Heuristics of Instability and Stabilization in Model Selection

By

Leo Breiman

Technical Report No. 416
June 1994

*Partially supported by NSF Grant No. DMS-9212419

Department of Statistics
University of California
Berkeley, California 94720

Heuristics of Instability and Stabilization in Model Selection

Leo Breiman*
University of California 94720

Abstract

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

If $\mathcal{L}$ is the data used to derive the sequence $\{\hat{\mu}(\cdot, s)\}$, the procedure is called unstable if a small change in $\mathcal{L}$ 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, etc. 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.

*Partially supported by NSF Grant No. DMS-9212419

1. Introduction

Given data $\mathcal{L} = \{(y_n, \mathbf{x}_n),\ n = 1, \ldots, N\}$ where the $\mathbf{x}_n$ take values in $E^{(M)}$, the goal of model selection is to use $\mathcal{L}$ to construct a function $\mu(\mathbf{x}, \mathcal{L})$ that will give accurate predictions of future y-values. One common approach is to define a large class $\mathcal{U}$ of predictors $\{\mu(\mathbf{x})\}$ and find that $\hat{\mu} \in \mathcal{U}$ which gives the best predictions on $\mathcal{L}$. Assuming that prediction error is mean-squared, this says – chose that $\hat{\mu} \in \mathcal{U}$ which minimizes

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

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

The strategy currently used to cope with this problem defines a class $\{\mathcal{U}_s\}$ of subspaces of $\mathcal{U}$ where s is a one-dimensional parameter and $s \leq s' \Rightarrow \mathcal{U}_s \subset \mathcal{U}_{s'}$. As s decreases, so does the effective dimensionality of $\mathcal{U}_s$. Define $\hat{\mu}(\cdot, s)$ as the minimizer in $\mathcal{U}_s$ of $\|y - \mu\|^2$. The final step is to form some estimate $\hat{PE}(\mu)$ of the "true" prediction error, $PE(\mu)$ of a predictor μ. Now define

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

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

select $\hat{s}$ as the minimizer of $\hat{PE}(s)$ and use the predictor $\hat{\mu}(\cdot, \hat{s})$. The usual $\hat{PE}$ estimates are gotten using test sets, cross-validation, bootstrap, etc.

The parameter s is a complexity parameter. The smaller s is, the simpler and smoother the predictors in $\mathcal{U}_s$. For instance, in subset selection in linear regression, s takes on non-negative integer values k, and $\mathcal{U}_k$ is the set of all linear regressions with at most k non-zero coefficients. In ridge regression, $\mathcal{U}_s$ is the set of all linear predictions $\boldsymbol{\beta} \cdot \mathbf{x}$ such that $\|\boldsymbol{\beta}\| \leq s$.

In binary tree regression ala CART, s is integer valued, and $\mathcal{U}_k$ is the set of all trees with k terminal nodes. Similarly, in MARS, $\mathcal{U}_k$ is the set of all μ that can be expressed as a sum of k or fewer basis functions (linear splines and products of linear splines). The current trend in neural networks is similar to ridge regression: minimize RSS using a constraint on the parameters of the form $\|\boldsymbol{\theta}\| \leq s$.

When it is not computationally feasible to find the $\mu \in \mathcal{U}_s$ minimizing $RSS(\mu)$, suboptimal methods are used. If $\mathcal{U}_k$ is the set of all linear predictors with k or fewer non-zero coefficients, and $\mathbf{x}$ is M dimensional, then finding the minimizing μ is slow if $M \geq 40$, and stepwise methods are used. Finding the k-terminal node tree μ which minimizes $RSS(\mu)$ is an NP-complete problem. The algorithm used in CART finds a suboptimal tree by stepwise splits.

The primary question we study is:

> *Let s^* be the crystal ball estimate of the best s, i.e. $s^* = \text{argmin} PE(s)$. Given an estimate $\hat{PE}(s)$ with $\hat{s} = \text{argmin}\hat{PE}(s)$, how much do we lose by not having a crystal ball? That is, how big is the predictive loss*
>
> $$PL = PE(\hat{s}) - PE(s^*)?$$

Obviously, the answer depends on how PE is estimated. But what we are trying to uncover is how the prediction loss is connected to the $\{\mathcal{U}_s\}$. A characteristic strongly related to predictive loss is the *instability* of the sequence $\{\hat{\mu}(\cdot, s, \mathcal{L})\}$. The intuitive definition of instability is that if the data $\mathcal{L}$ is changed slightly, then drastic changes can occur in $\{\hat{\mu}(\cdot, s, \mathcal{L})\}$.

Subset selection is unstable. Changing just one data case in $\mathcal{L}$ can cause a large change in the minimizer of $RSS(\mu)$ over $\mathcal{U}_k$. But ridge regression is stable. Eliminate one or a few cases and the minimizer of $RSS(\mu)$ over $\mathcal{U}_s$ is close to the original minimizer. Here is a list of well-known methods:

Unstable
CART
MARS
Neural Nets
Subset regression

Stable
k-nearest neighbors
ridge regression

On the surface, neural nets seem stable, since a gradient search is done to minimize $\|y - \mu(\cdot, \boldsymbol{\theta})\|^2$ under $\|\boldsymbol{\theta}\| \leq s$, where $\mu(\cdot, \boldsymbol{\theta})$ is a sum of sigmoid functions of linear functions. But the surface of $\|y - \mu(\cdot, \boldsymbol{\theta})\|^2 + \lambda\|\boldsymbol{\theta}\|^2$ generally has so many local minima that a small change in $\mathcal{L}$ may switch the minimizer from one local min to another some distance away.

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 1.1, 1.2 give illustrations of this. Figure 1.1 consists of $PE(k)$ plots for 3 runs of subset on 30-dimensional simulated data where the subsets are selected by forward stepwise addition. Figure 1.2 gives the plots of $PE(k)$ for ridge regression on the same data where k is the equivalent dimensionality. (See Section 5 for details of how the data were generated).

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 this 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.

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

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

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

2. Plan for Paper

Instead of trying to give rigorous definitions of instability, distance $\mathcal{L}$ to $\mathcal{L}'$, etc, we proceed by example. Linear regression is used as a paradigm. Four different methods with varying degrees of instability are studied. The most unstable is subset regression, the most stable is ridge regression. In a simple situation, analytical results can be derived. In more complex settings, we rely on simulations. The plan is:

Section 3 gives definitions of prediction error for X random and X controlled data. 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 $X'X = I$ case for PE estimated either by test set or little bootstrap. These illustrate the effects of instability

Figure 1.1

PE vs. No. Variables for Subset Selection

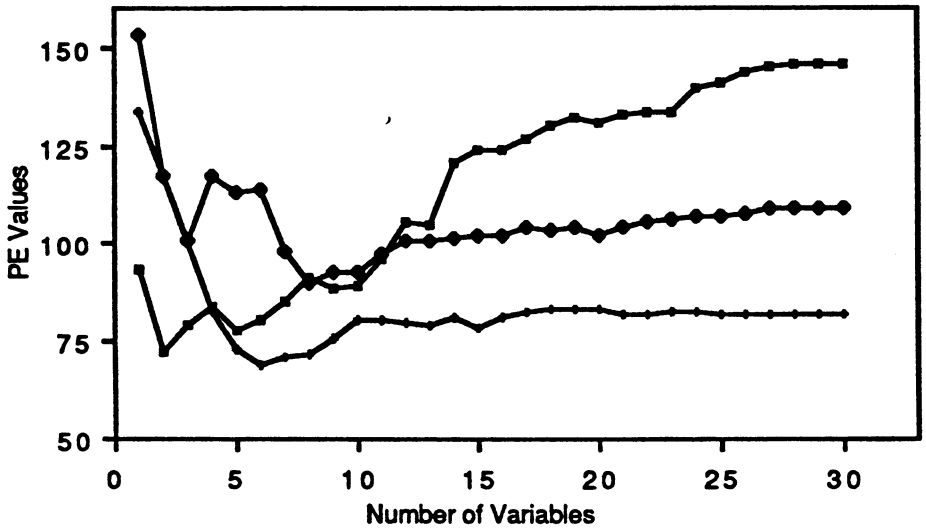
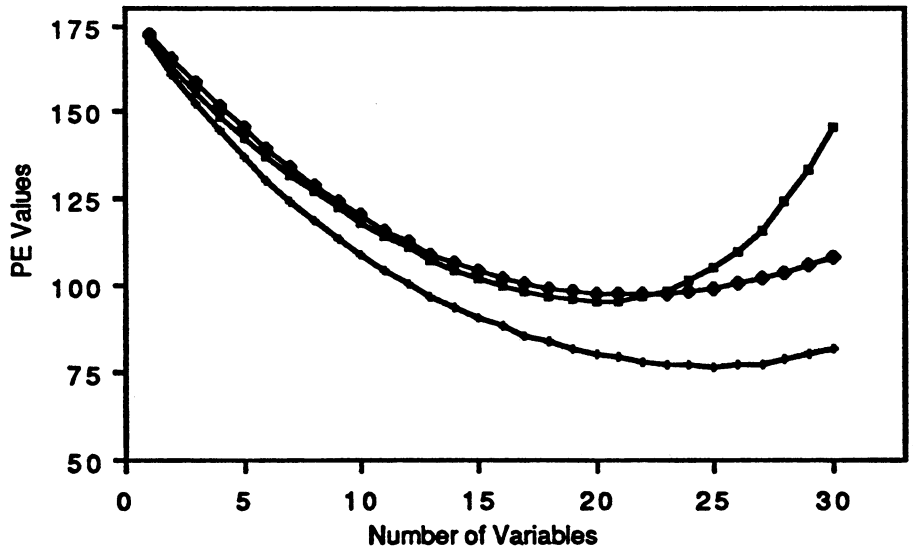


Figure 1.2

PE vs. No. Variables for Ridge



as the number of variables increases. The results of stabilization are made clear.

Section 5. Simulation results are given for X controlled using more complex $X'X$ designs. These again illustrate instability effects and the results of stabilization.

Section 6 gives results of a simulation study in the X-random case 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 has concluding remarks. We summarize the various threads in the preceding sections, and give some future research directions.

Appendix give details of the $X'X = I$ 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 covered. The two garrote 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 [1992a]) and its infinitesimal version, the tiny bootstrap (Breiman [1993]) are extended and strengthened as PE estimation methods.

The effects of instability first came up in connection with the study of one of the garrote methods compared to subset selection and ridge (Breiman [1993]). 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 $Y(x_n)$,

$$PE(\mu) = E(\sum_n (Y(x_n) - \mu(x_n))^2.$$

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

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

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

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

$$PE(\mu) = N\sigma^2 + NE(\mu^*(\mathbf{X}) - \mu(\mathbf{X}))^2$$

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

3.2 *PE Estimates*

a) *Test Sets.*

The simplest way to estimate PE is use of a test set. In the X-Random case this is data $\{(y'_n, \mathbf{x}'_n), n = 1, \ldots, N'\}$ iid from the same distribution as $\mathcal{L}$ and independent of $\mathcal{L}$. Then

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

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

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

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

b) *Cross-Validation.*

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

$$\hat{PE}(s) = \frac{N}{K \cdot N_{CV}} \sum_k \hat{PE}(\mathcal{F}_k).$$

The selection of the $\{\mathcal{F}_k\}$ 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 $\mathcal{F}_k$, each one consisting of the single case $(y_k, \mathbf{x}_k)$. Another selection is leave-many-out. Here the sizes of the $\mathcal{F}_k$ are fixed – usually some fraction of N, and the $\mathcal{F}_k$ selected at random. Another version of leave-many-out structures the $\{\mathcal{F}_k\}$ selection so that each $(y_n, \mathbf{x}_n)$ appears in exactly L of the $\{\mathcal{F}_k\}$. This is an extension of the V-Fold cross-validation used in CART.

c) *Little Bootstrap.*

In X-Controlled context, cross-validation is not appropriate. Consider

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

So

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

The term $N\bar{\sigma}^2 - \|\epsilon^*\|^2 - 2(\epsilon^*, \mu^*)$ has mean zero. 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 $\{\epsilon_n^*\}$ and $(\epsilon^*, \hat{\mu})$ does not usually have mean zero.

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

$$\hat{PE}(s) = RSS(\hat{\mu}(s) + 2B(s).$$

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

$$B_t(s) = \frac{1}{t^2} E_\epsilon(\epsilon, \hat{\mu}'). \tag{1.1}$$

Then

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*

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

Proof: In Appendix, Part II). 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 $\hat{\sigma}^2 = RSS/(N-M)$. The $\{\epsilon_n\}$ are generated from $N(0, \hat{\sigma}^2 t^2)$ using a rn generator, s_0 taken such that the corresponding $s_t = s$, and $\tilde{\mu}$ is the minimizer in $\mathcal{U}_{s_0}$ of $\|y + \epsilon - \mu\|^2$. Then $(\epsilon, \tilde{\mu})/t^2$ 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 $[.6, 1.0]$ seem to work best. The theorem states that for small t, B_t is an almost unbiased estimate of $(\epsilon^*, \hat{\mu})$. 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 $TB(s)$, and is an unbiased estimate of $(\epsilon^*, \hat{\mu})$.

For moderately unstable sequences, $TB(s)$ may exist, but be so noisy that more accurate estimates of $PE(s)$ are gotten by using B_t with $t > 0$. For nicely stable procedures, using $\hat{PE}(s) = RSS(s) + 2TB(s)$ gives accurate estimates.

d) *Others.*

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

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

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

$$\hat{PE}(k) = RSS(k)/(1 - \frac{k}{N})^2.$$

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], Breiman and Spector [1992].

3.3 *Four Linear Regression Methods*

a) *Best Subsets or Stepwise.*

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

b) *Ridge.*

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

c) *Non-negative Garotte.*

Let $\{\hat{\beta}\}$ be the full model OLS coefficients. Take the $(c_1, \ldots, c_M)$ to be the nonnegative minimizers of

$$\sum (y_n - \sum_m c_m \hat{\beta}_m x_{mn})^2 \tag{3.1}$$

under constraint $\sum c_m \leq s$. Then let $\hat{\mu}(\mathbf{x}, s) = \sum_m c_m \hat{\beta}_m x_m$.

d) *Garotte.*

Let $\{\hat{\beta}\}$ be the full model OLS coefficients and take $(c_1, \ldots, c_M)$ to minimize (3.1) under the constraint $\|\mathbf{c}\| \leq s$.

