model-dependent). A correct model with respect to Q is a model $M \in \mathscr{M}$ such that

$$Q\left(M\right) = \inf_{M' \in \mathscr{M}} Q\left(M'\right). \quad (6)$$

When M is a correct model with respect to Q, the corresponding θ_M^* is called a true parameter vector under M with respect to Q. Note that here a correct model is defined as a model that provides the best approximation, or best fit to the data (note that Q typically corresponds to a measure of lack-of-fit), which is not necessarily a correct model in the traditional sense. However, the above definitions are extensions of the traditional concepts in model selection (e.g., Jiang et al. [13]). The main difference is that, in the latter reference, the measure Q must satisfy a minimum requirement that $\mathrm{E}\{Q(M, y, \theta_M)\}$ is minimized when M is a correct model, and θ_M a true parameter vector under M. With the extended definition, the minimum condition is no longer needed, because it is automatically satisfied.

Let us now take another look at the fence. To be specific, assume that the minimum-dimension criterion is used to select the optimal model within the fence. In case there are ties (i.e., two models within the fence, both with the minimum dimension), the model with the minimum dimension and minimum $Q(M)$ will be chosen. Also recall the cut-off c in (2). As it turns out, whatever one does in choosing the cut-off (adaptively or otherwise), only a fixed small subset of models have nonzero chance to be selected; in other words, the majority of the candidate models do not even have a chance. We illustrate this with an example. Suppose that the maximum dimension of the candidate models is 3. Let $M_j^\dagger$ be the model with dimension j such that $c_j = Q(M_j^\dagger)$ minimizes $Q(M)$ among all models with dimension j, $j = 0, 1, 2, 3$. Note that $c_3 \le c_2 \le c_1 \le c_0$; assume no equality holds for simplicity. The point is that any $c \ge c_0$ does not make a difference in terms of the final model selected by the fence, which is $M_0^\dagger$. Similarly, any $c_1 \le c < c_0$ will lead to the selection of $M_1^\dagger$; any $c_2 \le c < c_1$ leads to the selection of $M_2^\dagger$; any $c_3 \le c < c_2$ leads to the selection of $M_3^\dagger$; and any $c < c_3$ will lead to nonselection, because no model is in the fence. In conclusion, any fence methods, adaptive or otherwise, will eventually select a model from one of the four: $M_j^\dagger$, $j = 0, 1, 2, 3$. The question is: Which one?

To solve this problem we use the AF idea by drawing bootstrap samples, say, under the full model. The idea is to select the model that has the highest empirical probability to best fit the data when controlling the dimension of the model. More specifically, for each bootstrap sample, we find the best-fitting model at each dimension, that is, $M_j^{*\dagger}$, such that $Q^*(M_j^{*\dagger})$ minimizes $Q^*(M)$ for all models with dimension j, where Q^* represents Q computed under the bootstrap sample. We then compute the relative frequency, out of the bootstrap samples, for different models selected, and the maximum relative frequency, say p_j^*, at each dimension j. Let $M_{j^*}^{*\dagger}$ be the model that corresponds to the maximum p_j^* (over different j's) and this is the model we select. In other words, if at a certain dimension we find a model that has the highest empirical probability to best fit the data, this is the model we select. As in Jiang et al. [13], some extreme cases (in which the relative frequencies always equal to one) need to be handled differently. Although the new procedure might look quite different from the fence, it actually uses implicitly the principle of the AF as explained above. For such a reason, the procedure is called invisible fence, or IF.

Computation is a major concern in high dimension problems. For example, for the 522 gene pathways developed by Subramanian et al. [19], at dimension $k = 2$ there are 135,981 different $Q(M)$'s to be evaluated; at dimension $k = 3$ there are 23,570,040, ... If one has to consider all possible k's, the total number of evaluations is 2^{522}, an astronomical number. Jiang et al. [18] proposed the following strategy, called the *fast algorithm*, to meet the computational challenge. Consider the situation where there are a (large) number of candidate elements (e.g., gene-sets, variables), denoted by $1, \ldots, m$, such that each candidate model corresponds to a subset of the candidate elements. A measure Q is said to be *subtractive* if it can be expressed as

$$Q\left(M\right) = s - \sum_{i \in M} s_i, \quad (7)$$

where s_i, $i = 1, \ldots, m$ are some nonnegative quantities computed from the data, M is a subset of $1, \ldots, m$, and s is some quantity computed from the data that does not depend on M. Typically, we have $s = \sum_{i=1}^m s_i$, but the definition does not impose such a restriction. For example, in gene-set analysis (GSA; e.g., Efron & Tibshirani [20]), s_i corresponds to the gene-set score for the ith gene-set. As another example, Mou [21] considered $s_i = |\widehat{\beta}_i|$, where $\widehat{\beta}_i$ is the estimate of the coefficient for the ith candidate variable under the full model, in selecting the covariates in longitudinal studies.

For a subtractive measure, the models that minimize $Q(M)$ at different dimensions are found almost immediately. Let $r_1, r_2, \ldots, r_m$ be the ranking of the candidate elements in terms of decreasing s_i. Then, the model that minimizes $Q(M)$ at dimension one is r_1; the model that minimizes $Q(M)$ at dimension two is $\{r_1, r_2\}$; the model that minimizes $Q(M)$ at dimension three is $\{r_1, r_2, r_3\}$, and so on.

Jiang et al. [18] implemented the IF with the fast algorithm, for which a natural choice of s_i is the gene-set score, as noted, and showed that IF significantly outperforms GSA in simulation studies. The authors also showed that IF has a nice theoretical property, called *signal-consistency*, and that GSA does not have this property. See Section 8 for details.

5. Predictive Model Selection

In many cases, the problem of interest is prediction rather than estimation. For example, problems of practical interest often arise in the context of mixed model prediction, in which the quantity of interest can be expressed in terms of a mixed effect. See, for example, Robinson [22], Jiang and Lahiri [23], for reviews on prediction of mixed effects and its applications. As noted (near the end of Section 1), the fence is flexible in choosing the measure of lack-of-fit, Q in (2), to incorporate the problem of interest. Now because prediction is of main interest, it makes sense to use a predictive measure of lack-of-fit rather than an estimation-based measure. Examples of the latter include the negative log-likelihood and the residual sum of squares.

To derive a predictive measure, let us consider a general problem of mixed model prediction (e.g., Robinson [22]). The assumed model is

$$y = X\beta + Zv + e, \tag{8}$$

where X, Z are known matrices; β is a vector of fixed effects; v, e are vectors of random effects and errors, respectively, such that $v \sim N(0, G)$, $e \sim N(0, \Sigma)$, and v, e are uncorrelated. An important issue for model-based statistical inference is the possibility of model misspecification. To take the latter into account, suppose that the true underlying model is

$$y = \mu + Zv + e, \tag{9}$$

where $\mu = \mathrm{E}(\mathbf{y})$. Here, E represents expectation with respect to the true distribution of y, which may be unknown but is not model-dependent. So, if $\mu = X\beta$ for some β, the model is correctly specified; otherwise, the model is misspecified. Our interest is prediction of a vector of mixed effects that can be expressed as

$$\theta = F'\mu + R'v, \tag{10}$$

where F, R are known matrices. We illustrate with an example.

