The Annals of Statistics
1996, Vol. 24, No. 6, 2350–2383

HEURISTICS OF INSTABILITY AND STABILIZATION IN MODEL SELECTION[1]

By Leo Breiman

University of California

In model selection, usually a "best" predictor is chosen from a collection $\{\hat{\mu}(\cdot, s)\}$ of predictors where $\hat{\mu}(\cdot, s)$ is the minimum least-squares predictor in a collection $\mathscr{U}_s$ of predictors. Here s is a complexity parameter; that is, the smaller s, the lower dimensional/smoother the models in $\mathscr{U}_s$.

If $\mathscr{L}$ is the data used to derive the sequence $\{\hat{\mu}(\cdot, s)\}$, the procedure is called unstable if a small change in $\mathscr{L}$ can cause large changes in $\{\hat{\mu}(\cdot, s)\}$. With a crystal ball, one could pick the predictor in $\{\hat{\mu}(\cdot, s)\}$ having minimum prediction error. Without prescience, one uses test sets, cross-validation and so forth. The difference in prediction error between the crystal ball selection and the statistician's choice we call predictive loss. For an unstable procedure the predictive loss is large.

This is shown by some analytics in a simple case and by simulation results in a more complex comparison of four different linear regression methods. Unstable procedures can be stabilized by perturbing the data, getting a new predictor sequence $\{\hat{\mu}'(\cdot, s)\}$ and then averaging over many such predictor sequences.

1. Introduction.

1.1. *The prediction problem.* In the "supervised" prediction problem, one has a "learning" set of data consisting of measurements on N cases, where each case consists of a response variable y_n and a prediction vector $\mathbf{x}_n = (x_{1n}, \dots, x_{Mn})$ taking its value in χ. That is, this learning set $\mathscr{L}$ consists of data $\{(y_n, \mathbf{x}_n),\ n = 1, \dots, N\}$. We assume here that the response variable is numerical. For notation, we use lower case letters for vectors with components indexed by case number $n = 1, \dots, N$, i.e., $y = (y_1, y_2, \dots, y_N)$. Lower case bold letters denote vectors indexed by dimension, i.e., $\mathbf{x}_n = (\mathrm{x}_{1_n}, \dots, \mathrm{x}_{M_n})$, and the inner product of two such vectors $\mathbf{u}, \mathbf{v}$ is written as $\mathbf{uv}$.

The goal of prediction is to use $\mathscr{L}$ to construct a function $\mu(\mathbf{x}, \mathscr{L})$, defined for $\mathbf{x}$ in χ such that $\mu(\cdot, \mathscr{L})$ gives accurate predictions of future responses. More specifically, suppose that we have a large test set $\mathscr{T} = \{(y'_n, \mathbf{x}'_n),\ n = 1, \dots, N'\}$ sampled (in some sense which we will make more specific later) from the same distribution as $\mathscr{L}$. Then we want $\mu(\mathbf{x}'_n, \mathscr{L})$ to be an accurate estimate of y'_n for all $(y'_n, \mathbf{x}'_n) \in \mathscr{T}$. Assuming squared error loss, we want the

Received January 1995; revised November 1995.

[1] Partially supported by NSF Grant DMS-92-12419.

AMS 1991 *subject classification.* 62H99.

Key words and phrases. Regression, subset selection, predictive loss, cross-validation, prediction error.

prediction error

$$\mathrm{PE}(\mu) \simeq \frac{1}{N'}\sum_n (y'_n - \mu(\mathbf{x}'_n, \mathscr{L}))^2$$

to be small.

The standard approach to the construction of a predictor $\mu(\mathbf{x}, \mathscr{L})$ goes like this: a large class of functions $\mathscr{U} = \{\mu(\mathbf{x})\}$ is defined. For instance, if all coordinates of $\mathbf{x}$ are numerical, $\mathscr{U}$ could be the set of all linear functions of $\mathbf{x}$. Or $\mathscr{U}$ could be some specified set of nonlinear functions. The usual next step is to select as the prediction function $\hat{\mu}(\mathbf{x})$ that $\mu \in \mathscr{U}$ which minimizes the prediction error on the *learning set*. That is, take $\hat{\mu}$ to be the minimizer in $\mathscr{U}$ of

$$\mathrm{RSS}(\mu) = \|y - \mu\|^2 = \sum_n (y_n - \mu(\mathbf{x}_n))^2.$$

The difficulties in this approach are well known. If $\mathscr{U}$ is high dimensional, then $\hat{\mu}$ "overfits" the data. It will have low mean-squared prediction error on $\mathscr{L}$ (low RSS) but higher prediction error on test sets. On the other hand, if $\mathscr{U}$ is too low dimensional, it may not contain a good fit to the data. That is, it "underfits" the data.

1.2. *An example.* As a simple example, take $\mathscr{L}$ to consist of 20 pairs (y_n, x_n), $n = 1, \dots, 20$, where the x_n are iid uniform on $[-1, +1]$ and

$$y_n = 2x_n^2 + \varepsilon_n,$$

where the $\{\varepsilon_n\}$ are $N(0,1)$. Suppose $\mathscr{U}$ is the class of all tenth-degree polynomials on $[-1, 1]$. The lowest RSS polynomial will give a good fit (small RSS) to the points in $\mathscr{L}$. But it is wriggly between the points in $\mathscr{L}$ and at the ends of the interval. It will have high prediction error on future data drawn from the same distribution as $\mathscr{L}$.

The tenth-degree polynomial predictor "overfits" the data. In statistical terms it has too large a variance—too many parameters are being estimated. Put another way, the space $\mathscr{U}$ is too large.

Now take $\mathscr{U}$ to be the class of all first-degree polynomials. Then the least-squares minimizer in $\mathscr{U}$ is linear and very smooth. But it gives a poor approximation to the underlying parabolic relation between y and x. In this case, $\hat{\mu}$ "underfits" the data. Variance is low but bias is high. The space $\mathscr{U}$ is too small.

Here is a way to get a compromise between overfitting and underfitting. Take $\mathscr{U}$ to be the space of all tenth-degree polynomials. Let $\mathscr{U}_k \subset \mathscr{U}$ be the space of all polynomials of degree less than or equal to k, where $k = 0, 1, 2, \dots, 10$. Note that $\mathscr{U}_k \subset \mathscr{U}_{k+1}$. For k large, $\mathscr{U}_k$ contains many wriggly functions—as k becomes smaller, the functions in $\mathscr{U}_k$ become smoother.

Let $\hat{\mu}(\cdot, k)$ be the polynomial in $\mathscr{U}_k$ that minimizes $\|y - \mu\|^2$. Thus, we now have a sequence of predictors $\hat{\mu}(\cdot, 0), \hat{\mu}(\cdot, 1), \dots, \hat{\mu}(\cdot, 10)$ which are polynomials of degree $0, 1, \dots, 10$. The problem now is to pick the best of these. Here are three ways to make this selection:

1. Knowing how the data was generated, pick $\hat{\mu}(\cdot, 2)$.
2. Conjure up a test set containing one million pairs (y'_n, x'_n), each drawn from the same distribution as $\mathscr{L}$. Use the test set to compute $\mathrm{PE}(\hat{\mu}(\cdot, k))$ and select
$$k^* = \arg\min_k \mathrm{PE}(\hat{\mu}(\cdot, k)).$$
Call k^* the *crystal ball* best value of k.
3. Estimate $\mathrm{PE}(\hat{\mu}(\cdot, k))$ by $\widehat{\mathrm{PE}}(\hat{\mu}(\cdot, k))$. For instance, one can use bootstrap or cross-validation to get estimates of $\mathrm{PE}(\hat{\mu}(\cdot, k))$. Then estimate the best k as
$$\hat{k} = \arg\min_k \widehat{\mathrm{PE}}(\hat{\mu}(\cdot, k)).$$
Call $\hat{k}$ the *fallible* estimate.

The question that we explore in this paper is how much accuracy do we lose by not having a crystal ball, that is, an infinite test set. With a crystal ball, we select the predictor whose true prediction error is

$$\mathrm{PE}(\hat{\mu}(\cdot, k^*)) = \min_k \mathrm{PE}(\hat{\mu}(\cdot, k)).$$

Without a crystal ball, the selected predictor has prediction error

$$\mathrm{PE}\big(\hat{\mu}(\cdot, \hat{k})\big).$$

Define the predictive loss PL as

$$\mathrm{PL} = \mathrm{PE}\big(\hat{\mu}(\cdot, \hat{k})\big) - \mathrm{PE}(\hat{\mu}(\cdot, k^*)).$$

This predictive loss is what we study in this article.

1.3. *Regularization procedures.* The context for our study of predictive loss is the sequence of predictors constructed from a regularization procedure. If we attempt to construct a predictor by defining a large class of functions $\mathscr{U}$ and defining $\hat{\mu}(\cdot)$ to be the minimizer in $\mathscr{U}$ of $\|y - \mu\|^2$, then overfitting will usually result. The currently used methods for compromising between overfitting and underfitting are similar to the strategy used in the simple example of subsection 1.2, and are referred to as *regularization procedures*.

DEFINITION 1.1. A regularization procedure consists of defining a sequence of subspaces $\mathscr{U}_s \subset \mathscr{U}$ indexed by a real parameter $s \ge 0$ such that

$$s \le s' \Rightarrow \mathscr{U}_s \subset \mathscr{U}_{s'}.$$

Let $\hat{\mu}(\cdot, s)$ be the function in $\mathscr{U}_s$ minimizing $\|y - \mu\|^2$. Then $\{\hat{\mu}(\cdot, s)\}$ is called the sequence of regularized predictors.

Regularization procedures are ubiquitous in prediction methods. Here are some examples:

Linear regression. Regularization by subset selection:

$$\mathscr{U}_k = \text{all linear functions with at most } k \text{ nonzero coefficients.}$$

Regularization by ridge:

$$\mathscr{U}_s = \text{all linear functions } \boldsymbol{\beta}\mathbf{x} \text{ such that } \|\boldsymbol{\beta}\| \le s.$$

Regression trees (*CART*).

$$\mathscr{U}_k = \text{all binary trees formed by univariate splits with at most } k \text{ terminal nodes.}$$

Multivariate splines (*MARS*).

$$\mathscr{U}_k = \text{all functions that are sums of at most } k \text{ products of linear splines.}$$

Neural nets. Regularization by limiting the hidden layer:

$$\mathscr{U}_k = \text{all functions expressible by } k \text{ or fewer units in the hidden layer.}$$

Regularization by weight decay:

$$\mathscr{U}_s = \text{all functions such that the norm of the weights is bounded by } s.$$

Given a regularized sequence of predictors $\{\hat{\mu}(\cdot, s)\}$, the standard procedure is to try to choose the most accurate in the sequence by forming estimates $\widehat{\text{PE}}$ of the "true" prediction error $\text{PE}(\hat{\mu}(\cdot, s))$. Define

$$\text{PE}(s) = \text{PE}(\hat{\mu}(\cdot, s)),$$
$$\widehat{\text{PE}}(s) = \widehat{\text{PE}}(\hat{\mu}(\cdot, s))$$

and

$$s^* = \arg\min_s \text{PE}(s),$$
$$\hat{s} = \arg\min_s \widehat{\text{PE}}(s).$$

The primary question we study is

How much do we lose by not having a crystal ball? That is, how big is the predictive loss

$$\text{PL} = \text{PE}(\hat{s}) - \text{PE}(s^*)?$$

1.4. *Predictive loss, instability and stabilization.* Obviously, the size of the predictive loss is related to how the prediction error is estimated. Poorer estimates will give larger PL. But what we are interested in is: given the best possible current methods of PE estimation, how is the predictive loss connected to the structure of the regularization procedure?

For instance, in linear regression there are two archetypical regularization procedures—subset selection and ridge. Both generate sequences of regularized predictors and we can try to select the best one in each sequence using cross-validation estimates of PE. It turns out that the PL for subset regression is considerably larger than the PL for ridge regression.

The difference between the two regularization procedures that is reflected in the PL is their relative *instability*.

Heuristic definition. A regularization procedure is unstable if a small change in the data $\mathscr{L}$ can make large changes in the regularized sequence $\{\hat{\mu}(\cdot, s, \mathscr{L})\}$.

In general (see Section 7):

Subset selection is unstable:

Changing just one data case in $\mathscr{L}$ can cause a large change in the minimizer of RSS(μ) over $\mathscr{U}_k$.

Ridge is stable:

Change $\mathscr{L}$ slightly and the minimizer of RSS(μ) over $\mathscr{U}_s$ is close to the original minimizer.

Many current regularizations are unstable. The list includes CART, MARS and neural nets. Besides ridge, the only other well-known stable method is k-nearest-neighbor regression.

The more unstable the procedure, the noisier PE(s) is, and the larger the predictive loss whatever method of PE estimation is used. With unstable procedures, we are less able to locate the best model, and the size of the predictive loss may be a substantial fraction of the prediction error.