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

4. The $X'X = I$ Case

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

$$\mathbf{y} = \beta^* \mathbf{x} + \epsilon^*$$

where the $\{\epsilon_n^*\}$ 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

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

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

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

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

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

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

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

$$PE = N + \|\beta^* - \hat{\hat{\beta}}\|^2$$

and we put

$$ME(\lambda) = \|\beta^* - \hat{\hat{\beta}}(\lambda)\|^2.$$

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

$$ME(\lambda) = \sum_m (\beta_m^* - \theta(\hat{\beta}_m, \lambda))^2$$

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 $ME(\lambda)$.

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

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

The $M \cdot A(\lambda)$ term is the dominant deterministic part of $ME(\lambda)$. The $\{W(\lambda)\}$ is a zero-mean, approximately Gaussian stochastic process with $0(1)$ variance. Efforts to locate the minimum of $ME(\lambda)$ 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 see how accurately the minimum $ME(\lambda)$ can be located and how good the $\hat{PE}$ estimates are for the selected models. Two PE estimation methods are used. One is based on a single replicate test set. The other on little and tiny bootstrap. Many calculations are necessary and most are relegated to the Appendix.

4.1 *Using a Test Set*

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

$$\begin{aligned} \hat{PE}(\lambda) &= \sum_n (y'_n - \hat{\hat{\boldsymbol{\beta}}}(\lambda)\mathbf{x}_n)^2 \\ &= \|\epsilon'\|^2 + ME(\lambda) + 2\sum \epsilon'_n(\mathbf{x}_n, \boldsymbol{\beta}^* - \hat{\hat{\boldsymbol{\beta}}}(\lambda)) \\ &= \|\epsilon'\|^2 + ME(\lambda) + 2\sum_m Z'_m(\beta^*_m - \hat{\hat{\boldsymbol{\beta}}}(\lambda)) \end{aligned}$$

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

$$\hat{PE}(\lambda) = V + ME(\lambda) + \sqrt{M}Z(\lambda), \tag{4.6}$$

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 is

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

The crystal ball model corresponds to

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

We want to estimate the expected size of the predictive loss

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

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

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

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

Furthermore, the bias is

$$E(PE(\hat{\lambda}) - \hat{PE}(\hat{\lambda})) \sim K_2 \tag{4.8}$$

where $K_2 > 0$ also depends on P, θ and

$$\text{Var}(PE(\hat{\lambda}) - \hat{PE}(\hat{\lambda})) \sim 2N + 4ME(\lambda_0). \tag{4.9}$$

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

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

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

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

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 $E(\epsilon^*, \hat{\mu})$ introduces another stochastic element into the PE estimate, i.e.

$$\hat{PE}(\lambda) = \|\epsilon^*\|^2 + 2(\epsilon^*, \mu^*) + ME(\lambda) + 2(B_t(\lambda) - (\epsilon^*, \hat{\mu})),$$

where

$$B_t(\lambda) = \frac{1}{t^2} E(\epsilon', \hat{\mu}(\cdot, y + \epsilon', \lambda_t)$$

and $\lambda_t = \lambda\sqrt{1+t^2}$ for all except ridge, where $\lambda_t = \lambda$. Let $tU_m = (\epsilon', \mathbf{x}_m)$, then

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

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

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

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

If $\theta(\hat{\beta}, \lambda)$ is differentiable in $\hat{\beta}$, then (4.12) limit exists and equals $\theta_1(\hat{\beta}, \lambda)$. Then $TB(\lambda) = \sum \theta_1(\hat{\beta}_m, \lambda)$ and $(\epsilon, \hat{\mu}) - TB(\lambda)$ is a mean-zero process. If $\theta_1(\hat{\beta}, \lambda)$ is differentiable in λ, then the process is differentiable near λ_0 and (4.6), (4.7), (4.8) hold. Thus, $E(PL)$ and $E(Bias)$ 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 $(\epsilon^*, \hat{\mu}) - TB(\lambda)$ 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 $E(PL) \sim M^{1/5}$.

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 vicinity at $\pm\lambda$, 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 zero as $M^{-1/7}$ gives

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

The bias in *nn*-garotte goes up like $M^{3/5}$ and in subset selection like $M^{5/7}$. *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 $\text{Var}(\delta) = \tau^2\sigma^2$. Form $\hat{\mu}(\cdot, y + \delta, s)$ and now repeat and average. This gives a new estimate sequence

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

which we call the stabilized sequence

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

$$\frac{1}{\tau^2} E_V V\theta(\hat{\beta} + V, \lambda). \tag{4.13}$$

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

5. Simulation Results for X-Controlled

To see how the results carry over to more complex situations, we constructed a simulation that used a variety of design matrices and coefficients. 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 $\{y\}$ generated as $y = \boldsymbol{\beta}^*\mathbf{x} + \epsilon^*$ with $\text{Var}(\epsilon^*) = 1$. On the average $R^2 \simeq .83$ and about 20 of the coefficients were non-zero.

Two runs of 500 repetitions each were done. One used a PE estimate based on a single replicate test set. The other used little or tiny bootstrap. Five procedures were compared. Four are the original regression methods. The fifth is stepwise forward stablized 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 (5.1) is a bar graph showing the average crystal ball MEs 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 $\hat{PE}$. Figure (5.2) shows two bars for each procedure. The upper bar is the average Bias as a percent of the average Prediction Error. The lower bar is the average of the percent error in $\hat{PE}$ as an estimate of PE.

5.2 *The Little and Tiny Bootstrap Results*

Little bootstrap with $t = .6$ was used for stepwise and nn-garotte, with 25 iterations averaged to get B_t. With ridge, garotte and stablized stepwise, tiny bootstrap was used.

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

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

The solution is

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

An easy computation (Breiman [1994]) yields

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

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

$$\|\mathbf{y} - \mathbf{c}\hat{\boldsymbol{\beta}} \cdot \mathbf{x}\|^2 + \lambda \|\mathbf{c}\|^2$$

where $(\mathbf{c}\hat{\boldsymbol{\beta}}\mathbf{x})_n = \sum_m c_m \hat{\beta}_m x_{mn}$. Let $W_{mk} = \hat{\beta}_m (X'X)_{mk} \hat{\beta}_k$ and $W_\lambda = W + \lambda I$. Then (see Apppendix).

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

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

$$TB(k) = AV_\delta(\delta, \hat{\mu}(\cdot, \mathbf{y} + \boldsymbol{\delta}, k)).$$

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

The simulation results are summarized in Figures (5.3) and (5.4). Figure (5.3) uses the Figure (5.1) format. Figure (5.4) uses the Figure (5.2) format.

6. The X-Random Simulation Results

The X-Random case differs from the X-Controlled 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 nearly as well as leave-many-out. We now understand this as a consequence of instability. Thus, with the cross-validation

Figure 5.1

Crystal Ball ME and Predictive Loss

X-Controlled Data--Test Set

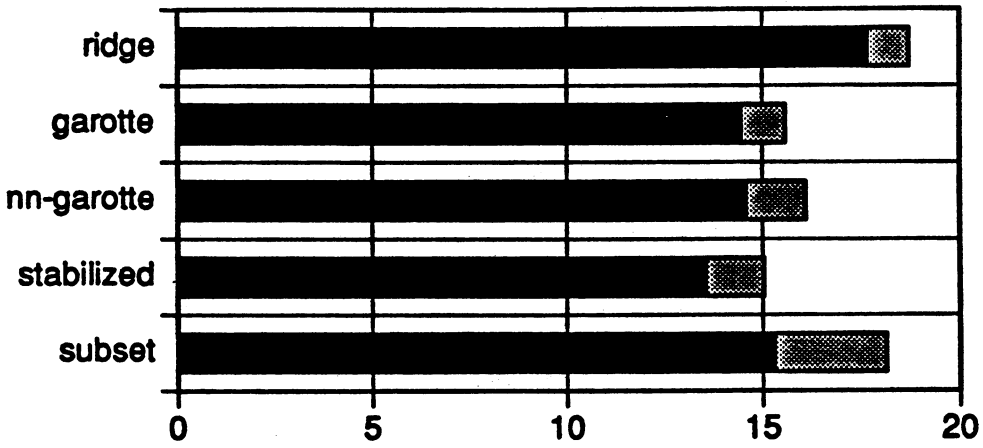
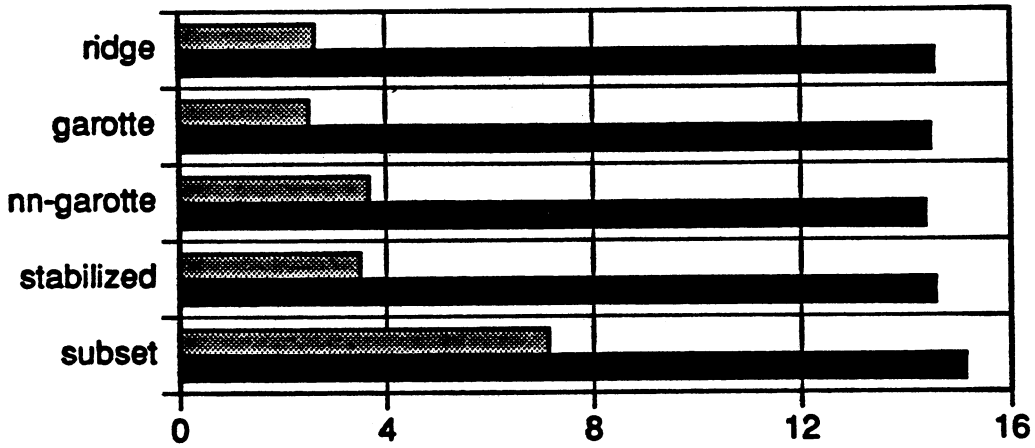


Figure 5.2

Percent Bias and Error

X-Controlled Data--Test Set



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 $(1/6)^{th}$ of the data (10 cases). Then the 2^{nd} $(1/6)^{th}$ was left out, etc. This was repeated five times.

Otherwise, the simulation has the same structure as in the X-Controlled case, i.e. 60 cases, 30 variables, some covariance and coefficients, etc. 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 6.1 and 6.2 using that same display format as in the X-Controlled figures. The output of the cross-validation run is in figures 6.3 and 6.4.

The symptoms of instability are that both $PE(s)$ and $\hat{PE}(s)$ are noisy and that $\hat{PE}(s)$ does not track $PE(s)$ accurately. In subset selection and stabilization PE and $\hat{PE}$ were computed for $k = 1,\ldots,30$ where k is the dimensionality. In the Ridge and garotte regressions, PE and $\hat{PE}$ were also computed at integer values $k = 1,\ldots,30$ where k is the dimensionality equivalent to the s parameter value. The values of $T_k = |\Delta_k - \hat{\Delta}_k|$, where $\Delta_k = PE(k+1) - PE(k)$ and $\hat{\Delta}_k = \hat{PE}(k+1) - \hat{PE}(k)$ were averaged over the 500 repetitions of the cross-validation simulation.

These are plotted vs k in Figure 6.5. The crucial parts of these curves are at the values of k of which $PE(k)$ 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 X-random 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 $\mathcal{F}$ of cases. Run the procedure on the remaining $\mathcal{L}-\mathcal{F}$ cases getting $\hat{\mu}(\cdot, s, \mathcal{L}-\mathcal{F})$. Now repeat this K times leaving out the subsets $\mathcal{F}_1,\ldots,\mathcal{F}_K$ and define

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

We implemented this using 10 cases in each $\mathcal{F}_K$, $K = 30$, with the random selection of the $\mathcal{F}_K$ structured so that each case occurred in exactly 5 of

