Better Subset Regression Using the Nonnegative Garrote

Leo Breiman

Technometrics, Vol. 37, No. 4 (Nov., 1995), 373-384.

Stable URL:
http://links.jstor.org/sici?sici=0040-1706%28199511%2937%3A4%3C373%3ABSRUTN%3E2.0.CO%3B2-3

Technometrics is currently published by American Statistical Association.

TECHNOMETRICS, NOVEMBER 1995, VOL. 37, NO. 4

Better Subset Regression Using the Nonnegative Garrote

Leo BREIMAN

Statistics Department
University of California, Berkeley
Berkeley, CA 94720

A new method, called the nonnegative (nn) garrote, is proposed for doing subset regression. It both shrinks and zeroes coefficients. In tests on real and simulated data, it produces lower prediction error than ordinary subset selection. It is also compared to ridge regression. If the regression equations generated by a procedure do not change drastically with small changes in the data, the procedure is called stable. Subset selection is unstable, ridge is very stable, and the nn-garrote is intermediate. Simulation results illustrate the effects of instability on prediction error.

KEY WORDS: Little bootstrap; Model error; Prediction; Stability.

1. INTRODUCTION

One of the most frequently used statistical procedures is subset-selection regression. That is, given data of the form $\{(y_n, x_{1n}, \ldots, x_{Mn}), n = 1, \ldots, N\}$, some of the predictor variables $x_1, \ldots, x_M$ are eliminated and the prediction equation for y is based on the remaining set of variables. The selection of the included variables uses either the best subset method or a forward/backward stepwise method. These procedures give a sequence of subsets of $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ of dimension $1, 2, \ldots, M$. Then some other method is used to decide which of the M subsets to use.

Subset selection is useful for two reasons, variance reduction and simplicity. It is well known that each additional coefficient estimated adds to the variance of the regression equation. The fewer coefficients estimated, the lower the variance. Unfortunately, using too few variables leads to increased bias. But, if a regression equation based on 40 variables, say, can be reduced (without loss of accuracy) to one based on 5 variables, then not only is the equation simpler but we may also have learned something about which variables are important in predicting y.

Using prediction accuracy as our "gold standard," the hope is that subset regression will produce a regression equation simpler and more accurate than the equation based on all variables. If M is large, it usually succeeds. But, if M is moderate to small—that is, $M \lesssim 10$—then there is evidence that often the full regression equation is more accurate than the selected subset regression. Roecker (1991) did recent work on this issue and gave references to relevant past work.

The prime competitor to subset regression in terms of variance reduction is ridge regression. Here the coefficients are estimated by $(X^tX + \lambda I)^{-1}X^tY$, where λ is a shrinkage parameter. Increasing λ shrinks the coefficient estimates, but none are set equal to zero. Gruber (1990) gave a recent overview of ridge methods. Some studies (i.e., Frank and Friedman 1993; Hoerl, Schuenemeyer, and Hoerl 1986) have shown that ridge regressions give more accurate predictions than subset regressions unless, assuming that y is of the form

$$y = \sum_k \beta_k x_k + \epsilon,$$

all but a few of the $\{\beta_k\}$ are nearly zero and the rest are large. Thus, although subset regression can improve accuracy if M is large, it is usually not as accurate as ridge.

Ridge has its own drawbacks. It gives a regression equation no simpler than the original ordinary least squares (OLS) equation. Furthermore, it is not scale invariant. If the scales used to express the individual predictor variables are changed, then the ridge coefficients do not change inversely proportional to the changes in the variable scales. The usual recipe is to standardize the $\{x_m\}$ to mean 0, variance 1 and then apply ridge. But the recipe is arbitrary; that is, interquartile ranges could be used to normalize instead, giving a different regression predictor. For a spirited discussion of this issue, see Smith and Cambell (1980).

Another aspect of subset regression is its instability with respect to small perturbations in the data. Say that $N = 100$, $M = 40$, and that using stepwise deletion of variables a sequence of subsets of variables $\{x_m; m \in \zeta_k\}$, of dimension k ($|\zeta_k| = k$), $k = 1, \ldots, M$, has been selected. Now remove a single data case $(y_n, \mathbf{x}_n)$, and use the same selection procedure, getting a sequence of subsets $\{x_m; m \in \zeta_k'\}$. Usually the $\{\zeta_k'\}$ and $\{\zeta_k\}$ are different so that for some k a slight data perturbation leads to a drastic change in the prediction equation. On the other hand, if one uses ridge estimates and deletes a single data case, the new ridge estimates, for the same λ, will be close to the old.

Much work and research have gone into subset-selection regression, but the basic method remains flawed

by its relative lack of accuracy and instability. Subset regression either zeroes a coefficient, if it is not in the selected subsets, or inflates it. Ridge regression gains its accuracy by selective shrinking. Methods that select subsets, are stable, and shrink are needed. Here is one: Let $\{\hat{\beta}_k\}$ be the original OLS estimates. Take $\{c_k\}$ to minimize

$$\sum_k \left(y_n - \sum_k c_k \hat{\beta}_k x_{kn} \right)^2$$

under the constraints

$$c_k \geq 0, \qquad \sum_k c_k \leq s.$$

The $\tilde{\beta}_k(s) = c_k \hat{\beta}_k$ are the new predictor coefficients. As the garrote is drawn tighter by decreasing s, more of the $\{c_k\}$ become zero and the remaining nonzero $\tilde{\beta}_k(s)$ are shrunken.

This procedure is called the *nonnegative* (nn) *garrote*. The garrote eliminates some variables, shrinks others, and is relatively stable. It is also scale invariant. I show that it is almost always more accurate than subset selection and that its accuracy is competitive with ridge. In general nn-garrote produces regression equations having more nonzero coefficients than subset regression. But the loss in simplicity is offset by substantial gains in accuracy.

The organization of this article is as follows: Section 2 on model selection gives definitions of prediction and model error together with a brief outline of useful estimates of these errors. These estimates are used to determine the value of the garrote parameter s, the ridge parameter λ, and the dimensionality of the subset regression. In Section 3, nn-garrote is compared to subset regression on two well-known data sets. The first is the stackloss data given by Daniel and Wood (1980). The second is an ozone data set used by Breiman and Friedman (1985). In Section 4, I assume that $X^t X = I$. The action of the nn-garrote becomes clear, and it can be compared to ridge and subset selection over an interesting range of $\{\beta_k\}$ distributions. Section 5 reports on a simulation comparison of methods. Conclusions and concluding remarks are given in Section 6.

In many regression problems the number of predictor variables is a substantial fraction of the sample size, and variable subset selection is used to reduce complexity and variance. The large ratio of variables to sample size often reflects the experimenters inclusion of nonlinear terms in search of a better fit. For instance, the stackloss data has three x variables and 17 cases (after removal of four outliers). To get a better fit, Daniel and Wood (1980) introduced quadratic and interaction terms, going from three to nine variables. Then subset selection was used to arrive at a three-variable model. The ozone data set has 330 cases and eight variables but is known to have strong nonlinearities. The analysis in Section 3 includes quadratic and interaction terms for a total of 44 variables.

Useful analytical results are not available when the number of variables is comparable to the sample size. In this area empirical results, good heuristics, and simulations are the only general tools available. Properly used, they can give valuable insights. For instance, the concept of stability was nurtured by the simulation results reported in Section 5 and previous work using simulations to study subset selection.

Sorting out how to reduce complexity and prediction error is a complicated problem. There are few relevant studies in the statistical literature. The book by Miller (1990) summarizes work on variable subset selection and gives an extensive bibliography, but it is primarily concerned with low-dimensional issues. The work in this article came mainly out of a combination of the ideas from Breiman (1993), who used nonnegativity and sum constraints in the context of combining regressions, and the previous explorations of subset selection of Breiman (1992) and Breiman and Spector (1992). Tibshirani (1994), stimulated by the results in a preprint of this article, devised another method for shrinking and subset selection.

The constrained least squares minimization used in the nn-garrote can be solved rapidly even for numerous x variables. I used a modification of the elegant nonnegative least squares algorithm given by Lawson and Hanson (1974). No stability problems were encountered and computation times increased only moderately as the number of x variables increased. A FORTRAN subroutine that outputs the values of the $\{c_k\}$ for any value of s, $0 < s < M$, is available by ftp to stat-ftp.berkeley.edu in the directory /pub/user/breiman.