Figures 1 and 2 give illustrations of this. Figure 1 consists of PE(k) plots for three runs of subset selection on 30-dimensional simulated data where the subsets are selected by forward stepwise addition. Figure 2 gives the plots of

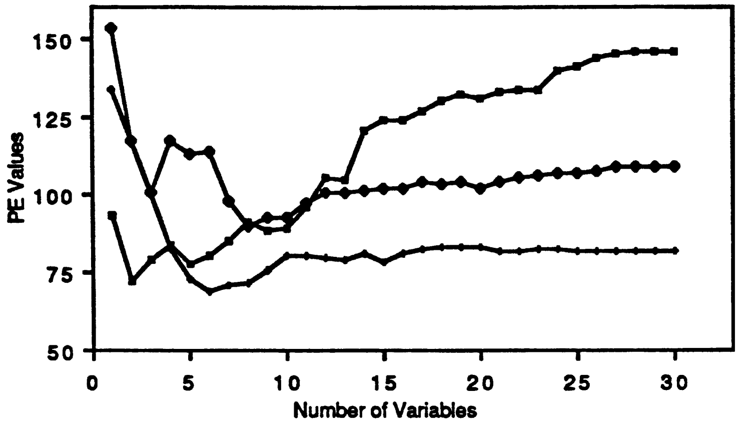

FIG. 1. *PE vs. number of variables for subset selection.*

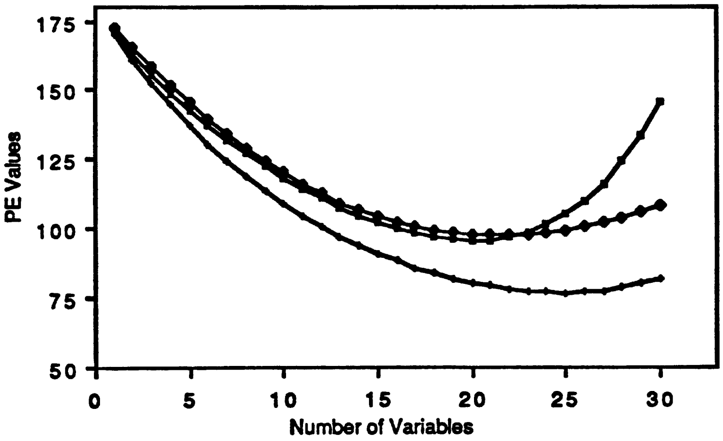

FIG. 2. *PE vs. number of variables for ridge.*

PE(k) for ridge regression on the same data where k is the equivalent dimensionality. (See Section 5 for details on how the data was generated.) The noisy behavior of the PE(k) values for subset selection makes estimating the minimum point more difficult than for the smooth ridge values.

There are other consequences of instability. One is that the estimates of the prediction error for the selected predictor $\hat{\mu}(\cdot, \hat{s})$ have large negative bias. Another is that "infinitesimal" methods for estimating PE do not work very well. An example of the latter is the discovery in Breiman and Spector (1992) that leave-one-out cross-validation is less accurate than leave-many-out in selecting the best subset dimension.

1.5. *Stabilization.* Given that instability has undesirable consequences, what can be done? *Unstable procedures can be stabilized*! Consider all data sets $\mathscr{L}'$ such that $d(\mathscr{L}, \mathscr{L}') \leq \delta$ in some (unspecified) metric d. Define

$$\hat{\mu}_{\mathrm{ST}}(\cdot, s) = \mathrm{Av}_{d(\mathscr{L}', \mathscr{L}) \leq \delta}\, \hat{\mu}(\cdot, s, \mathscr{L}').$$

Then the averaged predictors $\{\hat{\mu}_{\mathrm{ST}}(\cdot, s)\}$ are a more stable sequence with lower predictive loss and less biased PE estimates.

The implementation of this idea that worked the best among several alternatives (see subsection 6.1) is this: generate iid $N(0, \alpha^2)$ noise $\{\varepsilon_n\}$ and let

$$\mathscr{L}' = \{(y_n + \varepsilon_n, \mathbf{x}_n), n = 1, \dots, N\}.$$

That is, $\mathscr{L}$ is altered into $\mathscr{L}'$ by simply adding noise to each response value y_n. Using $\mathscr{L}'$, construct the regularized sequence $\{\mu(\cdot, s, \mathscr{L}')\}$. Now repeat many times, generating new noise $\{\varepsilon_n\}$ each time, and define

$$\hat{\mu}_{\mathrm{ST}}(\cdot, s) = \mathrm{Av}_{\mathscr{L}'}\, \hat{\mu}(\cdot, s, \mathscr{L}').$$

2. Organization of the article. Instead of trying to give rigorous definitions of instability, we proceed by example. Linear regression is used as a paradigm. Four different regularization methods with various degrees of instability are studied. The most unstable is subset selection; the most stable is ridge. The predictive losses for these four methods can be compared analytically in the $X^tX = I$ situation and through simulations in more realistic settings. Since PE definitions and estimates differ depending on whether X is considered random or controlled, both situations are studied. The article is organized as follows:

Section 3 gives definitions of prediction error for data with random X and with controlled X. The test set, cross-validation and little bootstrap methods for estimating PE are detailed. The four regression methods are defined.

Section 4. Analytic results are gotten in the $X^tX = I$ case for PE estimated either by test set or by little bootstrap. These illustrate the effects of instability as the number of variable increases. The results of stabilization are made clear.

Section 5. Simulation results are given for the case of controlled X using more complex $X'X$ designs with PE estimated either by test set or by little bootstrap. These again illustrate instability effects and the results of stabilization.

Section 6 gives results of a simulation study in the case of random X where PE is estimated either by test set or by cross-validation. Some perplexing aspects of stabilization are described.

Section 7. We look more closely at cross-validation estimates to see why leave-one-out cross-validation behaves poorly in selection from an unstable sequence.

Section 8 compares the performance of the stabilized subset predictors to the other prediction methods on a spectrum of simulated data.

Section 9 contains concluding remarks. We summarize the various threads in the preceding sections and give some future research directions.

Appendix gives details of the $X^tX = I$ computations, little bootstrap proof and the tiny bootstrap formula.

Although our main emphasis is on the effects of instability on predictive loss, some other new ground is covered. The two garotte regression methods and stabilization promise greater predictive accuracy than either subset selection on one extreme or ridge on the other. The limitations of ridge regression are seen. The little bootstrap [Breiman (1992)] and its infinitesimal version, the tiny bootstrap [Breiman (1995)], are extended and strengthened as PE estimation methods.

The effects of instability first came up in connection with the study of one of the garotte methods compared to subset selection and ridge [Breiman (1995)]. The simulations showed that, although subset selection often had a crystal ball model with lower PE than the best ridge model, it lost out because of higher predictive loss. The effort to understand this phenomenon better resulted in the present work.

3. Definitions.

3.1. *PE definitions.* Two definitions of prediction error are common and useful. Sometimes, the values of $\{\mathbf{x}_n\}$ are fixed in a controlled experiment. If the responses y_n are assumed iid selected from a distribution $Y(x_n)$,

$$\mathrm{PE}(\mu) = E\sum_n (Y(\mathbf{x}_n) - \mu(\mathbf{x}_n))^2.$$

If $Y(\mathbf{x}_n) = \mu^*(\mathbf{x}_n) + \varepsilon_n^*$, with $E\varepsilon_n^* = 0$, then $\mathrm{PE}(\mu) = N\sigma^2 + \|\mu - \mu^*\|^2$. We refer to μ^* as the "true" model and to $\|\mu - \mu^*\|^2$ as the model error.

In the situation of random X, the data is assumed iid from $Y, \mathbf{X}$. If the sample size is N, then the prediction error is

$$\mathrm{PE}(\mu) = N \cdot E(Y - \mu(\mathbf{X}))^2.$$

The N multiplier is used to get the PE measure for random X on the same scale as for controlled X. Defining $\mu^*(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$, then

$$\mathrm{PE}(\mu) = N\sigma^2 + N \cdot E(\mu^*(\mathbf{X}) - \mu(\mathbf{X}))^2,$$

where $\sigma^2 = E(Y - E(Y|\mathbf{X}))^2$. The model error is defined as the second term.

3.2. *PE estimates.*

3.2.1. *Test sets.* The simplest way to estimate PE is use of a test set. In the random X case this is data $\{(y_n', \mathbf{x}_n'),\ n = 1, \dots, N'\}$ iid from the same distribution as $\mathscr{L}$ and independent of $\mathscr{L}$. Then

$$\widehat{\mathrm{PE}}(\mu) = \frac{N}{N'} \sum_n (y_n' - \mu(\mathbf{x}_n'))^2.$$

In the controlled X situation, test sets are generated by replicating the experiment K times using the same set $\{\mathbf{x}_n\}$ of $\mathbf{x}$-values. Let the replicated outcomes at $\mathbf{x}_n$ be $y_{1,n}', \dots, y_{K,n}'$. Then

$$\widehat{\mathrm{PE}}(\mu) = \frac{1}{K} \sum_{k,n} (y_{k,n}' - \mu(\mathbf{x}_n))^2.$$

In practice, large test sets are usually not available and other PE estimation methods are used.

3.2.2. *Cross-validation.* In the random X situation, cross-validation reuses the data to get a PE estimate. Let $\mathscr{T} \subset \mathscr{L}$ contain N_{CV} cases and $\mathscr{L}_{\mathrm{CV}} = \mathscr{L} - \mathscr{T}$. Suppose that $\hat{\mu}$ is the minimizer of $\|y - \mu\|^2$ under the constraint $\mu \in \mathscr{U}_s$. Construct $\hat{\mu}_s^{(-\mathscr{T})}$ to be the minimizer of $\|y - \mu\|^2$ over the cases in $\mathscr{L}_{\mathrm{CV}}$ under the constraint $\mu \in \mathscr{U}_s$. Put $\widehat{\mathrm{PE}}(\mathscr{T}) = \sum_{(y_n, \mathbf{x}_n) \in \mathscr{T}} (y_n - \hat{\mu}_s^{(-\mathscr{T})}(\mathbf{x}_n))^2$. Do this for sets $\mathscr{T}_1, \mathscr{T}_2, \dots, \mathscr{T}_K$ and define

$$\widehat{\mathrm{PE}}(s) = \frac{N}{K \cdot N_{\mathrm{CV}}} \sum_k \widehat{\mathrm{PE}}(\mathscr{T}_k).$$

The selection of the $\{\mathcal{T}_k\}$ is usually structured so that they cover $\mathcal{L}$ more or less evenly. In leave-one-out cross-validation, there are $K = N$ left out sets $\mathcal{T}_k$, each one consisting of the single case $(y_k, \mathbf{x}_k)$. Another selection is leave-many-out. Here the sizes of the $\mathcal{T}_k$ are fixed—usually some fraction of N—and the $\mathcal{T}_k$ selected at random. Another version of leave-many-out structures the $\{\mathcal{T}_k\}$ selection so that each $(y_n, \mathbf{x}_n)$ appears in exactly L of the $\{\mathcal{T}_k\}$. This is an extension of the V-fold cross-validation used in CART.

3.2.3. *Little bootstrap.* In the controlled X context, cross-validation is not appropriate. Write

$$\begin{aligned}\mathrm{RSS}(\mu) &= \|\mu^* + \varepsilon^* - \mu\|^2\\ &= \|\varepsilon^*\|^2 + 2(\varepsilon^*, \mu^* - \mu) + \|\mu^* - \mu\|^2.\end{aligned}$$

Then

$$\mathrm{PE}(\mu) = \mathrm{RSS}(\mu) + N\sigma^2 - \|\varepsilon^*\|^2 - 2(\varepsilon^*, \mu^* - \mu).$$

The term $N\sigma^2 - \|\varepsilon^*\|^2 - 2(\varepsilon^*, \mu^*)$ has mean 0. If $\mu = \hat{\mu}(\cdot, s)$ is the minimizer of $\|y - \mu\|^2$, $\mu \in \mathcal{U}_s$, with RSS(s) denoting RSS$(\hat{\mu})$, then $\hat{\mu}$ is dependent on the $\{\varepsilon_n^*\}$ and $(\varepsilon^*, \hat{\mu})$ does not usually have mean 0.

What we would like to do is to find an estimate $B(s)$ of $E(\varepsilon^*, \hat{\mu})$ and put

$$\widehat{\mathrm{PE}}(s) = \mathrm{RSS}(s) + 2B(s).$$

Write $\hat{\mu}(\cdot, s) = \hat{\mu}(\cdot, \mu^* + \varepsilon^*, s)$. Suppose the $\{\varepsilon_n^*\}$ are iid $N(0, \sigma^2)$. Take $t > 0$ and generate $\{\varepsilon_n\}$ as iid $N(0, t^2\sigma^2)$. Define new $\{y_n'\}$ as $\{y_n + \varepsilon_n\}$, recalculate $\hat{\mu}'(\cdot, s)$ using the data $\{y_n', \mathbf{x}_n\}$ and consider the expression

$$B_t(s) = \frac{1}{t^2} E_\varepsilon(\varepsilon, \hat{\mu}'), \tag{3.1}$$

where E_ε denotes expectation over the $\{\varepsilon_n\}$ only. Then we have the following result.

THEOREM 3.1. *Suppose there is a* $\mathcal{U}_{s_t}$, *such that* $\mu \in \mathcal{U}_s \Leftrightarrow \mu/\sqrt{1 + t^2} \in \mathcal{U}_{s_t}$. *Then*

$$EB_t(s) = e\left(\frac{\mu^*}{\sqrt{1 + t^2}}, s_t\right),$$

where

$$e(\mu^*, s) = E(\varepsilon^*, \hat{\mu}).$$

The proof is given in the Appendix, Section A2. One version was proved in Breiman (1992).

For subset selection, $s_t = s$. For ridge, $s_t = s/\sqrt{1 + t^2}$. In practice, σ^2 is estimated from the full-variable OLS as $\hat{\sigma}^2 = \mathrm{RSS}/(N - M)$. The $\{\varepsilon_n\}$ are generated from $N(0, \hat{\sigma}^2 t^2)$ using a random number generator, s_0 taken such that the corresponding $s_t = s$, and $\tilde{\mu}$ is the minimizer in $\mathcal{U}_{s_0}$ of $\|y + \varepsilon - \mu\|^2$. Then $(\varepsilon, \tilde{\mu})/t^2$ is computed. This is repeated a number of times (usually 25 is enough) and the results averaged to give $B_t(s_0)$.