Figure 5.3

Crystal Ball ME and Predictive Loss

X-Controlled--Little Bootstrap

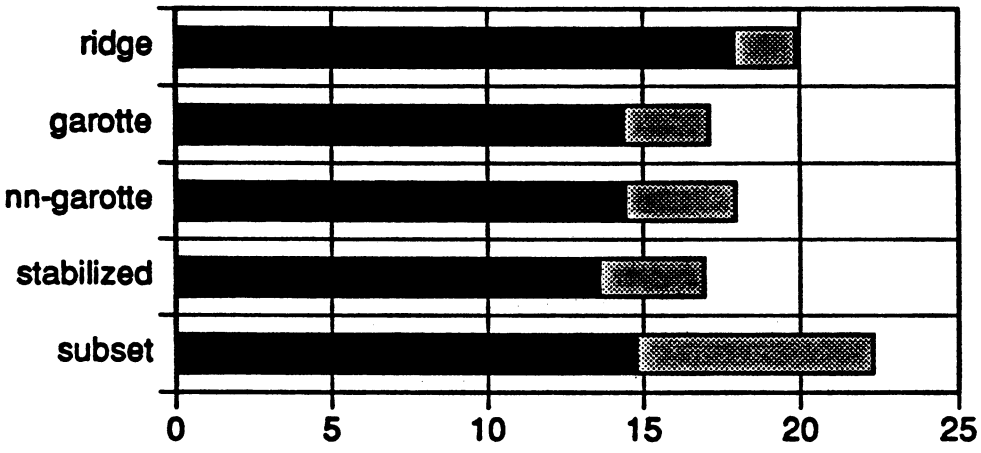


Table 5.4

Percent Bias and Error

X-Controlled--Little Bootstrap

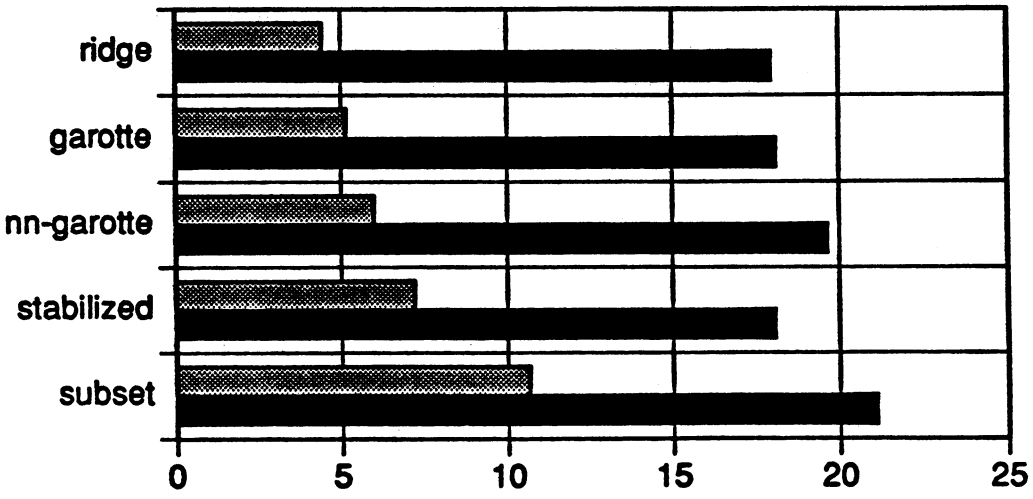


Figure 6.1

Crystal Ball ME and Predictive Loss

X-Random Data--Test Set

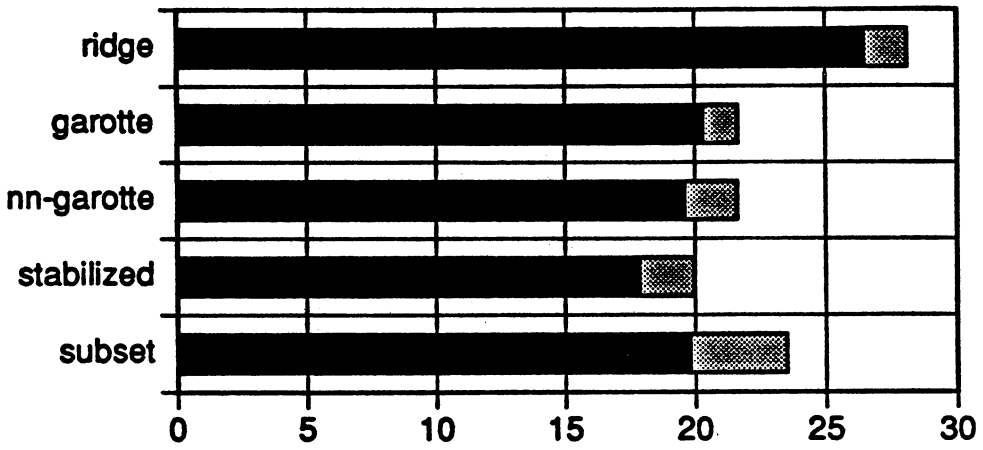


Table 6.2

Percent Bias and Error

X-Random Data--Test Set

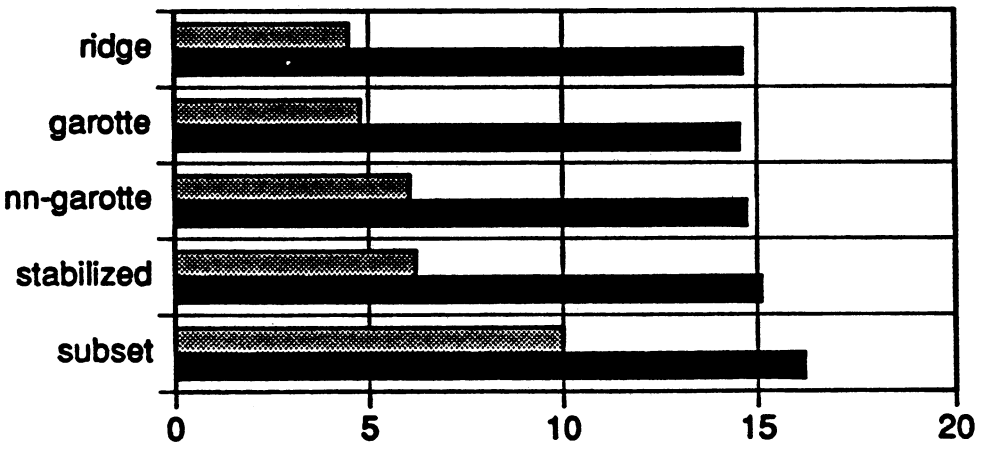
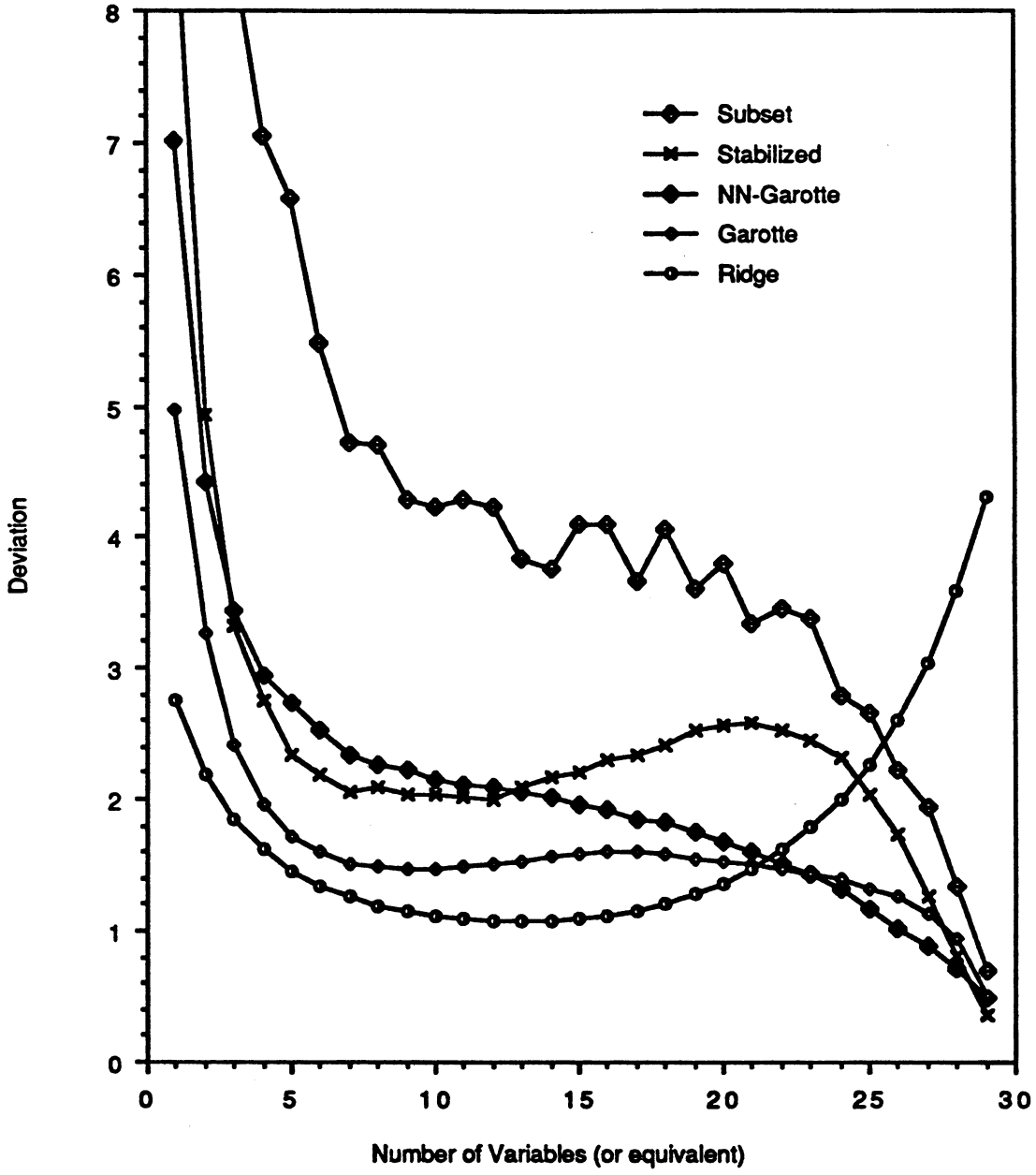


Figure 6.5

Tracking Deviations



the $\mathcal{F}_K$. The optimal value of s is selected using cross-validation. Another collection of sets $\{\mathcal{F}'_j\}$, $j = 1, \ldots, J$ is defined, set

$$\hat{\mu}_{ST}(\cdot, s, \mathcal{L} - \mathcal{F}'_j) = \frac{1}{K} \sum \hat{\mu}(\cdot, s, \mathcal{L} - \mathcal{F}_k - \mathcal{F}'_j),$$

$$\hat{PE}(\mathcal{F}'_j) = \sum_{(y_n, \mathbf{x}) \in \mathcal{F}'_j} (y_n - \hat{\mu}_{ST}(\mathbf{x}_n, s, \mathcal{L} - \mathcal{F}'_j))^2$$

and

$$\hat{PE}(s) = \frac{N}{\sum_j N_j} \sum_j \hat{PE}(\mathcal{F}'_j),$$

where $N_j = |\mathcal{F}'_j|$. Two definitions of the $\{\mathcal{F}'_j\}$ were used. One was $\mathcal{F}'_j = \{(y_j, \mathbf{x}_j)\}$, i.e. leave-one-out cross-validation was applied. In the second, $\{\mathcal{F}'_j\} = \{\mathcal{F}_k\}$, so leave-out-ten CV was used.

In applying stabilization to subset selection, the leave-out-ten 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 X-controlled situation, i.e. generate new y-values as $y' = y + \epsilon'$, rerun the procedure using the new y-values, repeat fifty times and average. This was combined with the use of leave-ten-out cross-validation to do PE estimation. $AV(PL)$ 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 don't 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.

Figure 6.3

Crystal Ball ME and Predictive Loss

X-Random Data-Cross Validation

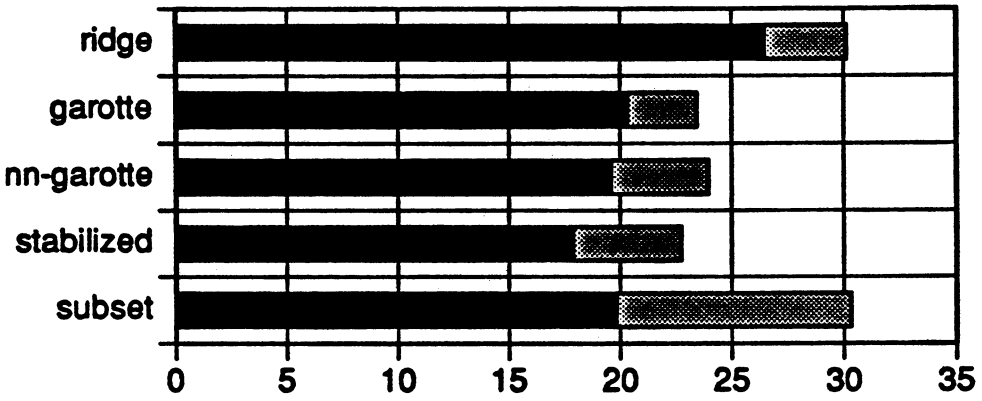
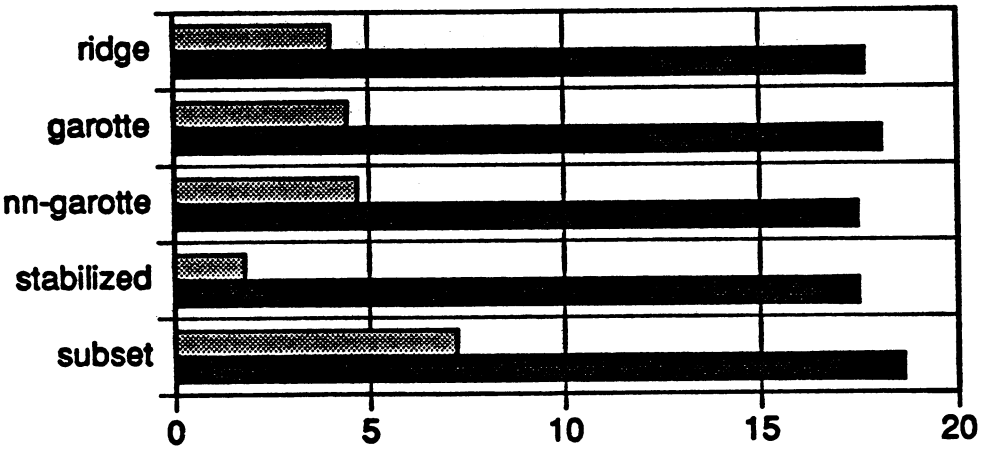


Figure 6.4

Percent Bias and Error

X-Random Data--Cross Validation



7. Leave-Out-One vs Leave-Out-Many

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

$$\hat{PE}(s)=\frac{N}{K\cdot N_{CV}}\sum_k\sum_{(y_n,\mathbf{x}_n)\in\mathcal{F}_k}(y_n-\hat{\mu}^{(-\mathcal{F}_k)}(\mathbf{x}_n))^2.$$

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

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

Figure 7.1 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 6-variable predictor. This gave 60 predictors. The RSS and ME were computed for each one. Figure 7.1 is a plot of RSS vs ME. The spread in ME is about ten times that in RSS. Figure 7.2 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 $\hat{PE}(s)$ is estimating some average of the values of $PE(\hat{\mu}^{(-\mathcal{F})})$. 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 the leave-out-ten cross-validation, the average prediction loss increases from 10.5 to 11.6. The average downward bias goes from 6.5 to 19.2 .

The large downward bias is an indicator of the problem. Figure 7.3 compares the average differences $|\hat{\Delta}|$ for leave-one-out and for leave-ten-out. Figure 7.4 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 leave-one-out $\hat{PE}$ estimate is noisier and tracks more poorly than the leave-ten-out estimate.

