Statistica Sinica **9**(1999), 273-287

BOOTSTRAP CHOICE OF COST COMPLEXITY FOR BETTER SUBSET SELECTION

J. Sunil Rao

The Cleveland Clinc Foundation

Abstract: Subset selection is a long-standing problem. One goal of a selection procedure is consistency. Consistency using Akaike's Final Prediction Error Criterion (FPE) as a selection procedure can be shown to be related to the cost complexity parameter in FPE. However, another goal of a selection procedure is accurate predictions. The consistency property does not necessarily guarantee this second objective. The issue can be thought of as a bias versus variance tradeoff for the procedure. We use the bootstrap to model this tradeoff and provide an objective way of choosing a procedure which attempts to balance the two objectives. This is done in the spirit of the cost complexity pruning algorithm of classification and regression trees. The methodology is described and illustrated on simulated and real data examples.

Key words and phrases: Adaptive estimation, Mallow's C_p, model selection, prediction error, resampling methods.

1. Introduction

One of the goals of subset selection procedures for linear regression models is consistent selection - i.e., picking the true underlying submodel with probability tending to 1 as the sample size gets large. Many procedures minimize estimates of prediction error (PE) for fixed subsets. The Final Prediction Error (FPE) criterion (Akaike (1970)) is one estimate that has been studied extensively. The conditions for consistent selection can be shown to be related to the cost complexity parameter (λ) used in FPE. While consistency seems a reasonable objective, we also want a procedure to produce accurate estimates in terms of quantities like mean squared model error (ME) or mean squared error of prediction (PE). (Note that PE is just a direct function of ME, see Breiman (1992).) The two objectives are not equivalent as overfitting (picking too large a subset) and underfitting (picking too small a subset) can have very different effects on prediction accuracy.

Rather than focus on a fixed subset, it seems more natural to focus the selection procedure on a cost complexity parameter. In particular, we use FPE as a cost complexity criterion. For a fixed cost parameter, we can assess the *prediction accuracy of a procedure* and choose that procedure which minimizes our

estimate of prediction error. To estimate the prediction accuracy of a procedure, we use the bootstrap (Efron (1979)), although other methods like leave-d-out cross-validation have also been used (Rao and Tibshirani (1997)). This cost complexity approach is the basis of the pruning procedure in the CART work of Breiman, Friedman, Olshen and Stone (1984). The article is organized as follows. Section 2 gives some background to the FPE. Section 3 introduces the new bootstrap-based approach termed BCC for *bootstrap choice of cost complexity*. Section 4 provides simulation results and Section 5 gives some concluding remarks and discussion.

2. Subset Selection and Prediction Error

In the linear regression setting, one assumes there is data of the form $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, where the $\mathbf{x}_i$ is the ith value p-variate predictor vector. The predictors are related to the response by $y_i = \mathbf{x}_i'\beta + e_i, \quad i = 1, \ldots, n$, where β is a $p \times 1$ vector of unknown parameters some of whose elements may be 0. The $e_i, i = 1, \ldots, n$, are assumed to be i.i.d. $N(0, \sigma^2)$, and the design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is assumed to be full rank.

If α is any non-empty subset of the p predictors, the subset selection problem is to select that subset $\tilde{\alpha}$ of k ($k \leq p$) predictors corresponding to the non-zero elements of β. For a given p candidate predictors, there are $2^p - 1$ possibilities for α so the search can become a prohibitive task.

Suppose there is a predictor for submodel α say $\hat{y}_\alpha$ for y and new data of the form $(y_i^{new}, \mathbf{x}_i^{new})$, $i = 1, \ldots, n$, exists. Then PE for $\hat{y}_\alpha$ can be defined by

$$PE_\alpha = \frac{1}{n} \sum_{i=1}^{n} E_F (y_i^{new} - \hat{y_i}^{new})^2, \tag{1}$$

where F is the population generating the operating or true model. Submodels can be classified into two categories - 1) A_c - those that contain at least all of the non-zero elements of β and 2) A_w - those that are incorrect, i.e., missing at least one non-zero element of β.

If the predictors are considered fixed, $\mathbf{x}_i^{new} = \mathbf{x}_i$ with $\mathbf{x}_i^{new}$ independent of y_i, then the above expectation is taken with respect to the new and original responses. Thus if a submodel $\alpha \in A_c$ the unconditional prediction error is

$$\begin{aligned} PE_\alpha &= \sigma^2 + \frac{1}{n} \sum_{i=1}^{n} E_F (\mathbf{x}_i^{'new} \beta - \mathbf{x}_{i\alpha}^{'new} \hat{\beta}_\alpha)^2, \\ &= \sigma^2 + \frac{1}{n} d(\alpha) \sigma^2, \end{aligned} \tag{2}$$

where $\hat{\beta}_\alpha$ is the OLS estimator for submodel α based on the original data, and $d(\alpha)$ is the number of predictors in α. If $\alpha \in A_w$ the unconditional PE_α is given by

$$PE_\alpha = \sigma^2 + \Delta_\alpha + \frac{1}{n} d(\alpha)\sigma^2, \tag{3}$$

where $\Delta_\alpha = n^{-1}\beta' \mathbf{X}'(I - H_\alpha)\mathbf{X}\beta$ with $H_\alpha = \mathbf{X}_\alpha(\mathbf{X}'_\alpha \mathbf{X}_\alpha)^{-1}\mathbf{X}'_\alpha$. Further, I is an $n \times n$ identity matrix and $\mathbf{X}$, $\mathbf{X}_\alpha$ are $n \times p$ and $n \times d(\alpha)$ design matrices for the full model and submodel α respectively.