Example 2 (Fay-Herriot model). Fay and Herriot [24] assumed the following model to estimate the per capita income of small places with population size less than 1,000: $y_i = x_i'\beta + v_i + e_i$, $i = 1, \ldots, m$, 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 are sampling errors. It is assumed that v_i's, e_i's 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. The assumed model can be expressed as (8) with $X = (x_i')_{1\le i\le m}$, $Z = I_m$, $G = AI_m$, and $\Sigma = \mathrm{diag}(D_1, \ldots, D_m)$. The problem of interest is estimation of the small area means. Let $\mu_i = \mathrm{E}(y_i)$. Then, the small area means can be expressed as $\theta_i = \mathrm{E}(y_i \mid v_i) = \mu_i + v_i$, under the true underlying model (9). Thus, the quantity of interest can be expressed as (10) with $\theta = (\theta_i)_{1\le i\le m}$ and $F = R = I_m$.

For simplicity, assume that both G and Σ are known. Then, under the assumed model, the best predictor (BP) of θ, in the sense of minimum mean squared prediction error (MSPE), is the conditional expectation

$$\begin{aligned}\mathrm{E}_M(\theta \mid y) &= F'\mu + R'\mathrm{E}_M(v \mid y)\\ &= F'X\beta + R'GZ'V^{-1}(y - X\beta),\end{aligned} \tag{11}$$

where $V = \Sigma + ZGZ'$ and β is the true vector of fixed effects (e.g., Jiang [9, page 75]). The E_M in (11) denotes conditional expectation under the assumed model (8) rather than the true model (9). Write $B = R'GZ'V^{-1}$ and $\Gamma = F' - B$. Let $\widetilde{\theta}$ denote the right side of (11). The predictive performance of $\widetilde{\theta}$ is typically measured by the MSPE, defined as $\mathrm{MSPE}(\widetilde{\theta}) = \mathrm{E}(|\widetilde{\theta} - \theta|^2)$. Here, again, E denotes expectation under the true model. Jiang et al. [25] showed that the MSPE has another expression, which is the key to our derivation:

$$\mathrm{MSPE}\left(\widetilde{\theta}\right) = \mathrm{E}\left\{(y - X\beta)'\Gamma'\Gamma(y - X\beta) + \cdots\right\}, \tag{12}$$

where $\cdots$ does not depend on β. Note that, unlike $\mathrm{E}(|\widetilde{\theta} - \theta|^2)$, the first term inside the expectation on the right side of (12) is a function of the observed data and β, and $\cdots$ is unrelated to the model, or the parameters. Thus, a natural predictive measure of lack-of-fit, $Q(M)$, is the minimizer, over β, of the expression inside the expectation in (12) without $\cdots$. Clearly, this measure is designed specifically for the mixed model prediction problem. Also, when it comes to model selection, it is important that the measure of lack-of-fit is "fair" to every candidate model. The above measure $Q(M)$ has this feature, because the expectation in (12) is under an objective true model. Once we have the measure Q, we can use it in (2) for the fence.

The idea of the derivation can be generalized in several ways. For example, in case that G is completely unknown, it can be shown that essentially the same derivation of Jiang et al. [25] goes through, and the resulting measure of lack-of-fit, $Q(M)$, is the minimizer of $(y - X\beta)'\Gamma'\Gamma(y - X\beta) - 2tr(\Gamma'\Sigma)$, assuming that Σ is known. It is also possible to extend the idea of predictive model selection to the generalized linear mixed models (GLMMs; e.g., Jiang [9]). Consider, for example, a problem of small area estimation (e.g., Rao [26]) with count data. The observation, y_i, is at the area-level (similar to the Fay-Herriot model) along with a vector of covariates, x_i, also at the area-level. The model of interest assumes that, given the area-specific random effects, v_i, $y_i \sim \mathrm{Poisson}(\mu_i)$ such that $\log(\mu_i) = x_i'\beta + v_i$. The BP of μ_i, under the assumed model, can be expressed as $\mathrm{E}_{M,\psi}(\mu_i \mid y) = g_{M,i}(\psi, y_i)$, where $\mathrm{E}_{M,\psi}$ denotes conditional expectation under the assumed model, M, and parameter vector, ψ, under M, and $g_{M,i}(\cdot,\cdot)$ is a known function which may not have an analytic expression. Nevertheless, the g function can be evaluated numerically fairly easily. Following a similar idea, we evaluate the performance of the BP under a broader model, which is the same as the assumed model except without the log-linear model $\log(\mu_i) = x_i'\beta + v_i$. In other words, under the broader model, the μ_i's are completely unspecified. Let $\mu = (\mu_i)_{1\le i\le m}$ and

$\widetilde{\mu} = (\widetilde{\mu}_i)_{1\le i\le m}$. Chen et al. [27] showed that, similarly, MSPE = $\mathrm{E}(|\widetilde{\mu} - \mu|^2)$ =

$$\mathrm{E}\left\{\sum_{i=1}^{m} g_{M,i}^2(\psi, y_i) - 2\sum_{i=1}^{m} g_{M,i}(\psi, y_i - 1)\, y_i + \cdots\right\}. \quad (13)$$

A predictive measure of lack-of-fit, $Q(M)$, is then given by the minimizer, over ψ, of the expression inside the expectation in (13) without $\cdots$.

6. Model Selection with Incomplete Data

The missing-data problem has a long history (e.g., Little and Rubin [28]). While there is an extensive literature on statistical analysis with missing or incomplete data, the literature on model selection in the presence of missing data is relatively sparse. See, Jiang et al. [29] for a review of literature on model selection with incomplete data. Existing model selection procedures face special challenges when confronted with missing or incomplete data. Obviously, the naive complete-data-only strategy is inefficient, sometimes even unacceptable by the practitioners due to the overwhelmingly wasted information. For example, in a study of backcross experiments (e.g., Lander and Botstein [30]), a dataset was obtained by researchers at UC-Riverside (personal communications; see Zhan et al. [31] for a related work). Out of the 150 or so subjects, only 4 have complete data record. Situations like this are, unfortunately, the reality that we often have to deal with.