Figure 7.1

RSS vs ME Subset Selection (k=6)

Sixty Leave-One-Out Models

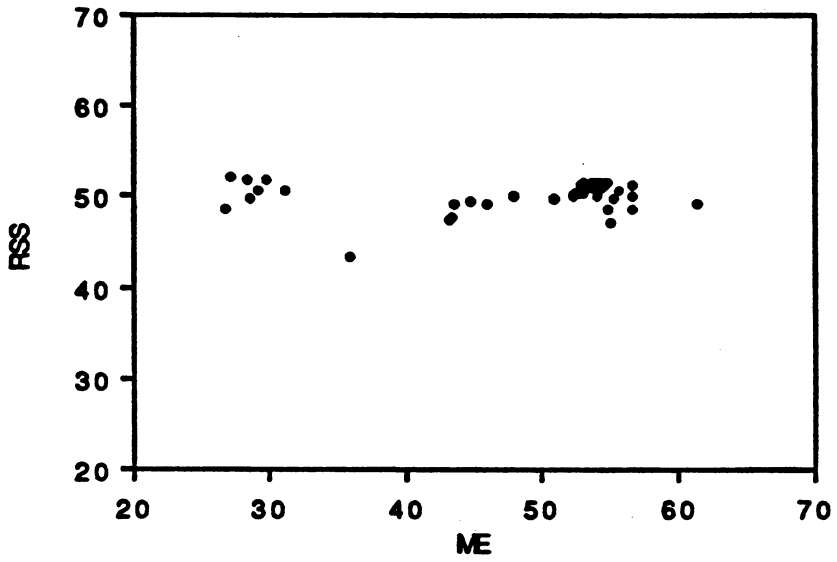
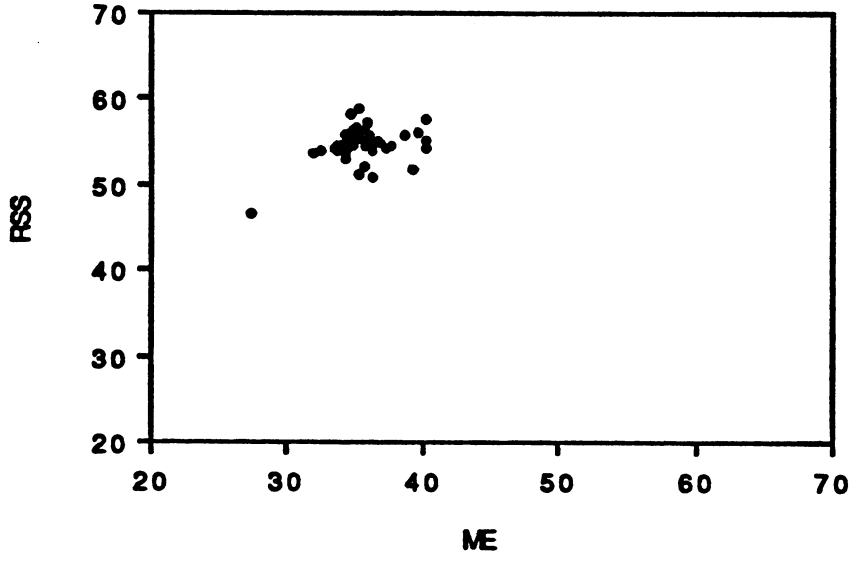


Figure 7.2

RSS vs ME Garotte (equiv. k=6)

Sixty Leave-One-Out Models



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

$$\frac{1}{2}(\hat{PE}(\hat{k}-1) + \hat{PE}(\hat{k}+1)) - \hat{PE}(\hat{k}).$$

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

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

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. Its fairly well known that if there are only a few non-zero coefficients, then subset selection gives good prediction. With many non-zero coefficients, ridge does better.

To compare methods, we generated five sets of simulated data that ranged from a few non-zero coefficients to many non-zero 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 non-zero 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

$$\beta^*_{10+j} = (h-j)^2, \quad |j| \leq h.$$

The clusters at 15 and 25 had the same shape. All other coefficients were zero. The coefficients were then multiplied by a common constant to give an $R^2 \simeq .75$ 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 non-zero coefficients. For $h = 1$ there were three strong, virtually independent variables. At the other extreme, for $h = 5$ each cluster contained 9 weak variables. This simulation structure is almost identical to that used in Breiman and Spector

Figure 7.3

Differences in Successive PE Estimates

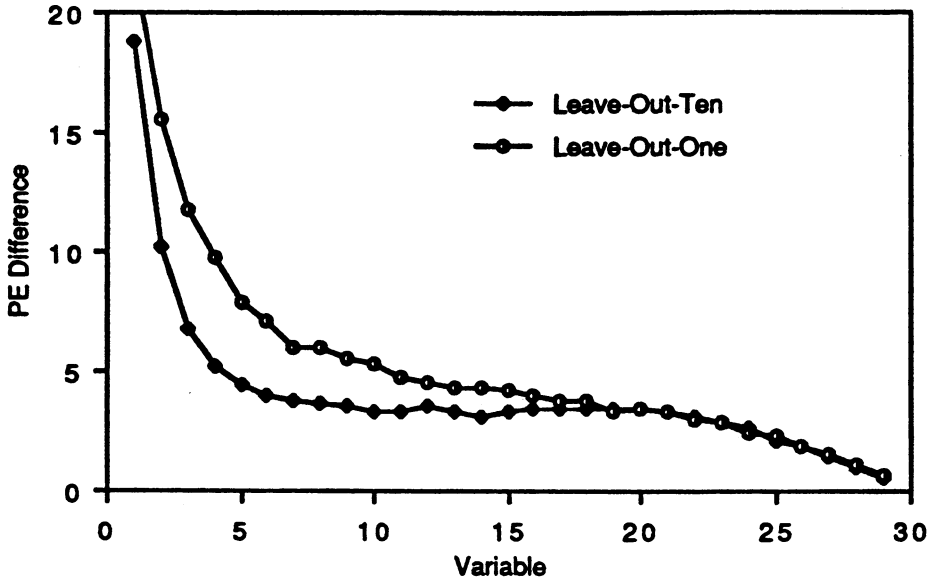
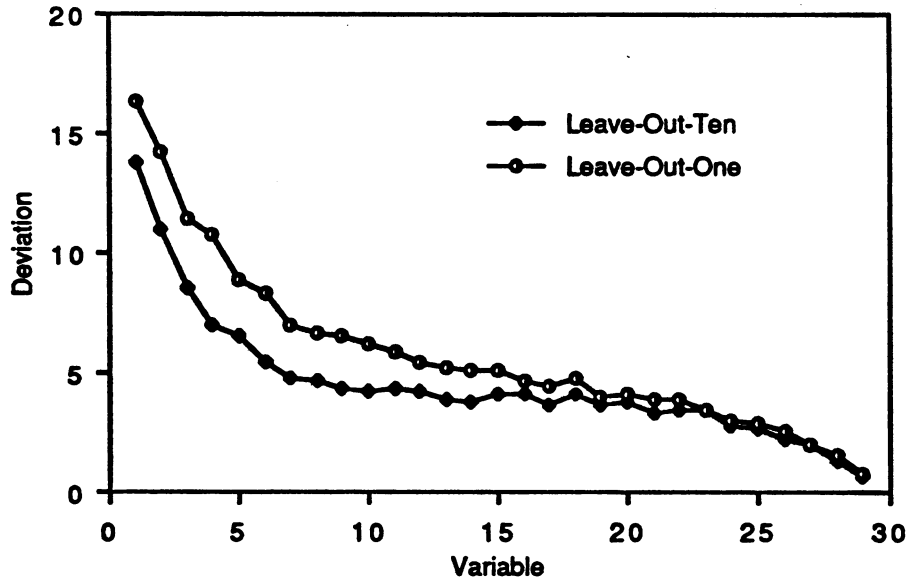


Figure 7.4

Tracking Deviations



[1992]. Two PE estimation methods were used. Six-fold cross-validation repeated five times was used in subset selection stabilization, and nn-garotte. Leave-one-out cross-validation was used in garotte and ridge.

Figure 8.1 is a graph of the average ME's vs h for the various prediction methods. Figure 8.2 is a graph of the average crystal ball ME's vs h, and Figure 8.3 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 non-zero 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.

9. Concluding Remarks

We have studied the effects of instability on predictive loss and on the bias 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 6-variable subset predictor generally gives a predictor with many more than 6 non-zero 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 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 [1993]). An interesting research issue we are exploring is whether there is a more stable version of CART.

A possible alternative is the idea of stacking predictors (Wolpert [1990], Breiman [1994]). In a 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 non-negative) are determined by a linear regression of the y-values on the $\{\hat{\mu}_k\}$ using the cross-validated values of the $\{\hat{\mu}_k\}$.

While averaging trees produces a muddy predictor, stacking a sequence of nested subtrees produces a single tree. Limited experiments showed appreciable decreases in PE over use of M_k single best tree. Simulations on linear regressions similar to those in this paper show that stacking subset regressions produces significant PE decreases. A question left for future work is whether stacking can produce as universal a stabilization as averaging.

References

Breiman, L. [1993] "Better subset selection using the non-negative garotte", Technical Report #405, Department of Statistics, University of California, Berkeley.

Breiman, L. [1992a] "The little bootstrap and other methods for dimensionality selection in regression: x-fixed prediction error", *J. Amer. Statist. Assoc.* **87**, 738-754.

Breiman, L. and Spector, P. [1992] "Submodel selection and evaluation in regression. The X-random case", *International Statistical Review* **60**, 291-319.

Breiman, L. [1992b] "Stacking Regressions". Technical Report No. 367, Statistics Department, University of California, Berkeley.

Wolpert, D. [1992] "Stacked generalization", *Neural Networks* **5**, 241-259.