2. MODEL SELECTION

2.1 Prediction and Model Error

The prediction error is defined as the average error in predicting y from $\mathbf{x}$ for future cases not used in the construction of the prediction equation. The data on hand are of the form $\{(y_n, \mathbf{x}_n), n = 1, \ldots, N\}$, where $\mathbf{x}_n = (x_{1n}, \ldots, x_{Mn})$ and the symbols $y, x_1, \ldots, x_M$ are used as generic notation for the response and M predictor variables.

There are two regression situations, *X-controlled* and *X-random*. In the controlled situation, the $\{\mathbf{x}_n\}$ are selected by the experimenter and only y is random. In the X-random situation, both y and $\mathbf{x}$ are randomly selected. Different definitions of prediction error are appropriate.

In the controlled situation, future data are assumed gathered using the same $\{\mathbf{x}_n\}$ as in the present data and thus have the form $\{(y_n^{\text{new}}, \mathbf{x}_n), n = 1, \ldots, N\}$. If $\hat{\mu}(\mathbf{x})$ is the prediction equation derived from the present data, then define the prediction error as

$$\text{PE}(\hat{\mu}) = E \sum_n \left(y_n^{\text{new}} - \hat{\mu}(\mathbf{x}_n) \right)^2,$$

where the expectation is over $\{y_n^{\text{new}}\}$.

If the data are generated by the mechanism $y_n = \mu(\mathbf{x}_n) + \epsilon_n$, where the $\{\epsilon_n\}$ are mean-zero uncorrelated with average variance σ^2, then

$$\mathrm{PE}(\hat{\mu}) = N\sigma^2 + \sum_n (\mu(\mathbf{x}_n) - \hat{\mu}(\mathbf{x}_n))^2.$$

The first component is the inherent prediction error due to the noise. The second component is the prediction error due to lack of fit to the underlying model. This component is called *model error* and denoted by $\mathrm{ME}(\hat{\mu})$. The size of the model error reflects different methods of model estimation. If $\mu = \Sigma\beta_m x_m$ and $\hat{\mu} = \Sigma\hat{\beta}_m x_m$, then

$$\mathrm{ME}(\hat{\mu}) = (\hat{\beta} - \beta)^t (X^t X)(\hat{\beta} - \beta).$$

If the $\{\mathbf{x}_n\}$ are random, then it is assumed that the $(y_n, \mathbf{x}_n)$ are iid selections from the parent distribution $(Y, \mathbf{X})$. Then if $\hat{\mu}(\mathbf{x})$ is the prediction equation constructed using the present data, $\mathrm{PE}(\hat{\mu}) = E(Y - \hat{\mu}(\mathbf{X}))^2$. Assuming that $Y = \mu(\mathbf{X}) + \epsilon$, where $E(\epsilon \mid \mathbf{X}) = 0$, then $\mathrm{PE}(\hat{\mu}) = E\epsilon^2 + E(\mu(\mathbf{X}) - \hat{\mu}(\mathbf{X}))^2$. Again, the relevant error is the second component. To put model error in this situation on the same scale as in the X-controlled case, define $\mathrm{ME}(\hat{\mu}) = N \cdot E(\mu(\mathbf{X}) - \hat{\mu}(\mathbf{X}))^2$, and similarly for $\mathrm{PE}(\hat{\mu})$. If $\mu = \Sigma\beta_m x_m$ and $\hat{\mu} = \Sigma\hat{\beta}_m x_m$, then $\mathrm{ME}(\hat{\mu}) = (\hat{\beta} - \beta)^t (N \cdot \Gamma)(\hat{\beta} - \beta)$, when $\Gamma_{ij} = EX_i X_j$.

2.2 Estimating Error

Each regression procedure that we study produces a sequence of models $\{\hat{\mu}_k(\mathbf{x})\}$. Variable selection gives a sequence of subsets of variables $\{x_m, m \in \zeta_k\}$, $|\zeta_k| = k, k = 1, \ldots, M$, and $\hat{\mu}_k(\mathbf{x})$ is the OLS linear regression based on $\{x_m, m \in \zeta_k\}$. In nn-garrote, a sequence of s-parameter values $s_1, \ldots, s_K$ is selected and $\hat{\mu}_k(\mathbf{x}), k = 1, \ldots, K$, is the prediction equation using parameter s_k. In ridge, a sequence of λ-parameter values $\lambda_1, \ldots, \lambda_K$ is selected, and $\hat{\mu}_k(\mathbf{x})$ is the ridge regression based on λ_k.

If we knew the true value of $\mathrm{PE}(\hat{\mu}_k)$, the model selected would be the minimizer of $\mathrm{PE}(\hat{\mu}_k)$. We refer to these selections as the *crystal-ball* models. Otherwise, the selection process constructs an estimate $\widehat{\mathrm{PE}}(\hat{\mu}_k)$ and selects that $\hat{\mu}_k$ that minimizes $\widehat{\mathrm{PE}}$. The estimation methods differ for X-controlled and X-random.

2.2.1 X-Controlled Estimates. The most widely used estimate in subset selection is Mallows C_p. If k is the number of variables in the subset, $\mathrm{RSS}(k)$ is the residual sum of squares using $\hat{\mu}_k$, and $\hat{\sigma}^2$ is the noise variance estimate derived from the full model (all variables), then the C_p estimate is $\widehat{\mathrm{PE}}(\hat{\mu}_k) = \mathrm{RSS}(k) + 2k\hat{\sigma}^2$. But Breiman (1992) showed that this estimate is heavily biased and does poorly in model selection.

It was shown in the same article that a better estimate for $\mathrm{PE}(\hat{\mu}_k)$ is

$$\mathrm{RSS}(k) + 2B_t(k),$$

where $B_t(k)$ is defined as follows:

Let σ^2 be the noise variance, and add iid $N(0, t^2\sigma^2)$, $0 < t \leq 1$, noise $\{\tilde{\epsilon}_n\}$ to the $\{y_n\}$, getting $\{\tilde{y}_n\}$. Using the data $\{(\tilde{y}_n, \mathbf{x}_n)\}$, repeat the subset-section process getting a new sequence of OLS predictors $\{\tilde{\mu}_k, k = 1, \ldots, M\}$. Then

$$B_t(k) = \frac{1}{t^2} E\left(\sum_n \tilde{\epsilon}_n \tilde{\mu}_k(\mathbf{x}_n)\right),$$

where the expectation is on the $\{\tilde{\epsilon}_n\}$ only.

This is made computable by replacing σ^2 by the noise variance estimate $\hat{\sigma}^2$ and the expectation over the $\{\tilde{\epsilon}_n\}$ by the average over many repetitions. This procedure is called the little bootstrap.

Little bootstrap can also be applied to nn-garrote and ridge. Suppose that the nn-garrote predictor $\hat{\mu}_k$ has been computed for parameter values s_k with resulting residual sum of squares $\mathrm{RSS}(s_k), k = 1, \ldots, K$. Now add $\{\tilde{\epsilon}_n\}$ to the $\{y_n\}$, getting $\{\tilde{y}_n\}$, where the $\{\tilde{\epsilon}_n\}$ are iid $N(0, t^2\hat{\sigma}^2)$. Using the $(\tilde{y}_n, \mathbf{x}_n)$ data, derive the nn-garrote predictor $\tilde{\mu}_k$ for the parameter value s_k, and compute the quantity $\frac{1}{t^2}\sum \tilde{\epsilon}_n \tilde{\mu}_k(\mathbf{x}_n)$. Repeat several times, average these quantities, and denote the result by $B_t(s_k)$. The PE estimate is $\widehat{\mathrm{PE}}(\hat{\mu}_k) = \mathrm{RSS}(s_k) + 2B_t(s_k)$.

In ridge regression, denote by $\hat{\mu}_k$ the predictor using parameter λ_k. The little bootstrap estimate is $\mathrm{RSS}(\lambda_k) + 2B_t(\lambda_k)$, where $B_t(\lambda_k)$ is computed just as in subset selection and nn-garrote. It was shown by Breiman (1992) that, for subset selection, the bias of the little bootstrap estimate is small for t small. The same proof holds, almost word for word, for the nn-garrote and ridge. But what happens in subset selection is that as $t \downarrow 0$, the variance of B_t increases rapidly, and B_t has no sensible limiting value. Experiments by Breiman (1992) indicated that the best range for t is [.6, .8] and that averaging over 25 repetitions to form B_t is usually sufficient.