If the predictors are random, then the new observations $(y_i^{new}, \mathbf{x}_i^{new})$, $i = 1, \ldots, n$, are assumed to come from the joint distribution of $(y, \mathbf{x})$, but are assumed independent from the original data set. The expectation in (1) is done with respect to the distribution over the new observations and the original data set. The form of the resulting PE definitions is similar but an extra remainder $R = o(n^{-1})$ is added on to the above two equations. The reader is referred to Shao (1993) for a more complete discussion.

A common assumption made is to assume that for $\alpha \in A_w$

$$\liminf_{n\to\infty} \Delta_\alpha > 0. \tag{4}$$

This amounts to saying that a wrong model cannot have asymptotically minimal PE. Equations (2), (3) and the above assumption (4) imply that the search for the correct model with minimum size then amounts to finding the subset with minimum PE_α. One goal of a subset selection procedure is to provide consistent selection defined as

$$\lim_{n\to\infty} P(\hat{\alpha} = \tilde{\alpha}) = 1, \tag{5}$$

where $\hat{\alpha}$ is a submodel chosen by some procedure. Another is to have accurate predictions given by $ME = E_F(\mu - \hat{\mu})'(\mu - \hat{\mu})$ where $\mu = \mathbf{X}\beta$ and $\hat{\mu} = \mathbf{X}_{\hat{\alpha}}\hat{\beta}_{\hat{\alpha}}$.

2.1. Subset selection using FPE

Since PE_α is typically unknown, it must be estimated. The Final Prediction Error (FPE) criterion estimates PE_α as

$$FPE_\alpha(\lambda) = RSS_\alpha + \lambda d(\alpha)\sigma^2 \tag{6}$$

for a fixed positive value of a cost complexity parameter λ. The term σ^2 is estimated by s^2, the unbiased estimate of σ^2 under the full model of degree p. The term RSS_α is the residual sum of squares for submodel α given by $RSS_\alpha = \sum_{i=1}^n (y_i - \mathbf{x}'_i\hat{\beta}_\alpha)^2$. The submodel picked by minimizing (6) for fixed λ is then $\alpha_\lambda = \text{argmin}_\alpha\, FPE_\alpha(\lambda)$. If $\lambda = 2$, (6) is the traditional Mallow's C_p

statistic; if $\lambda = \log n$, (6) is the Bayesian Information Criterion (BIC) (Schwarz (1978)); if $\lambda = \log\log n$, (6) it is the criterion proposed by Hannan and Quinn (1979) in the context of autoregressive model order determination.

The conditions for satisfying (5), using $FPE_\alpha(\lambda)$, have been previously studied by Zhang (1993) for fixed $\mathbf{X}$. Under $E_F e = 0$, $E_F e^2 = \sigma^2$, $E_F e^r < \infty$ for $r > 2$ and condition (4), we simply require that $\lambda/n \to 0$, and $\lambda \to \infty$ as $n \to \infty$. The random $\mathbf{X}$ case was studied by Pötscher (1991).

3. Bootstrap Choice of λ

The trouble with the asymptotic results is that they give very little guidance on the behaviour of a procedure in finite samples (Bickel and Zhang (1992)). Consistency of a selection procedure can be thought of as elimination of underfitting and overfitting probabilities. These probabilities - expressed as $P(\alpha_\lambda \in A_w)$ and $P(\alpha_\lambda \in A_c)$ with $d(\alpha_\lambda) > d(\tilde{\alpha})$ respectively, both tend to 0 as $n \to \infty$ but at very different rates. The underfitting probability vanishes much more quickly than the overfitting probability (Zhang (1993)). Thus in finite samples, one has a bias (caused by underfitting) versus variance (caused by overfitting) tradeoff captured in the prediction accuracy of the procedure. The goal of this paper is to introduce a method that minimizes this tradeoff over a set of candidate procedures by estimating prediction accuracy using the bootstrap (Efron (1979)). By restricting candidate procedures to be those satisfying Zhang's (1993) conditions, this provides a compromise method of choosing a procedure with good selection and prediction accuracy. First however we provide some background to PE estimation using the bootstrap.

3.1. Bootstrapping regression models and subsequent PE estimation

The material in this subsection can also be found in other references, including Efron (1982) and Shao (1996), but is included here for completeness of discussion. Under the setup of Section 2, in the fixed predictor case, the bootstrap resamples the normalized residuals under the full model, with replacement. If $r_i = y_i - \mathbf{x}_i'\hat{\beta}, i = 1, \ldots, n$, are residuals from the estimated full model of dimension p, then a bootstrap sample consists of $e_i^*,\ i = 1, \ldots, n$, sampled with replacement from $\hat{F}_n$, the distribution putting mass $1/n$ on each of the normalized residuals $(1 - p/n)^{-1/2}(r_i - \bar{r}), i = 1, \ldots, n$, where $\bar{r} = \sum_{i=1}^n r_i/n$. Then let $y_{i\alpha}^* = \mathbf{x}_{i\alpha}'\hat{\beta}_\alpha + e_i^*, i = 1, \ldots, n$, and define the bootstrap estimator $\hat{\beta}_\alpha^* = (\mathbf{X}_\alpha'\mathbf{X}_\alpha)^{-1} \sum_{i=1}^n \mathbf{x}_{i\alpha} y_{i\alpha}^*$. The bootstrap estimate of PE_α is then

$$\widehat{PE}_\alpha^{fB} = \frac{1}{n}\sum_{i=1}^n E_*(y_i - \mathbf{x}_{i\alpha}'\hat{\beta}_\alpha^*)^2, \tag{7}$$

where the expectation E_* is with respect to $\hat{F}_n$.