Verbeke et al. [32] offered a review of formal and informal model selection strategies with incomplete data, but the focus is on model comparison, instead of model selection. As noted by Ibrahim et al. [33], while model comparisons "demonstrate the effect of assumptions on estimates and tests, they do not indicate which modeling strategy is best, nor do they specifically address model selection for a given class of models." The latter authors further proposed a class of model selection criteria based on the output of the E-M algorithm. Jiang et al. [29] point out a potential drawback of the E-M approach of Ibrahim et al. [33] in that the conditional expectation in the E-step is taken under the assumed (candidate) model, rather than an objective (true) model. Note that the complete-data log-likelihood is also based on the assumed model. Thus, by taking the conditional exception, again, under the assumed model, it may bring false supporting evidence for an incorrect model. Similar problems have been noted in the literature, which are sometimes referred to as "double-dipping" (e.g., Copas and Eguchi [34]).

On the other hand, the AF idea (see Section 2) works naturally with the incomplete data. For the simplicity of illustration, let us assume, for now, that the candidate models include a correct model as well as a full model. It then follows that the full model is, at least, a correct model, even though it may not be the most efficient one. Thus, in the presence of missing data, we can run the E-M to obtain the MLE of the parameters, under the full model. Note that, here, we do not have the double-dipping problem, because the conditional expectation (under the full model) is "objective." Once the parameter estimates are obtained, we can use the model-based (or parametric) bootstrap to draw samples, under the full model, as in the AF. The best part of this strategy is that, when one draws the bootstrap samples, one draws samples of complete data, rather than data with the missing values. Therefore, one can apply any existing model selection procedure that is built for the complete-data situation, such as the fence, to the bootstrap sample. Suppose that B bootstrap samples are drawn. The model with the highest frequency of being selected (as the optimal model), out of the bootstrap samples, is the (final) optimal model. We call this procedure the EMAF algorithm due to its similarity to the AF idea. We can extend the EMAF idea to situations where a correct model may not exist, or exists but not among the candidates. In such a case, the bootstrap samples may be drawn under a model $\widetilde{M}$, which is the model with the minimum Q in the sense of the second paragraph of Section 4.

7. Shrinkage Model Selection

Simultaneous variable selection and estimation by penalized likelihood methods have received considerable interest in the recent literature on statistical learning (see Fan and Lv [35] for a review). The developments followed the original idea of Tibshirani [36], who proposed the least absolute shrinkage and selection operator (Lasso) to automatically select variables via continuous shrinkage. The method has gained popularity thanks to the developments of several computation algorithms (e.g., Fu [37], Efron et al. [38]). Fan and Li [39] proposed the smoothly clipped absolute deviation method (SCAD) which has the *oracle property* under certain conditions. Zou [40] proposed the adaptive Lasso and showed that it also has the oracle property.

The selection of the regularization parameter in the penalized likelihood approach is an important problem which determines the dimension of the selected model. Consider, for example, a linear regression model: $y = X\beta + \epsilon$, where $y = (y_1, \ldots, y_n)'$ is an $n \times 1$ vector of responses, $X = (x_1, \ldots, x_p)$ is an $n \times p$ matrix of predictors, $\beta = (\beta_1, \ldots, \beta_p)'$ is the vector of regression coefficients, and ϵ is a vector of independent errors with mean 0 and variance σ^2. In a high-dimensional selection problem, β is supposed to be sparse, which means that most of the predictors do not have relationship with the response. The penalized likelihood is designed to identify which components of β are to be shrunk to zero and estimate the rest. The penalized likelihood estimator is

$$\widehat{\beta} = \arg\min_{\beta} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \sum_{j=1}^{p} \rho\left(\lambda_j, \left|\beta_j\right|\right), \quad (14)$$

where $\|\cdot\|$ is the Euclidean norm and $\rho(\lambda_j, \cdot)$ is a given nonnegative penalty function that depends on the regularization parameter, λ_j. For Lasso, the ℓ_1 penalty is used with $\rho(\lambda_j, |\beta_j|) = \lambda|\beta_j|$; for SCAD, the penalty is a continuous differentiable function whose derivative is defined by $(\partial/\partial\theta)\rho(\lambda_j, \theta) = \lambda\{1_{(\theta\le\lambda)} + \{(a-1)\lambda\}^{-1}(a\lambda - \theta)_+ 1_{(\theta>\lambda)}\}$ for $a > 2$ and $\theta > 0$. The parameter a is shown to be fairly stable for the SCAD and is usually set as 3.7 (Fan and Li [39]). Thus, for both Lasso and SCAD, one needs to determine λ in order

to find an optimal model. It should be noted that the choice of λ is critically important because, for example, depending on the value of λ, it is possible that none of the β's are shrunk to zero, meaning that all of the predictors are selected, or all of the β's are shrunk to zero, meaning that none of the predictors are selected.

One approach to selecting regularization parameter is to use the information criteria such as the AIC (Akaike [2]) and BIC (Schwarz [5]). Some data-driven approaches, such as the delete-d CV (Shao [15]), have also been proposed. More recent work can be found in Wang et al. [41] and Zhang et al. [42]. These procedures are consistent if the regularization parameter satisfies some order conditions. However, there may be many functions satisfying these conditions and their finite-sample behaviors may differ. Pang et al. [43] developed a data-driven approach to choosing the regularization parameter using an idea very similar to the AF. To be more specific, consider the Lasso and assume, for now, that $p < n$. In this case, one can fit the full model, which corresponds to (14) with $\lambda = 0$. One then draws bootstrap samples under the full model. For each bootstrap sample, and a given value of λ among a collection of candidate λ values, run the Lasso and a model is selected. One then computes the relative frequencies, out of the bootstrap samples, for different models that are selected, and the maximum of those frequencies (maximum frequency). Note that the maximum frequency depends on λ, which is exactly what one wants. The optimal λ is the one corresponding to the largest maximum frequency among the candidate λ's. Pang et al. [43] showed, via simulation studies, that their procedure performs significantly better than the information criteria and CV for choosing the regularization parameter, λ. The authors also developed an accelerated algorithm to improve the computational efficiency. Using a similar idea, Melcon [44] developed a procedure or choosing the regularization parameter in the penalized likelihood method for generalized linear models (GLM).

Here is another look at the bootstrap procedures developed by Pang et al. [43] and Melcon [44]: not only are the procedures similar to AF, they are, indeed, also special cases of the AF. To see this, let $\widehat{\lambda}(M)$ be the smallest regularization parameter, λ, for which model M is selected. It is easy to see that $\widehat{\lambda}(M)$ is a measure of lack-of-fit in the sense that "larger" model has smaller $\widehat{\lambda}(M)$. More specifically, if M_1 is a submodel of M_2, then one must have $\widehat{\lambda}(M_1) \geq \widehat{\lambda}(M_2)$. In particular, we have $\widehat{\lambda}(M_{\mathrm{f}}) = 0$, where M_{f} is the full model. Thus, if we let $Q(M) = \widehat{\lambda}(M)$ and $\widetilde{M} = M_{\mathrm{f}}$ in (2), the fence inequality becomes $\widehat{\lambda}(M) \leq \mathrm{c}$. It follows that the selection of the optimal λ is equivalent to selecting the optimal cut-off c; hence the procedures of Pang et al. [43] and Melcon [44] may be viewed as special cases of AF with $Q = \widehat{\lambda}$.