For unstable sequences, values of t in the range $[0.6, 1.0]$ seem to work best. The theorem states that for small t, B_t is an almost unbiased estimate of $E(\varepsilon^*, \hat{\mu})$. But we will see that for unstable sequences, as $t \to 0$, the variance of $B_t \to \infty$. If the limit B_t as $t \to 0$ exists in some nice way, this limit is called the tiny bootstrap, denoted by $\mathrm{TB}(s)$, and is an unbiased estimate of $E(\varepsilon^*, \hat{\mu})$. For moderately unstable sequences, $\mathrm{TB}(s)$ may exist, but be so noisy that more accurate estimates of $\mathrm{PE}(s)$ are gotten by using B_t with $t > 0$. For nicely stable procedures, using $\widehat{\mathrm{PE}}(s) = \mathrm{RSS}(s) + 2\mathrm{TB}(s)$ gives accurate estimates.

3.2.4. *Others.* The literature and popular usage contain other (and simpler) methods for PE estimation. For instance, in subset selection, if $\mathscr{U}_k$ consists of all regressions with k or fewer nonzero coefficients, then the C_P estimate

$$\widehat{\mathrm{PE}}(k) = \mathrm{RSS}(k) + 2\hat{\sigma}^2 k$$

is often used in the controlled X case. For random X, the corresponding estimate is

$$\widehat{\mathrm{PE}}(k) = \mathrm{RSS}(k) \Big/ \left(1 - \frac{k}{N}\right)^2.$$

None of these Akaike-type PE estimates work very well in model selection where the sample size is moderate compared to the number of variables. See Breiman (1992) and Breiman and Spector (1992).

3.3. *Four linear regression methods.*

3.3.1. *Best subsets or stepwise.* Here $\mathscr{U}_k$ is the set of all linear $\mu = \boldsymbol{\beta}\mathbf{x}$, where $\boldsymbol{\beta}$ has at most k nonvanishing coordinates. Minimizing $\|y - \mu\|^2$ over $\mathscr{U}_k$ is called the best subsets method, and may be computationally expensive. In our simulations, the suboptimal stepwise forward addition of variables is used.

3.3.2. *Ridge.* Ridge regression minimizes $\|y - \boldsymbol{\beta}\mathbf{x}\|^2$ under the constraint $\|\boldsymbol{\beta}\| \le s$. Usually, the $\mathbf{x}$-coordinates are prenormalized, since ridge is not scale invariant.

3.3.3. *Nonnegative garotte.* Let $\{\hat{\beta}_m\}$ be the full-model OLS coefficients. Take the $(c_1, \dots, c_M)$ to be the nonnegative minimizers of

$$\sum \Big(y_n - \sum_m c_m \hat{\beta}_m x_{mn} \Big)^2 \tag{3.2}$$

under the constraint $\Sigma c_m \le s$. Then let $\hat{\mu}(\mathbf{x}, s) = \Sigma_m c_m \hat{\beta}_m x_m$ [Breiman (1995)].

3.3.4. *Garotte*. Let $\{\hat{\beta}_m\}$ be the full-model OLS coefficients and take $(c_1, \dots, c_M)$ to minimize (3.1) under the constraint $\|\mathbf{c}\| \le s$.

These methods cover an instability range, with subset selection the most unstable to the very stable ridge procedure.

4. The $X^tX = I$ case.

The case $X^tX = I$ is simple enough to provide some analytic insights into the instability problem. Assume that

$$y = \boldsymbol{\beta}^* \mathbf{x} + \varepsilon^*,$$

where the $\{\varepsilon_n^*\}$ are iid $N(0,1)$. The OLS coefficients are $\hat{\beta}_m = (x_m, y) = \beta_m^* + Z_m$, where the $\{Z_m\}$ are iid $N(0,1)$.

The best subset of k variables consists of those variables $\{x_m\}$ corresponding to the k largest values of $|\hat{\beta}_m|$. Thus, the family of best subset regressions is given by coefficients of the form

$$\hat{\hat{\beta}}_m(\lambda) = I\big(|\hat{\beta}_m| \ge \lambda\big)\hat{\beta}_m. \tag{4.1}$$

The ridge coefficients are of the form

$$\hat{\hat{\beta}}_m(\lambda) = \frac{\hat{\beta}_m}{1+\lambda}. \tag{4.2}$$

The *nn*-garotte coefficients are

$$\hat{\hat{\beta}}_m(\lambda) = \left(1 - \frac{\lambda^2}{\hat{\beta}_m^2}\right)^+ \hat{\beta}_m \tag{4.3}$$

and the garotte $\hat{\hat{\beta}}$ are

$$\hat{\hat{\beta}}_m(\lambda) = \frac{\hat{\beta}_m^2}{\hat{\beta}_m^2 + \lambda^2}\hat{\beta}_m. \tag{4.4}$$

Thus, all methods perform a shrinkage on the OLS $\hat{\boldsymbol{\beta}}$ and are of the form $\hat{\hat{\beta}}_m(\lambda) = \theta(\hat{\beta}_m, \lambda)$. The best subset θ is discontinuous. The *nn*-garotte θ is continuous but has discontinuous first partial derivatives. The θ's for garotte and ridge are $C^{(\infty)}$ in λ, $\hat{\beta}$.

The PE(μ) for $\mu = \hat{\hat{\boldsymbol{\beta}}}(\lambda)\mathbf{x}$ is

$$\text{PE}(\lambda) = N + \left\|\boldsymbol{\beta}^* - \hat{\hat{\boldsymbol{\beta}}}(\lambda)\right\|^2$$

and we put

$$\text{M}E(\lambda) = \left\|\boldsymbol{\beta}^* - \hat{\hat{\boldsymbol{\beta}}}(\lambda)\right\|^2.$$

To simplify further, take M large and the $\{\beta_m^*\}$ iid from a distribution $P(d\beta^*)$. Then the $\{\hat{\beta}_m\}$ are also iid and

$$\text{M}E(\lambda) = \sum_m \left(\beta_m^* - \theta\left(\hat{\beta}_m, \lambda\right)\right)^2$$

is a sum of iid terms. The best crystal ball model in the family $\mu(\cdot, \lambda) = \hat{\hat{\boldsymbol{\beta}}}(\lambda)\mathbf{x}$ corresponds to the λ that minimizes M$E(\lambda)$.

Set $A(\lambda) = E(\beta^* - \theta(\hat{\beta}, \lambda))^2$ so that

$$\mathrm{M}E(\lambda) = M \cdot A(\lambda) + \sqrt{M}\, W(\lambda). \tag{4.5}$$

The $MA(\lambda)$ term is the dominant deterministic part of $\mathrm{M}E(\lambda)$. The $\{W(\lambda)\}$ is a zero-mean, approximately Gaussian stochastic process with $O(1)$ variance. Efforts to locate the minimum of $\mathrm{M}E(\lambda)$ will depend on the smoothness of $W(\lambda)$. [Note: $A(\lambda)$ is smooth and $C^{(\infty)}$ in λ for all $\theta(\beta, \lambda)$ used.]

In the following subsections we get rates of growth for $E(\mathrm{PL})$ as a function of M for the four regression types under study. These rates depend on the PE estimate used.

Estimating PE using a test set of the same size as $\mathscr{L}$ gives

$$\begin{aligned} E(\mathrm{PL}) &= O(1), \quad \text{ridge} \\ &= O(1) \quad \text{garotte} \\ &= O(1) \quad nn\text{-garotte} \\ &= O(M^{1/3}) \quad \text{subset selection.} \end{aligned}$$

Using little or tiny bootstrap estimates, we get

$$\begin{aligned} E(\mathrm{PL}) &= O(1) \quad \text{ridge} \\ &= O(1) \quad \text{garotte} \\ &= O(M^{1/5}) \quad nn\text{-garotte} \\ &= O(M^{3/7}) \quad \text{subset selection.} \end{aligned}$$

The rate computations show that the results depend on the smoothness of $\theta(\beta, \lambda)$. It is illuminating that in this simple case the causes of predictive loss show up so clearly. Along the way, we also derive rates of growth for the bias and variance of the prediction error estimates for the fallible selections. These rates also increase as $\theta(\beta, \lambda)$ becomes less smooth.

4.1. *Using a test set.* The test set consists of $\{(y'_n, \mathbf{x}_n),\ n = 1, \dots, N\}$ with the same values of the $\{\mathbf{x}_n\}$ as in the original data. Then $y'_n = \sum_m \beta^*_m x_{mn} + \varepsilon'_n$, $\{\varepsilon'_n\}$ iid $N(0,1)$ and the $\{\varepsilon'_n\}$ are independent of the $\{\varepsilon^*_n\}$. For estimates $\hat{\hat{\boldsymbol{\beta}}}(\lambda)$ of the $\boldsymbol{\beta}^*$, the test set PE estimate is

$$\begin{aligned} \widehat{\mathrm{PE}}(\lambda) &= \sum_n \Big(y'_n - \hat{\hat{\boldsymbol{\beta}}}(\lambda)\mathbf{x}_n \Big)^2 \\ &= \|\varepsilon'\|^2 + \mathrm{M}E(\lambda) + 2\sum \varepsilon'_n \Big(\mathbf{x}_n, \boldsymbol{\beta}^* - \hat{\hat{\boldsymbol{\beta}}}(\lambda) \Big) \\ &= \|\varepsilon'\|^2 + \mathrm{M}E(\lambda) + 2\sum_m Z_m \Big(\beta^*_m - \hat{\hat{\beta}}_m(\lambda) \Big), \end{aligned}$$

where the $\{Z_m\}$ are iid $N(0,1)$ independent of the $\{\hat{\hat{\boldsymbol{\beta}}}(\lambda)\}$. Therefore, $\widehat{\mathrm{PE}}(\lambda)$ can be written as

$$\widehat{\mathrm{PE}}(\lambda) = V + \mathrm{M}E(\lambda) + \sqrt{M}\, Z(\lambda), \tag{4.6}$$

where $\{Z(\lambda)\}$ is an approximately Gaussian, mean-zero process, and V is a fixed r.v. not depending on λ.

The model selected using the test set PE estimate corresponds to

$$\hat{\lambda} = \arg\min \widehat{\mathrm{PE}}(\lambda).$$

The crystal ball model corresponds to

$$\lambda^* = \arg\min \mathrm{PE}(\lambda).$$

We want to estimate the expected size of the predictive loss

$$E(\mathrm{PL}) = E\big[\mathrm{PE}(\hat{\lambda}) - \mathrm{PE}(\lambda^*)\big].$$

Now λ^* is the minimum of $M \cdot A(\lambda) + \sqrt{M}\, W(\lambda)$ and $\hat{\lambda}$ is the minimizer of $M \cdot A(\lambda) + \sqrt{M}\,(W(\lambda) + Z(\lambda))$. Let $\lambda_0 = \arg\min A(\lambda)$. Then, if $W(\lambda), Z(\lambda)$ are differentiable at λ_0, simple calculations (see the Appendix) give the result that, for M large,

$$E(\mathrm{PL}) \sim K_1, \tag{4.7}$$

where K_1 is a constant depending on the distribution $P(d\beta^*)$ of the $\{\beta_m^*\}$ and θ.

Furthermore, the bias is

$$E\big(\mathrm{PE}(\hat{\lambda}) - \widehat{\mathrm{PE}}(\hat{\lambda})\big) \sim K_2, \tag{4.8}$$

where $K_2 > 0$ also depends on P, θ and the variance grows as

$$\mathrm{Var}\big(\mathrm{PE}(\hat{\lambda}) - \widehat{\mathrm{PE}}(\hat{\lambda})\big) \sim 2N + 4\mathrm{M}E(\lambda_0). \tag{4.9}$$

If $\theta(\beta, \lambda)$ is continuous and differentiable in λ, then $\{Z(\lambda)\}, \{W(\lambda)\}$ are differentiable processes and (4.7), (4.8) and (4.9) hold. That is, the expected predictive loss and bias are bounded as $M \to \infty$.

For subset regression θ is not continuous. The $\{W(\lambda)\}, \{Z(\lambda)\}$ processes are approximately Brownian motions in a neighborhood of λ_0. More complicated computations give (see the Appendix)

$$E(\mathrm{PL}) \sim M^{1/3}, \tag{4.10}$$

$$E\big(\mathrm{PE}(\hat{\lambda}) - \widehat{\mathrm{PE}}(\hat{\lambda})\big) \sim M^{1/3}, \tag{4.11}$$

with the same dominant variance terms as in (4.9). Thus, there is a sharp increase in predictive loss and bias for large M.

4.2. *Little and tiny bootstrap.* Using little and tiny bootstrap to approximate $E(\varepsilon^*, \hat{\mu})$ introduces another stochastic element into the PE estimate, that is,

$$\widehat{\mathrm{PE}}(\lambda) = \|\varepsilon^*\|^2 + 2(\varepsilon^*, \mu^*) + \mathrm{M}E(\lambda) + 2\big(B_t(\lambda) - (\varepsilon^*, \hat{\mu})\big),$$

where

$$B_t(\lambda) = \frac{1}{t^2} E_\varepsilon(\varepsilon, \hat{\mu}(\cdot, y + \varepsilon, \lambda_t))$$

and $\lambda_t = \lambda\sqrt{1+t^2}$ for all except ridge, where $\lambda_t = \lambda$. Let $tU_m = (\varepsilon, \mathbf{x}_m)$. Then

$$B_t(\lambda) = \frac{1}{t} E_U\left[\sum_m U_m \theta\left(\hat{\beta}_m + tU_m, \lambda_t\right)\right],$$

where the $\{U_m\}$ are iid $N(0,1)$.