On the other hand, in ridge regression the variance of B_t does not increase appreciably as $t \downarrow 0$, and taking this limit results in the more unbiased estimate

$$\widehat{\mathrm{PE}}(\hat{\mu}_k) = \mathrm{RSS}(\lambda_k) + 2\hat{\sigma}^2 \mathrm{tr}\left(X^t X (X^t X + \lambda_k I)^{-1}\right). \tag{2.1}$$

This turns out to be an excellent PE estimate that selects regression equations $\hat{\mu}_k$ with $\mathrm{PE}(\hat{\mu}_k)$ close to $\min_{k'} \mathrm{PE}(\hat{\mu}_{k'})$. The estimate (2.1) was proposed on other grounds by Mallows (1973). See also Hastie and Tibshirani (1990).

The situation in nn-garrote is intermediate between subset selection and ridge. The variance of B_t increases as t gets small, but a finite variance limit exists. It does not perform as well as using t in the range [.6, .8], however. Therefore, our preferred PE estimates for subset selection and nn-garrote use $t \in [.6, .8]$ and (2.1) for the ridge PE estimate. The behavior of B_t for t small is a reflection of the stability of the regression procedures used. This was explored further by Breiman (1994).

2.2.2 X-Random Estimates. For subset regressions $\{\hat{\mu}_k\}$ in the X-random situations, the most frequently

encountered PE estimate is

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

The results of Breiman and Spector (1992) show that this estimate can be strongly biased and does poorly in selecting accurate models. What does work is cross-validation.

V-fold CV is used to estimate $\mathrm{PE}(\widehat{\mu}_k)$ for subset selection and nn-garrote. The data $\mathcal{L} = \{(y_n, \mathbf{x}_n), n = 1, \ldots, N\}$ are split into V subsets $\mathcal{L}_1, \ldots, \mathcal{L}_V$. Let $\mathcal{L}^{(v)} = \mathcal{L} - \mathcal{L}_v$. Using subset selection (nn-garrote) and the data in $\mathcal{L}^{(v)}$, form the predictors $\{\mu_k^{(v)}(\mathbf{x})\}$. The CV estimate is

$$\widehat{\mathrm{PE}}(\widehat{\mu}_k) = \sum_v \sum_{(y_n, \mathbf{x}_n) \in \mathcal{L}_v} \left(y_n - \mu_k^{(v)}(\mathbf{x}_n)\right)^2$$

and $\widehat{\mathrm{ME}}(\widehat{\mu}_k) = \widehat{\mathrm{PE}}(\widehat{\mu}_k) - N\widehat{\sigma}^2$. Taking V in the range 5 to 10 gives satisfactory results.

To get an accurate PE estimate for the ridge regression $\widehat{\mu}_\lambda$, remove the nth case $(y_n, \mathbf{x}_n)$ from the data, and recompute $\widehat{\mu}_\lambda(\mathbf{x})$ getting $\widehat{\mu}_\lambda^{(-n)}(\mathbf{x})$. Then the estimate is

$$\widehat{\mathrm{PE}}(\lambda) = \sum_n \left(y_n - \widehat{\mu}_\lambda^{(-n)}(\mathbf{x}_n)\right)^2.$$

This is the leave-one-out CV estimate. If $r_n(\lambda) = y_n - \widehat{\mu}_\lambda(\mathbf{x}_n)$ and $h_n(\lambda) = \mathbf{x}_n'(X^tX + \lambda I)^{-1}\mathbf{x}_n$, then

$$\widehat{\mathrm{PE}}(\lambda) = \sum_n (r_n(\lambda)/1 - h_n(\lambda))^2.$$

Usually, $h_n(\lambda) \simeq \overline{h}(\lambda)$ is a good approximation, where $\overline{h}(\lambda) = \mathrm{tr}(X^tX(X^tX + \lambda I)^{-1})/N$. With this approximation

$$\widehat{\mathrm{PE}}(\lambda) = \mathrm{RSS}(\lambda)/(1 - \overline{h}(\lambda))^2. \tag{2.2}$$

This estimate was first derived by Golub, Heath, and Wahba (1979) and is called the GCV (generalized cross-validation) estimate of PE. Its accuracy is confirmed in the simulation in Section 7.

Breiman and Spector (1992) found that the "infinitesimal" version of CV—that is, leave-one-out—gave poorer results in subset selection than five- or tenfold CV [for theoretical work on this issue, see Shuo (1993) and Zhang (1992)]. But leave-one-out works well in ridge regression. Simulation results show that tenfold CV is slightly better for nn-garrote than leave-one-out. This again reflects the relative stabilities of the three procedures.

3. TWO EXAMPLES

The use of the nn-garrote is illustrated in two well-known data sets. One is X-controlled data, and the other I put into the X-random context.

3.1 The Stackloss Data

These data are the three-variable stackloss data studied in chapter 5 of Daniel and Wood (1980). By including quadratic terms along with the linear, it becomes a nine-variable problem. Eliminating the outliers identified by Daniel and Wood leaves 17 cases.

I compare nn-garrote to subset selection using backward deletion. Daniel and Wood gave two possible fitting equations, stating that there is little to choose between them. Backward deletion and 250 repetitions of little bootstrap pick the second of these equations,

$$\widehat{y} = 14.1 + .71x_1 + .51x_2 + .0254x_1x_2. \tag{3.1}$$

Garrote picks an equation using the same variables,

$$\widehat{y} = 14.1 + .77x_1 + .40x_2 + .0152x_1x_2. \tag{3.2}$$

The estimated model errors are 3.0 and 1.0, respectively (with estimated prediction errors 41.4 and 39.3). The two equations appear similar, but each pair of coefficients differs by almost .5 if the x variables are put on standardized scales.

The value of s selected is $.25M$ ($M = 9$). Because $s = 9$ corresponds to the full OLS regression, this could be interpreted as meaning that the coefficients were shrunk to 25% of the OLS values. The sum of the coefficients in the garrote equation (3.2) is a bit smaller than those in (3.1), but the major effect is the redistribution of emphasis on the three variables included.

3.2 Ozone Data

The ozone data were also used by Friedman and Silverman (1989), Hastie and Tibshirani (1990), and Cook (1993). It consists of daily ozone and meteorological data for the Los Angeles Basin in 1976. There are 330 cases with no missing data. The dependent variable is ozone. There are eight meterological predictor variables:

x_1: 500 mb height
x_2: wind speed
x_3: humidity
x_4: surface temperature
x_5: inversion height
x_6: pressure gradient
x_7: inversion temperature
x_8: visibility

These data are known to be nonlinear in some of the variables, so, after subtracting means, interactions and quadratic terms were added, giving 44 variables. Subset selection was done using backward deletion of variables. To get the estimates of the best subset size, garrote parameter, and prediction errors, tenfold CV was used. The tenfold CV was repeated five times using different random divisions of the data and the results averaged.

Subset selection chooses the five-variable equation

$$\widehat{y} = 6.2 + 4.6x_6 + 2.4x_2x_4 - 1.3x_2x_5 + 5.5x_4^2 - 4.2x_6^2, \tag{3.3}$$

whereas nn-garrote chooses the seven-variable equation

$$\widehat{y} = 6.2 + 3.9x_1 - 1.7x_5 - .3x_2^2 + .6x_2x_4 + 5.2x_4^2 + .8x_5x_7 - .4x_6^2. \tag{3.4}$$

(All variables, including interactions and quadratic terms are standardized to mean 0, variance 1.)

The estimated mean prediction error for the subset equation (3.3) is 10.0, with mean model error 3.3. The nn-garrote equation (3.4) has an estimated mean prediction error of 9.0 with mean model error of 2.3. Each equation has a strong temperature term x_4^2 with about the same coefficient. Otherwise, they are dissimilar, and include different variables. All of the coefficients in the subset-selection equation (3.3) are substantial in size. But due to the shrinking nature of nn-garrote, some of the coefficients in (3.4) are small.