Figure 8.1

Total ME for Five Methods

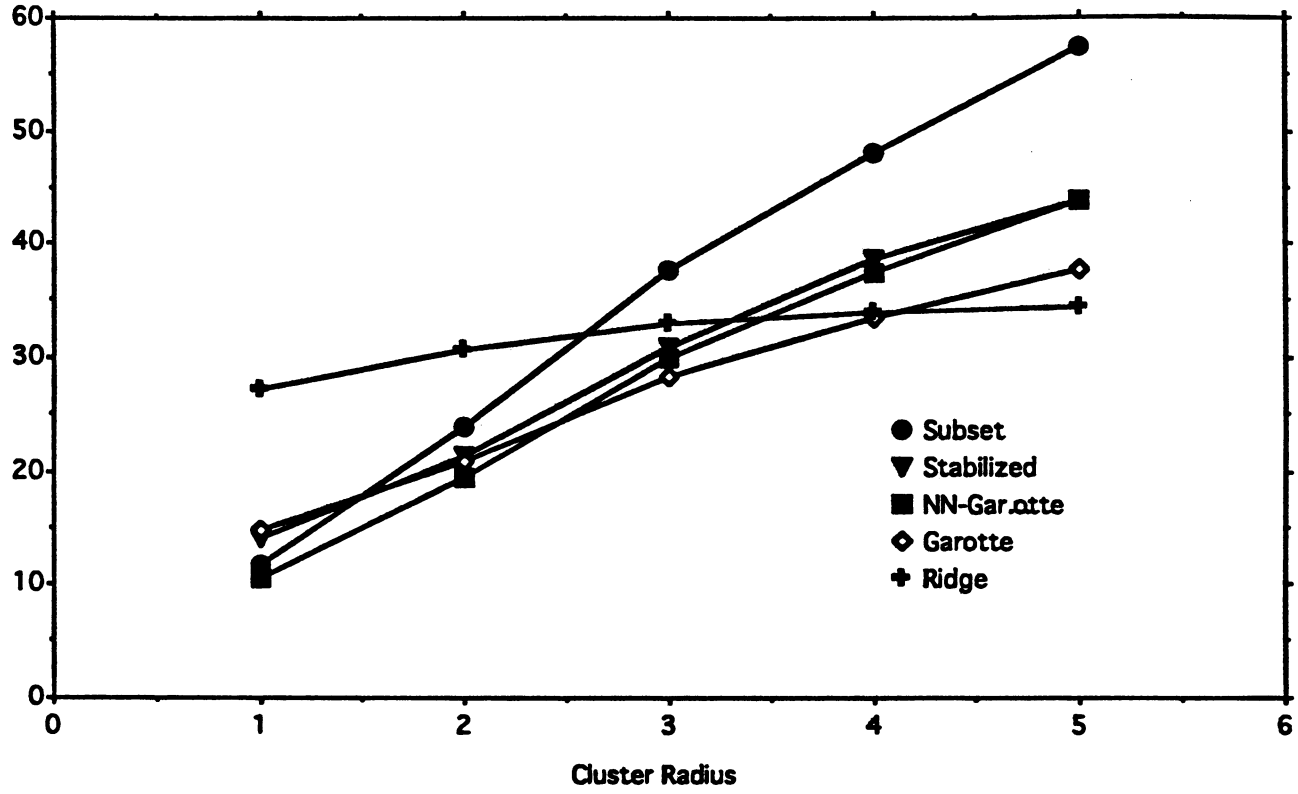


Figure 8.2

Crystal Ball ME for Five Methods

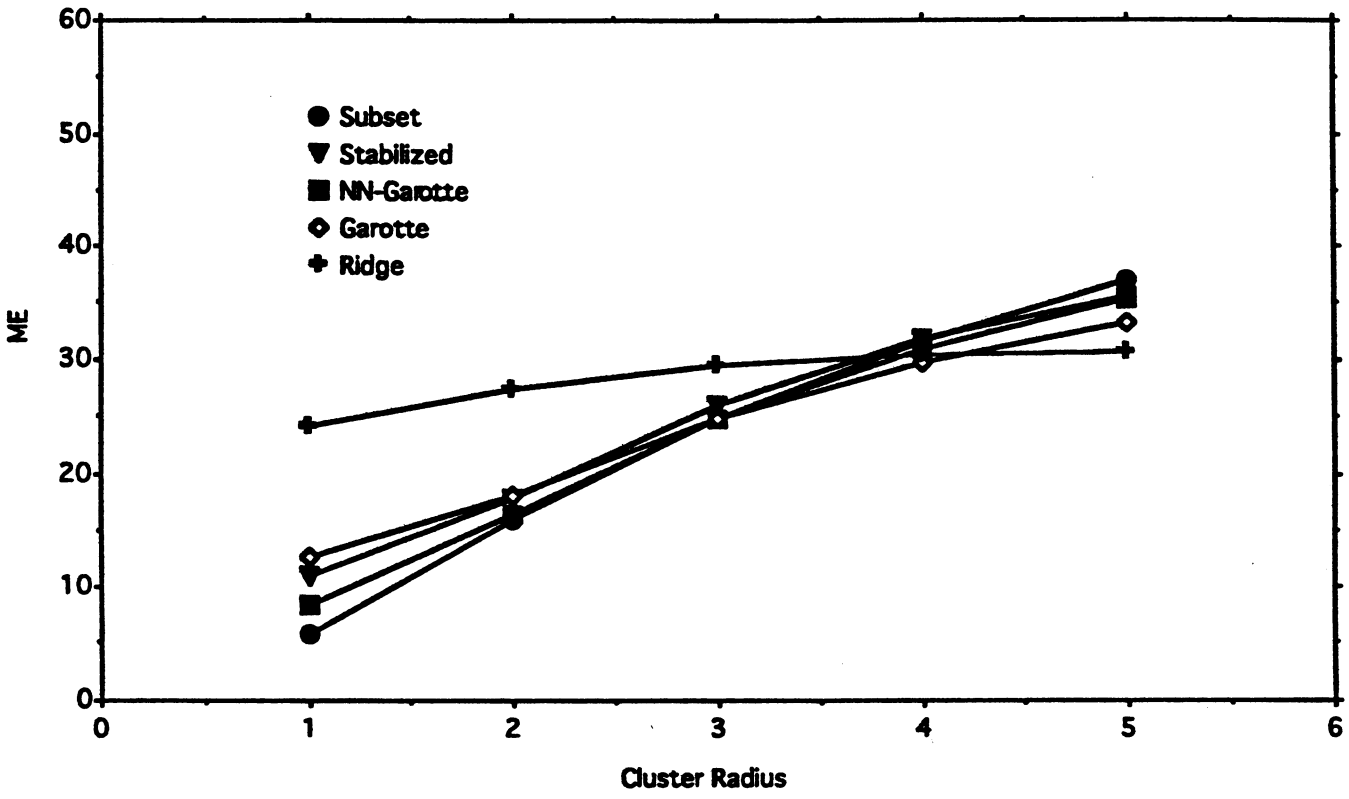
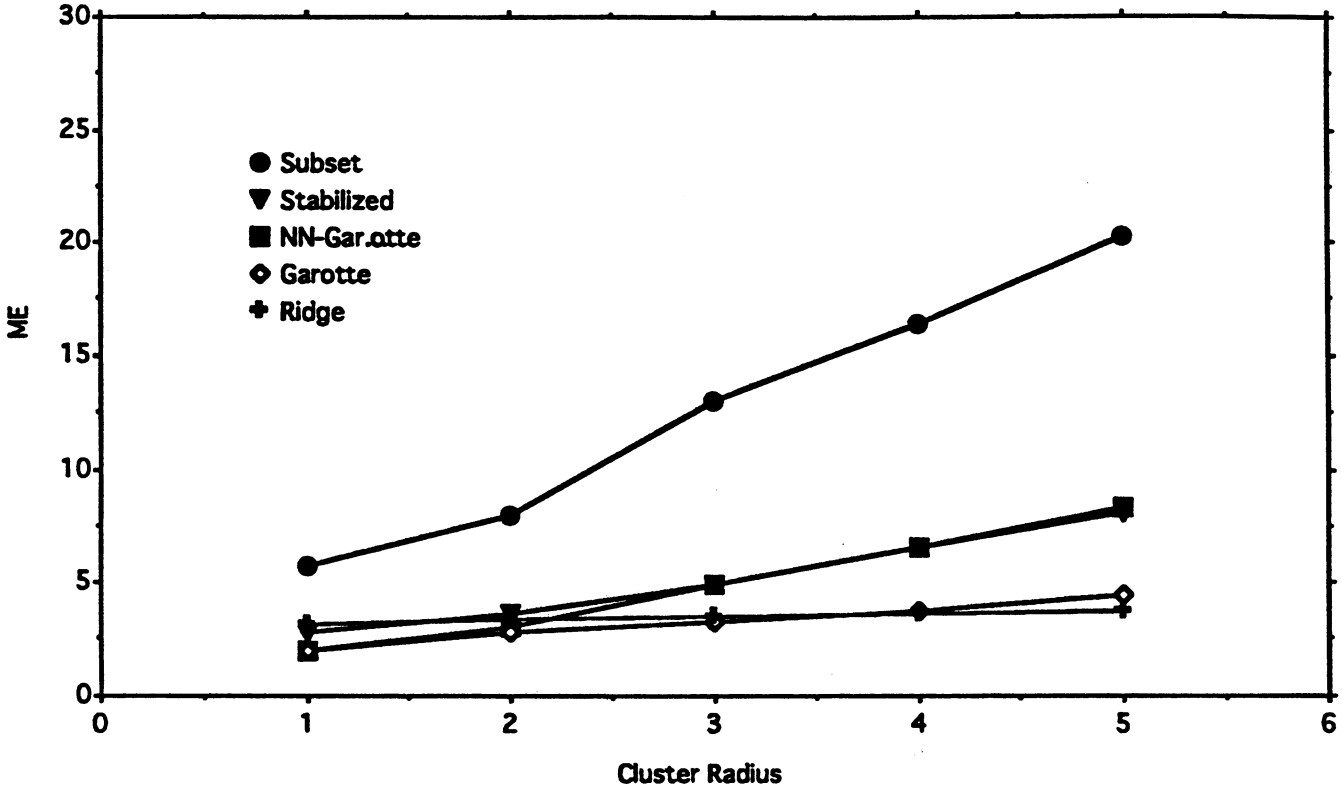


Figure 8.3

Predictive Loss for Five Methods



Appendix

Part I. Computations in the $X'X = I$ Results

From (4.5)

$$ME(\lambda) = M \cdot A(\lambda) + \sqrt{M}W(\lambda).$$

Let $\lambda_0 = \arg\min A(\lambda)$, and $\Delta = \lambda - \lambda_0$, so

$$ME(\lambda_0+\Delta) \simeq M \cdot A(\lambda_0)+\sqrt{M}W(\lambda_0)+M\Delta^2 A''(\lambda_0)/2+\sqrt{M}(W(\lambda_0+\Delta)-W(\lambda_0)).$$

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

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

i) *Test Set*

Using a replicate test set gives (see (4.6))

$$\hat{PE}(\lambda) = V + ME(\lambda) + \sqrt{M}Z(\lambda)$$

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

$$\hat{\Delta} = \arg\min(\frac{1}{2}M\Delta^2 A''(\lambda_0)+\sqrt{M}(W(\lambda_0+\Delta)-W(\lambda_0))+\sqrt{M}(Z(\lambda_0+\Delta)-Z(\lambda_0)) \tag{A.1}$$

If both W and Z are differentiable

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

and

$$ME(\hat{\lambda}) = M \cdot A(\lambda_0) + \sqrt{M}W(\lambda_0) + \frac{1}{2}M\hat{\Delta}^2 A''(\lambda_0) + \hat{\Delta}\sqrt{M}W'(\lambda_0).$$

resulting in

$$E(ME(\hat{\lambda}) - ME(\lambda^*)) \simeq \frac{1}{2}\frac{EZ'(\lambda_0)^2}{A''(\lambda_0)} \tag{A.3}$$

which verifies our assertion that $E(PL)$ 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 zero. Then $Y(\lambda)$ is mean-square differentiable if $\lim E[X(\lambda+\Delta)-X(\lambda))/\Delta]^2$ exists as $\Delta \to 0$. Its straightforward to verify that all methods except subset regression are ms 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 $\hat{\beta} = \beta^* + Z$ and Z, β^* are independent, $Z \in N(0,1)$. Let $X_m(\lambda) = I(|\hat{\beta}_m| < \lambda)(\beta_m^{*2} - Z_m^2)$ so

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

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 $\Delta_1, \Delta_2 > 0$

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

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

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

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 $ME(\lambda)$ as

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

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

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

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

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

and

$$\min_{\lambda} ME(\lambda) \simeq V + b \min_{t}[t^2 + B(t)]$$

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

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

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,

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

Since

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

is also a two-sided B-motion, then

$$\hat{\Delta} = \arg\min\{\frac{1}{2}\Delta^2 MA'' + r\sqrt{M}B_0(\Delta)]$$

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

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

So,

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

The resulting approximation is

$$ME(\lambda_0 + \hat{\Delta}) = V + \frac{1}{2}M\hat{\Delta}^2 A'' + c\sqrt{M}B_1(\hat{\Delta}).$$

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

$$E(cB_1(\hat{\Delta})|\{cB_1(\Delta) + dB_2(\Delta), -\infty < \Delta < \infty) =$$

$$\frac{c^2}{c^2 + d^2}(cB_1(\hat{\Delta}) + dB_2(\hat{\Delta})).$$

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

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

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

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

$$E(ME(\hat{\lambda}) - ME(\lambda^*)) \simeq \frac{1}{2}\frac{d^2}{c^2 + d^2}MA''E\hat{\Delta}^2 + (\frac{ec^2}{c^2 + d^2} - b)Ey \qquad (A.4)$$

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

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

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

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

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

ii) *Errors in PE-Estimates Using a Test Set*

Here we look at the mean and variance of $\hat{PE}(\hat{\lambda}) - PE(\hat{\lambda})$. This difference equals

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

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

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

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

To get the variance, note that $\|\epsilon'\|^2 - N$ is χ^2_N independent of $Z(\lambda_0)$ and that $EZ^2(\lambda_0) = 4\sum_m(\beta_m - \hat{\beta}_m(\lambda_0))^2 \simeq 4ME(\lambda_0)$. Thus

$$\mathrm{Var}(\hat{PE}(\hat{\lambda}) - PE(\hat{\lambda})) \sim 2N + 4ME(\lambda_0).$$

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

In subset selection,

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

Using calculations similar to those in the $E(PL)$ calculations gives

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

and

$$\mathrm{Var}(\hat{PE}(\hat{\lambda}) - PE(\hat{\lambda})) \simeq 2N + 4ME(\lambda_0) + K_4M^{2/3}.$$

iii) *Little Bootstrap*

For $\theta(\beta, \lambda)$ differentiable in β, $B_t(\lambda) \to TB(\lambda)$ as $t \to 0$,

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

and

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

where V is a 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

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

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, $TB(\lambda)$ is not differentiable in λ. Now $TB(\lambda_0+\Delta)-TB(\lambda_0)$ can be approximated by a Brownian motion, leading to $E(PL)\sim M^{1/3}$. However, there is an alternative strategy leading to lower $E(PL)$, i.e. take $t>0$ going to zero as $M\to\infty$.

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

Let

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

so

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

Put

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

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

$$\hat{PE}(\lambda)=V+ME(\lambda)+\sqrt{M}Z_t(\lambda)+\sqrt{M}EY_t(\lambda).$$

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

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

Therefore

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

In consequence,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

In simulations (Breiman [1993], [1994]) we found that in subset selection and *nn*-garrote, using $t \in [.6, 1.0]$ gave better results than smaller t values. Now we can begin to understand the reason.

iv) *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 $\hat{PE}(\hat{\lambda})$ equals

$$2N + 4ME[\beta^* Z - Z\theta(\hat{\beta}, \lambda_0) + \theta_1(\hat{\beta}, \lambda_0)]^2.$$

With some integration by parts, the expectation term equals

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

giving the variance approximation

$$2N + 4ME(\lambda_0) + 4ME\theta_1^2(\hat{\beta}, \lambda_0).$$

Thus, use of Tiny Bootstrap adds an $O(M)$ term to the $\hat{PE}$ variance as compared to the test set $\hat{PE}$.

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

Part II. Proof of Theorem 3.1

Using the identity

$$E(\epsilon_k | \{\epsilon_n^* + \epsilon_n, n = 1, \ldots, N\}) = \frac{t^2}{1+t^2}(\epsilon_k^* + \epsilon_k)$$

gives

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

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

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

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

$$\|\frac{\mu^*}{\sqrt{1+t^2}} + \epsilon^* - \mu\|^2$$

so $\hat{\mu}(\cdot, \mu^* + \delta^*, s) = \sqrt{1+t^2}\hat{\mu}(\cdot, \frac{\mu^*}{\sqrt{1+t^2}} + \epsilon^*, s_t)$. Putting things together

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

which is equivalent to the statement of the theorem.

Part III. 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 $\sum C_m^2 \leq s^2$.

Take $\{\epsilon_n\}$ to be iid $N(0, \sigma^2)$, and put $y'_n = y_n + t\epsilon_n$. Denote $S = X'X$. The new OLS $\hat{\boldsymbol{\beta}}(t) = S^{-1}X'y'$. Put $\mathbf{Z} = (\epsilon, \mathbf{X})$, so $\hat{\boldsymbol{\beta}}(5) = \hat{\boldsymbol{\beta}} + tS^{-1}\mathbf{Z}$. The altered $\{c_m(t)\}$ minimize

$$\|\mathbf{y} + t\boldsymbol{\epsilon} - \sum_m c_m(t)\hat{\beta}_m(t)\mathbf{x}_m\|^2$$

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

Little bootstrap equals

$$\frac{1}{t}E\sum_m Z_m c_m(t)\hat{\beta}_m(t)$$

so

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

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 $W_{mk} = \hat{\beta}_m(t)S_{mk}\hat{\beta}_k(t)$. Then the Lagrangian equation for determining $\mathbf{C}(t)$ is

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

Differentiating gives

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

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

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

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

$$E(\sum_m Z_m\hat{\beta}_m\dot{c}_m(0)) = \sigma^2(\sum_{m,k} W_{mn}^{-1}(\lambda)W_{mk}(1-c_m) + \sum_{m,k} W_{mm}^{-1}(\lambda)W_{mk}(1-c_k).$$

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

with $\{e_i^2(m)\}_1^q$ given by (6.11). This measures the error squared (averaged over the responses) of each method relative to the corresponding minimum over all of the methods. For each replication (7.12) will have the value 1.0 for the best (minimum average error squared) method and larger values for the other two methods. The results of this simulation study are summarized by the average of (7.12) over the 100 replications for each of the 144 situations.

Figure 10 shows box plots for each method of the distribution of the 144 averages of (7.12) over all situations. C&W is seen to produce the best average error (squared), or within a few percent of the best, in every situation. The corresponding quantity for separate ridge regressions is typically 22% larger than the best, and that for two-block PLS is 30% larger. However, the dispersion of values for two-block PLS about its median is somewhat less than that for separate ridge regressions.

Figure 11 divides the 144 situations into three subsets corresponding to each of the three levels of population predictor variable collinearity: low ($r = 0.0$), medium ($r = 0.90$), and high ($r = 0.99$), in (6.1) (6.2). The three box plots shown for each method summarize the 48 averages of (7.12) over the situations in each if the three respective subsets. One can see that for low (population) collinearity all three methods perform comparably, C&W holding a slight edge. This is due to the fact that for $p \prec N$ and low collinearity none of the three methods is able to produce predictions that are much more accurate than simply the response means. In higher (population) collinearity settings more accurate prediction is possible and the C&W procedure is seen to be much more dominant over the other two. This is especially the case for the highest collinearity ($r = 0.99$) where it is typically 42% better than two-block PLS and 75% better than separate ridge regressions. It is also interesting to note that this is the only setting in which two-block PLS appears to perform better than separate ridge regressions.