In the random predictor case, the bootstrap resamples the pairs $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$ with replacement from the empirical distribution $\hat{F}_n^r$, putting mass $1/n$ on each data pair. Let $(\mathbf{x}_i^*, y_i^*), i = 1, \ldots, n$ be the bootstrap sample of size n resampled from $\hat{F}_n^r$. Then define the bootstrap estimated coefficient vector to be

$$\hat{\beta}_\alpha^* = \begin{cases} (\mathbf{X}_\alpha^{*'}\mathbf{X}_\alpha^*)^{-1} \sum_{i=1}^n \mathbf{x}_{i\alpha}^* y_i^* & \text{if } \gamma_{(n)}^* \geq \gamma_{(n)}/2 \\ \hat{\beta}_\alpha & \text{otherwise,} \end{cases}$$

where $\mathbf{X}_\alpha^*$ is the bootstrap analog of $\mathbf{X}_\alpha$ obtained by bootstrapping pairs, and $\gamma_{(n)}^*$ and $\gamma_{(n)}$ are the smallest eigenvalues of $(\mathbf{X}_\alpha^{*'}\mathbf{X}_\alpha^*)/n$ and $(\mathbf{X}_\alpha'\mathbf{X}_\alpha)/n$ respectively (Shao (1996)). The bootstrap estimate of PE_α in the random predictor case is then

$$\widehat{PE}_\alpha^{rB} = \frac{1}{n}\sum_{i=1}^n E_*(y_i - \mathbf{x}_{i\alpha}^{*'}\hat{\beta}_\alpha^*)^2, \tag{8}$$

where the expectation is taken with respect to $\hat{F}_n^r$.

Efron (1982) showed that in fact (8) leads to an underestimation of PE_α of the order n^{-1}. He gave a less biased form making use of the bootstrap estimate of optimism. This is an attempt at correcting the underestimation of PE_α by the apparent error rate given by $n^{-1}RSS_\alpha = n^{-1}\sum_{i=1}^n (y_i - \mathbf{x}'_{i\alpha}\hat{\beta}_\alpha)^2$. The correction known as the optimism in RSS_α is given by $\omega_\alpha = E_F(PE_\alpha - n^{-1}RSS_\alpha)$. The natural bootstrap estimate of this is $\hat{\omega}_\alpha^B = E_*(n^{-1}\Sigma_{i=1}^n (y_i - \mathbf{x}_{i\alpha}^{'}\hat{\beta}_\alpha^*)^2 - n^{-1}RSS_\alpha^*)$, with $n^{-1}RSS_\alpha^* = \frac{1}{n}\Sigma_{i=1}^n (y_i^* - \mathbf{x}_{i\alpha}^{*'}\hat{\beta}_\alpha^*)^2$. This estimator of PE_α can then be written as

$$\widehat{PE}_\alpha^{r2B} = n^{-1}RSS_\alpha + \hat{\omega}_\alpha^B. \tag{9}$$

This estimator is almost unbiased (Shao (1996)). Some similar estimates were provided by Bunke and Droge (1984). Bootstrap estimates of Kullback-Leibler information are presented in Shibata (1995). Note that either bootstrapping recipe can be used whether $\mathbf{X}$ is fixed or random (Shao (1996)).

3.2. Choosing λ

We generalize the above use of the bootstrap to estimate PE associated with a cost parameter λ in FPE. For a fixed $\lambda > 0$ we can find the subset minimizing $FPE_\alpha(\lambda)$. In practice, we use a procedure like the bootstrap to find the value $\hat{\lambda}$ producing the smallest estimated prediction error, and then for our final model we choose the α minimizing $FPE_\alpha(\hat{\lambda})$. The parameter λ roughly indexes model size. This cost-complexity approach is the basis of the pruning procedure in the CART work (Breiman, Friedman, Olshen and Stone (1984)).

Define the true prediction error for λ as

$$PE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E_F(y_i^{new} - \hat{y}_{i\lambda}^{new})^2, \tag{10}$$

where $\hat{y}_{i\lambda}^{new} = \mathbf{x}_{i\alpha_\lambda}^{'new}\hat{\beta}_{\alpha_\lambda}$, $i = 1, \ldots, n$. The subtle difference in (10) versus (1) is that the emphasis is now being placed on the cost parameter λ and not a subset. Our estimate, $\widehat{PE}(\lambda)^B$ applies the Efron-type modification in a fashion similar to (9). Define the submodel minimizing (6) for the bootstrap data and the same λ as $\alpha_\lambda^* = \text{argmin}_\alpha\, FPE_\alpha(\lambda)^*$. where $FPE_\alpha(\lambda)^*$ is (6) evaluated on the bootstrap dataset. The analog of (9) is given by

$$\widehat{PE}(\lambda)^B = n^{-1}RSS(\lambda) + \hat{\omega}_B(\lambda), \tag{11}$$