The value .26 is selected for the nn-garrote parameter. This is surprisingly small because $s = 44$ corresponds to full OLS regression. Thus the coefficients have been shrunk to less than 1% of their OLS value.

Equation (3.4) includes some quadratic terms without the corresponding linear terms. One referee objected to the nonhierarchical form of this model, and an associate editor asked me to comment, noting that many statisticians prefer hierarchical regression models. My model-selection approach is based on minimizing prediction error, and if a nonhierarchical model has smaller prediction error, so be it. I agree, however, that in some situations hierarchical models make more physical sense.

4. X ORTHONORMAL

In the X-controlled case, assume that $X'X = I$ and that y is generated as

$$y_n = \sum_n \beta_m x_{mn} + \epsilon_n,$$

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

Then OLS $\widehat{\beta}_m = \beta_m + Z_m$, where the Z_m are iid $N(0, 1)$. Although this is a highly simplified situation, it can give interesting insights into the comparative behavior of subset selection, ridge regression, and the nn-garrote regressions. The best subset of k variables consists of those x_m corresponding to the k largest $|\widehat{\beta}_m|$ so that the coefficients of a best subset regression are

$$\widehat{\beta}_m^{(S)} = I(|\widehat{\beta}_m| \geq \lambda)\widehat{\beta}_m, \qquad m = 1, \ldots, M, \tag{4.1}$$

for some $\lambda \geq 0$, where $I(\cdot)$ is the indicator function.

In nn-garrote, the expression

$$\sum \left(y_n - \sum_m c_m \widehat{\beta}_m x_{mn} \right)^2$$

is minimized under the constraints $c_m \geq 0$, all m, $\sum_m c_m = s$. The solution is of the form

$$c_m = \left(1 - \frac{\lambda^2}{\widehat{\beta}_m^2}\right)^+,$$

where λ is determined from s by the condition $\sum c_m = s$ and the superscript $+$ indicates the positive part of the expression. The nn-garrote coefficients are

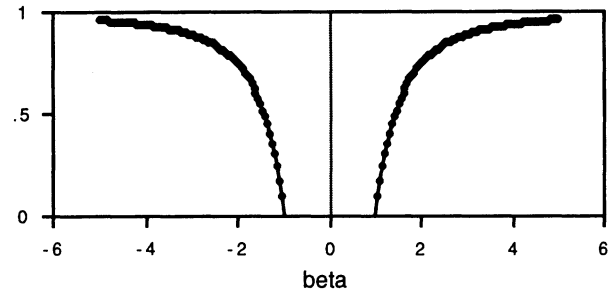

Figure 1. Shrinkage Factor for nn-Garrote.

$$\widehat{\beta}_m^{(G)} = \left(1 - \frac{\lambda^2}{\widehat{\beta}_m^2}\right)^+ \widehat{\beta}_m. \tag{4.2}$$

The ridge coefficients are

$$\widehat{\beta}_m^{(R)} = \frac{1}{1+\lambda}\widehat{\beta}_m. \tag{4.3}$$

All three of these estimates are of the form $\widehat{\beta}' = \theta(\widehat{\beta}, \lambda)\widehat{\beta}$, where θ is a shrinkage factor. OLS estimates correspond to $\theta \equiv 1$. Ridge regression gives a constant shrinkage, $\theta = 1/(1+\lambda)$. Subset selection is 0 for $|\widehat{\beta}| \leq \lambda$ and 1 otherwise. The nn-garrote shrinkage is continuous, 0 if $|\widehat{\beta}| \leq \lambda$ and then increasing to 1. The nn-garrote shrinkage factor is graphed in Figure 1 for $\lambda = 1$.

If the $\{\widehat{\beta}_m'\}$ are any estimates of the $\{\beta_m\}$, then the model error is

$$\mathrm{ME}(\widehat{\beta}') = \sum_m (\beta_m - \widehat{\beta}_m')^2.$$

For estimates of the form $\theta\widehat{\beta}$,

$$\mathrm{ME}(\lambda) = \sum_m (\beta_m - \theta(\widehat{\beta}_m, \lambda)\widehat{\beta}_m)^2.$$

I denote the minimum loss by $\mathrm{ME}^* = \min_\lambda \mathrm{ME}(\lambda)$.

Assume that M is large and that the $\{\beta_m\}$ are iid selections from a distribution $P(d\beta)$. Then

$$\begin{aligned}\mathrm{ME}(\lambda) &= \sum_m (\beta_m - \theta(\beta_m + Z_m, \lambda)(\beta_m + Z_m))^2 \\ &\simeq M \cdot E(\beta - \theta(\beta + Z, \lambda)(\beta + Z))^2,\end{aligned}$$

giving the approximation

$$\mathrm{ME}^* \simeq M \min_\lambda E[\beta - \theta(\beta + Z, \lambda)(\beta + Z)]^2, \tag{4.4}$$

where Z has an $N(0, 1)$ distribution independent of β. To simplify notation, put $\mathrm{ME}^* = \mathrm{ME}^*/M$. For the ridge shrink,

$$\begin{aligned}\mathrm{ME}^* &= \min_\lambda E\left[\beta - \frac{\beta + Z}{1+\lambda}\right]^2 \\ &= \frac{E\beta^2}{E\beta^2 + 1}.\end{aligned}$$

The other minimizations are not analytically tractable but are easy to compute numerically.

I wanted to look at the ME* values for a "revealing" family of distributions of β. It is known that ridge is "optimal" if β has an $N(0, \sigma^2)$ distribution. Subset selection is best if many of the coefficients are 0 and the rest large. This led to use of the family $P(d\beta) = p\delta(d\beta) + qQ(d\beta, \sigma)$, where $\delta(d\beta)$ is a mass concentrated at 0 and $Q(d\beta, \sigma)$ is $N(0, \sigma^2)$. The range of p is [0, 1], and $\sigma \in [0, 5]$.

Figure 2 plots ME* versus σ for $p = 0, .3, .6, .9$ for subset selection, ridge, and nn-garrote. The scaling is such that the OLS ME* is 1. Note that the ME* for nn-garrote is always lower than the subset selection ME* and is usually lower than the ridge ME* except at $p = 0$.

Another question is how many variables are included in the regressions by subset selection compared to nn-garrote. If λ_S and λ_G are the values of λ that minimize the respective model errors, then the proportions P_S and P_G of β's zeroed are

$$P_S = P(|\beta + Z| \leq \lambda_S)$$
$$P_G = P(|\beta + Z| \leq \lambda_G).$$

Figure 3 gives plots of P_S, P_G versus p for $\sigma = 1.0, 1.5, 3.0$.