Whether $B_t \to$ TB is equivalent to the existence of

$$\lim_{t\downarrow 0}\int u\left[\frac{\theta\left(\hat{\beta}+tu, \lambda_t\right) - \theta\left(\hat{\beta}, \lambda_t\right)}{t}\right] f(u)\,du. \tag{4.12}$$

If $\theta(\hat{\beta}, \lambda)$ is differentiable in $\hat{\beta}$, then the (4.12) limit exists and equals $\theta_1(\hat{\beta}, \lambda)$. Then $\mathrm{TB}(\lambda) = \sum\theta_1(\hat{\beta}_m, \lambda)$ and $(\varepsilon^*, \hat{\mu}) - \mathrm{TB}(\lambda)$ is a mean-zero process. If $\theta_1(\beta, \lambda)$ is differentiable in λ, then the process is differentiable near λ_0 and (4.6), (4.7) and (4.8) hold. Thus, E(PL) and the bias are bounded for ridge and garotte. But an $O(M)$ term is added to the variance.

A change occurs in *nn*-garotte. The limit in (4.12) exists but $\theta_1(\beta, \lambda)$ is discontinuous. In this case, $(\varepsilon^*, \hat{\mu}) - \mathrm{TB}(\lambda)$ is a zero-mean process, but resembles a Brownian motion near λ_0. The resulting E(PL) is in the $M^{1/3}$ range. Smaller values of E(PL) can be gotten by taking t to decrease as $M^{-1/5}$. Then $E(\mathrm{PL}) \sim M^{1/5}$.

For subset regression, the integral in (4.12) converges weakly (in $\hat{\beta}$) to $\hat{\beta}(\delta(\hat{\beta}-\lambda) - \delta(\hat{\beta}+\lambda))$, where δ is the Dirac delta. In fact, if $P(d\hat{\beta})$ has mass in the vicinity of $\pm\lambda$, then the expected square of the integral in (4.11) goes to ∞ as $t \downarrow 0$. Thus, $B_t \to$ TB is not possible. Taking t to go to 0 as $M^{-1/7}$ gives

$$E(\mathrm{PL}) \sim M^{3/7}.$$

The bias in *nn*-garotte goes up like $M^{3/5}$ and in subset selection like $M^{5/7}$. Also, *nn*-garotte adds both an $O(M)$ and an $O(M^{4/5})$ term to the variance and subset selection adds an $O(M)$ and $O(M^{8/7})$ term.

4.3. *Stabilization.* Consider generating new data $y' = y + \delta$, where $\mathrm{var}(\delta) = \tau^2\sigma^2$. Form $\hat{\mu}(\cdot, y+\delta, s)$ and now repeat and average. This gives a new estimate sequence

$$\hat{\mu}_{\mathrm{ST}}(\cdot, y, s) = E_\delta\,\hat{\mu}(\cdot, y+\delta, s), \tag{4.13}$$

which we call the stabilized sequence.

For $\hat{\hat{\beta}}(\lambda) = \theta(\hat{\beta}, \lambda)$, the stabilized coefficients are $\theta_{\mathrm{ST}}(\hat{\beta}, \lambda) = E_V\theta(\hat{\beta} + V, \lambda)$, where $V \in N(0, \tau^2\sigma^2)$. Thus, stabilization smooths θ. For θ discontinuous, θ_{ST} is nicely differentiable. The limit of the integral (4.12) is

$$\frac{1}{\tau^2}E_V V\theta\left(\hat{\beta}+V, \lambda\right). \tag{4.14}$$

This is differentiable in λ, so the tiny bootstrap works and gives bounded E(PL) and bias.

5. Simulation results for controlled X. To see how the results carry over to more complex situations, we constructed a simulation that used a variety of design matrices and coefficients. The sample size was 60 with 30 variables. The $\{\mathbf{x}_n\}$ data was sampled from the covariance matrix $\Gamma_{mk} = \rho^{|m-k|}$, where ρ was selected from a uniform $[0,1]$ distribution in each repetition and the coefficients occurred in random clusters with random sizes. The response values $\{y\}$ were generated as $y = \boldsymbol{\beta}^* \mathbf{x} + \varepsilon^*$ with ε^* iid $N(0,1)$. On the average, $R^2 \simeq 0.83$ and about 20 of the coefficients were nonzero.

Two runs of 500 repetitions each were done. One used a PE estimate based on a single replicate test set. The other used the little or tiny bootstrap. Five procedures were compared. Four are the regression methods defined in subsection 3.3. The fifth is stepwise forward stabilized by 40 repetitions of adding $N(0,1)$ noise to the $\{y_n\}$ and averaging the results. In each repetition of the simulation, PL, bias and some other characteristics were computed and then averaged over all repetitions.

5.1. *Test set results.* Figure 3 is a bar graph showing the average crystal ball ME's and the average PL's for the five procedures. The crystal ball ME's are in black, the PL's in white. The total bar height is the ME for the models selected by the test set $\widehat{\text{PE}}$. Figure 4 shows two bars for each procedure. The upper bar is the average bias as a percentage of the average PE. The lower bar is the average of the percentage error in $\widehat{\text{PE}}$ as an estimate of PE.

5.2. *The little and tiny bootstrap results.* The little bootstrap with $t = 0.6$ was used for stepwise and *nn*-garotte, with 25 iterations averaged to get B_t. With ridge, garotte and stabilized stepwise, the tiny bootstrap was used. To get the expressions for TB, we start with ridge.

Ridge turns the constrained minimization problem into the problem of locating the stationary points of the Lagrangian

$$\|\mathbf{y} - \boldsymbol{\beta}\mathbf{x}\|^2 + \lambda\|\boldsymbol{\beta}\|^2.$$

The solution is

$$\hat{\hat{\boldsymbol{\beta}}}(\lambda) = (X^t X + \lambda I)^{-1} X^t y.$$

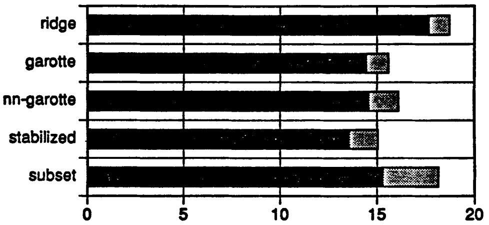

FIG. 3. *Crystal ball* ME *and predictive loss-controlled X data—test set.*

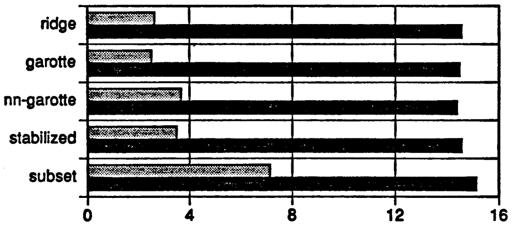

FIG. 4. *Percentage bias and error-controlled X data—test set.*

An easy computation [Breiman (1995)] yields

$$\mathrm{TB}(\lambda) = \hat{\sigma}^2 \operatorname{Tr}\Big((X^t X + \lambda I)^{-1} X^t X \Big).$$

For garotte, the restricted minimization over $(c_1, \dots, c_M)$ is converted into Lagrangian form as

$$\|y - \hat{\mu}(\mathbf{c})\|^2 + \lambda \|\mathbf{c}\|^2,$$

where $\hat{\mu}_n(\mathbf{c}) = \sum_m c_m \hat{\beta}_m x_{mn}$. Let $W(m,k) = \hat{\beta}_m (X'X)_{mk} \hat{\beta}_k$ and $W_\lambda = W + \lambda I$. Then (see the Appendix)

$$\mathrm{TB}(\lambda) = \hat{\sigma}^2 \bigg(M + \sum_{m,k} W_\lambda^{-1}(m,k) W(m,k)(1 - c_k) \\ - \lambda \sum_k W_\lambda^{-1}(k,k)(1 - c_k) \bigg).$$

With the stabilized stepwise procedures, if the $\{\delta\}$ are the noise added in stabilization, then

$$\mathrm{TB}(k) = \mathrm{Av}_\delta(\delta, \hat{\mu}(\cdot, y + \delta, k)).$$

Thus, TB(k) can be computed at the same time as the stabilized predictor is computed.

The simulation results are summarized in Figures 5 and 6. Figure 5 uses the Figure 3 format. Figure 6 uses the Figure 4 format.

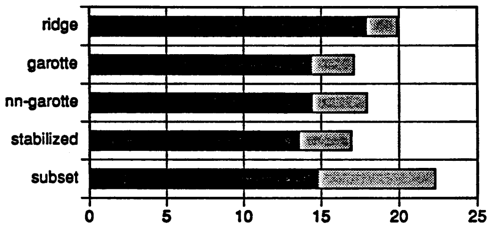

FIG. 5. *Crystal ball ME and predictive loss-controlled X data—little bootstrap.*

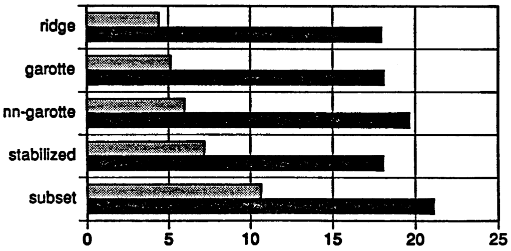

FIG. 6. *Percentage bias and error-controlled X data—little bootstrap.*

6. The random X simulation results. The random X case differs from the controlled X case in two aspects: first, the definition of prediction error; second, the methods for getting PE estimates. PE estimates can be gotten using a test set. The other common method is cross-validation. Somewhat to our surprise, Breiman and Spector (1992) found that, for selecting the best dimension in a stepwise procedure, leave-one-out did not work as well as leave-many-out. We now understand this as a consequence of instability. Thus, with the cross-validation run, we used leave-one-out for ridge and garotte, and leave-many-out for the others. In leaving-many-out, 30 left-out sets were constructed as follows: the data was randomly permuted. The first left-out set was the first 1/6 of the data (10 cases). Then the second 1/6 was left out and so on. This was repeated five times.

Otherwise, the simulation has the same structure as in the controlled X case, that is, 60 cases, 30 variables, some covariance and coefficients and so forth. In the test set run, a test set of the same size (60) as the learning set was used. The results are summarized in Figures 7 and 8 using that same display format as in the controlled X figures. The output of the cross-validation run is shown in Figures 9 and 10.

The symptoms of instability are that both PE(s) and $\widehat{\text{PE}}(s)$ are noisy and that $\widehat{\text{PE}}(s)$ does not track PE(s) accurately. In subset selection and stabilization PE and $\widehat{\text{PE}}$ were computed for $k = 1, \dots, 30$, where k is the dimensionality. In the ridge and garotte regressions, PE and $\widehat{\text{PE}}$ were also computed at integer values $k = 1, \dots, 30$, where k is the dimensionality equivalent to the

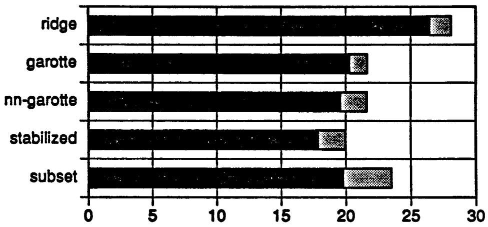

FIG. 7. *Crystal ball* M*E and predictive loss-random X data—test set.*

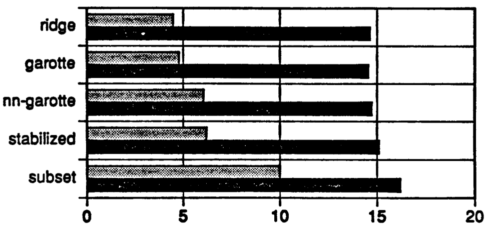

FIG. 8. *Percentage bias and error-random X data—test set.*

s parameter value. The values of $T_k = |\Delta_k - \hat{\Delta}_k|$, where $\Delta_k = \mathrm{PE}(k+1) - \mathrm{PE}(k)$ and $\hat{\Delta}_k = \widehat{\mathrm{PE}}(k+1) - \widehat{\mathrm{PE}}(k)$, were averaged over the 500 repetitions of the cross-validation simulation.

These are plotted vs. k in Figure 11. The crucial parts of these curves are at the values of k of which $\mathrm{PE}(k)$ is a minimum. The average value of the minimizing k is about 5 for subset selection, stabilization and *nn*-garotte; 9 for garotte; and 18 for ridge.