where $RSS(\lambda)$ is the residual sum of squares in the original data for the submodel α_λ, and $\hat{\omega}_B(\lambda)$ is the bootstrap estimate of optimism in $n^{-1}RSS(\lambda)$ given by $\hat{\omega}_B(\lambda) = E_*(n^{-1}\sum_{i=1}^n (y_i - \hat{y}_{i\lambda}^*)^2 - n^{-1}RSS(\lambda)^*)$, with $\hat{y}_{i\lambda}^* = \mathbf{x}_{i\alpha_\lambda^*}^{'}\hat{\beta}_{\alpha_\lambda^*}^*$, $i = 1, \ldots, n$ and $n^{-1}RSS(\lambda)^* = \frac{1}{n}\sum_{i=1}^n (y_{i\alpha_\lambda}^* - \mathbf{x}_{i\alpha_\lambda^*}^{*'}\hat{\beta}_{\alpha_\lambda^*}^*)^2$. If $\mathbf{X}$ is random, $y_{i\alpha_\lambda}^* = y_i^*, i = 1, \ldots, n$. If $\mathbf{X}$ is fixed, then $\mathbf{x}_{i\alpha_\lambda^*}^{*'} = \mathbf{x}_{i\alpha_\lambda^*}^{'}$ and $y_{i\alpha_\lambda}^* = \mathbf{x}_{i\alpha_\lambda}^{'}\hat{\beta}_{\alpha_\lambda} + e_i^*$, $i = 1, \ldots, n$ where the e_i^*, $i = 1, \ldots, n$, are the normalized residuals from the full model of degree p. The steps of the algorithm are summarized as follows.

BCC algorithm

1. For a fixed value of λ find $\alpha_\lambda = \text{argmin}_\alpha\, FPE_\alpha(\lambda)$ in the orginal data and generate B bootstrap samples. If resampling the residuals, each sample is of the form $(x_{i\alpha_\lambda}, y_{i\alpha_\lambda}^*)$, $i = 1, \ldots, n$, where $y_{i\alpha_\lambda}^* = \mathbf{x}_{i\alpha_\lambda}^{'}\hat{\beta}_{\alpha_\lambda} + e_i^*$, $i = 1, \ldots, n$, where the e_i^*, $i = 1, \ldots, n$, are normalized residuals from the full model of degree p sampled with replacement from $\hat{F}_n$. If resampling pairs, the bootstrap sample is (x_i^*, y_i^*), $i = 1, \ldots, n$, sampled with replacement from $\hat{F}_n^r$. The rest of the algorithm will be provided for bootstrapping pairs.
2. For each bootstrap sample, find the submodel minimizing $FPE_\alpha(\lambda)^*$, say $\alpha_\lambda^*(b)$.
3. Evaluate the optimism in $\alpha_\lambda^*(b)$, say $\hat{\omega}_b(\lambda)$. Specifically, $\hat{\omega}_b(\lambda) = (n^{-1}\Sigma_{i=1}^n (y_i - \hat{y}_{i\lambda}^*(b))^2 - n^{-1}RSS_b(\lambda)^*)$, where $\hat{y}_{i\lambda}^*(b) = \mathbf{x}_{i\alpha_\lambda^*(b)}^{'}\hat{\beta}_{\alpha_\lambda^*(b)}^*$, $i = 1, \ldots, n$, and $n^{-1}RSS_b(\lambda)^* = n^{-1}\sum_{i=1}^n (y_i^* - \mathbf{x}_{i\alpha_\lambda^*(b)}^{*'}\hat{\beta}_{\alpha_\lambda^*(b)}^*)^2$.
4. Average the $\hat{\omega}_b(\lambda)$, $b = 1, \ldots, B$, over the B bootstrap samples to derive $\hat{\omega}_B(\lambda)$. Add $\hat{\omega}_B(\lambda)$ to $n^{-1}RSS(\lambda)$ to obtain $\widehat{PE}^B(\lambda)$.
5. Repeat steps 1-4 for a grid of λ values.
6. Minimize $\widehat{PE}(\lambda)^B$ over λ giving $\widehat{\lambda}_B$.
7. The selected model is then $\alpha_{\widehat{\lambda}_B}$.

The BCC algorithm requires a grid of λ values. We would like the BCC to be consistent in the sense of (5). This provides some minimal asymptotic optimality in terms of selection accuracy. Note that $\widehat{\lambda}_B$ is a random variable and hence the conditions of Zhang (1993) do not guarantee consistency. Stronger conditions need to be imposed as in Rao and Wu (1989). If $\mathbf{X}$ is fixed, then under $E_F e = 0$, $E_F e^2 = \sigma^2$, $E_F e^r < \infty$ for $r > 2$ and condition (4), we need $\widehat{\lambda}_B/n \to 0$ *a.s.* and $(\log\log n)^{-1}\widehat{\lambda}_B \to \infty$ *a.s.* when $n \to \infty$. For $\mathbf{X}$ random, we can use Pötscher (1991) and impose the additional mild conditions that $x_{n\alpha}^2 = o(n^\gamma)$ for some $0 < \gamma < 1$ and all α, where $x_{n\alpha}$ correspond to the elements of the nth row of $\mathbf{X}_\alpha$. In practice, this is simply accomplished by searching over a range like $[\log n, n/\log n]$. A potential drawback of this is to have poorer small sample accuracy. A modified version of the BCC was studied in Rao and Tibshirani (1993) where an unrestricted range was used.