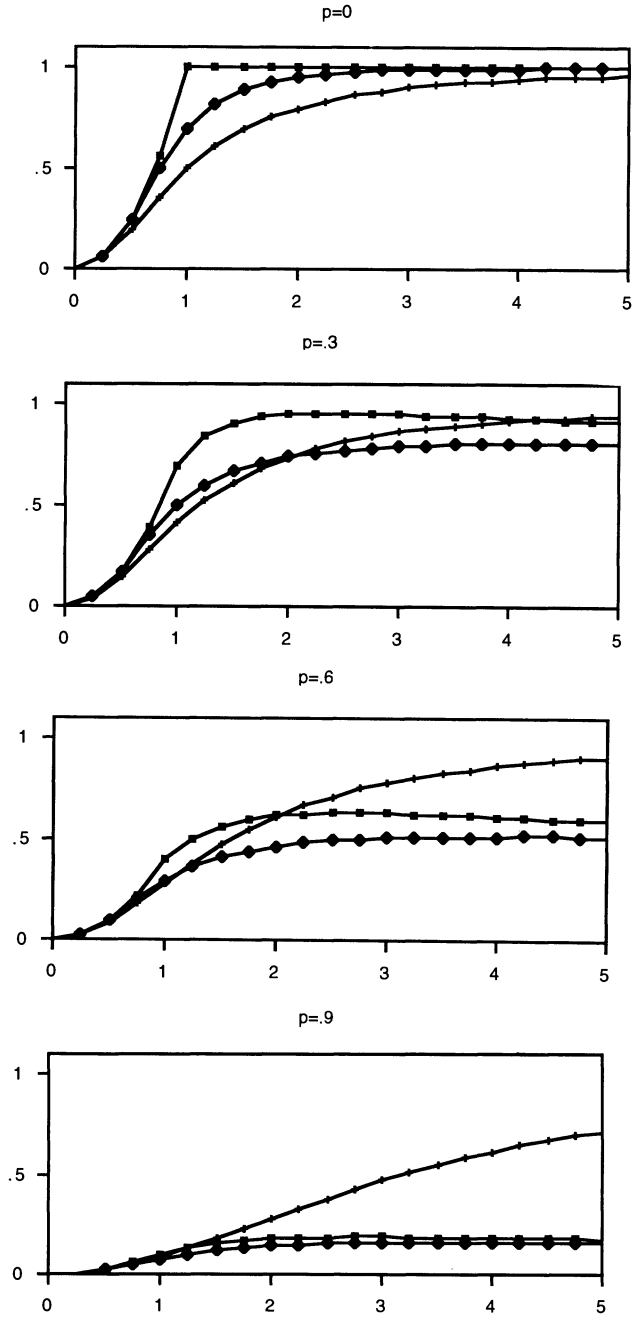

Figure 2. ME Versus Sigma: —♦—, Garrote; —■—, Subset; —|—, Ridge.*

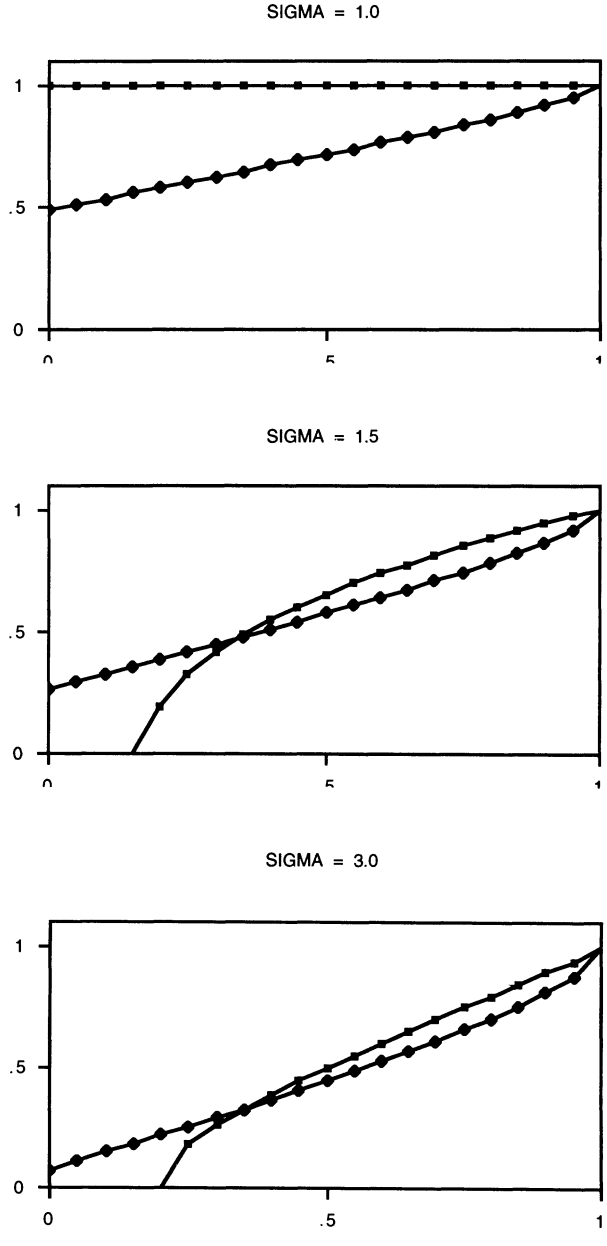

Figure 3. Proportion Zeroed by Procedure Versus Proportion Zero in Distribution: —♦—, Garrote; —■—, Subset.

In regard to simplicity—that is, how many variables are included in the regression—Figure 3 shows that nn-garrote is comparable to subset selection. Subset selection has a discontinuity at $\sigma = 1$. For $\sigma \leq 1$, it deletes all variables and $P_S \equiv 1$. For $\sigma > 1$, it settles down to the behavior shown in the $\sigma = 1.5$ and 3.0 graphs.

5. SIMULATION RESULTS

Because analytic results are difficult to come by in this area, the major proving ground is testing on simulated data.

5.1 Simulation Structure

I did two simulations, one, with 20 variables and 40 cases in the X-controlled case and the other with 40 variables and 80 cases in the X-random case. The major purpose was to compare the accuracies of subset selection, nn-garrote, and ridge regression. The secondary purpose was to learn as much as possible about other interesting characteristics.

The data were generated by

$$\mathbf{y} = \sum \beta_m \mathbf{x}_m + \epsilon$$

with $\{\epsilon\}$ iid $N(0, 1)$. The ticklish specifications are the coefficients $\{\beta_m\}$ and the X design. What is clear is that the results are sensitive to the proportion of the $\{\beta_m\}$ that are 0. To create a "level playing field," five different sets of coefficients were used in each simulation. At one extreme, almost all of the coefficients are 0. At the other, most are nonzero.

A coefficient cluster centered at j of radius rc is defined as

$$\begin{aligned} \beta(j+i) &= (rc - |i|)^2, && |i| \leq rc \\ &= 0, && \text{otherwise.} \end{aligned}$$

Each cluster has $2rc - 1$ nonzero coefficients. In the X-controlled case with 20 variables, the coefficients were in two clusters centered at 5 and 15, in the X-random case, in three clusters centered at 10, 20, and 30. The values of the coefficients were renormalized so that in the X-controlled case, the average R^2 was around .85, in the X-random, about .75.

Each simulation consisted of five runs with 250 iterations in each run. Each of the five runs used a different cluster radius with $rc = 1, 2, 3, 4, 5$. This gave the results shown in Table 1. The X distribution was generated by sampling from $N(0, \Gamma)$, where $\Gamma_{ij} = \rho^{|i-j|}$. In each iteration ρ was selected at random from $[-1, 1]$.

In each X-controlled iteration, subset selection was done using backward variable deletion. nn-garrote used s values $1, 2, \ldots, 20$, and ridge regression searched over λ values such that $\text{tr}(X^tX(X^tX + \lambda I)^{-1}) = 1, 2, \ldots, 20$. The ME values for subset selection and nn-garrote were estimated using the average of 25 repetitions of little bootstrap with $t = .6$. The ME values for ridge were estimated using (2.1). The true ME for each predictor was computed as $(\hat{\beta}' - \beta)X^tX(\hat{\beta}' - \beta)$.

Table 1. Results of Simulation Consisting of Five Runs, 250 Iterations Each

Cluster radius	*X-controlled #nonzero coeff.*	*X-random #nonzero coeff.*
1	2	3
2	6	9
3	10	15
4	14	21
5	18	27

The X-random runs had a similar structure, using backward deletion, s values $= 1, \ldots, 40$, and λ values such that $\text{tr}(X^tX(X^tX + \lambda I)^{-1}) = 1, \ldots, 40$. The ME values for subset selection and nn-garrote were estimated using tenfold CV. The ME values for ridge regression were estimated using GCV (2.2). The true ME was computed as $N(\hat{\beta}' - \beta)^t\Gamma(\hat{\beta}' - \beta)$.

5.2 Simulation Results

In each run, various summary statistics for the 250 iterations were computed.

5.2.1 Accuracy. The most important results were the average true model errors for the predictors selected by the various methods. Figure 4(a) plots these values versus the cluster radius for the X-controlled simulation. Figure 5(a) plots the average true ME values versus the cluster radius for the X-random case (the nn-garrote estimate is chosen using tenfold CV). The two graphs give the same message: nn-garrote is always more accurate than variable selection. If there are many nonzero coefficients, ridge is more accurate. If there are only a few, nn-garrote wins.

An important issue is how much of the differences between the ME values for subset selection, garrote, and ridge [plotted in Figs. 4(a) and 5(a)] can be attributed to random fluctuation. In the simulation, standard errors are estimated for these differences at each cluster radius. Table 2 gives these estimates averaged over the five cluster radii.