6.1. *Stabilization.* The stabilization story for random X is somewhat perplexing. Our first idea was to perturb the data by a mechanism similar to that used in cross-validation. That is, leave out a set $\mathcal{T}$ of cases. Run the procedure on the remaining $\mathcal{L} - \mathcal{T}$ cases, getting $\hat{\mu}(\cdot, s, \mathcal{L} - \mathcal{T})$. Then repeat this K times, leaving out the subsets $\mathcal{T}_1, \dots, \mathcal{T}_K$, and define

$$\hat{\mu}_{\mathrm{ST}}(\cdot, s) = \frac{1}{K}\sum_k \hat{\mu}(\cdot, s, \mathcal{L} - \mathcal{T}_k).$$

We implemented this using 10 cases in each $\mathcal{T}_K$, $K = 30$, with the random selection of the $\mathcal{T}_K$ structured so that each case occurred in exactly 5 of the $\mathcal{T}_K$. The optimal value of s is selected using cross-validation. Another collection of sets $\{\mathcal{T}_j'\}$, $j = 1, \dots, J$, is defined. Set

$$\hat{\mu}_{\mathrm{ST}}(\cdot, s, \mathcal{L} - \mathcal{T}_j') = \frac{1}{K}\sum \hat{\mu}(\cdot, s, \mathcal{L} - \mathcal{T}_k - \mathcal{T}_j'),$$

$$\widehat{\mathrm{PE}}(\mathcal{T}_j') = \sum_{(y_n, \mathbf{x}) \in \mathcal{T}_j'} \left(y_n - \hat{\mu}_{\mathrm{ST}}(\mathbf{x}_n, s, \mathcal{L} - \mathcal{T}_j')\right)^2$$

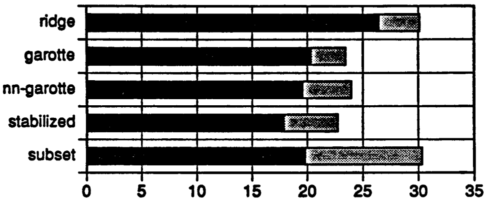

FIG. 9. *Crystal ball* M*E and predictive loss-random X data—cross-validation.*

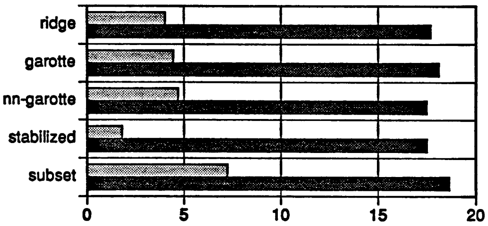

FIG. 10. *Percentage bias and error-random X data—cross-validation.*

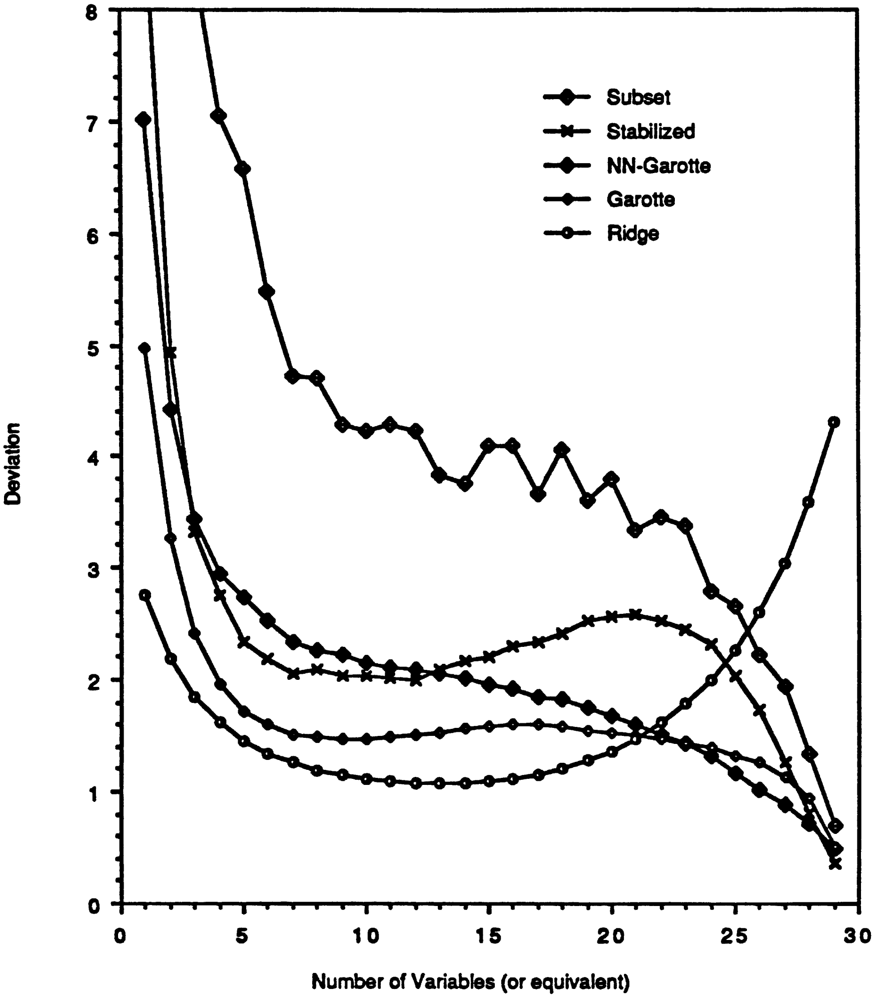

FIG. 11. *Tracking deviations.*

and

$$\widehat{\mathrm{PE}}(s) = \frac{N}{\Sigma_j N_j} \sum_j \widehat{\mathrm{PE}}(\mathscr{T}'_j),$$

where $N_j = |\mathscr{T}'_j|$. Two definitions of the $\{\mathscr{T}'_j\}$ were used. One was $\mathscr{T}'_j = \{(y_j, \mathbf{x}_j)\}$; that is, leave-one-out cross-validation was applied. In the second, $\{\mathscr{T}'_j\} = \{\mathscr{T}_k\}$, so leave-ten-out cross-validation was used.

In applying stabilization to subset selection, the leave-ten-out estimate did better. It gave Av(PL) of 7.1. Subset selection itself had Av(PL) = 10.5, so stabilization did give a 32% reduction in average PL. However, since Av(PL) for the two garottes and ridge were 4.3, 3.1 and 3.6, we questioned whether the results could be improved.

Two avenues seemed open. One was to increase the amount of averaging in the stabilization. We went from 30 sets to averaging over 60 sets. The same sets were used for averaging and cross-validation PE estimates. The results improved a little, with Av(PL) = 6.7.

The other possibility was to change the method of stabilization. One candidate was the method used in the controlled X situation, that is, generate new y-values as $y' = y + \varepsilon'$, rerun the procedure using the new y-values, repeat 50 times and average. This was combined with the use of leave-ten-out cross-validation to do PE estimation. In this case, Av(PL) dropped to 4.9.

Thus, perturbing the y-values and averaging does better at stabilization than perturbing by leaving out some portion of the data and averaging. This also suggests that we do not know yet what the best stabilization method is. Our intuition is that some method which perturbs both the y- and $\mathbf{x}$-values will probably do better than a perturbation of the y-values only.

7. Leave-one-out vs. leave-many-out.

Recall that in cross-validation a set $\mathscr{T}$ of N_{CV} cases is left out, and $\hat{\mu}^{(-\mathscr{T})}$ is defined as the minimizer in $\mathscr{U}_s$ of $\|y - \mu\|^2$ for the data $\mathscr{L} - \mathscr{T}$. Then repeating this for sets $\mathscr{T}_1, \dots, \mathscr{T}_K$,

$$\widehat{\mathrm{PE}}(s) = \frac{N}{K \cdot N_{\mathrm{CV}}} \sum_k \sum_{(y_n, \mathbf{x}_n) \in \mathscr{T}_k} \left(y_n - \hat{\mu}^{(-\mathscr{T}_k)}(\mathbf{x}_n)\right)^2.$$

The relevant question is: *what is* $\widehat{\mathrm{PE}}(s)$ *an estimate for*? In particular, let $\hat{\mu}$ be the minimizer of $\|y - \mu\|^2$ in $\mathscr{U}_s$. How is $\widehat{\mathrm{PE}}(s)$ connected to $\mathrm{PE}(\hat{\mu})$?

If N_{CV} is small and the procedure stable, then $\hat{\mu}^{(-\mathscr{T})} \simeq \hat{\mu}$ and $\widehat{\mathrm{PE}}(s)$ resembles a test set estimate of $\mathrm{PE}(\hat{\mu})$. But if the procedure is unstable, then, even for N_{CV} small, $\hat{\mu}^{(-\mathscr{T})}$ may be considerably different from $\hat{\mu}$. The $\hat{\mu}$, $\hat{\mu}^{(-\mathscr{T})}$ are chosen to be minimum RSS predictors in $\mathscr{U}_s$. Then, usually, $\mathrm{RSS}(\hat{\mu}) \simeq \mathrm{RSS}(\hat{\mu}^{(-\mathscr{T})})$, but $\mathrm{PE}(\hat{\mu})$ may differ considerably from $\mathrm{PE}(\hat{\mu}^{(-\mathscr{T})})$.

Figure 12 illustrates the last point. In the data generated in the first repetition of the cross-validation simulation, one case at a time was left out and the forward stepwise procedure applied to get a six-variable predictor. This gave 60 predictors. The RSS and ME were computed for each one. Fig-

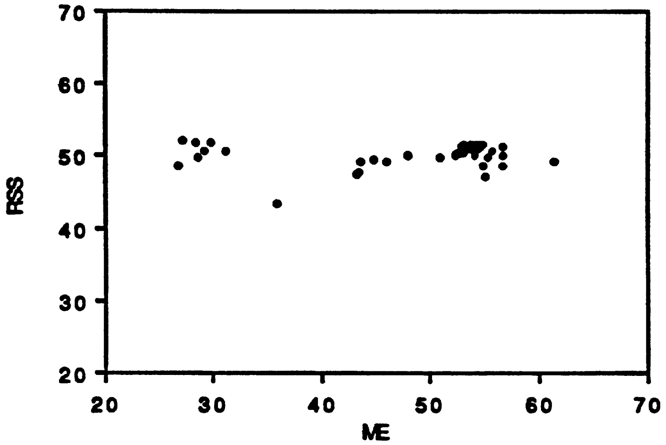

FIG. 12. *RSS vs.* M*E subset selection* ($k = 6$) *in* 60 *leave-one-out models.*

ure 12 is a plot of RSS vs. ME. The spread in ME is about 10 times that in RSS. Figure 13 is a similar plot for the same data using the garotte method. Here the ME spread is about equal to the RSS spread. The cross-validation $\widehat{\mathrm{PE}}(s)$ is estimating some average of the values of PE($\hat{\mu}^{(-\mathscr{T})}$). For an unstable procedure, there is no guarantee that this average is close to PE($\hat{\mu}$).

In Breiman and Spector (1992), simulation results showed that leave-one-out cross-validation was inferior to 10-fold cross-validation in subset selection. The simulation structure here is different, but the results are similar. When leave-one-out cross-validation is used on the same data as leave-ten-out cross-validation, the average prediction loss increases from 10.5 to 11.6. The average downward bias goes from 6.5 to 19.2.

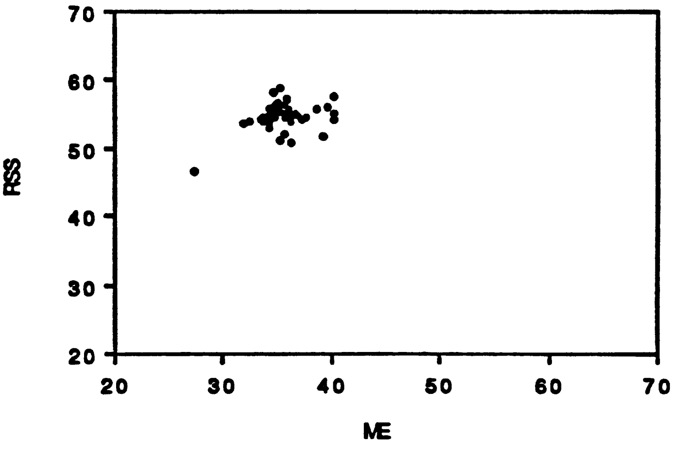

FIG. 13. *RSS vs.* M*E garotte* (*equiv.* $k = 6$) *in* 60 *leave-one-out models.*

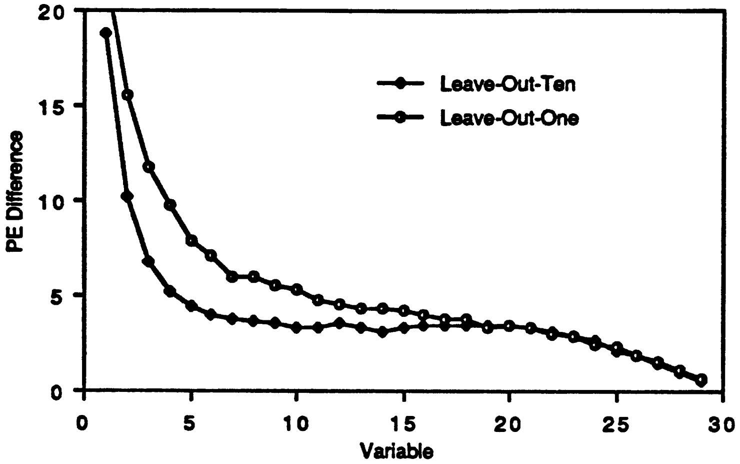

FIG. 14. *Differences in successive PE estimates.*

The large downward bias is an indicator of the problem. Figure 14 compares the average differences $|\hat{\Delta}|$ for leave-one-out and leave-ten-out. Figure 15 compares the average values of the tracking differences T. The average dimension of the minimum PE subset is $k \simeq 5$, and this is in the vicinity where the leave-one-out $\widehat{\mathrm{PE}}$ estimate is noisier and tracks more poorly than the leave-ten-out estimate.