8. Conclusion

The results presented in this paper strongly suggest that the conventional (statistical) wisdom, that one should avoid combining multiple responses and treating them in a multivariate manner, may not be the best advice. Our simulation studies indicate that the best of the multiple response procedures considered here can provide large gains in expected prediction accuracy (for each response), over separate single response regressions, with surprisingly little risk of making things worse. In the fields of neural networks and chemometrics, by contrast, the conven-

tional wisdom has always been in favor of multivariate multiple regression. The results of this paper generally validate that intuition, but it is not clear that the respective recommended approaches in each of those fields best serve that purpose. For example, the two-block PLS approach commonly used in chemometrics was seen in our simulation studies to provide generally lower accuracy than separate ridge regressions.

Our results suggest the intriguing prospect that even when there is only a single response of interest, if there are variables available that are correlated with it, then the prediction for the response of interest may be improved by introducing the other variables as additional responses and using the C&W procedure. Of course, if the values of these variables will also be available for (future) prediction, they should be regarded as predictors (rather than responses) and included in the regression equation. In some circumstances however, the (training) data may include measurements of variables whose values will not be available in the prediction setting.

In the neural network literature such variables are known as "hints". These are variables whose values are available for use during training but not available for future prediction. Examples might be expensive or difficult to obtain medical measurements that were available at the hospital where the training data were collected, but not available in the field or at smaller hospitals where the predictions are made. In financial forecasting, "future" values of other quantities, thought to be correlated with the response, might be included as hints. The results presented in this paper suggest that the inclusion of such hint variables as extra responses during training using C&W can indeed improve prediction accuracy.

9. References.

Anderson, T. W. (1957). *An Introduction to Multivariate Analysis.* Wiley.

Brown, P. J. and Zidek, J. V. (1980). Adaptive multivariate ridge regression. *Annals of Statist.* **8**, 64-74.

Brown, P, J. and Zidek, J. V. (1982). Multivariate regression shrinkage estimators with unknown covariance matrix. *Scad. J. Statist.* **9**, 209-215.

Copas, J. B. (1983). Regression, prediction, and shrinkage (with discussion). *J. Roy. Statist. Soc.* **B45**, 311-354.

Copas, J. B. (1987). Cross-validation shrinkage of regression predictors. *J. Roy. Statist. Soc.* **B49**, 175-183.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**, 317-403.

Efron, B. and Morris, C. (1972). Empirical Bayes on vector valued observations - An extension of Stein's rule. *Biometrika* **59**, 335-347.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109-148.

Garthwaite, P. H. (1994). An interpretation of partial least squares. *J. Amer. Statist. Assoc.* **89**, 122-127.

Golub, G. H. and van Loan, C. F. (1989). *Matrix Computations.* Johns Hopkins Univ. Press.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **8**, 27-51.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *J. Multiv. Anal.* **5**, 248-264.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium (Vol. I)*, ed. J. Neyman, Berkeley: University of Calif. Press, 361-379.

Massey, W. F. (1965). Principal components regression in exploratory statistical research. *J. Amer. Statist. Assoc.* **60**, 234-246.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statis. Soc.* **B36**, 111-147.

van der Merwe, A. and Zidek, J. V. (1980). Multivariate regression analysis and canonical variates. *Canadian J. Statist.* **8**, 27-39.

Wold, H. (1975). Soft modeling by latent variables; the nonlinear iterative partial least squares approach. In *Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett,* ed. J. Gani. Academic Press.

10. Figure captions.

Figure 1: Population canonical coordinate shrinkage factors (2.26) as a function of squared (population) canonical correlation, for various ratios of parameter to observation count.

Figure 2: Sample based canonical coordinate shrinkage factors (3.12) (3.13) as a function of squared (sample) canonical correlation, for the same ratios of parameter to observation count as in Fig. 1.

Figure 3: Distribution over all 120 situations ($p \prec N$) of the overall average response mean-squared error relative to OLS (6.12) for each biased method.

Figure 4: Distribution over all 120 situations ($p \prec N$) of the average individual response mean-squared error relative to OLS (6.13) for each biased method.

Figure 5: Distribution over all 120 situations ($p \prec N$) of the ratio of overall average response mean-squared error for each method, to that of the best method (6.15).

Figure 6: Distribution over all 120 situations ($p \prec N$) of the ratio of average individual response mean-squared error (relative to OLS) for each method, to that of the best method (6.16).

Figure 7: Distributions over all the 30000 replications ($p \prec N$) of the fraction of responses in each, for which the respective biased methods were less accurate than the corresponding OLS estimate.

Figure 8: Distribution ($p \prec N$) of the worst individual response mean-squared error relative to OLS (6.14) of each of the six biased methods, for each of the two error covariance matrix structures (6.5) (ERRVAR1, ERRVAR2, respectively).

Figure 9: Distribution of the overall average response mean-squared error relative to that of OLS (6.12) of C&W-CV for various subsets of the 120 ($p \prec N$) situations (RESP = number of responses, SS = sample size, S/N = signal to noise ratio).

Figure 10: Distribution over all 144 ($p \succ N$) situations of the ratio of the overall average response mean-squared error for each method, to that of the best method (7.12).

Figure 11: Distribution ($p \succ N$) of the ratio of the overall average response mean-squared error for each method, to that of the best method (7.12), separately for each of the three levels of predictor variable collinearity ("XCORR") (6.1) (6.2) (LOW: $r = 0.0$, MED: $r = 0.9$, HIGH: $r = 0.99$).

Figure 1

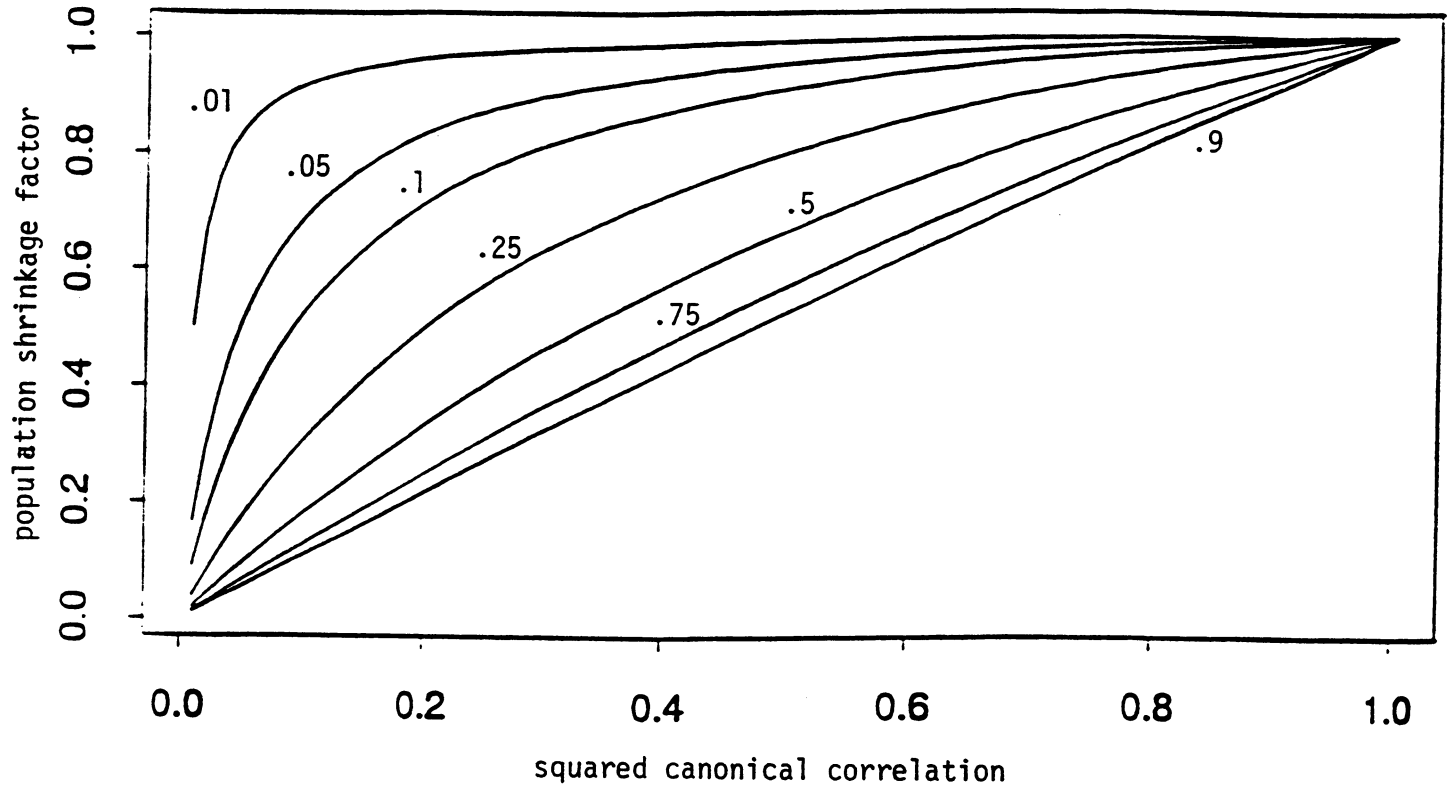


Figure 2

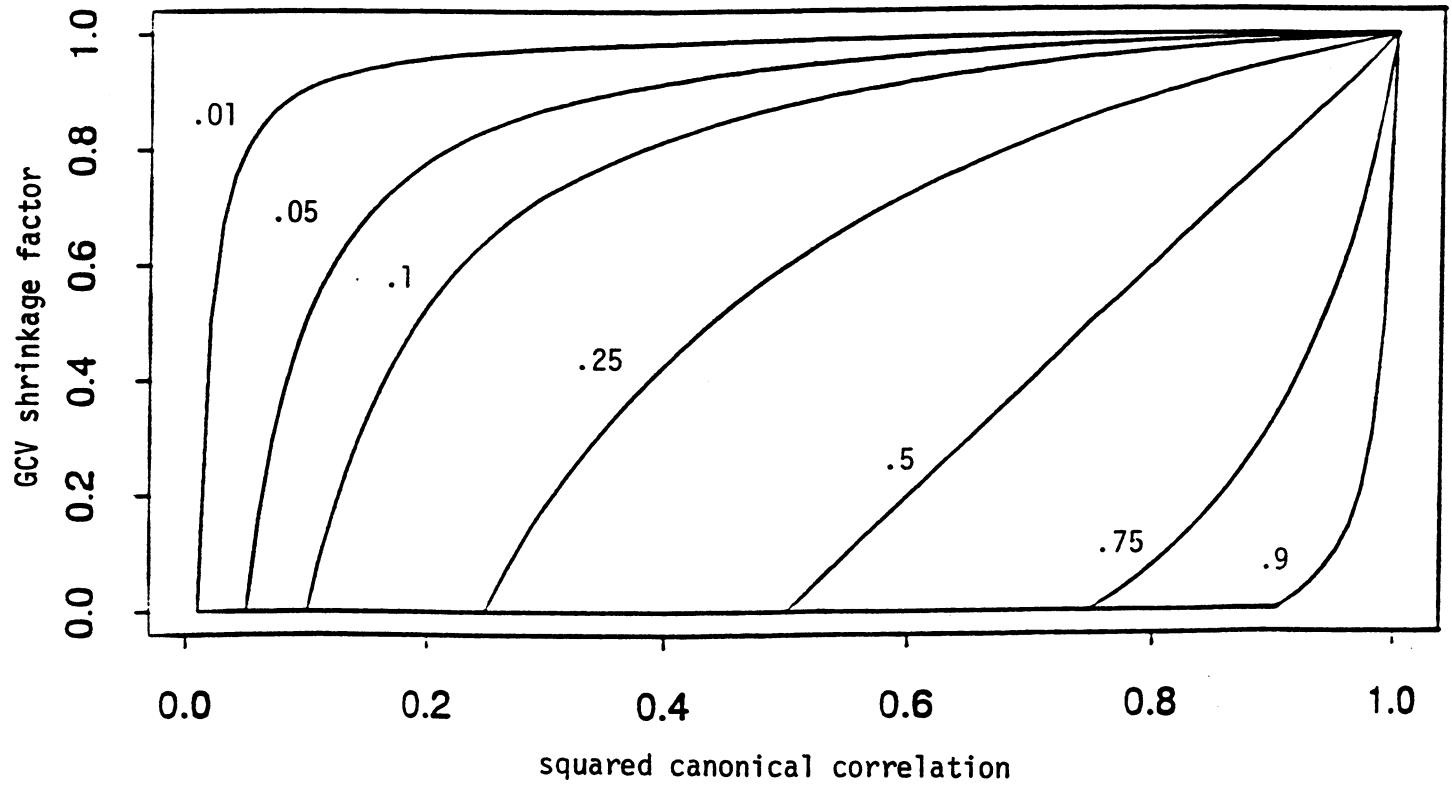


Figure 3

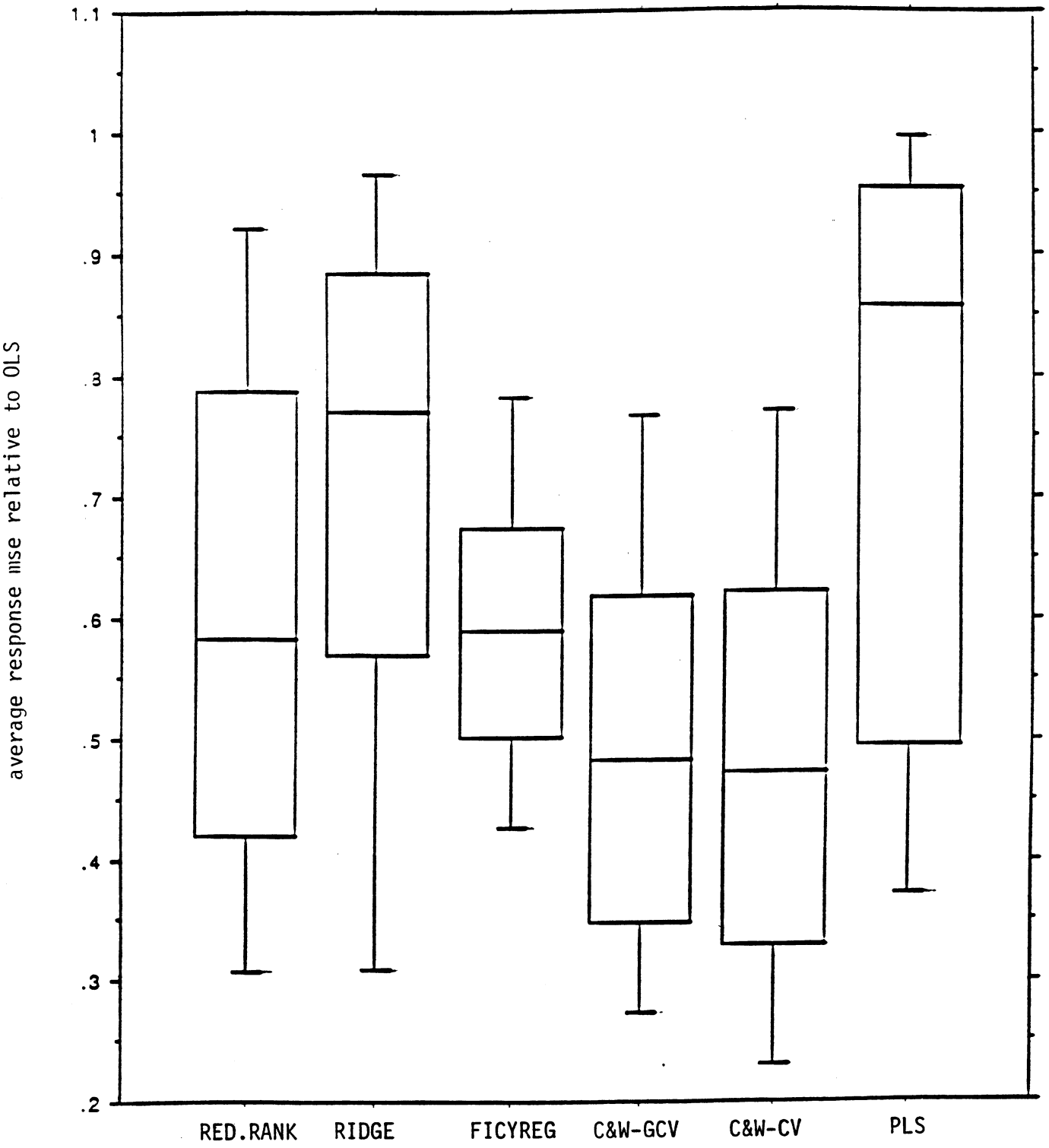

average response mse relative to OLS
1.1
1
.9
.8
.7
.6
.5
.4
.3
.2
RED.RANK
RIDGE
FICYREG
C&W-GCV
C&W-CV
PLS