5.2.2 Using a Crystal Ball. I have been comparing the estimated best of M subset regressions to the estimated best of M nn-garrote and ridge regressions. That is, PE estimates are constructed and the prediction equation having minimum estimated PE selected. A natural question is what would happen if we had a crystal ball—that is, if we selected the best predictor based on "inside" knowledge of the true ME? For instance, what is the minimum ME among all subset regression predictors? Among all nn-garrote predictors? Among ridge predictors?

This separates the issue of how good a predictor there is among the set of M candidates from the issue of how well we can recognize the best. Figure 4(b) gives a plot of the minimum true ME's for the subset selection, nn-garrote, and ridge predictors versus cluster radius for the

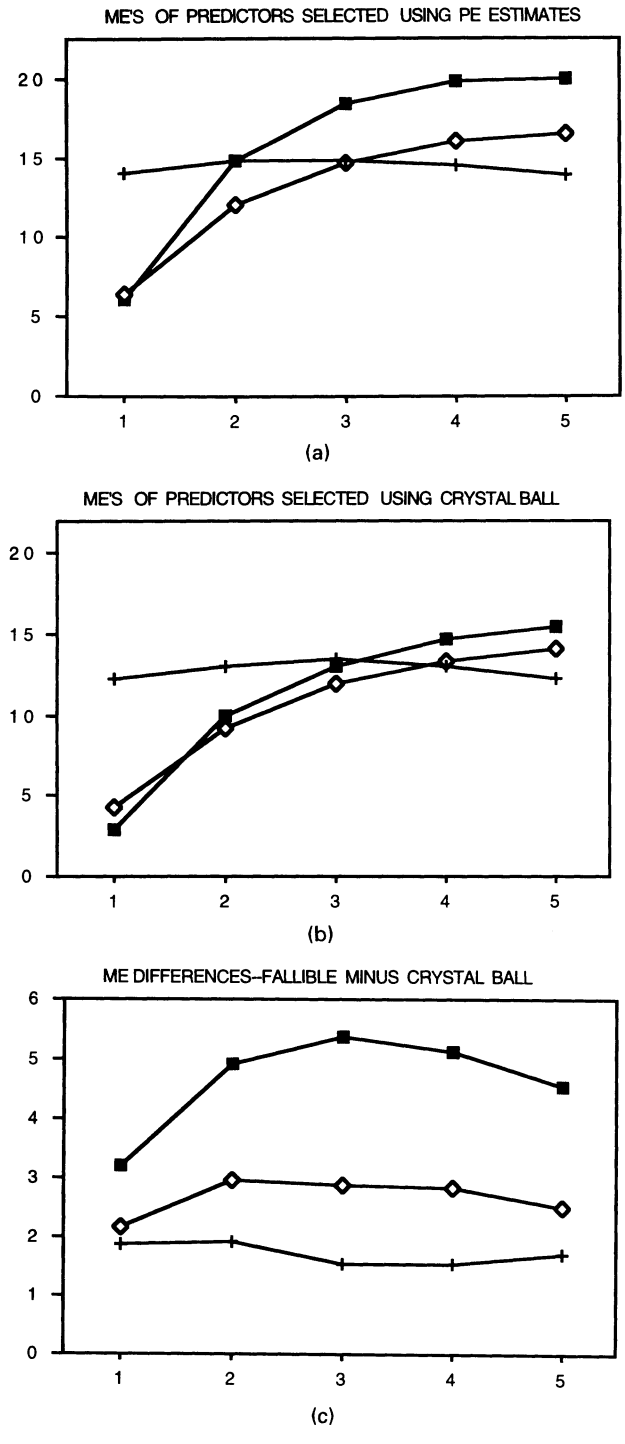

Figure 4. ME in X-Controlled Simulation Versus Cluster Radius: —■—, Subset; —◇—, Garrote; —|—, Ridge.

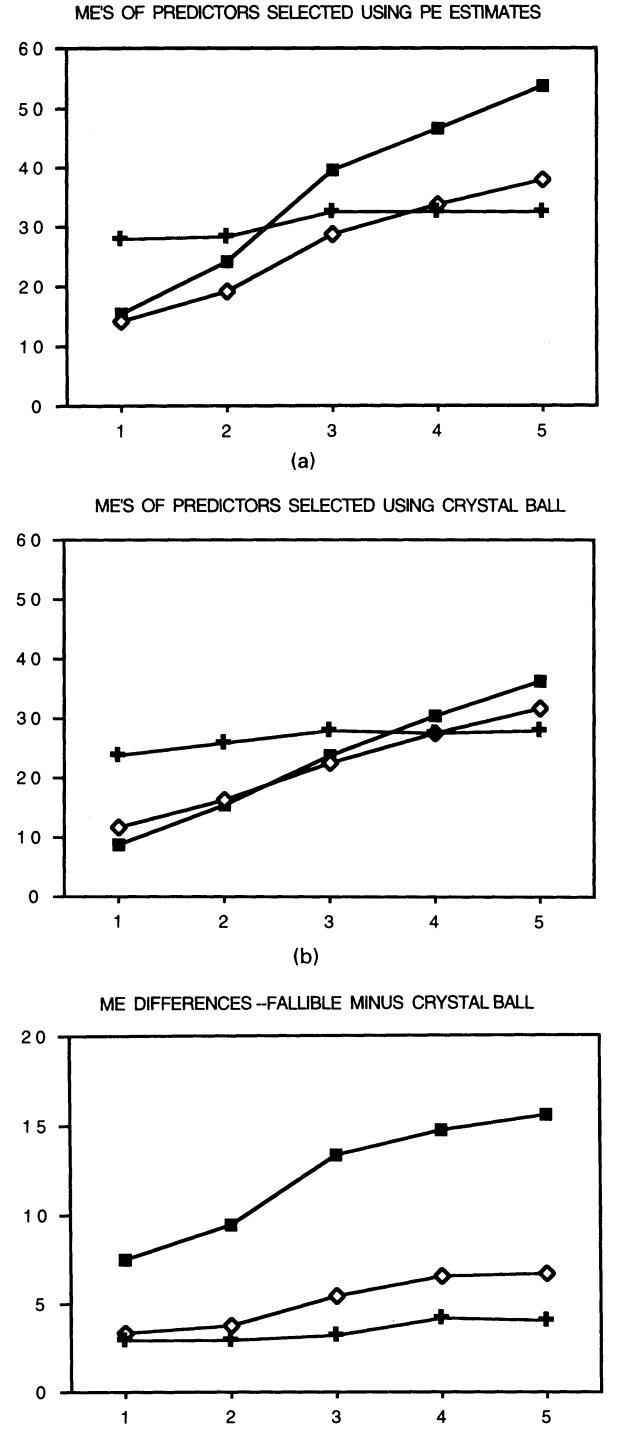

Figure 5. ME in X-Random Simulation Versus Cluster Radius: —■—, Subset; —◇—, Garrote; —+—, Ridge.

X-controlled simulation. Figure 5(b) gives the analogous plot for the X-random simulation. Figures 4(c) and 5(c) show how much larger the fallible knowledge ME is than the crystal-ball ME. Table 3 gives the estimated SE's for the differences of the crystal-ball ME's plotted in Figures 4(b) and 5(b) (averaged over the cluster radii).

The differences between the minimum true ME's for the three methods are smaller than the ME differences using the predictors selected by the ME estimates. The implications are interesting. The crystal-ball subset-selection predictors are close (in ME) to the crystal-ball nn-garrote predictors. The problem is that it is difficult to find the minimum ME subset-selection model. On the other hand, the crystal-ball ridge predictors are not as good as the other two, but the ridge ME estimates do better selection.

Table 2. Estimated SE's for ME Differences

Difference	*X-controlled*	*X-random*
Subset–garrote	.3	.9
Subset–ridge	.4	1.2
Garrote–ridge	.3	.8

Better methods to select low ME subset regressions could make that procedure more competitive in accuracy. But I believe that the intrinsic instability of the method will not allow better selection.

5.2.3 Accuracy of the ME Estimates. The ME estimates are used both to select a prediction equation and to supply an ME estimate for the selected predictor. For the selected predictors I computed the average absolute value of the difference between the estimated and true ME's. The results are graphed in Figure 6(a) versus cluster radius for the X-controlled simulation and in 6(b) for the X-random simulation.

The ME estimates for subset selection are considerably worse than those for nn-garrote or ridge regression. Part of the lack of accuracy is downward bias, given in Table 4 as averaged over all values of cluster radius. But downward bias is not the only source of error. The standard deviation of the ME estimates for subset regression is also considerably larger than for nn-garrote and ridge regression.