We also computed the following value: in each repetition, let $\hat{k} = \arg\min \widehat{\mathrm{PE}}(k)$. Then the hole size in that repetition is defined as

$$\tfrac{1}{2}\big(\widehat{\mathrm{PE}}(\hat{k}-1)+\widehat{\mathrm{PE}}(\hat{k}+1)\big)-\widehat{\mathrm{PE}}(\hat{k}).$$

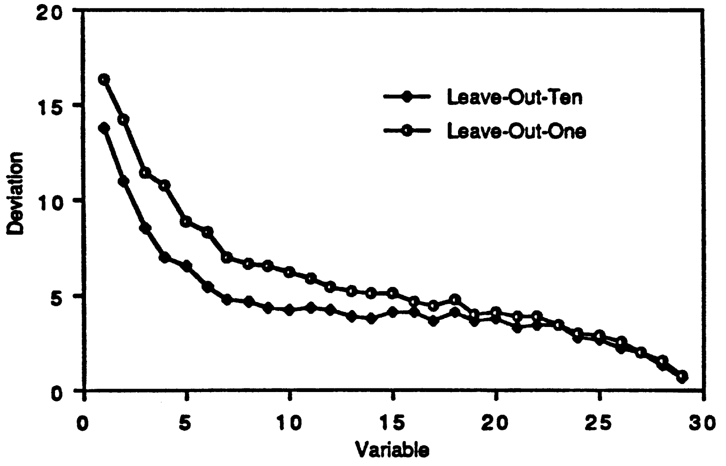

FIG. 15. *Tracking deviations.*

The average hole size in leave-one-out is 18.1 compared to 6.7 in leave-ten-out. Thus, in leave-one-out $\widehat{\mathrm{PE}}(s)$, the minimums occur at a place where there are deep local downward excursions.

The root of the problem is that while the leave-one-out estimate $\widehat{\mathrm{PE}}(k)$ has lower bias for fixed k, it is degraded by its higher variance. This is illustrated more concretely by the fact that the variance of the little bootstrap estimate in subset selection went to ∞ as t decreased to 0.

8. Comparing predictors. We were curious to see how stabilization of subset selection would compare with the other prediction methods across a spectrum of simulated data. It is fairly well known that if there are only a few nonzero coefficients, then subset selection gives good prediction. With many nonzero coefficients, ridge does better.

To compare methods, we generated five sets of simulated data that ranged from a few nonzero coefficients to many nonzero coefficients. The **X**-distribution was mean-zero 30-variable multivariate normal with $\Gamma_{ij} = \rho^{|i-j|}$. In each repetition, ρ was chosen from the uniform distribution on $[-1, 1]$.

The nonzero coefficients were in three clusters of adjacent variables with clusters centered at the 5th, 15th and 25th variables. For the variables clustered around the 5th variable, the initial coefficient values were given by

$$\beta^*_{5+j} = (h - j)^2, \qquad |j| \leq h,$$

where h is a fixed integer governing the cluster width. The clusters at 15 and 25 had the same shape. All other coefficients were 0. The coefficients were then multiplied by a common constant to give an $R^2 \simeq 0.75$ when $N(0,1)$ noise was added to $\sum \beta^*_m x_m$ to give y.

The h-values $1, 2, 3, 4, 5$ were used. This gave $3, 9, 15, 21, 27$ nonzero coefficients. For $h = 1$ there were three stong, virtually independent variables. At the other extreme, for $h = 5$ each cluster contained nine weak variables. This simulation structure is almost identical to that used in Breiman and Spector (1992). Two PE estimation methods were used. Sixfold cross-validation repeated five times was used in subset selection, subset selection stabilization and *nn*-garotte. Leave-one-out cross-validation was used in garotte and ridge.

Figure 16 is a graph of the average ME's vs. h for the various prediction methods. Figure 17 is a graph of the average crystal ball ME's vs. h, and Figure 18 is a plot vs. h of the differences. The conclusions are clear and interesting.

All methods except ridge have similar crystal ball MEs. Ridge has high ME except when there are many small nonzero coefficients. This reflects its inability to fit equations with a mixture of large and small underlying coefficients. Predictive loss separates the methods. Subset regression's predictive loss is large. Stabilization and *nn*-garotte have lower and similar losses. Lowest are garotte and ridge. In total ME, subset regression is a loser due to its high predictive loss. Ridge loses due to its high crystal ball ME. The garottes and stabilization do well.

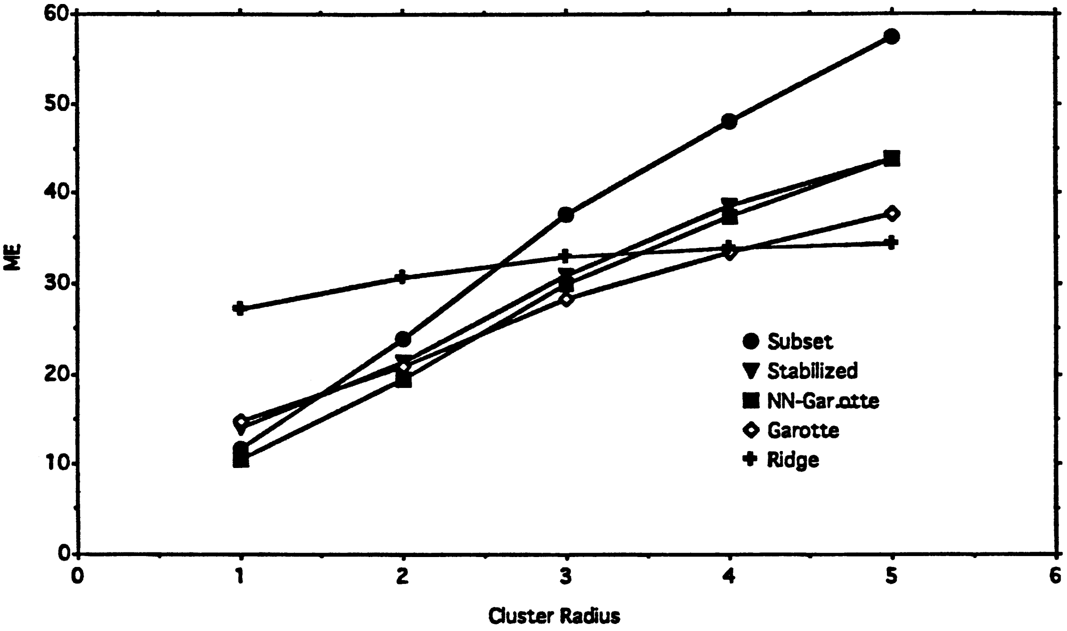

FIG. 16. *Total* M*E for five methods.*

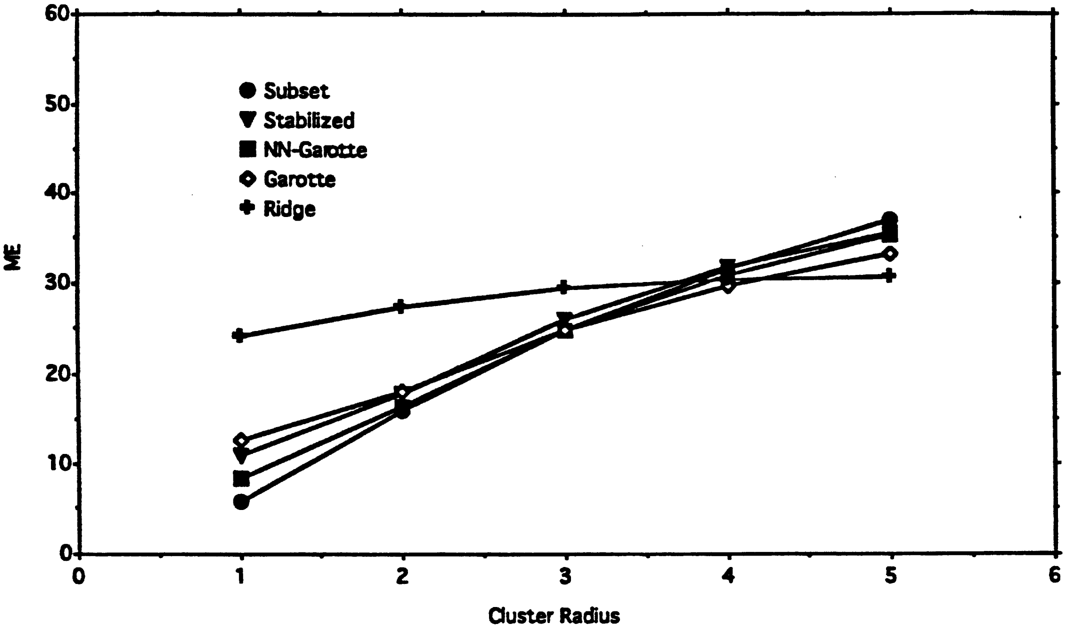

FIG. 17. *Crystal ball* M*E for five methods.*

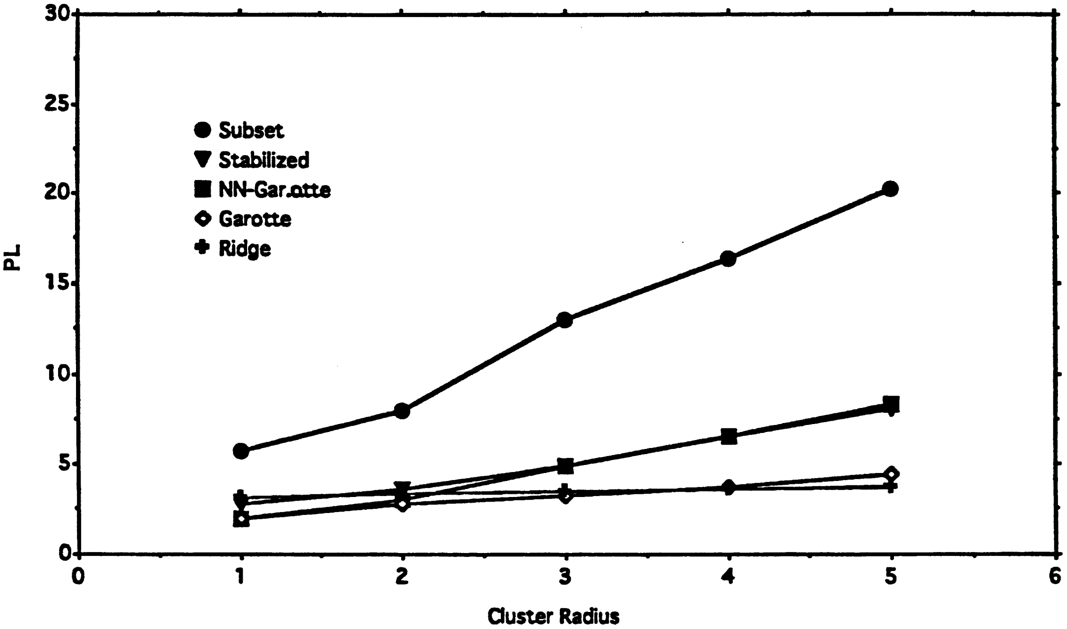

FIG. 18. *Predictive loss for five methods.*

9. Concluding remarks. We have studied the effects of instability on predictive loss and on the bias and error of PE estimates. Stabilization works, but within limits. In our implementation (altered y's) it reduces the level of instability sharply, but not to the level of garotte and ridge. This may be because our stabilization method is not sufficiently optimized.

Stabilized predictors lack simplicity. For instance, stabilizing the six-variable subset predictor generally gives a predictor with many more than six nonzero coefficients. Stabilization is computationally intensive and, in our context, does no better than the garotte methods. Why use it then?

The answer lies outside of the domain of linear regression predictors. When using nonlinear predictors there are usually no simple and effective stable alternatives. There are no known stable versions of CART, MARS or neural networks. Stabilizing these methods can give nonlinear predictors with improved accuracy. In particular, Breiman (1996b, c) shows that stabilizing CART leads to dramatic improvements in accuracy.

In the interesting linear regression sideshow the garotte methods show up as uniformly better than subset selection or ridge. Subset selection loses because of its large predictive loss. Ridge loses because its best models cannot fit the data as well as the other methods when there is a mix of large and small coefficients. The best methods combine stability with a better range of fits.

While stable procedures have desirable properties, stabilization by averaging is not a panacea. An area that needs exploration is the possibility of stabilization of procedures by changing their structure instead of averaging. For instance, the *nn*-garotte is a more stable alternative to subset selection [Breiman (1995)]. An interesting research issue we are exploring is whether there is a more stable single-tree version of CART.

A possible alternative is the idea of stacking predictors [Wolpert (1992) and Breiman (1996a)]. A stacking collection $\{\hat{\mu}_k\}$ of predictors are combined to form a predictor

$$\mu = \sum \alpha_k \hat{\mu}_k,$$

where the $\{\alpha_k\}$ (constrained to be nonnegative) are determined by a linear regression of the y-values on the $\{\hat{\mu}_k\}$ using the cross-validated values of the $\{\hat{\mu}_k\}$.