Figure 4

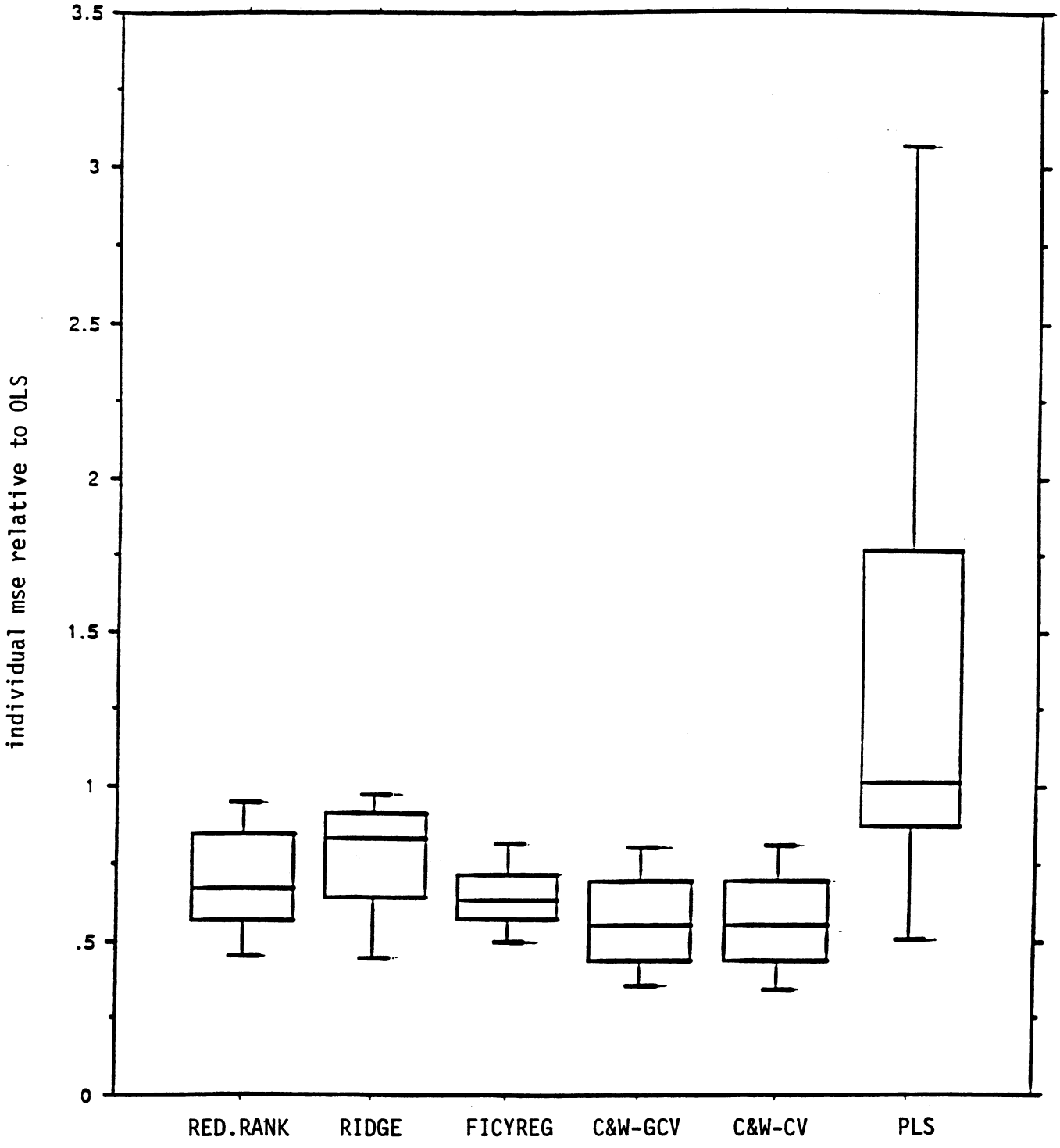

3.5
3
2.5
2
1.5
1
.5
0
individual mse relative to OLS
RED.RANK
RIDGE
FICYREG
C&W-GCV
C&W-CV
PLS

Figure 5

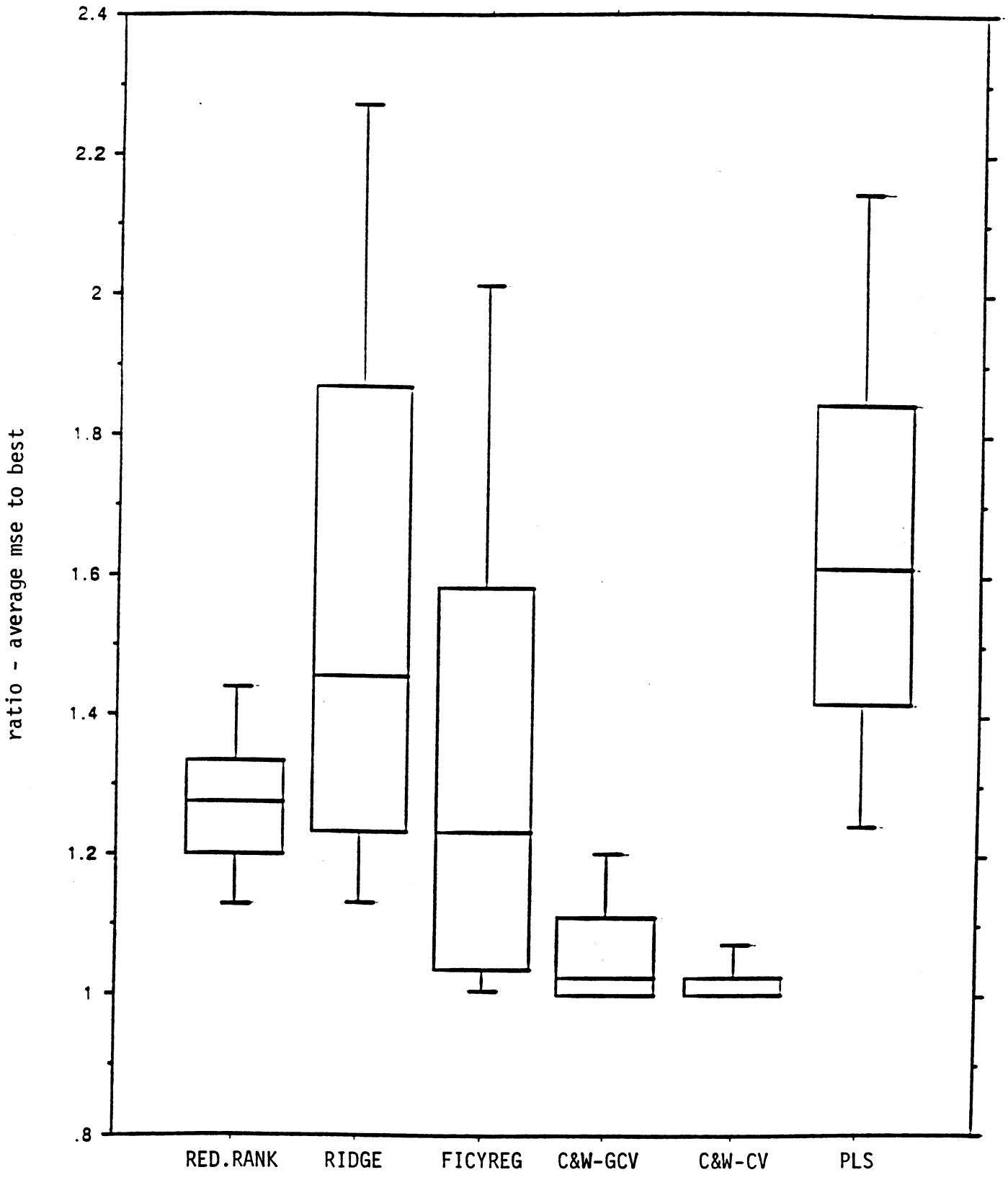

2.4
2.2
2
1.8
1.6
1.4
1.2
1
.8
ratio - average mse to best
RED.RANK
RIDGE
FICYREG
C&W-GCV
C&W-CV
PLS

Figure 6

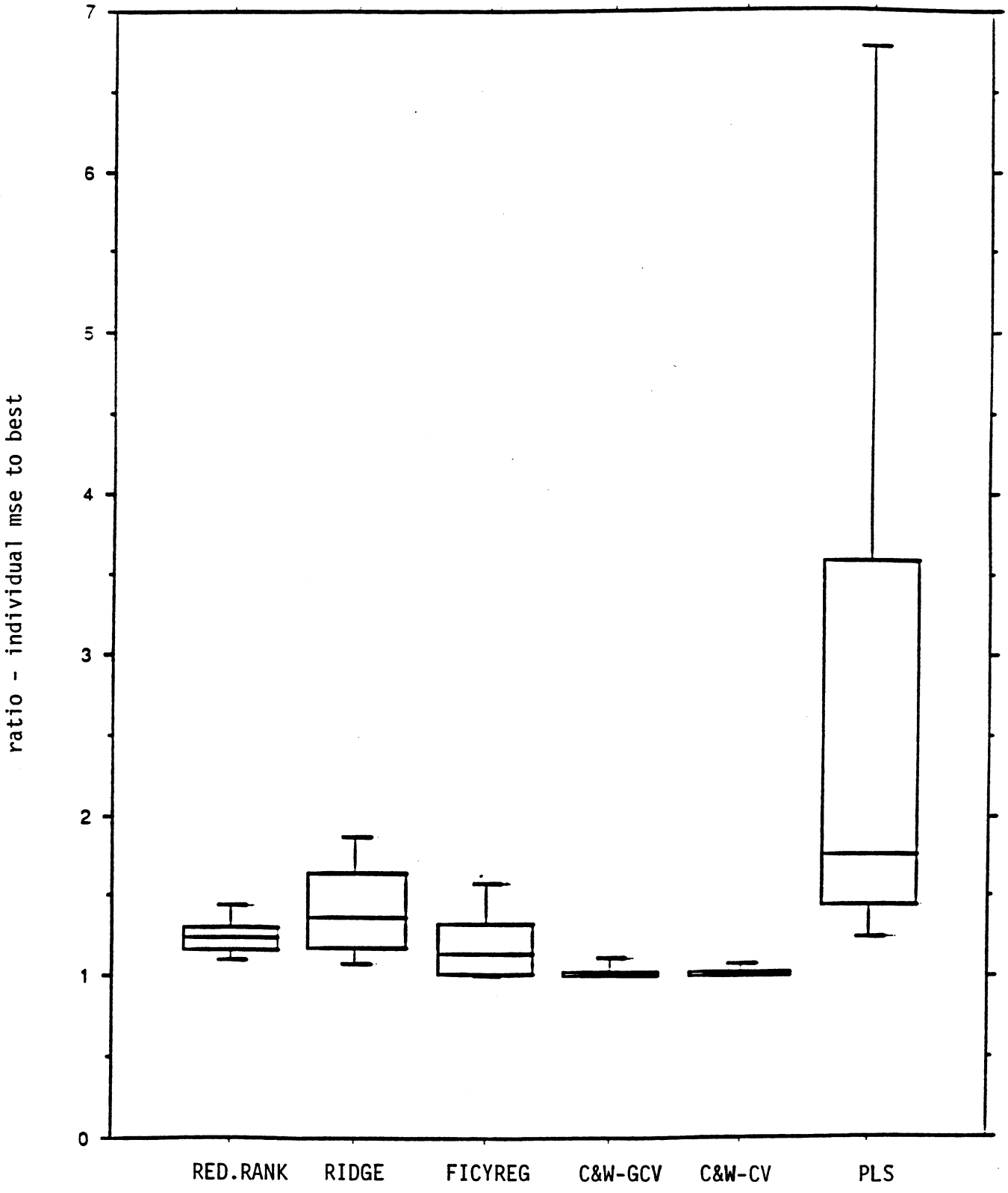

ratio - individual mse to best
0
1
2
3
4
5
6
7
RED.RANK
RIDGE
FICYREG
C&W-GCV
C&W-CV
PLS