3.3. $\widehat{PE}(\lambda)^B$ as an estimator of $PE(\lambda)$

We need to show that our bootstrap estimate of $PE(\lambda)$ is indeed tracking the true value. This is a very difficult question to answer in general and, as a result, we choose to look at a special case as done in Breiman (1992). Consider the special case where $X_{ii} = 1, i = 1, \ldots, n$, and $X_{ij} = 0, i \neq j$, and all $\beta_j = 0, j = 1 \ldots p$. Then the ordinary least square estimates can be written as $\hat{\beta}_j = \beta_j + Z_j, j = 1, \ldots, p$, where $\{Z_j\}$ are i.i.d. $N(0, \sigma^2)$. In fact, $Z_j = e_j, j = 1, \ldots, p$. The estimated coefficients in the bootstrap samples can then be written as $\hat{\beta}_j^* = \hat{\beta}_j + Z_j^*, j = 1, \ldots, p$, where $\{Z_j^*\}$ are i.i.d. $N(0, \sigma^2)$ and the $\{Z_j\}$ are independent of the $\{Z_j^*\}$.

It is well known in the orthogonal design case that minimizing $FPE_\alpha(\lambda)$ is equivalent to choosing a submodel with $\hat{\beta}_j^2 \geq \lambda\sigma^2$, assuming σ^2 known. Under the same assumption, minimizing the corresponding criterion in the bootstrap data is equivalent to picking those $\hat{\beta}_j^*$ such that $\hat{\beta}_j^{*2} \geq \lambda\sigma^2$. Take β_λ and $\mathbf{x}_\lambda$ and their bootstrap versions to indicate corresponding quantities from α_λ and α_λ^* respectively. Then $PE(\lambda)$ becomes

$$\begin{aligned} PE(\lambda) &= \frac{1}{n}\sum_{i=1}^{n} E_F(y_i^{new} - \mathbf{x}'_{i\lambda}\hat{\beta}_\lambda)^2, \\ &= \frac{1}{n}\sum_{i=1}^{n} E_F\Big(e_i^{new} - \sum_{j=1}^{d(\alpha_\lambda)} Z_j + \sum_{j=d(\alpha_\lambda)+1}^{p} \beta_j\Big)^2. \end{aligned}$$

If we take all $\beta_j = 0$, then

$$PE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E_F\Big[e_i^{new} - \sum_{j=1}^{p} Z_j I(Z_j^2 \geq \lambda\sigma^2)\Big]^2.$$

The bootstrap estimate of this, following (10), is

$$\begin{aligned}\widehat{PE}^B(\lambda) &= \frac{1}{n}\sum_{i=1}^{n} E_*(y_i - \mathbf{x}'_{i\lambda}\hat{\beta}^*_\lambda)^2,\\ &= \frac{1}{n}\sum_{i=1}^{p} E_*\left[\beta_i+e_i-(\beta_i+e_i+Z^*_i)I[(\beta_i+e_i+Z^*_i)^2\geq\lambda\sigma^2]\right]^2+\frac{1}{n}\sum_{i=p+1}^{n} e_i^2\\ &= \frac{1}{n}\sum_{i=1}^{p} E_*\left[(\beta_i+e_i)I[(\beta_i+e_i+Z^*_i)^2<\lambda\sigma^2]-Z^*_i I[(\beta_i+e_i+Z^*_i)^2\geq\lambda\sigma^2]\right]^2\\ &\quad+\frac{1}{n}\sum_{i=p+1}^{n} e_i^2.\end{aligned}$$

If we again take all $\beta_j = 0$ then this becomes

$$\widehat{PE}^B(\lambda)=\frac{1}{n}\sum_{i=1}^{p} E_*\left[e_iI[(e_i+Z^*_i)^2<\lambda\sigma^2]-Z^*_i I[(e_i+Z^*_i)^2\geq\lambda\sigma^2]\right]^2+\frac{1}{n}\sum_{i=p+1}^{n} e_i^2.$$

Since the e_i's and the Z^*_i's are i.i.d. $N(0,\sigma^2)$, the $e_i + Z^*_i$'s are i.i.d $N(0, 2\sigma^2)$. So $PE(\lambda)$ decreases as $\lambda \to \infty$ and identifies the best submodel as the empty one. The (unconditional) expected value of $\widehat{PE}^B(\lambda)$ is such that the values for $\lambda = 0$ and $\lambda \to \infty$ are equivalent but identify full and empty models as best ones respectively. But by definition (6), $\lambda = 0$ is inadmissible and hence the best subset is also identified as the empty one. In practice we use the refined estimate (11) instead of the above bootstrap estimate. Admittedly this illustration is a special situation, but it does provide some insight into the behaviour of the BCC procedure.

3.4. An all-subsets bootstrap approach

As a point of comparison in the simulations, we make reference to a result found by Shao (1996). He found that the submodel selected by minimizing the bootstrap estimate of PE_α given by (7), (8) or (9) does not provide consistent selection in the sense of (5). For the random predictor case, we must take a bootstrap sample of size m, $(\mathbf{x}^*_i, y^*_i), i = 1, \ldots, m$ where m is chosen to satisfy $m/n \to 0$, as m and $n \to \infty$. In the fixed predictor case, Shao proposes that instead of resampling the $e^*_i, i = 1, \ldots, n$, from $\hat{F}_n$, we should resample the $\tilde{e^*_i}, i = 1, \ldots, n$, from the distribution $\tilde{F}_{n,m}$ putting mass n^{-1} on each $(n/m)^{1/2}(1 - p/n)^{-1/2}(r_i - \bar{r}), i = 1, \ldots, n$, where m is chosen in such a way so as $m \to \infty$, $m/n \to 0$ as $n \to \infty$. The issue is one of determining an appropriate value of m which may depend on unknown model parameters. Not only is this an issue with regard to subset selection accuracy, but also with prediction accuracy.