So far the illustration has focused on the case $p < n$ (otherwise, fitting the full model may be problematic). However, Pang et al. [43] actually applied their method to what they call divergent and NP-dimensionality situations, where p is much larger than n. To do so, the authors first used the sure independence screening (SIS; Fan and Lv [45]), to reduce the dimension to less than n, then applied the method described above. Alternatively, one could first run the Lasso with a positive λ (which would allow $p > n$) to reduce the dimension to $p < n$ and then apply the methods of Pang et al. [43] or Melcon [44].

8. Asymptotic Properties

A standard large sample property for model selection is consistency in model selection. From a classical point of view, this means that, as the sample size increases, the probability that the selected model is the true underlying model, also known as the optimal model, goes to one. See, for example, Jiang et al. [13]. The latter authors established consistency of various fence procedures, including the AF, under this classical setting. Also see Jiang et al. [14]. The consistency of AF requires that both the sample size, n, and the bootstrap sample size, B, go to infinity. For the most part, there are two regularity conditions. The first condition states that there is a separation in terms of the measure of lack-of-fit between the correct models and the incorrect ones. The second condition states that, as both n and B increase, the bootstrap approximation to the probability distribution P in (3) is accurate. These conditions, although stated in a technical way, are reasonable.

A key assumption for the classical consistency is that the sample size goes to infinity. However, this assumption is not always practical. For example, in genome-wide association study (e.g., Hindorff et al. [46]), the number of genetic variants in the genome that are potentially associated with certain traits of interest is usually much larger than the number of individuals involved in the study. In other words, the number of parameters is much larger than the sample size. Jiang et al. [18] introduced a different type of consistency property, called *signal-consistency*. The idea is to consider an "ideal" situation where the "strength" of the parameters—the signals—increases, and shows that, in such a case, the probability of selecting the optimal model goes to one. Of course, in real life, one may not be able to increase the signal, but the same can be said regarding increasing the sample size, in real life. The point for signal consistency is as fundamental as consistency in the classical sense; that is, a procedure would work perfectly well, at least, in an ideal situation. The latter authors argued that signal-consistency is a more reasonable asymptotic property than the classical consistency in a situation where the sample size is (much) smaller than the number of parameters. The authors further established signal consistency of the IF (see Section 4), under some regularity conditions.

9. Examples

We illustrate empirical performance and application of the fence via a few examples.

9.1. Fay-Herriot Model Revisited: A Simulation Study. The Fay-Herriot model was introduced in Section 5 (see Example 2). A simulation study was carried out in Jiang et al. [13] to compare the AF with several other nonadaptive

Table 1: Fence with different choice of c in the F-H model.

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

choices of the cut-off, c, in (2). The measure Q was taken as the negative log-likelihood, and the sample size was $n = m = 30$. The candidate predictors are $x_1, \ldots, x_5$, generated from the $N(0, 1)$ distribution and then fixed throughout the simulations. The candidate models included all possible models with at least an intercept. Five cases were considered, in which the data y were generated under the model $y = \sum_{j=1}^{5} \beta_j x_j + \nu + e$, 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 Models 1, 2, 3, 4, and 5, respectively. The authors considered the simple case $D_i = 1$, $1 \le i \le n$. The true value of A is 1 in all cases. The number of bootstrap samples for the evaluation of the p^*'s is 100. In addition to the AF, five different nonadaptive choices of $c = c_n$ were considered, which satisfy the consistency requirements given in Jiang et al. [13], namely, $c_n \to \infty$ and $c_n/n \to 0$ in this case. The results are presented in Table 1, which were the percentage of times, out of the 100 simulations, that the optimal model was selected by each method. It is seen that the performances of the fence with $c = \log(n)$, $\sqrt{n}$, or $n/\log(n)$ are fairly close to that of the AF. Of course, in any particular case, one might get lucky to find a good c value, but one cannot be lucky all the time. Regardless, the AF always seems to pick up the optimal value, or something close to the optimal value, of c in terms of the finite-sample performance.

9.2. Ultrahigh Dimensional Selection Problem: A Simulation Study. Pang et al. [43] carried out a simulation study involving what they called ultrahigh dimensional data, in which p is much larger than n. The example was originally given by Fan and Lv [45]. The data were generated under the model $y_i = x_i'\beta + \epsilon_i$, where $\beta = (\beta_1', 0_{p_n-8})'$, β_1 being an 8-dimensional vector with each component being of the form $(-1)^u(a_n + |z|)$ with $a_n = 4\log(n)/\sqrt{n}$, u was generated from the Bernoulli(0.4) distribution, and z was generated from the $N(0, 1)$ distribution. Furthermore, $\epsilon_i, i = 1, \ldots, n$ were generated independently from the $N(0, 1.5^2)$ distribution. The predictors x_i, $i = 1, \ldots, n$ were independent $N(0, I)$ vectors. The authors set $n = 200$ and $p_n = 1000$ in their simulation study. Using the SIS (Fan and Lv [45]), the authors first reduced the dimensionality from 1000 to nsis $= [5.5n^{2/3}] = 188$, before applying the Lasso with the regularization parameter chosen by the method based on the AF idea (see Section 7). Note that 188 is only slightly smaller than the sample size n. The authors compared their method (SIS + Lasso/AF) with the SIS + Lasso with the regularization parameter chosen by BIC (SIS + Lasso/BIC).

Table 2: SIS + shrinkage methods with selection of regularization parameter.

Method	NC	NIC	TP	UF	OF
SIS + Lasso/AF	7.99	0.00	0.99	0.01	0.00
SIS + Lasso/BIC	8.00	14.87	0.01	0.00	0.99
SIS + ALsso/AF	7.99	0.00	0.99	0.01	0.00
SIS + ALsso/BIC	8.00	2.62	0.37	0.00	0.63
SIS + SCAD/AF	8.00	0.00	1.00	0.00	0.00
SIS + SCAD/BIC	8.00	1.48	0.62	0.00	0.38

Similar comparisons were made for adaptive Lasso (ALasso) and SCAD as well. The results based on 100 simulation runs are summarized in Table 2. Here, a few measures of performance are considered: NC denotes average (over the simulations) number of correctly identified nonzero coefficients; NIC denotes average number of incorrectly identified "nonzero coefficients"; TP denotes the empirical probability (again, over the simulations) of true positive (i.e., the selected variables are exactly those with the nonzero coefficients); UF stands for the empirical probability of underfitting (i.e., at least one variable with nonzero coefficient is missing); and OF stands for overfitting (i.e., the selected variables include all of those with the nonzero coefficients, plus something else). It is seen that the AF-based methods perform nearly perfectly by all measures.