APPENDIX

A1. Computations in the $X^tX = I$ case. From (4.5),

$$\mathrm{M}E(\lambda) = M\cdot A(\lambda) + \sqrt{M}\,W(\lambda).$$

Let $\lambda_0 = \arg\min A(\lambda)$ and $\Delta = \lambda - \lambda_0$, using prime to indicate derivatives,

$$\begin{aligned}\mathrm{M}E(\lambda_0+\Delta) \simeq M\cdot A(\lambda_0) + \sqrt{M}\,W(\lambda_0) + M\Delta^2 A''(\lambda_0)/2\\ +\sqrt{M}\,(W(\lambda_0+\Delta) - W(\lambda_0)).\end{aligned}$$

If $W(\lambda)$ is differentiable at λ_0, then $W(\lambda_0+\Delta) - W(\lambda_0) \simeq \Delta W'(\lambda_0)$ and

$$\mathrm{M}E(\lambda^*) = \min \mathrm{M}E(\lambda) \simeq M\cdot A(\lambda_0) + \sqrt{M}\,W(\lambda_0) - \frac{W'(\lambda_0)^2}{2A''(\lambda_0)}.$$

Test set. Using a replicate test set gives [see (4.6)]

$$\widehat{\mathrm{PE}}(\lambda) = V + \mathrm{M}E(\lambda) + \sqrt{M}\,Z(\lambda),$$

where V is an r.v. not depending on λ and $\sqrt{M}\,Z(\lambda) = -2\Sigma_m Z_m \hat{\hat{\beta}}_m(\lambda)$. Put $\tilde{\lambda} = \arg\min \widehat{\mathrm{PE}}(\lambda) = \lambda_0 + \hat{\Delta}$, where

$$\begin{aligned}\hat{\Delta} = \arg\min\big\{\tfrac{1}{2}M\Delta^2 A''(\lambda_0) + \sqrt{M}\,(W(\lambda_0+\Delta) - W(\lambda_0))\\ +\sqrt{M}\,(Z(\lambda_0+\Delta) - Z(\lambda_0))\big\}.\end{aligned} \tag{A.1}$$

If both W and Z are differentiable then,

$$\hat{\Delta} = -\frac{(W'(\lambda_0) + Z'(\lambda_0))}{\sqrt{M}\,A''(\lambda_0)} \tag{A.2}$$

and

$$\mathrm{M}E(\hat{\lambda}) = M \cdot A(\lambda_0) + \sqrt{M}\,W(\lambda_0) + \tfrac{1}{2}M\,\hat{\Delta}^2 A''(\lambda_0) + \hat{\Delta}\sqrt{M}\,W'(\lambda_0),$$

resulting in

$$E\big(\mathrm{M}E(\hat{\lambda}) - \mathrm{M}E(\lambda^*)\big) \simeq \tfrac{1}{2}\frac{EZ'(\lambda_0)^2}{A''(\lambda_0)}, \tag{A.3}$$

which verifies our assertion that $E(\mathrm{PL})$ is bounded in M.

Consider a process

$$Y(\lambda) = \frac{1}{\sqrt{M}}\sum_m X_m(\lambda),$$

where the $\{X_m(\lambda)\}$ are iid mean 0. Then $Y(\lambda)$ is mean-square differentiable if $\lim E[X(\lambda + \Delta) - X(\lambda)/\Delta]^2$ exists as $\Delta \to 0$. It is straightforward to verify that all methods except subset regression are mean-square differentiable which is enough to justify (A.3).

In subset selection $\hat{\hat{\beta}}(\lambda) = I(|\hat{\beta}| \geq \lambda)\hat{\beta}$, so $(\beta^* - \hat{\hat{\beta}})^2 = I(|\hat{\beta}| < \lambda)(\beta^{*2} - Z^2) + Z^2$, where $\hat{\beta} = \beta^* + Z$; Z, β^* are independent and $Z \in N(0,1)$. Let $X_m(\lambda) = I(|\hat{\beta}_m| < \lambda)(\beta_m^{*2} - Z_m^2)$, so

$$W(\lambda) = \frac{1}{\sqrt{M}}\sum_m \big(X_m(\lambda) - EX_m(\lambda)\big) + \frac{1}{\sqrt{M}}\sum_m \big(Z_m^2 - 1\big).$$

Let $H(\Delta) = W(\lambda_0 + \Delta) - W(\lambda_0)$ and $D(\Delta) = X(\lambda_0 + \Delta) - X(\lambda_0)$. Then

$$EH(\Delta_1)H(\Delta_2) = ED(\Delta_1)D(\Delta_2) - ED(\Delta_1)ED(\Delta_0).$$

The second term is $O(\Delta_1\Delta_2)$. Write $X(\lambda) = I(|\hat{\beta}| < \lambda)Y$, $Y = (\beta^{*2} - Z^2)$. For $\Delta > 0$, $X(\lambda_0 + \Delta) - X(\lambda_0) = YI(\lambda_0 \leq |\hat{\beta}| < \lambda_0 + \Delta)$ and, for $\Delta_1, \Delta_2 > 0$,

$$EH(\Delta_1)H(\Delta_2) \simeq E\big(Y^2|\,|\hat{\beta}| = \lambda_0\big)P\big(\lambda_0 \leq |\hat{\beta}| \leq \lambda_0 + \min(\Delta_1, \Delta_2)\big).$$

Denoting the density of $|\hat{\beta}|$ by f,

$$EH(\Delta_1)H(\Delta_2) \simeq E\big(Y^2|\,|\hat{\beta}| = \lambda_0\big)f(\lambda_0)\min(\Delta_1, \Delta_2) + O(\Delta_1\Delta_2).$$

If $\Delta_1, \Delta_2 < 0$, the same result follows with $\min(\Delta_1, \Delta_2)$ replaced by $\min(|\Delta_1|, |\Delta_2|)$. If Δ_1, Δ_2 have opposite signs, then $EH(\Delta_1)H(\Delta_2) = O(\Delta_1\Delta_2)$. Thus,

$$H(\Delta) \simeq c(\lambda_0)B_1(\Delta),$$

where $B_1(\Delta)$ is a two-sided Brownian motion.

We can write $\mathrm{M}E(\lambda)$ as

$$\begin{aligned}\mathrm{M}E(\lambda_0 + \Delta) &= V + \tfrac{1}{2}M\,\Delta^2 A'' + \sqrt{M}\,cB_1(\Delta)\\ &= V + Q(\Delta),\end{aligned}$$

where V is an r.v. not depending on Δ. Put $\Delta = \alpha t$, where α is determined by

$$b = \sqrt{M\alpha}\,c = \tfrac{1}{2}\alpha^2 MA''.$$

Use the fact that, for fixed α, $B(t) = B_1(\alpha t)/\sqrt{\alpha}$ is a two-sided B-motion to get

$$Q(\Delta) = b\left[t^2 + B(t)\right]$$

and

$$\min_{\lambda} \mathrm{M}E(\lambda) \simeq V + b\min_{t}\left[t^2 + B(t)\right].$$

The processes $\{Z(\lambda)\}, \{W(\lambda)\}$ are independent, and, for $G(\Delta) = Z(\lambda_0 + \Delta) - Z(\lambda_0)$, another straightforward computation gives

$$E(G(\Delta_1)G(\Delta_2)) = 4\lambda_0^2 f(\lambda_0)\min(|\Delta_1|, |\Delta_2|)$$

if Δ_1, Δ_2 have the same sign and $O(\Delta_1\Delta_2)$ if not. Put $d(\lambda_0) = 4\lambda_0^2 f(\lambda_0)$. Then $G(\Delta) = d(\lambda_0)B_2(\Delta)$, where $\{B_2(\Delta)\}$ is a two-sided B-motion independent of $\{B_1(\Delta)\}$.

Thus,

$$\hat{\Delta} = \arg\min\left[\tfrac{1}{2}\Delta^2 MA'' + \sqrt{M}\,(cB_1(\Delta) + dB_2(\Delta))\right].$$

Since

$$B_0(\Delta) = \frac{cB_1(\Delta) + dB_2(\Delta)}{\sqrt{c^2 + d^2}}$$

is also a two-sided B-motion, then

$$\hat{\Delta} = \arg\min\left\{\tfrac{1}{2}\,\Delta^2 MA'' + r\sqrt{M}\,B_0(\Delta)\right\},$$

where $r = \sqrt{c^2 + d^2}$. Put $\Delta = \gamma t$, where γ is determined by

$$e = \tfrac{1}{2}\gamma^2 MA'' = r\sqrt{M\gamma}\,.$$

So

$$\hat{\Delta} = \gamma \arg\min_{t}\left[t^2 + B_0(t)\right].$$

The resulting approximation is

$$\mathrm{M}E\left(\lambda_0 + \hat{\Delta}\right) = V + \tfrac{1}{2}M\,\hat{\Delta}^2 A'' + c\sqrt{M}\,B_1(\hat{\Delta}).$$

Because $\hat{\Delta}$ is measurable on $\{cB_1(\Delta) + dB_2(\Delta),\ -\infty < \Delta < \infty\}$, another argument gives

$$E\left(cB_1(\hat{\Delta})\middle|\{cB_1(\Delta) + dB_2(\Delta),\ -\infty < \Delta < \infty\}\right)$$
$$= \frac{c^2}{c^2 + d^2}\left(cB_1(\hat{\Delta}) + dB_2(\hat{\Delta})\right).$$

The conditional expectation of $\mathrm{M}E(\hat{\lambda})$ given $\{cB_1(\Delta) + dB_2(\Delta),\ -\infty < \Delta < \infty\}$ is

$$\frac{1}{2}M\,\hat{\Delta}^2A'' + \frac{c^2}{c^2+d^2}\sqrt{M}\left(cB_1(\hat{\Delta}) + dB_2(\hat{\Delta})\right)$$

$$= \frac{1}{2}M\frac{d^2}{c^2+d^2}\,\hat{\Delta}^2A'' + \frac{c^2}{c^2+d^2}\left(\frac{1}{2}M\,\hat{\Delta}^2A'' + \sqrt{M}\left(cB_1(\hat{\Delta}) + dB_2(\hat{\Delta})\right)\right).$$

The last term equals $ec^2y/(c^2+d^2)$, where $y = \min[t^2 + B_0(t)] < 0$. Thus,

$$\text{(A.4)}\quad E\left(\mathrm{M}E(\hat{\lambda}) - \mathrm{M}E(\lambda^*)\right) \simeq \frac{1}{2}\frac{d^2}{c^2+d^2}MA''E\hat{\Delta}^2 + \left(\frac{ec^2}{c^2+d^2} - b\right)Ey.$$

To see how big (A.4) is, note that

$$b = c\left(\frac{2cM}{A''}\right)^{1/3}, \qquad e = r\left(\frac{2rM}{A''}\right)^{1/3}, \qquad \gamma = M^{-1/3}\left(\frac{2}{A''}\right)^{2/3}.$$

Put $\hat{t} = \arg\min(t^2 + B_0(t))$. Then

$$E(\mathrm{PL}) = M^{1/3}\left(\frac{2}{A''}\right)^{1/3}\left[d^2r^{-2/3}E\hat{t}^2 + (1 - R^{2/3})c^{4/3}|Ey|\right],$$

where $R^2 = c^2/c^2 + d^2$. Thus, $E(\mathrm{PL}) \sim M^{1/3}$.

Errors in PE *estimates using a test set.* Here we look at the mean and variance of $\widehat{\mathrm{PE}}(\hat{\lambda}) - \mathrm{PE}(\hat{\lambda})$. This difference equals

$$\|\varepsilon'\|^2 - N + \sqrt{M}Z(\lambda_0) + \sqrt{M}\left(Z\left(\lambda_0 + \hat{\Delta}\right) - Z(\lambda_0)\right).$$

In the differentiable case, the last term equals $+\sqrt{M}\,\hat{\Delta}Z'(\lambda_0)$. From (A.2), we get

$$E\left(\widehat{\mathrm{PE}}(\hat{\lambda}) - \mathrm{PE}(\hat{\lambda})\right) = -\frac{EZ'(\lambda_0)^2}{A''(\lambda_0)},$$

so that $\widehat{\mathrm{PE}}(\hat{\lambda})$ has an $O(1)$ downward bias.

To get the variance, note that $\|\varepsilon'\|^2$ is χ_N^2 independent of $Z(\lambda_0)$ and that $M\cdot EZ^2(\lambda_0) = 4\Sigma_m(\beta_m^* - \hat{\hat{\beta}}_m(\lambda_0))^2 \simeq 4\mathrm{M}E(\lambda_0)$. Thus,

$$\mathrm{Var}\left(\widehat{\mathrm{PE}}(\hat{\lambda}) - \mathrm{PE}(\hat{\lambda})\right) \sim 2N + 4\mathrm{M}E(\lambda_0).$$

A more memorable version of this result is that the SE of $\widehat{\mathrm{PE}}(\hat{\lambda})$ is the range $\sqrt{2\widehat{\mathrm{PE}}(\hat{\lambda})}$ to $\sqrt{4\widehat{\mathrm{PE}}(\hat{\lambda})}$.

In subset selection,

$$\sqrt{M}\left(Z\left(\lambda_0 + \hat{\Delta}\right) - Z(\lambda_0)\right) \simeq d(\lambda_0)B_2(\hat{\Delta}).$$

Using calculations similar to those in the $E(\mathrm{PL})$ calculations gives

$$E\left(\widehat{\mathrm{PE}}(\hat{\lambda}) - \mathrm{PE}(\tilde{\lambda})\right) \simeq -K_3M^{1/3}$$

and

$$\operatorname{Var}\big(\widehat{\mathrm{PE}}(\hat{\lambda}) - \mathrm{PE}(\hat{\lambda})\big) \simeq 2N + 4\mathrm{M}E(\lambda_0) + K_4 M^{2/3}.$$

Little bootstrap. For $\theta(\beta, \lambda)$ differentiable in β, $B_t(\lambda) \to \mathrm{TB}(\lambda)$ as $t \to 0$,

$$\mathrm{TB}(\lambda) = \sum_m \theta_1\big(\hat{\beta}_m, \lambda\big)$$

and

$$\begin{aligned}\widehat{\mathrm{PE}}(\lambda) &= \mathrm{RSS}(\lambda) + 2\mathrm{TB}(\lambda)\\ &= \|\varepsilon^*\|^2 + 2\sum Z_m \beta_m^* + \mathrm{M}E(\lambda) - 2\sum\big(Z_m\theta\big(\hat{\beta}_m, \lambda\big) - \theta_1\big(\hat{\beta}_m, \lambda\big)\big)\\ &= V + \mathrm{M}E(\lambda) + \sqrt{M}Z(\lambda),\end{aligned}$$

where V is an r.v. not depending on λ and $Z(\lambda)$ is a zero-mean, approximately Gaussian process. If both $W(\lambda)$ and $Z(\lambda)$ are differentiable, then, as in the test set case,

$$E\big(\mathrm{M}E(\hat{\lambda}) - \mathrm{M}E(\lambda^*)\big) = \frac{1}{2}\frac{EZ'(\lambda_0)^2}{A''(\lambda_0)}.$$

For $Z(\lambda)$ to be differentiable, the existence of $\partial^2\theta/\partial\beta\,\partial\lambda$ is necessary. This is violated both by subset selection and *nn*-garotte, but holds for garotte and ridge. In the *nn*-garotte case, $\mathrm{TB}(\lambda)$ is not differentiable in λ. Now $\mathrm{TB}(\lambda_0 + \Delta) - \mathrm{TB}(\lambda_0)$ can be approximated by a Brownian motion, leading to $E(\mathrm{PL}) \sim M^{1/3}$. However, there is an alternative strategy leading to lower $E(\mathrm{PL})$, that is, take $t > 0$ going to 0 as $M \to \infty$.