An insufficient correction can lead to excess overfitting and too drastic a correction can cause undue underfitting. Shao (1996) shows that this is equivalent to determining an appropriate λ in $FPE_\alpha(\lambda)$.

4. Simulation Examples

4.1. Example 1

The first example is taken from Shao (1993). Note that for all simulations we used $B = 20$ for the Monte Carlo approximations to keep computations to a minimum. The subset selection algorithms were from the S-plus statistical software package and, in particular, the leaps and bounds procedure was used. All computations were done on a Sparc workstation. We will assume a true model of the form $y = \mathbf{x}'\beta + e$ where $e \in N(0,1)$. The predictors $\mathbf{x}$ are assumed to be a 5-variate vector and the total sample size is 40. The 40×5 design matrix from Shao (1993) will be considered fixed with $x_{i1} = 1,\ i = 1, \ldots, n$. The true values of β are given in Table 1 with varying degrees of model complexity. The results of the simulation are described in Table 1 based on 100 realizations. The following selection methods were examined: full model, Mallow's C_p, the ordinary bootstrap over all subsets (Btm ($m = n$)), Shao's corrected bootstrap (Btm) for $m = 20, 15, 10$, the BIC, FPE with $\lambda = n/\log n$, and the BCC with a grid search done over λ values in $[\log n, n/\log n]$. We have included in our table the number of underfit and overfit models and the true model error $ME = E_F(\mu - \hat{\mu})'(\mu - \hat{\mu})$ where $\mu = \mathbf{X}\beta$ and $\hat{\mu} = \mathbf{X}_{\hat{\alpha}}\hat{\beta}_{\hat{\alpha}}$. This is estimated by the empirical average over the 100 realizations. Note that the same $e_i, i = 1, \ldots, 40$, were used for all methods in a given realization. Thus the ME's are directly comparable. We opt for presenting ME instead of PE since ME contains the most important information about the performance of the selection procedures. Since we know the true model in the simulations, we can directly estimate this quantity. This is a slight variation from the earlier presentation, but does not alter the conclusions.

The following results are apparent.

1. Methods with high correct selection probabilities (consistent) do not always give accurate predictions. In particular, many of the analytically corrected bootstrap methods and $FPE_\alpha(\lambda = n/\log n)$ sometimes underfit badly, resulting in high model error. Note that BCC and BIC perform almost the same in this example (with BCC being slightly better in terms of selection accuracy and model error), but poor BIC performance has been previously noted (see Bickel and Zhang (1992) for example).
2. In the last case the full model beats all model selectors. We feel that the full model should be included in any comparison of such procedures.
3. The BCC method performs well in terms of both correct model selection and model error.

4. Figure 1 shows the range of cost parameters chosen by the BCC for each of the models. It is clear that no one value is preferred over the 100 realizations. The range of cost parameters chosen narrows as the number of predictors in the true model increases.

Table 1. Shao's Example 1 ($n = 40$, 100 realizations) indicating overfitting, underfitting counts, number of times true model selected, and ME for the selection procedure. Signal to noise ratios (snr) are indicated below each model.

True Model	Procedure	number overfit	number underfit	number correct	ME
$\beta = (2, 0, 0, 4, 0)'$	full model	100	0	0	5.231
(snr = 4.24)	C_p	41	0	59	3.943
	BIC	17	0	83	3.031
	$FPE_\alpha(\lambda = n/\log n)$	1	0	99	2.174
	Btm($m = n$)	70	0	30	4.279
	Btm($m = 20$)	53	0	47	3.349
	Btm($m = 15$)	43	0	57	2.964
	Btm($m = 10$)	42	0	58	2.740
	BCC	12	0	88	2.815
$\beta = (2, 0, 0, 4, 8)'$	full model	100	0	0	5.231
(snr = 11.62)	C_p	28	0	72	4.373
	BIC	10	0	90	3.792
	$FPE_\alpha(\lambda = n/\log n)$	0	0	100	3.309
	Btm($m = n$)	58	0	42	4.419
	Btm($m = 20$)	47	0	53	4.099
	Btm($m = 15$)	31	0	69	3.825
	Btm($m = 10$)	37	3	60	5.629
	BCC	7	0	93	3.693
$\beta = (2, 9, 0, 4, 8)'$	full model	100	0	0	5.231
(snr = 14.10)	C_p	18	0	82	4.734
	BIC	5	0	95	4.389
	$FPE_\alpha(\lambda = n/\log n)$	0	1	99	4.412
	Btm($m = n$)	48	0	52	4.813
	Btm($m = 20$)	31	0	69	4.642
	Btm($m = 15$)	19	5	76	5.681
	Btm($m = 10$)	12	34	54	17.637
	BCC	5	0	95	4.389
$\beta = (2, 9, 6, 4, 8)'$	full model	0	0	100	5.231
(snr = 16.69)	C_p	0	0	100	5.231
	BIC	0	0	100	5.231
	$FPE_\alpha(\lambda = n/\log n)$	0	6	94	6.608
	Btm($m = n$)	0	2	98	5.653
	Btm($m = 20$)	0	8	92	7.659
	Btm($m = 15$)	0	18	82	11.083
	Btm($m = 10$)	0	68	32	34.089
	BCC	0	0	100	5.231
$\beta = (1, 2, 3, 2, 3)'$	full model	0	0	100	5.231
(snr = 6.72)	C_p	0	68	32	5.717
	BIC	0	85	15	6.238
	$FPE_\alpha(\lambda = n/\log n)$	0	100	0	13.689
	Btm($m = n$)	0	56	44	6.392
	Btm($m = 20$)	0	84	16	10.232
	Btm($m = 15$)	0	91	9	13.516
	Btm($m = 10$)	0	100	0	22.600
	BCC	0	85	15	6.748