9.3. Analysis of Cytokines Data. 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 includes 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 three time points: baseline, after 6 and 12 months. A problem of interest is to model the Cytokines (IL1BBLLA, TNFABLLA, and IL6BLLA)—inflammatory markers—over time on gene expression for IFNb and cFos, along with other variables. In other words, one is interested in finding a subset of relevant variables/covariates that contribute to the variation of Cytokines. There are a total of 36 covariate variables. See Appendix A.3 of Nguyen and Jiang [16] for details. Due to the relatively large number of candidate variables, the latter authors used RF to analyze the data. The results for IL1BBLLA were reported. The covariate variables were

Table 3: Modeling IL1BBLLA. Symbol × indicates variable selected; variables not listed were not selected by any of the comparing methods.

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

grouped into four groups according to biological interest. More specifically, one of the authors worked closely with an expert scientist in 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 Nguyen and Jiang [16], Supplementary Appendix, for details). The RF results were compared with other procedures, reported in Table 3, where BIC, CAIC, HQ, and AIC refer to the forward-backward (F/B) BIC (Broman and Speed [12]) and similarly for the other information criteria (see Nguyen and Jiang [16] description). The F/B procedures were used due to the large number of candidate models (therefore all-subset selection is computationally too expensive).

The main objective of the study was to examine whether Soy Isoflavones treatment affects the bone metabolism. This treatment effect was selected by RF, AIC, and SCAD, but not by the other methods. The weight variable was picked up by AIC and HQ, but not by other procedures; however, the BMI variable, which is a function of weight and height, was picked up by RF and SCAD. As also seen in the table, BMD for lumbar and spine measures (LSTBMD) was picked up by RF, but not by any other methods. Apparently in this analysis, BIC, CAIC, HQ, and the adaptive Lasso had overpenalized; as a result, their optimal models did not pick up some of the relevant covariates, such as BMD and BMC (ALasso did not pick up any of the variables). As for AIC, it was 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, it was confirmed that BMD and BMC are more important variables than area measures in this case. Thus, the results of the RF data analysis were more clinically relevant. Although SCAD had selected the most variables, it had missed the important variable LSTBMD. As for the total body area (WBodArea) that was uniquely picked up by SCAD, the variable is relatively less important, compared to the BMD and BMC, as noted. In fact, simulation studies (Nguyen and Jiang [16]) have suggested that SCAD has the tendency of selecting extraneous variables as well as missing relevant ones.

Some sample R codes under a very similar setting to the analysis presented here are vailable at http://www.stat.ucdavis.edu/jiang/fenc.ecodes/fencecodes.zip.

10. Challenges and Future Directions

We have discussed a number of variations of the fence as well as some areas where the fence idea seems to be at play. It is seen that some of these variations (AF, RF, and IF) are more developed than others, leaving rooms for further research. In fact, some ideas and thoughts are being explored. Furthermore, there are some general questions regarding the fence methods. Let us start with the latter.

10.1. Connections to other Fields. It may be argued that, at least, some part of the fence idea has already been used in other fields, possibly in different forms. It would be extremely beneficial, and interesting, to explore such connections. For example, Ethan Anderes (personal communication) has pointed out the following apparent connection between the information criteria, (1), and the fence. In numerical analysis, two forms of optimization problems are often equivalent: (i) minimizing $f(x)$ subject to $g(x) \leq c$ for some constant c and (ii) minimizing $f(x) + bg(x)$ for some constant b. For example, Tibshirani [36] originally proposed the Lasso in the form of (i); it is nowadays usually treated in the form of (ii), as in (14). The point is, if one lets the $Q(M)$ in (2) be the same as the $\widehat{Q}_M$ in (1), the connection between the fence and information criteria is quite obvious, in view of the equivalence noted above, provided that the minimum-dimension criterion is used to select the model within the fence.

As another example, Iain Johnstone (personal communications) has pointed out the following connection of the IF to orthogonal thresholding. Consider the simplest orthogonal model $y_i = \mu_i + z_i$, $i = 1, \ldots, n$, where the μ_i's are unknown means and z_i's are independent $N(0, 1)$ errors. For $M \subset \{1, \ldots, n\}$, the LS estimator of μ_i is the projection, $(P_M y)_i = y_i$, if $i \in M$, and 0 if $i \notin M$. The RSS is $Q(M) = |y - P_M y|^2 = \sum_{i \notin M} y_i^2$. Here, the full model, $\widetilde{M} = \{1, \ldots, n\}$, has $Q(\widetilde{M}) = 0$,

so the fence criterion (2) yields Fence$(c) = \{M : \sum_{i\notin M} y_i^2 \le c\}$. To see the connection to thresholding, define the order statistics, $y_{(1)}^2 \le \cdots \le y_{(n)}^2$, and $\widehat{K} = \widehat{K}(c)$ as the random integer $\widehat{K}$ satisfying $\sum_{i=1}^{\widehat{K}} y_{(i)}^2 \le c < \sum_{i=1}^{\widehat{K}+1} y_{(i)}^2$. The smallest model inside Fence(c) is obtained by having as many small y_i^2 out of the model as possible; hence, the model selected by the fence with the given c is $M_0(c; y) = \{i : |y_i| > \widehat{t}(c) = |y_{(\widehat{K})}|\}$. It follows that $M_0(c; y)$ is obtained by hard thresholding at $\widehat{t}(c) = |y_{(\widehat{K})}|$. Furthermore, some calculations are motivated by the AF idea. Consider a particular class of configurations under the model. For a fixed (small) K and a value μ, set $\mu_i = \mu, i = 1, \ldots, K$, and $\mu_i = 0, i = K + 1, \ldots, n$. Consider the set of variables chosen by one-sided thresholding at λ, $M_0(\lambda) = \{i : y_i > \lambda\}$. Note that the true model under this configuration is $M_0 = \{1, \ldots, K\}$. Suppose that one wishes to study

$$p(\lambda) = \max_M \mathrm{P}_{\mu,K}\{M_0(\lambda) = M\}, \tag{15}$$

where $\mathrm{P}_{\mu,K}$ denotes probability under the true model. It is conjectured that the model M_0 achieves the maximum in (15). Given that this conjecture is true, then, we have

$$p(\lambda) = \overline{\Phi}(\lambda - \mu)^K \Phi(\lambda)^{n-K}, \tag{16}$$

where Φ is the standard normal cdf and $\overline{\Phi} = 1 - \Phi$. Let λ^* be the maximizer of (16); that is, $p(\lambda^*) = \max_\lambda \mathrm{P}_{\mu,K}\{M_0(\lambda) = M_0\}$. It is claimed that, with all of these assumptions, one can show the following facts. (I) If $\mu = 0$, we have $\lambda^* = z_{K/n} \approx \sqrt{2\log(n/K)}$, where z_α is the α-critical value of the standard normal distribution. (II) For $\mu \ge 0$, the corresponding $\lambda^* = \lambda^*(\mu)$ is increasing in μ. The claims are supported by the results of a small simulation study with $n = 200$ and $K = 5$ (omitted).