Figure 7

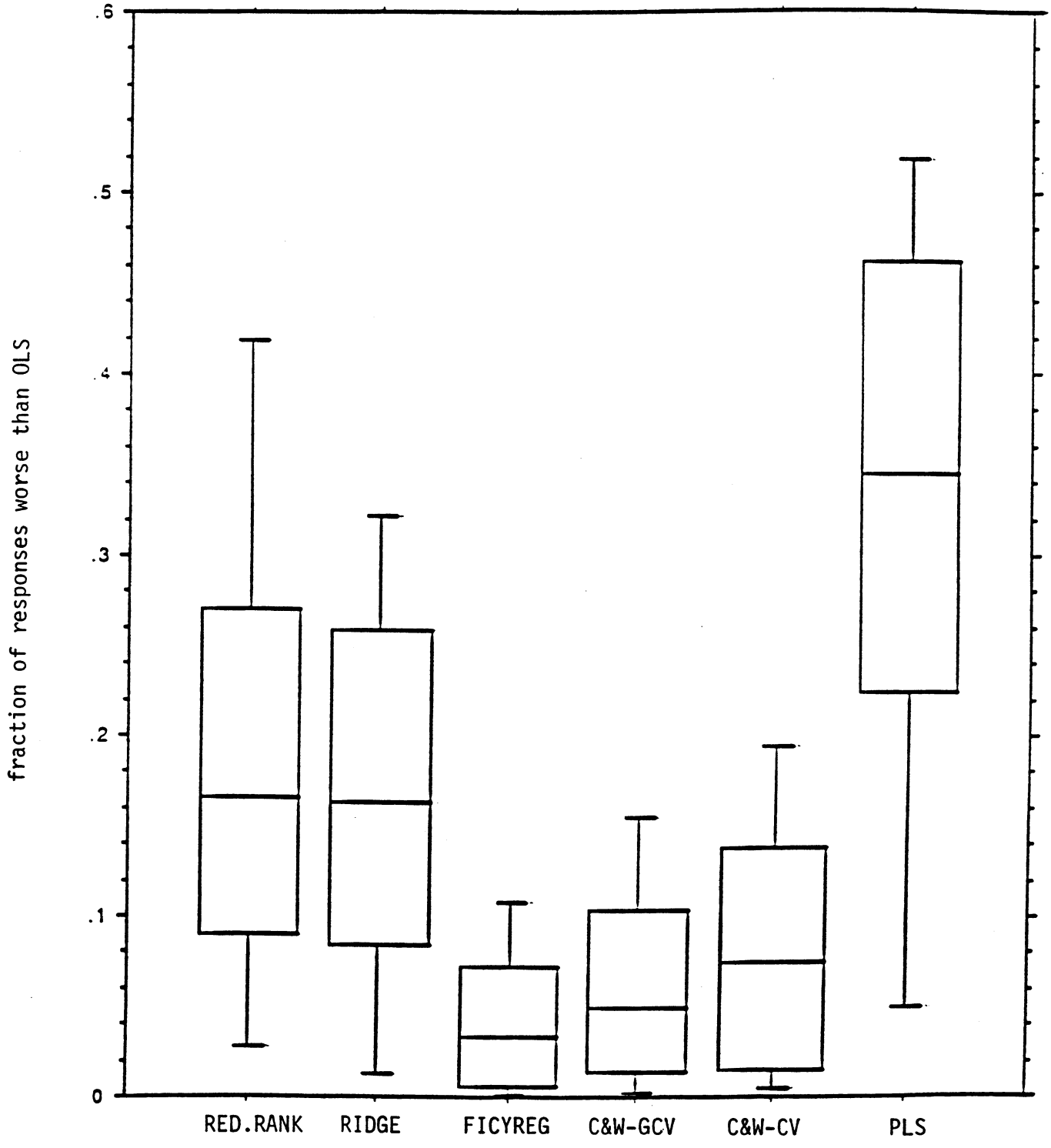

fraction of responses worse than OLS
.6
.5
.4
.3
.2
.1
0
RED.RANK
RIDGE
FICYREG
C&W-GCV
C&W-CV
PLS

Figure 8

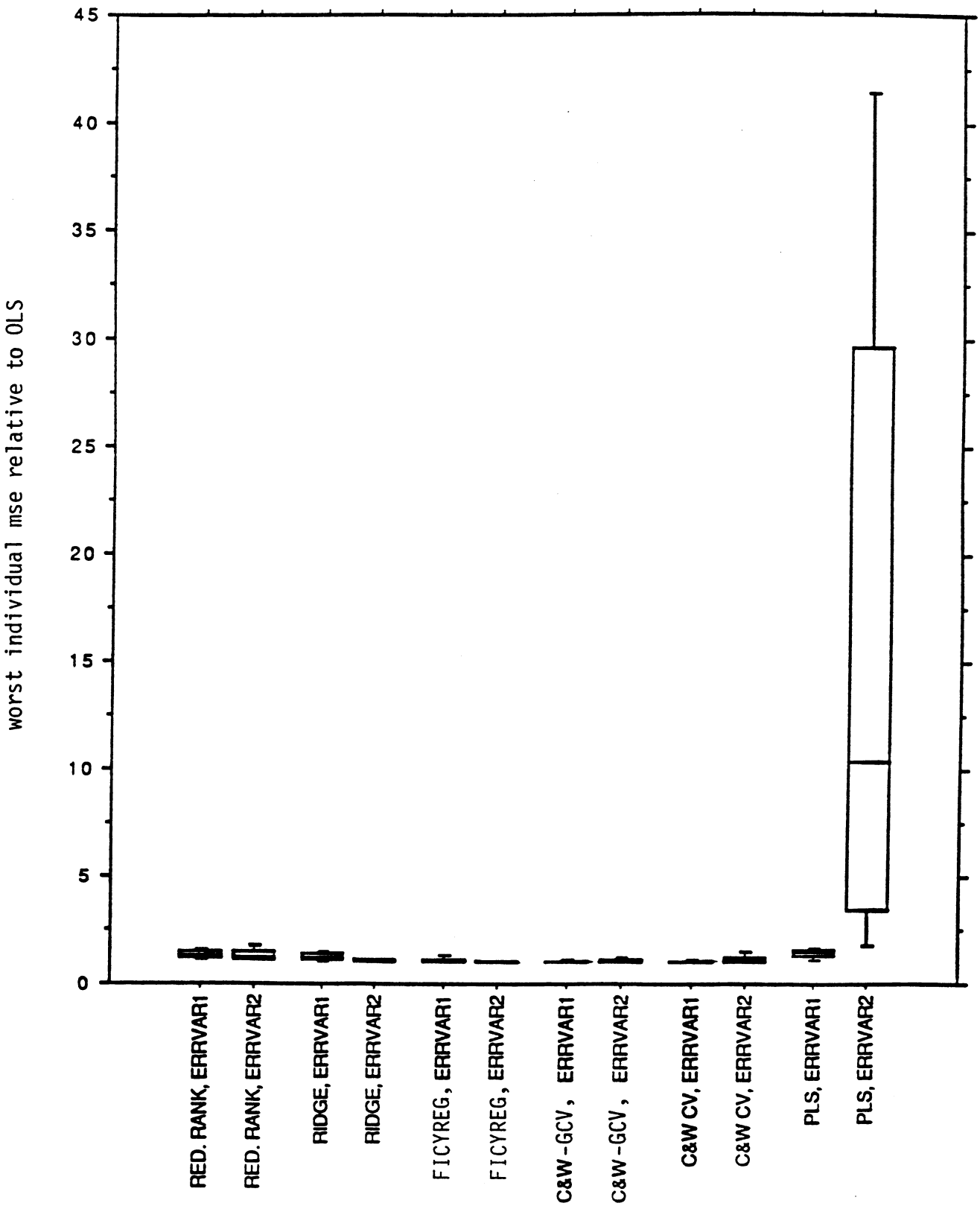

worst individual mse relative to OLS
45
40
35
30
25
20
15
10
5
0
RED. RANK, ERRVAR1
RED. RANK, ERRVAR2
RIDGE, ERRVAR1
RIDGE, ERRVAR2
FICYREG, ERRVAR1
FICYREG, ERRVAR2
C&W-GCV, ERRVAR1
C&W-GCV, ERRVAR2
C&W CV, ERRVAR1
C&W CV, ERRVAR2
PLS, ERRVAR1
PLS, ERRVAR2

Figure 9

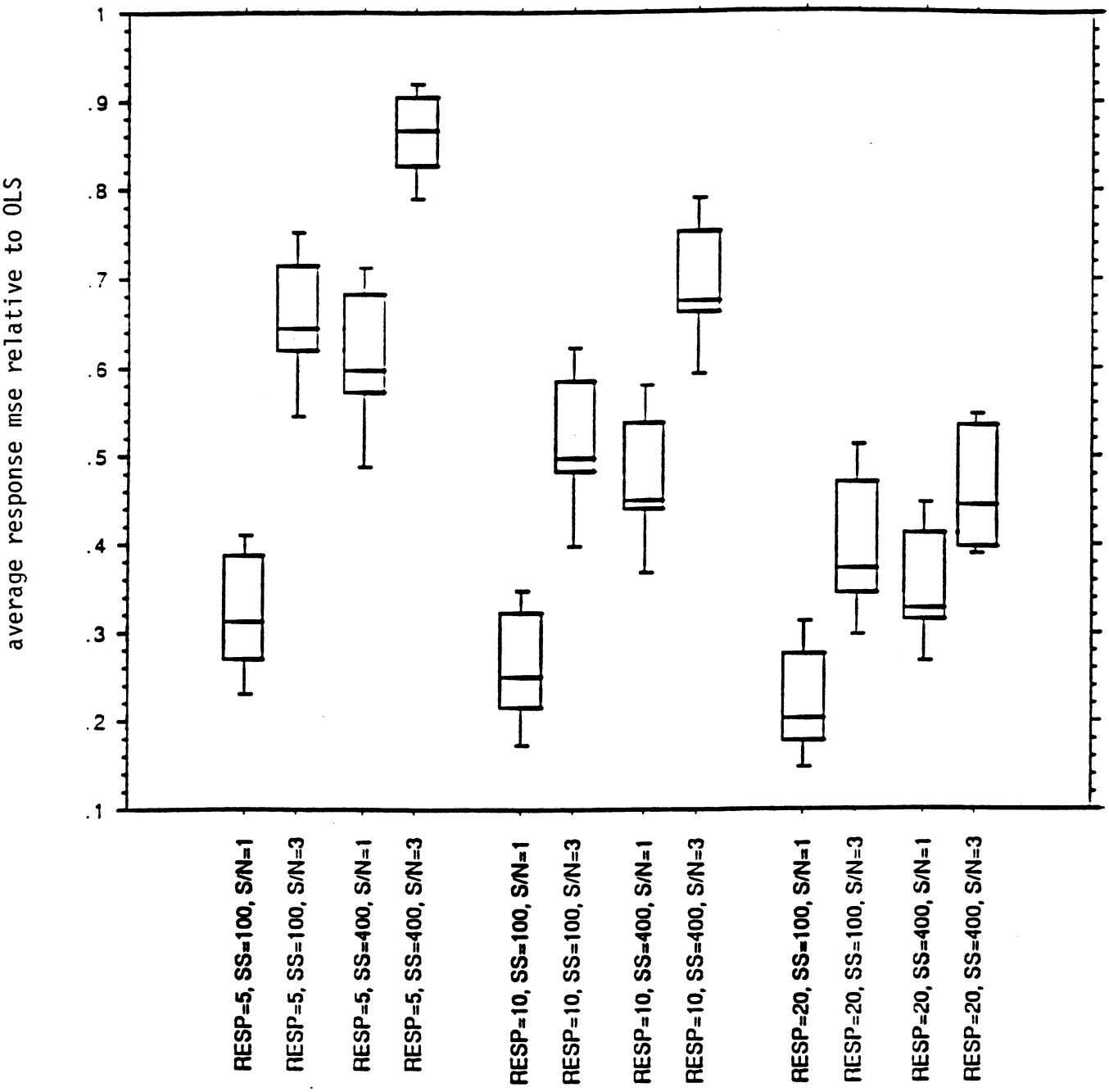

average response mse relative to OLS
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
RESP=5, SS=100, S/N=1
RESP=5, SS=100, S/N=3
RESP=5, SS=400, S/N=1
RESP=5, SS=400, S/N=3
RESP=10, SS=100, S/N=1
RESP=10, SS=100, S/N=3
RESP=10, SS=400, S/N=1
RESP=10, SS=400, S/N=3
RESP=20, SS=100, S/N=1
RESP=20, SS=100, S/N=3
RESP=20, SS=400, S/N=1
RESP=20, SS=400, S/N=3

Figure 10

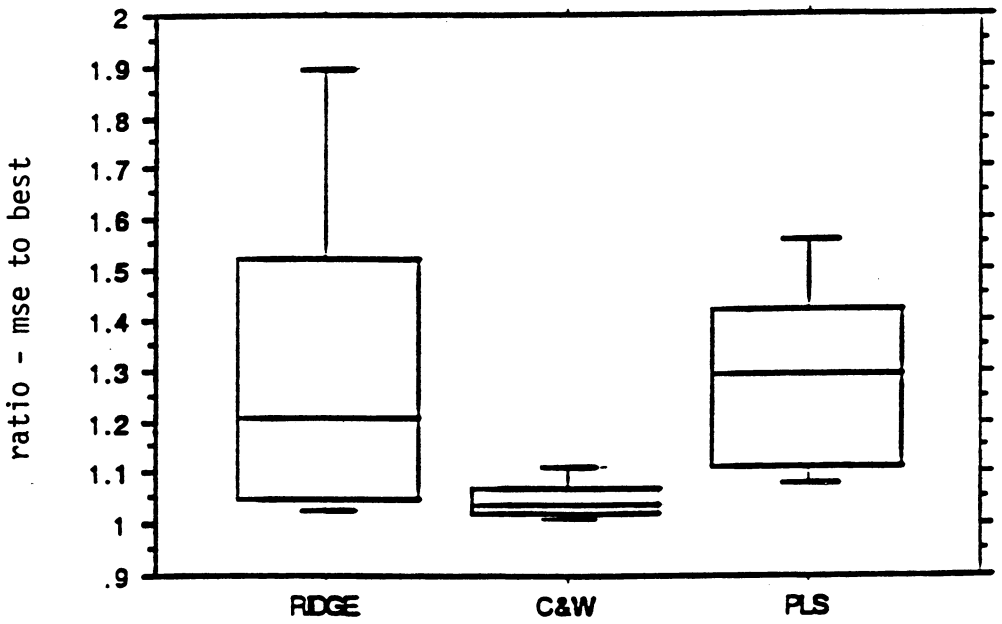


Figure 11