For $t > 0$, $B_t(\lambda)$ is smooth and differentiable in λ. The problem is that $EB_t(\lambda) \neq E(\varepsilon^*, \hat{\mu})$. Trade off by taking t small enough so that $EB_t(\lambda)$ is not far from $E(\varepsilon^*, \hat{\mu})$, but positive enough so that $B_t(\lambda)$ is nicely differentiable.

Let

$$\eta_t(\beta, \lambda) = \frac{1}{t}EU\theta\big(\beta + tU, \lambda\sqrt{1 + t^2}\big), \qquad U \in N(0, 1),$$

so

$$B_t(\lambda) = \sum_m \eta_t\big(\hat{\beta}_m, \lambda\big).$$

Put

$$Y_t(\lambda) = -\frac{2}{\sqrt{M}}\sum_m\big(Z_m\theta\big(\hat{\beta}_m, \lambda\big) - \eta_t\big(\hat{\beta}_m, \lambda\big)\big)$$

and $Z_t(\lambda) = Y_t(\lambda) - EY_t(\lambda)$. Then

$$\widehat{\mathrm{PE}}(\lambda) = V + \mathrm{M}E(\lambda) + \sqrt{M}Z_t(\lambda) + \sqrt{M}EY_t(\lambda).$$

Define $h(\beta, \lambda) = E_Z Z\theta(\beta + Z, \lambda)$. A conditional expectation computation gives

$$E_Z \eta_t(\beta + Z, \lambda) = h\left(\frac{\beta}{\sqrt{1+t^2}}, \lambda\right)$$
$$= h(\beta, \lambda) - \frac{t^2}{2}\beta h_1(\beta, \lambda) + o(t^2).$$

Therefore,

$$\sqrt{M}\,EY_t(\lambda) = -Mt^2 E\beta^* h_1(\beta^*, \lambda) + o(t^2)$$
$$= -Mt^2 D(\lambda) + o(t^2).$$

In consequence,

$$\hat{\Delta} = \arg\min\left[\tfrac{1}{2}M\,\Delta^2 A''(\lambda_0) + Mt^2\,\Delta D'(\lambda_0) + \Delta\sqrt{M}\left(Z_t'(\lambda_0) + W'(\lambda_0)\right)\right],$$

so

$$\hat{\Delta} = -\frac{Z_t'(\lambda_1) + W'(\lambda_0)}{\sqrt{M}\,A''(\lambda_0)} - \frac{t^2 D'(\lambda_0)}{A''(\lambda_0)},$$

resulting in

$$E(\mathrm{PL}) = \frac{1}{2A''}\left[E(Z_t')^2 + Mt^4 D'^2\right]. \tag{A.5}$$

Now t is selected to minimize (A.5). The dominant term in $E(Z_t')^2$ is

$$4E\left[\frac{\partial}{\partial\lambda}\eta_t(\hat{\beta}, \lambda)\right]^2_{\lambda=\lambda_0}. \tag{A.6}$$

Put $\alpha = \lambda\sqrt{1+t^2}$. Then

$$\frac{\partial}{\partial\lambda}\eta_t(\hat{\beta}, \lambda) = \frac{\sqrt{1+t^2}}{\sqrt{2\pi}t}\left(\exp\left[-\frac{1}{2}\left(\frac{\hat{\beta}-\alpha}{t}\right)^2\right] - \exp\left[-\frac{1}{2}\left(\frac{\hat{\beta}+\alpha}{t}\right)^2\right]\right).$$

For small t, (A.6) is given by

$$\frac{2}{\sqrt{\pi}}\frac{1}{t}f(\lambda_0).$$

Then the minimizing t in (A.5) is approximately $M^{-1/5}$ and $E(\mathrm{PL}) \sim M^{1/5}$.

In subset selection, the rates are different. Equation (A.5) holds and we need to evaluate $E((\partial/\partial\lambda)\eta_t(\hat{\beta}, \lambda))^2$ for small t. Direct integration leads to the expression

$$\frac{\lambda_0^2}{\sqrt{\pi}}\frac{f(\lambda_0)}{t^3} + o(t^{-2}).$$

Minimizing (A.5) leads to $t \sim M^{-1/7}$ and

$$\mathrm{M}E(\mathrm{PL}) \sim M^{3/7}.$$

In simulations [Breiman (1992)] we found that in subset selection using $t \in [0.6, 1.0]$ gave better results than smaller t-values. Now we can begin to understand the reason.

Errors in PE *estimates using little bootstrap.* If the second mixed partial derivative of $\theta(\beta, \lambda)$ exists, then the bias is

$$-EZ'(\lambda_0)^2/A''(\lambda_0).$$

Ignoring the $O(1)$ bias term, the variance of $\widehat{\mathrm{PE}}(\hat{\lambda})$ equals

$$2N + 4\mathrm{M}E\left[\beta^* Z - Z\theta\left(\hat{\beta}, \lambda_0\right) + \theta_1\left(\hat{\beta}, \lambda_0\right)\right]^2.$$

With some integration by parts, the expectation term equals

$$E\left(\beta^* - \theta\left(\hat{\beta}, \lambda_0\right)\right)^2 + E\theta_1^*\left(\hat{\beta}, \lambda_0\right),$$

giving the variance approximation

$$2N + 4\mathrm{M}E(\lambda_0) + 4\mathrm{M}E\theta_1^2\left(\hat{\beta}, \lambda_0\right).$$

Thus, use of the tiny bootstrap adds an $O(M)$ term to the $\widehat{\mathrm{PE}}$ variance as compared to the test set $\widehat{\mathrm{PE}}$.

The situation differs for *nn*-garotte and subset selection. In both of these, the dominant term in $E(\widehat{\mathrm{PE}}(\hat{\lambda}) - \mathrm{PE}(\hat{\lambda}))$ is $\sqrt{M}EY_t(\lambda_0) \simeq Mt^2D(\lambda_0)$. The resulting bias in *nn*-garotte is approximately $M^{3/5}$ and in subset selection approximately $M^{5/7}$. Besides an additional $O(M)$ term in the variance of $\widehat{\mathrm{PE}}(\hat{\lambda})$, more computations show another additional $O(M^{4/5})$ term in *nn*-garotte and an $O(M^{8/7})$ term in subset selection.

A2. Proof of Theorem 3.1. Using the identity

$$E\left(\varepsilon_k | \{\varepsilon_n^* + \varepsilon_n, n = 1, \dots, N\}\right) = \frac{t^2}{1 + t^2}(\varepsilon_k^* + \varepsilon_k)$$

gives

$$EB_t(s) = \frac{1}{1 + t^2} E(\varepsilon^* + \varepsilon, \hat{\mu}(\cdot, \mu^* + \varepsilon^* + \varepsilon, s)).$$

Let $\delta^* = \varepsilon^* + \varepsilon$. Now $\hat{\mu}(\cdot, \mu^* + \delta^*, s)$ is the minimizer in $\mathscr{U}_s$ of

$$\|\mu^* + \delta^* - \mu\|^2 = (1 + t^2)\left\|\frac{\mu^* + \delta^*}{\sqrt{1 + t^2}} - \frac{\mu}{\sqrt{1 + t^2}}\right\|^2.$$

Since $\delta^*/\sqrt{1 + t^2}$ has the same distribution as ε^*, denote it so. Now $\mu \in \mathscr{U}_s \Leftrightarrow \mu/\sqrt{1 + t^2} \in \mathscr{U}_{s_t}$. Thus, $\mu\sqrt{1 + t^2}$ is the minimizer in $\mathscr{U}_{s_t}$ of

$$\left\|\frac{\mu^*}{\sqrt{1 + t^2}} + \varepsilon^* - \mu\right\|^2,$$

so

$$\hat{\mu}(\cdot, \mu^* + \delta^*, s) = \sqrt{1 + t^2}\, \hat{\mu}\left(\cdot, \frac{\mu^*}{\sqrt{1 + t^2}} + \varepsilon^*, s_t\right).$$

Putting things together,

$$EB_t(s) = E\left(\varepsilon^*, \hat\mu\left(\cdot, \frac{\mu^*}{\sqrt{1+t^2}}, s_t\right)\right),$$

which is equivalent to the statement of the theorem.

A3. Tiny bootstrap formula for garotte. The garotte coefficients are determined by minimizing $\|y - \sum_m c_m \hat\beta_m x_m\|^2$, where the $\{\hat\beta_m\}$ are the full-model OLS coefficients and the $\{c_m\}$ are restricted by $\sum c_m^2 \le s^2$.

Take $\{\varepsilon_n\}$ to be iid $N(0, \sigma^2)$, and put $y_n' = y_n + t\varepsilon_n$. Denote $S = X^tX$. The new OLS $\hat{\boldsymbol\beta}(t) = S^{-1}X'\mathbf{y}'$. Put $\mathbf{Z} = (\varepsilon, x)$, so $\hat{\boldsymbol\beta}(t) = \hat{\boldsymbol\beta} + tS^{-1}\mathbf{Z}$. The altered $\{c_m(t)\}$ minimize

$$\left\| y + t\varepsilon - \sum_m c_m(t)\hat\beta_m(t)\mathbf{x}_m \right\|^2$$

under $\sum c_m^2(t) \le s^2$.

The little bootstrap equals

$$\frac{1}{t}E\sum_m Z_m c_m(t)\hat\beta_m(t),$$

so

$$\text{(A.7)} \qquad \text{TB}(s) = E\sum_m Z_m \dot{\hat\beta}_m(0)c_m(0) + E\sum_m Z_m \hat\beta_m(0)\dot c_m(0),$$

where $\cdot$ above is d/dt. Note that $\dot{\hat{\boldsymbol\beta}}(0) = S^{-1}\mathbf{Z}$, so the first term in (A.7) is $\sigma^2\sum_m c_m$. Let $W_{mk} = \hat\beta_m(t)S_{mk}\hat\beta_k(t)$ and use $u \otimes \mathbf{v}$ to denote the vector with components $u_m v_m$. Then the Lagrangian equation for determining $\mathbf{c}(t)$ is

$$W\mathbf{c} + \lambda\mathbf{c} = \hat\beta(t) \otimes (Xy + t\mathbf{Z}).$$

Differentiating gives

$$(W + \lambda I)\dot{\mathbf{c}} + (\dot W + \dot\lambda I)\mathbf{c} = \dot{\hat{\boldsymbol\beta}} \otimes X\mathbf{y} + \mathbf{Z} \otimes \hat{\boldsymbol\beta}.$$

After numerical experiments, we concluded that the $\dot\lambda$ term was negligible. Thus, putting $t = 0$ and letting $W_\lambda = W + \lambda I$,

$$\dot{\mathbf{c}} = W_\lambda^{-1}\left[-\dot W\mathbf{c} + (Xy) \otimes \dot{\hat{\boldsymbol\beta}} + \mathbf{Z} \otimes \hat{\boldsymbol\beta}\right].$$

Using $EZ_m \dot{\hat\beta}_k(0) = \sigma^2\delta_{mk}$ and $EZ_mZ_k = \sigma^2S_{mk}$ gives

$$\begin{aligned} &E\left(\sum_m Z_m \hat\beta_m \dot c_m(0)\right) \\ &\qquad = \sigma^2\left(\sum_{m,k} W_{mn}^{-1}(\lambda)W_{mk}(1 - c_m) + \sum_{m,k} W_{mm}^{-1}(\lambda)W_{mk}(1 - c_k)\right). \end{aligned}$$

Finally, use $\sum_k W_{mk}^{-1}(\lambda)W_{mk} = 1 - \lambda W_{mm}^{-1}(\lambda)$ to get the result stated in Section 3.

REFERENCES

BREIMAN, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: x-fixed prediction error. *J. Amer. Statist. Assoc.* **87** 738–754.

BREIMAN, L. (1995). Better subset selection using the non-negative garotte. *Technometrics* **37** 373–384.

BREIMAN, L. (1996a). Stacked regressions. *Machine Learning* **24** 41–64.

BREIMAN, L. (1996b). Bagging predictors. *Machine Learning* **26** 123–140.

BREIMAN, L. (1996c). Bias, variance and arcing classifiers. Report 460, Dept. Statistics, Univ. California.

BREIMAN, L. and SPECTOR, P. (1992). Submodel selection and evaluation in regression. The random X case. *Internat. Statist. Rev.* **60** 291–319.

WOLPERT, D. (1992). Stacked generalization. *Neural Networks* **5** 241–259.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
367 EVANS HALL 3860
BERKELEY, CALIFORNIA 94720-3860