5.2.4 Number of Variables. I kept track of the average number of variables in the selected predictors for subset selection and nn garrote. Figure 7(a) plots these values versus cluster radius for the X-controlled simulation. Figure 7(b) is a plot for the X-random simulation. In the X-controlled situation, not many more variables are used by nn-garrote than subset selection. In the X-random simulation, nn-garrote uses almost twice the number of variables as subset selection.

5.2.5 Best Subsets Versus Variable Deletion. In the best-subsets procedure, the selected subset ζ_k of k variables is such that the regression of y on $\{x_m, m \in \zeta_k\}$ has minimum RSS among all k variable regressions. Our simulations did subset selection using backward deletion of variables. The question (raised by an associate editor) is how much our results reflect the use of deletion rather than best subsets.

Certainly, the subsets selected by the best-subsets procedure have lower RSS than those found using deletion. But, as exemplified in Section 5.2.7, lower RSS

Table 3. Estimated SE's for Crystal-Ball ME Differences

Difference	*X-controlled*	*X-random*
Subset–garrote	.2	.4
Subset–ridge	.3	.8
Garrote–ridge	.3	.6

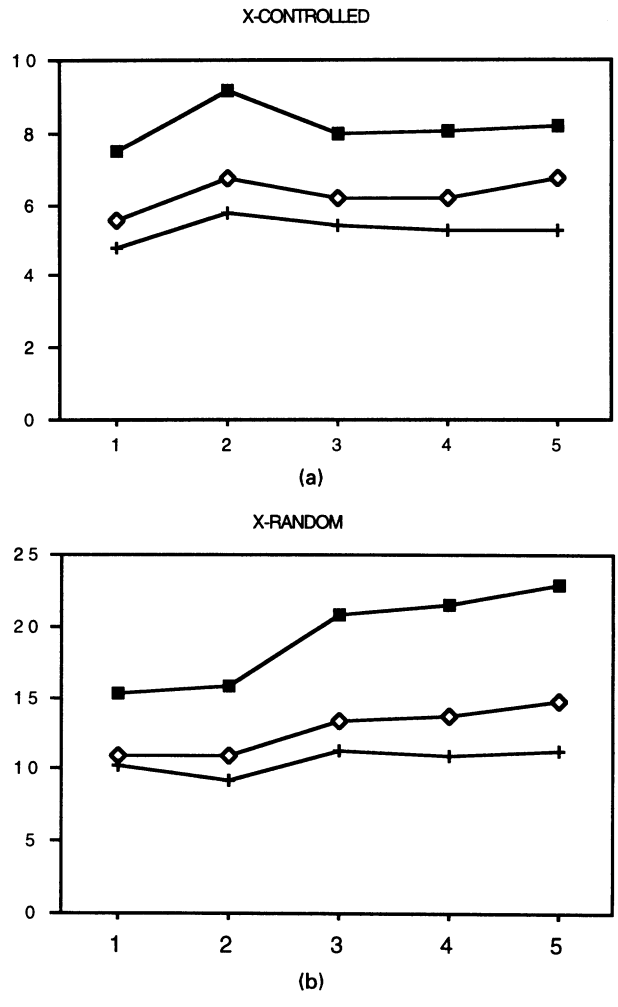

Figure 6. Average Absolute Error in ME Estimate for Selected Predictor Versus Cluster Radius: —■—, Subset; —◇—, Garrote; —|—, Ridge.

does not necessarily translate into lower prediction error. To explore the difference, I used the same data as in the 20-variable X-controlled simulation. In each of 250 iterations, two sequences of subsets were formed, one by deletion, the other by the Furnival and Wilson (1974) best-subsets algorithm, Leaps.

Then 25 repetitions of little bootstrap were done using deletion and the result used to select one subset out of the deletion sequence. Another 25 little bootstraps were done using Leaps and the results used to select one of the best subsets. The ME's were computed for each of the selected subsets and then averaged over the 250 repetitions. The results are plotted in Figure 8. The differences are small.

Table 4. Downward Bias as Averaged Over all Values of Cluster Radius

Downward bias	*X-controlled*	*X-random*
Subset selection	5.8	13.0
nn-Garrote	3.2	5.0
Ridge	1.9	4.1

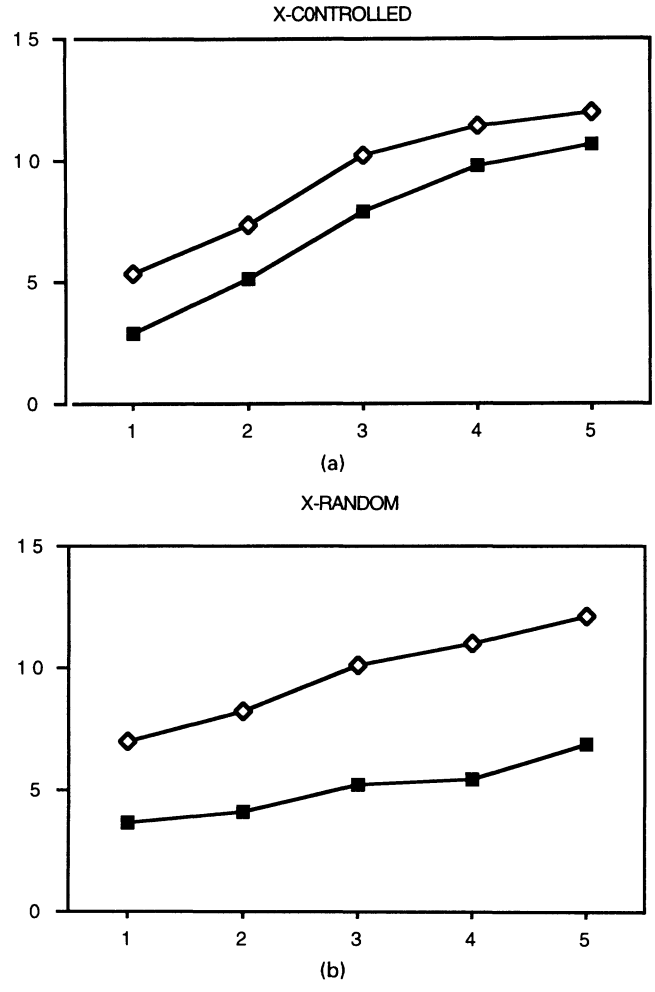

Figure 7. Average Number of Variables Used in Predictors Selected Versus Cluster Radius: —■—, Subset; —♦—, Garrote.

5.2.6 Nesting of the nn-Garrote Subsets. Stepwise variable deletion or addition produces nested subsets of variables. But the sequence of best (lowest RSS) subsets of dimension $1, 2, \ldots, M$ are generally not nested. A natural question is whether the subsets of variables

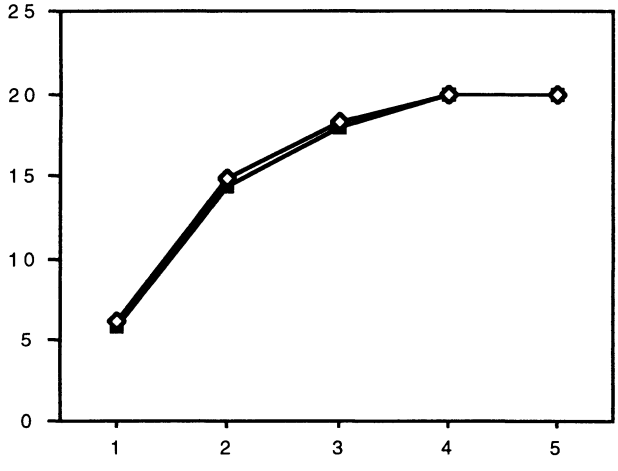

Figure 8. ME in X-Controlled Simulation, Leaps Versus Deletion: —■—, Best Subsets; —◇—, Deletion.