10.2. Software Development. Software development for AF has been a main challenge in the implementation of the fence. Recall that the procedure requires finding a peak in the middle of the plot of the p^* against c (see Section 2). This may be easier said than done. For example, Figure 1(a) shows a "good" case of such a plot. It is a good case because there is a clear high peak in the middle, which corresponds to the optimal c. In fact, the plots in Figure 1 show some standard patterns, from which a decision is often relatively easy to make. For example, Figure 1(c) has a "flat" area where the corresponding p^* is equal to 1. In such a case, one may choose c as the median of the interval over which the p^* is equal to 1. A similar rule may be applied to Figure 1(b). As for Figure 1(d), it may correspond to a case where the minimum model, M_*, would be selected. In many cases, however, one is not so lucky to have such a standard plot, and this would especially be the case when the sample size is limited, or when the "signals" are weak (see Section 8).

It is not uncommon to have multiple peaks, such as those in Figure 1, but the difficulty is that the highest peak (such as the one to the left in Figure 1(a)) is not necessarily the best choice. Once again, this often happens when the signal strength is relatively weak. Nguyen and Jiang [16] presented a number of such plots. The latter authors found that the "first significant peak" may be a better strategy in situations of moderate sample size, or relatively weak signal. The latter observation is supported by Melcon [44] in the context of shrinkage GLM selection. Müller et al. [47] find a simple rule that was "surprisingly successful" in identifying the optimal model. In another related work, Jiang et al. [14] suggest to use a confidence lower bound to identify the cut-off c for the AF in cases of relatively small sample size, or weak signal.

Furthermore, the plot of p^* versus c typically has small bumps due to the bootstrap errors, and one does not want to count these bumps as the peaks, especially when it "counts" (i.e., has impact on the decision). Nguyen and Jiang [16] used the **loess**( ) in R to "smooth out" the plot and showed that the smoothing is helpful, especially in the case of weak signals.

In a way, the process of inspecting the plot and finding the "optimal" c is easier to do with human eyes than, say, with a computer (so, in this regard, there is no doubt that a human being is "smarter" than a computer). However, identifying the c with human eyes raises concerns of subjectiveness and opens the door for potential bias. Can a software package be developed that would allow the computer to do exactly what the human eyes would do? For example, the computer may "match" the observed plot to one of the standard plots, such as those shown in Figure 1. The software package would also be very useful in simulation studies—it is impractical, although not impossible, to inspect hundreds, or even thousands of plots with the human eyes in order to summarize the results. Clearly, such a "computer visualization" is likely to have revolutionary impact not only on the fence but also on statistical computing in general.

10.3. Grouping in RF. One problem associated with the RF (see Section 3) is how to divide the candidate variables into the groups. Nguyen and Jiang [16] suggests that the grouping "should be based on biological information." However, it is not clear what to do if no biological information is available, or the information does not help, say, in reducing the group sizes to something that is computationally manageable.

Consider, for simplicity, the problem of regression variable selection. Suppose that the candidate variables are $x_1, \ldots, x_K$ and that some initial estimates of the corresponding regression coefficients, $\widehat{\beta}_1, \ldots, \widehat{\beta}_K$ are obtained. We can sort the estimates according to decreasing order of their absolute values, say, $|\widehat{\beta}_{(1)}| \ge \cdots \ge |\widehat{\beta}_{(K)}|$. Let $\widetilde{x}_1, \ldots, \widetilde{x}_K$ be the candidate variables corresponding to $\widehat{\beta}_{(1)}, \ldots, \widehat{\beta}_{(K)}$. It has been observed that RF performs better when the true underlying model is neither the minimum model nor the full model. Based on this observation, a potential grouping strategy would be the following. For simplicity, suppose that all of the groups have equal size. It is conjectured that the way to divide the variables into the groups so that it is most likely to happen that, for each group, the true underlying model is neither the minimum model nor the full model is the following: start from the top of the order, assign the variables $\widetilde{x}_1, \widetilde{x}_2, \ldots$ to different groups, one for each group; once all of the groups have their first assignments, start over again from the top of the order among the remaining variables, and so on.

The idea discussed here is based on a suggestion by Samuel Müller (personal communications).

10.4. Relative IF. The IF (see Section 4) has been shown to have impressive performance in simulation studies (e.g., Jiang et al. [18]). On the other hand, in the analyses of some real datasets, the method appears to be reluctant in picking up a higher dimensional model. In other words, unless there is a strong evidence, IF seems to prefer a lower dimensional model.

To see why this is happening, recall that, at each dimension d, the maximum empirical (i.e., bootstrap) frequency, p_d^*, is computed. The p_d^*'s of different d's are compared and the dimension, d^*, corresponding to the maximum p_d^* is determined. Suppose that there are K candidate variables. It might seem a lot easier for one variable to "stand out" among all of the K variables than for a pair of variables to stand out among all pairs of the K variables (just counting the numbers, there are $K-1$ "competitors to beat" in the former case, and $K(K-1)/2-1$ competitors in the latter).

One idea for a possible improvement is to compare the different d's not according to their p_d^*'s, which are based on the direct counts out of the bootstrap samples, but according to their relative "strangeness." Of course, we have to define what the strangeness means here. Roughly speaking, the strangeness is with respect to what is expected to happen when "nothing is going on," meaning that no candidate variable is relevant (to the response). Thus, in a way, the quantities to compare with are the "P values" with respect to a "null distribution," or reference distribution. One may call such a procedure *relative IF*.

10.5. EMAF versus E-MS. Jiang et al. [48] proposed another method for model selection with incomplete data, known as the E-MS algorithm. The idea is to combine the parameters with the model under which the parameters are defined, so that both the model and the parameters under the model are included in the E-M iteration. The convergence of E-MS as well as consistency has been established (Jiang et al. [48]). Although the E-MS is not restricted to the fence, it would be interesting to compare it with the EMAF method, discussed in Section 6, in terms of performance in model selection with incomplete data. Some initial comparison was made in Jiang et al. [48] via a limited simulation study.