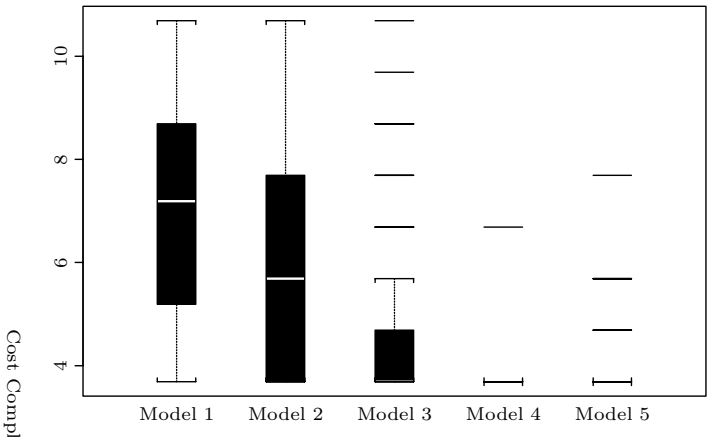

Figure 1. Boxplot of cost parameters selected by the BCC over the 100 realizations of Example 1.

4.2. Example 2 - BCC versus fixed λ methods

An example will now be examined to compare performance to fixed λ methods - i.e., those using fixed λ values that satisfy Zhang's (1993) conditions for consistency. The example is from Hurvich and Tsai (1990) with sample size $n = 50$ and the true value of $\beta' = (1, 2, 3, 0.6, 0, 0, 0)$. The $x_j, j = 1, \ldots, 7$ are generated once from a $N(0, 1)$ distribution and fixed for all subsequent realizations across all methods. The $e_i, 1 = 1, \ldots, 50$, were generated from a $N(0, 1)$ and fixed across all methods as in Example 1. A total of 200 realizations were run.

Table 2. Example 2 ($n = 50$, 200 realizations) for fixed values of λ and the BCC applied over a range $[\log n, (n/\log n)]$. The true model is $\beta = (1, 2, 3, 0.6, 0, 0, 0)'$ and the signal to noise ratio is 3.48.

Method	Number overfit	Number underfit	Number correct	ME
Full Model	0	200	0	7.183
C_p	85	8	107	6.100
BIC	24	20	156	5.679
$FPE_\alpha(4)$	24	20	156	5.679
$FPE_\alpha(5)$	20	24	156	5.734
$FPE_\alpha(6)$	10	38	152	6.284
$FPE_\alpha(7)$	7	47	146	6.626
$FPE_\alpha(8)$	2	56	142	6.936
$FPE_\alpha(9)$	1	73	126	7.826
$FPE_\alpha(10)$	0	82	118	8.297
$FPE_\alpha(11)$	0	93	107	8.926
$FPE_\alpha(12)$	0	106	94	9.599
BCC	22	24	154	5.821
Random λ	4	60	136	7.306

The results are given in Table 2. Once again number of overfit, underfit and correct models are included as well as the ME. Many of the fixed λ values produce serious underfitting which inflates the ME. The BCC does as well as "the best" fixed λ procedure (BIC in this case) in terms of selection accuracy and ME. Note that randomly choosing among the λ values clearly is not optimal in terms of both accuracy and ME. Figure 2 gives the distribution of cost parameters chosen by BCC.

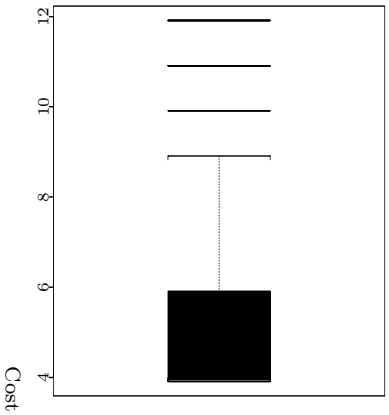

Figure 2. Boxplot of cost parameters selected by the BCC over the 200 realizations of Example 2.

4.3. Real data example

A dataset from Weisberg (1985) relates automobile accident rates in accidents per million vehicle miles (Y) to 13 potential predictor variables. It is an example of Normal regression. The dataset includes 39 sections of highways in Minnesota in 1973. The legend for the variables is as follows.

1. Y = RATE = 1973 accident rate per million vehicle miles.
2. LEN = length of the segment in miles.
3. ADT = average daily traffic count in thousands (estimated).
4. TRKS = truck volume as a percent of total volume.
5. SLIM = speed limit (in 1973).
6. LWID = lane width in feet.
7. SHLD = width in feet of outer shoulder on the roadway.
8. ITG = number of highway-type interchanges per mile in the segment.
9. SIGS = number of signalized interchanges per mile in the segment.
10. ACPT = number of access points per mile in the segment.
11. LANE = total number of lanes of traffic in both directions.
12. FAI = 1 if federal aid highway, 0 otherwise.
13. PA = 1 if principal arterial highway, 0 otherwise.
14. MA = 1 if major arterial highway, 0 otherwise.

Two of the segments, numbers 38 and 39 were neither federal, principal arterial, nor major arterial, but were classified as major collectors (MC) coded here by FAI = PA = MA = 0. Using a separate variable would have resulted in exact collinearity.