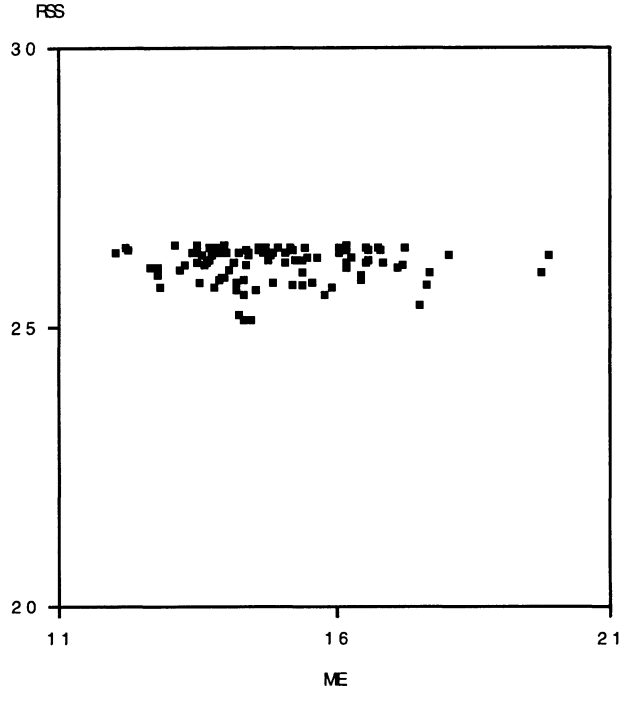

Figure 9. RSS Versus ME for 100 "Best" 10-Variable Subset Regressions.

produced by nn-garrote as s decreases are nested. The answer is "almost always, but not always." For instance, in the 1,250 iterations of nn-garrote in the 20-variable X-controlled simulation, 17 were not nested. Of the 1,250 in the 40-variable X-random simulation, 68 were not nested.

5.2.7 RSS Versus ME Instability in Subset Selection. To illustrate the instability of subset selection, I generated an X-controlled data set with $rc = 3$ and $\rho = .7$. Leaps was used to find the 100 subset regressions based on 10 variables having lowest RSS. For each of these regressions, the true ME was computed. Figure 9 is a graph of RSS versus ME for the 100 equations. The 100 lowest RSS values are tightly packed. But the ME spreads over wide range. Shifting from one of these models to another would result in only a small RSS difference but could give a large ME change.

6. CONCLUSIONS AND REMARKS

I have given evidence that the nn-garrote is a worthy competitor to subset selection methods. It provides simple regression equations with better predictive accuracy. Unless a large proportion of the "true" coefficients are nonnegligible, it gives accuracy better than or comparable to ridge methods. Data reuse methods such as little bootstrap or V-fold CV do well in estimating good values of the garrote parameter.

Some simulation results can be viewed as intriguing aspects of stability. Each regression procedure chooses from a collection of regression equations. Instability is intuitively defined as meaning that small changes in the data can bring large changes in the equations selected. If, by use of a crystal ball, we could choose the lowest PE equations among the subset section, nn-garrote, and ridge collections, the differences in accuracy between the three procedures are sharply reduced. But the more unstable a procedure is, the more difficult it is to accurately estimate PE or ME. Thus, subset-selection accuracy is severely affected by the relatively poor performance of the PE estimates in picking a low PE subset.

On the other hand, ridge regression, which offers only a small diversity of models but is very stable, sometimes wins because the PE estimates are able to accurately locate low PE ridge predictors. nn-garrote is intermediate. Its crystal-ball selection is usually somewhat better than the crystal-ball subset selection, but its increased stability allows a better location of low PE nn-garrote predictions, and this increases its edge.

The work in this article raises interesting questions. For instance, can the concept of stability be formalized and applied to the general issue of selection from a family a predictors? Can one use a formal definition to get a numerical measure of stability for procedures such as the three dealt with here? In another area, why is it that the nn-garrote produces "almost always, but not always" nested sequences of variable subsets?

The nn-garrote results may have profitable application to tree-structured classification and regression. The present method for finding the "best" tree resembles stepwise variable deletion using V-fold CV. Specifically, a large tree is grown and pruned upward using V-fold CV to estimate the optimal amount of pruning. I am experimenting with the use of a procedure analogous to nn-garrote to replace pruning. The results, to date, have been as encouraging as in the linear regression case.

Another possible application is to selection of more accurate autoregressive models in time series. Picking the order of the autoregressive scheme is similar to estimating the best subset regression. The nn-garrote methodology should carry over to this area and may provide increased prediction accuracy.

The ideas used in the nn-garrote can be applied to get other regression shrinkage schemes. For instance, let $\{\hat{\beta}_k\}$ be the original OLS estimates. Take $\{c_k\}$ to minimize

$$\sum_n \left(y_n - \sum_k c_k \hat{\beta}_k x_{kn} \right)^2$$

under the constraint $\sum c_k^2 \le s$. This version leads to a procedure intermediate between nn-garrote and ridge regression. In the $X'X = I$ case, its shrinkage factor is

$$\theta(\hat{\beta}, \lambda) = \frac{\hat{\beta}^2}{\hat{\beta}^2 + \lambda^2}.$$

Unlike ridge, it is scale invariant. Our expectation is that it will be uniformly more accurate than ridge regression while being almost as stable. Like ridge regression, it does not zero coefficients and produce simplified predictors. Study of this version of the garrote is left to future research.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge the many illuminating conversations on regression regularization that I have had with Jerry Friedman over the years and particularly during our recent collaboration on methods for predicting multiple correlated responses. This work doubtless stimulated some of my thinking about the nn-garrote. Phil Spector did the S run that produced the data used in Figure 9, and I gratefully acknowledge his assistance. Research was supported by National Science Foundation Grant DMS-9212419.

[*Received December 1993. Revised March 1995.*]

REFERENCES

Breiman, L. (1992), "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error," *Journal of the American Statistical Association*, 87, 738–754.

——— (1993), "Stacked Regressions," Technical Report 367, University of California, Berkeley, Statistics Dept.

——— (1994), "The Heuristics of Instability in Model Selection," Technical Report 416, University of California, Berkeley, Statistics Dept.

Breiman, L., and Friedman, J. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation" (with discussion), *Journal of the American Statistical Association*, 80, 580–619.

Breiman, L., and Spector, P. (1992), "Submodel Selection and Evaluation in Regression: The X-Random Case," *International Statistical Review*, 60, 291–319.

Cook, R. (1993), "Exploring Partial Residual Plots," *Technometrics*, 35, 351–362.

Daniel, C., and Wood, F. (1980), "Fitting Equations to Data," New York: John Wiley.

Frank, E., and Friedman, J. (1993), "A Statistical View of Some Chemometrics Regression Tools" (with discussion), *Technometrics*, 35, 109–148.

Friedman, J., and Silverman, B. (1989), "Flexible Parsimonious Smoothing and Additive Modeling" (with discussion), *Technometrics*, 31, 3–40.

Furnival, G., and Wilson, R. (1974), "Regression by Leaps and Bounds," *Technometrics*, 16, 499–511.

Golub, G., Heath, M., and Wahba, G. (1979), "Generalized Cross-validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21, 215–224.

Gruber, M. (1990), *Regression Estimators: A Comparative Study*, Boston: Academic Press.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*. New York: Chapman and Hall.

Hoerl, R., Schuenemeyer, J., and Hoerl, A. (1986), "A Simulation of Biased Estimation and Subset Selection Regression Technique," *Technometrics*, 28, 369–380.

Lawson, C., and Hanson, R. (1974), *Solving Least Squares Problems*, Englewood Cliffs, NJ: Prentice-Hall.

Mallows, C. (1973), "Some Comments on Cp," *Technometrics*, 15, 661–675.

Miller, A. (1990), *Subset Selection in Regression*, London: Chapman and Hall.

Roecker, E. (1991), "Prediction Error and its Estimation of Subset-Selected Models," *Technometrics*, 33, 459–468.

Shao, J. (1993), "Linear Model Selection via Cross-validation," *Journal of the American Statistical Association*, 88, 486–494.

Smith, G., and Cambell, F. (1980), "A Critique of Some Ridge Regression Methods" (with discussion), *Journal of the American Statistical Association*, 75, 74–103.

Tibshirani, R. (1994), "Regression Shrinkage and Selection via the Lasso," Technical Report 9401, University of Toronto, Dept. of Statistics.

Zhang, P. (1992), "Model Selection via Multifold Cross-validation," Technical Report 257, University of California, Berkeley, Statistics Dept.