10.6. Shrinkage Mixed Model Selection. There has been some recent work on joint selection of the fixed and random effects in mixed effects models. Bondell et al. [49] considered such a selection problem in a certain type of linear mixed models, which can be expressed as

$$y_i = X_i\beta + Z_i\alpha_i + \epsilon_i, \quad i = 1,\ldots,m, \tag{17}$$

where y_i is an $n_i \times 1$ vector of responses for subject i, X_i is an $n_i \times p$ matrix of explanatory variables, β is a $p \times 1$ vector of regression coefficients (the fixed effects), Z_i is an $n_i \times q$ known design matrix, α_i is a $q \times 1$ vector of subject-specific random effects, ϵ_i is an $n_i \times 1$ vector of errors, and m is the number of subjects. It is assumed that the α_i, ϵ_i, $i = 1,\ldots,m$ are independent with $\alpha_i \sim N(0, \sigma^2\Psi)$ and $\epsilon_i \sim N(0, \sigma^2 I_{n_i})$, where Ψ is an unknown covariance matrix. The problem of interest, using the terms of shrinkage model selection (see Section 7), is to identify the nonzero components of β and α_i, $1 \le i \le m$. For example, the components of α_i may include a random intercept and some random slopes. To take advantage of the idea of shrinkage variable selection, Bondell et al. [49] adopted a modified Cholesky decomposition. Note that the covariance matrix, Ψ, can be expressed as $\Psi = D\Omega\Omega'D$, where $D = \mathrm{diag}(d_1,\ldots,d_q)$ and $\Omega = (\omega_{kj})_{1\le k,j\le q}$ is a lower triangular matrix with 1's on the diagonal. Thus, one can express (17) as

$$y_i = X_i\beta + Z_iD\Omega\xi_i + \epsilon_i, \quad i = 1,\ldots,m, \tag{18}$$

where the ξ_i's are independent $N(0, \mathrm{I})$ random vectors. The idea is to apply shrinkage estimation to both β_j, $1 \le j \le p$ and d_k, $1 \le k \le q$. Note that setting $d_k = 0$ is equivalent to setting all the elements in the kth column and kth row of Ψ to zero, and thus creating a new submatrix by deleting the corresponding row and column, or the exclusion of the kth component of α_i. However, direct implementation of this idea is difficult, because the ξ_i's are unobserved, even though their distribution is known. To overcome this difficulty, Bondell et al. [49] used the E-M algorithm (Dempster et al. [50]). A similar approach was taken by Ibrahim et al. [51] for joint selection of the fixed and random effects in GLMMs.

Suppose that the purpose of the joint selection is for prediction of mixed effects, as discussed in Section 5. We may combine the predictive measure of lack-of-fit developed in Section 5 with the shrinkage idea of Bondell et al. [49] and Ibrahim et al. [51]. Namely, we replace the negative log-likelihood, which is first term in the penalized log-likelihood, by a predictive measure, which can be derived in a similar way as in Section 5 (see the beginning of the last paragraph of the section). The tuning parameter may be chosen using a similar procedure as discussed in Section 7. We refer to this method as *predictive shrinkage selection*, or PSS. The idea has been explored by Hu et al. [52] except that the choice of the tuning parameter was not considered. Nevertheless, the latter authors showed that PSS performs better than the method of Bondell et al. [49] not only in terms of the predictive performance, but also in terms of *parsimony*. Hu et al. [52] also extended the PPS to Poisson mixed models. There is another apparent advantage of the PSS over Bondell et al. [49] and Ibrahim et al. [51]. Unlike the latter approaches, PSS does not need to run the E-M algorithm and thus potentially is computationally more efficient.

10.7. Theoretical Properties. The consistency of AF was established in Jiang et al. [13]. Although the technical conditions have intuitive interpretations, these conditions are not "clean." In other words, some of the so-called "regularity conditions" ought to be proved, rather than assumed. However, it seems difficult to prove these conditions, which likely involves asymptotic analysis of the joint distribution of the observed data and bootstrap samples.

In a way, consistency is a first-order asymptotic property. A second-order asymptotic property would be efficiency in

model selection. Although efficiency is a standard asymptotic property for parameter estimation, it is rarely considered in the context of model selection. On the other hand, in order to compare, from a theoretical standpoint, different model selection procedures that are consistent, some kind of efficiency property needs to be considered. However, although model selection could be viewed as parameter estimation (see Section 10.5), there is usually no asymptotic normality for model selection. This makes straightforward extension of the efficiency in estimation to model selection difficult. As an alternative, one may consider the Kullback-Leibler divergence between the selected model and the optimal model (e.g., Hershey and Olsen [53]).

It is also of interest to establish some nonasymptotic theoretical properties, such as prediction error bounds for comparing different methods, and do so objectively.

Having said these, a potential drawback of focusing on some well-known statistical properties, such as the K-L divergence, or any other measures of the "distance" between models, is that the latter tends to be based purely on statistical/mathematical considerations, while model selection is rarely a (purely) statistical problem. This leads to the next subsection, our concluding remark.

10.8. Concluding Remark. In delivering his 2013 R. A. Fisher Lecture, Peter Bickel offered what he called a "humble view of future" for the role of statistician. "(1) Applied statistics must engage theory and conversely. (2) More significantly it should be done in full ongoing collaboration with scientists and participation in the gathering of data and full knowledge of their nature." As noted, the fence allows flexibility of choosing the optimal model within the fence. Namely, the criterion of optimality for the selection within the fence does not have to be statistical; it can incorporate scientific, economic, political, or even legal concerns. However, so far this potentially important aspect of the fence has yet to be explored (with, perhaps, the exception of Jiang et al. [54]). Of course, there may be a concern that, with a "right" choice of the optimality measure, a practitioner can always "legitimize" the selection in his/her favor, under the fence framework. Nevertheless, it is our strong belief that the practical problem, not the statistical one, is the only gold standard. The practical problem could be scientific, economic, political, legal, or of other natures. As long as the practitioner is faithful to the best interest of his/her problem, the practitioner can surely legitimize his or her choice. Therefore, in view of Professor Bickel's note, we expect this remarkable feature of the fence to be fully explored in the near future.

The future of the fence looks promising.

Conflict of Interests

The author declares that there is no conflict of interests regarding to the publication of this paper.

Acknowledgments

The author wishes to thank Professor Peter Bickel for sharing the slides of his 2013 Fisher Lecture and Professors Ethan Anderes, Iain Johnstone, Samuel Müller, and Alan Welsh for helpful discussion and comments. The author is grateful to the valuable comments from an editor and two referees.

References

[1] H. Akaike, "Information theory as an extension of the maximum likelihood principle," in *Proceedings of the 2nd International Symposium on Information Theory*, B. N. Petrov and F. Csaki, Eds., pp. 267–281, Akademiai Kiado, Budapest, Hungary, 1973.

[2] H. Akaike, "A new look at the statistical model identification," *IEEE Transaction on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[3] H. Bozdogan, "Editor's general preface," in *Proceedings of the 1st US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, H. Bozdogan, Ed., vol. 3 of *Engineering and Scientific Applications*, pp. 9–12, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.

[4] J. de Leeuw, "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., vol. 1, pp. 599–609, Springer, London, UK, 1992.