Weisberg (1985) notes that when fitting the full model, none of the covariates have t-values in absolute magnitude greater than 2. Thus there does not seem to be any obviously strong predictors in the full model. For subject matter reasons the variable LEN must be included in all candidate models, and that the three dummy variables coding it must be treated as a group. Table 3 shows the results of using various subset selection procedures. Included are a naive all-subsets bootstrap (Btm($m = n$)), Shao's corrections with $m = n^{3/4}, n^{1/2}, n^{1/3}, n^{1/4}$, C_p, BIC and BCC. Results for any of the procedures involving bootstrapping have been summarized over 25 replications of the bootstrap process (each bootstrapping process involves using $B = 20$ bootstrap samples). The BIC gives slightly different conclusions than the other consistent methods (Shao's corrections and the BCC). BIC picks the same model as C_p (predictors LEN,SLIM,ACPT) since the difference in cost parameters is quite small (log(39) = 3.66 for this dataset). The BCC and the analytically corrected bootstrap tend to pick models with predictors LEN,ACPT and LEN,LWID,ACPT respectively. Note that with $m = n^{1/2}$ we pick the model LEN,ACPT slightly more often, and this can be attributed to some Monte Carlo variation in the bootstrapping process (different sets of bootstrap samples used for different values of m and then averaged to reduce variation).

Table 3. Results comparing models selected by various methods for the highway accident rate example. Any method superscripted with a † indicates that the bootstrap process was repeated 25 times, and the number in parentheses is the frequency of times this model was selected.

Method	Model Selected
Bt($m = n$) †	LEN,SLIM,SIGS,ACPT,FAI,PA,MA (7/25)
Btm($m = n^{3/4}$)†	LEN,SLIM,ACPT (9/25)
Btm($m = n^{1/2}$)†	LEN,ACPT (10/25), LEN,LWID,ACPT (7/25)
Btm($m = n^{1/3}$)†	LEN,LWID,ACPT (7/25)
Btm($m = n^{1/4}$)†	LEN,ACPT (6/25), LEN,LWID,ACPT (7/25)
C_p	LEN,SLIM,ACPT
BIC	LEN,SLIM,ACPT
BCC †	LEN,ACPT (8/25), LEN,LWID,ACPT (7/25)

5. Summary

This paper has presented a bootstrap-based method to balance two objectives of a subset selection procedure - accurate selection and accurate predictions.

This is done by using a cost complexity approach similar to that of the CART work from Breiman, Friedman, Olshen and Stone (1984). The bootstrap was used to implement the algorithm, although as previously stated, other methods like leave-d-out cross-validation can also be used (Rao and Tibshirani (1997)). However the bootstrap can easily be extended to more complicated problems such as nonlinear regression models and generalized linear models (Shao (1996)). By applying candidate selection procedures to bootstrap datasets, we can assess the inherent instability of a procedure to slight perturbations of the data. Then averaging predictions gives an indication of the average prediction accuracy of the procedure. This has a flavor somewhat similar to the bagging procedure (for bootstrap aggregation) of Breiman (1994) but with bagging, submodels themselves are averaged over the bootstrap datasets to produce hybrid models with improved predictions. Extensions of the methodology to generalized linear models have also be developed. Similar extensions can also be made to generalized estimating equation models for longitudinal data (Liang and Zeger (1986)). These results will be reported in a separate communication.

References

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203-217.

Bickel, P. and Zhang, P. (1992). Variable selection in c regression with categorical covariates. *J. Amer. Statist. Assoc.* **87**, 90-97.

Breiman, L. (1992). Little bootstrap and other methods for dimensionality selection in regression - X fixed prediction error. *J. Amer. Statist. Assoc.* **87**, 738-754.

Breiman, L. (1996). Bagging predictors. *Machine Learning.* To appear.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth Publishers, Belmont, California.

Bunke, O. and Droge, B. (1984). Bootstrap and cross-validation estimates of prediction error for linear regression models. *Ann. Statist.* **12**, 1400-1424.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1-26.

Efron, B. (1982). *The Jackknife, Bootstrap and Other Resampling Plans.* SIAM, Philadelphia.

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

Hurvich, C. M. and Tsai, C. L. (1990). The impact of model selection on inference in linear regression. *Amer. Statist.* **44**, 214-217.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Pötscher, B. M. (1991). Model selection under nonstationarity autoregressive models and stochastic linear regression models. *Ann. Statist.* **17**, 1257-1274.

Rao, J. S. and Tibshirani, R. (1993). Bootstrap model selection via the cost complexity parameter in regression. Technical Report, Department of Statistics, University of Toronto.

Rao, J. S. and Tibshirani, R. (1997). Discussion to "An asymptotic theory for model selection" by Jun Shao. *Statist. Sinica* **7**, 249-252.

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

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

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

Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.* **91**, 655-665.

Shibata, R. (1995). Bootstrap estimate of Kullback-Leibler information for model selection. Technical report, Department of Statistics, University of California, Berkeley.

Weisberg, S. (1985). *Applied Linear Regression*. John Wiley, New York.

Zhang, P. (1993). On the convergence rate of model selection criteria. *Comm. Statist., Part A - Theory Methods* **22**, 2765-2775.

Department of Biostatistics, The Cleveland Clinc Foundation, Cleveland, Ohio, 44195, USA.
E-mail: srao@bio.ri.ccf.org

(Received July 1996; accepted March 1998)