[5] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[6] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *Journal of the Royal Statistical Society B*, vol. 41, no. 2, pp. 190–195, 1979.

[7] R. Nishii, "Asymptotic properties of criteria for selection of variables in multiple regression," *The Annals of Statistics*, vol. 12, no. 2, pp. 758–765, 1984.

[8] R. Shibata, "Approximate efficiency of a selection procedure for the number of regression variables," *Biometrika*, vol. 71, no. 1, pp. 43–49, 1984.

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

[10] J. Ye, "On measuring and correcting the effects of data mining and model selection," *Journal of the American Statistical Association*, vol. 93, no. 441, pp. 120–131, 1998.

[11] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.

[12] K. W. Broman and T. P. Speed, "A model selection approach for the identification of quantitative trait loci in experimental crosses," *Journal of the Royal Statistical Society B*, vol. 64, no. 4, pp. 641–656, 2002.

[13] J. Jiang, J. S. Rao, Z. Gu, and T. Nguyen, "Fence methods for mixed model selection," *The Annals of Statistics*, vol. 36, no. 4, pp. 1669–1692, 2008.

[14] J. Jiang, T. Nguyen, and J. S. Rao, "A simplified adaptive fence procedure," *Statistics & Probability Letters*, vol. 79, no. 5, pp. 625–629, 2009.

[15] J. Shao, "Linear model selection by cross-validation," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 486–494, 1993.

[16] T. Nguyen and J. Jiang, "Restricted fence method for covariate selection in longitudinal data analysis," *Biostatistics*, vol. 13, no. 2, pp. 303–314, 2012.

[17] T. Nguyen, J. Peng, and J. Jiang, "Fence methods for backcross experiments," *Journal of Statistical Computation and Simulation*, vol. 84, no. 3, pp. 644–662, 2014.

[18] J. Jiang, T. Nguyen, and J. S. Rao, "Invisible fence methods and the identification of differentially expressed gene sets," *Statistics and its Interface*, vol. 4, no. 3, pp. 403–415, 2011.

[19] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[20] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 107–129, 2007.

[21] J. Mou, *Two-stage fence methods in selecting covariates and covariance for longitudinal data [Ph.D. dissertation]*, Department of Statistics, University of California, Davis, Calif, USA, 2012.

[22] G. K. Robinson, "That BLUP is a good thing: the estimation of random effects," *Statistical Science*, vol. 6, no. 1, pp. 15–51, 1991.

[23] J. Jiang and P. Lahiri, "Mixed model prediction and small area estimation," *Test*, vol. 15, no. 1, pp. 1–96, 2006.

[24] R. E. Fay and R. A. Herriot, "Estimates of income for small places: an application of James-Stein procedures to census data," *Journal of the American Statistical Association*, vol. 74, pp. 269–277, 1979.

[25] J. Jiang, T. Nguyen, and J. S. Rao, "Best predictive small area estimation," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 732–745, 2011.

[26] J. N. K. Rao, *Small Area Estimation*, Wiley, New York, NY, USA, 2003.

[27] S. Chen, J. Jiang, and T. Nguyen, "Observed best prediction for small area counts," *Journal of Survey Statistics and Methodology*. Revised.

[28] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2002.

[29] J. Jiang, T. Nguyen, and J. S. Rao, "The E-MS algorithm: model selection with incomplete data," *Journal of the American Statistical Association*. In Press.

[30] E. S. Lander and S. Botstein, "Mapping mendelian factors underlying quantitative traits using RFLP linkage maps," *Genetics*, vol. 121, no. 1, p. 185, 1989.

[31] H. Zhan, X. Chen, and S. Xu, "A stochastic expectation and maximization algorithm for detecting quantitative trait-associated genes," *Bioinformatics*, vol. 27, no. 1, Article ID btq558, pp. 63–69, 2011.

[32] G. Verbeke, G. Molenberghs, and C. Beunckens, "Formal and informal model selection with incomplete data," *Statistical Science*, vol. 23, no. 2, pp. 201–218, 2008.

[33] J. G. Ibrahim, H. Zhu, and N. Tang, "Model selection criteria for missing-data problems using the EM algorithm," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1648–1658, 2008.

[34] J. Copas and S. Eguchi, "Local model uncertainty and incomplete-data bias," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 67, no. 4, pp. 459–513, 2005.

[35] J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, vol. 20, no. 1, pp. 101–148, 2010.

[36] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B. Methodological*, vol. 58, no. 1, pp. 267–288, 1996.

[37] W. J. Fu, "Penalized regressions: the bridge versus the lasso," *Journal of Computational and Graphical Statistics*, vol. 7, no. 3, pp. 397–416, 1998.

[38] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[39] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[40] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[41] H. Wang, R. Li, and C. Tsai, "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, vol. 94, no. 3, pp. 553–568, 2007.

[42] Y. Zhang, R. Li, and C. Tsai, "Regularization parameter selections via generalized information criterion," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 312–323, 2010.

[43] Z. Pang, B. Lin, and J. Jiang, "Regularization parameter selections with divergent and NP-dimensionality via bootstrapping," *Australian & New Zealand Journal of Statistics*. In press.

[44] E. K. Melcon, "On optimized shrinkage variable selection in generalized linear models," *Journal of Statistical Computation and Simulation*. Revised.

[45] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society B*, vol. 70, no. 5, pp. 849–911, 2008.

[46] L. A. Hindorff, P. Sethupathy, H. A. Junkins et al., "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 23, pp. 9362–9367, 2009.

[47] S. Müller, J. L. Scealy, and A. H. Welsh, "Model selection in linear mixed models," *Statistical Science*, vol. 28, no. 2, pp. 135–167, 2013.

[48] J. Jiang, T. Nguyen, and J. S. Rao, "The E-MS algorithm: model selection with incomplete data," *Journal of the American Statistical Association*, 2013.

[49] H. D. Bondell, A. Krishna, and S. K. Ghosh, "Joint variable selection for fixed and random effects in linear mixed-effects models," *Biometrics*, vol. 66, no. 4, pp. 1069–1077, 2010.

[50] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B: Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[51] J. G. Ibrahim, H. Zhu, R. I. Garcia, and R. Guo, "Fixed and random effects selection in mixed effects models," *Biometrics*, vol. 67, no. 2, pp. 495–503, 2011.

[52] K. Hu, J. Choi, J. Jiang, and A. Sim, "Best predictive regularized modelling for mixed effects in high-speed network data," Tech. Rep., Department of Statistics, University of California, Davis, Calif, USA, 2013.

[53] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 4, pp. IV317–IV320, Honolulu, Hawaii, USA, April 2007.

[54] J. Jiang, T. Nguyen, and J. S. Rao, "Fence method for nonparametric small area estimation," *Survey Methodology*, vol. 36, no. 1, pp. 3–11, 2010.