A Statistical View of Some Chemometrics Regression Tools

Author(s): Ildiko E. Frank and Jerome H. Friedman

Source: *Technometrics*, May, 1993, Vol. 35, No. 2 (May, 1993), pp. 109-135

Published by: Taylor & Francis, Ltd. on behalf of American Statistical Association and American Society for Quality

Stable URL: https://www.jstor.org/stable/1269656

A Statistical View of Some Chemometrics Regression Tools

Ildiko E. Frank

Jerll, Inc.
Stanford, CA 94305

Jerome H. Friedman

Department of Statistics
and
Stanford Linear Accelerator Center
Stanford University
Stanford, CA 94305

Chemometrics is a field of chemistry that studies the application of statistical methods to chemical data analysis. In addition to borrowing many techniques from the statistics and engineering literatures, chemometrics itself has given rise to several new data-analytical methods. This article examines two methods commonly used in chemometrics for predictive modeling—partial least squares and principal components regression—from a statistical perspective. The goal is to try to understand their apparent successes and in what situations they can be expected to work well and to compare them with other statistical methods intended for those situations. These methods include ordinary least squares, variable subset selection, and ridge regression.

KEY WORDS: Multiple response regression; Partial least squares; Principal components regression; Ridge regression; Variable subset selection.

1. INTRODUCTION

Statistical methodology has been successfully applied to many types of chemical problems for some time. For example, experimental design techniques have had a strong impact on understanding and improving industrial chemical processes. Recently the field of chemometrics has emerged with a focus on analyzing observational data originating mostly from organic and analytical chemistry, food research, and environmental studies. These data tend to be characterized by many measured variables on each of a few observations. Often the number of such variables p greatly exceeds the observation count N. There is generally a high degree of collinearity among the variables, which are often (but not always) digitizations of analog signals.

Many of the tools employed by chemometricians are the same as those used in other fields that produce and analyze observational data and are more or less well known to statisticians. These tools include data exploration through principal components and cluster analysis, as well as modern computer graphics. Predictive modeling (regression and classification) is also an important goal in most applications. In this area, however, chemometricians have invented their own techniques based on heuristic reasoning and intuitive ideas, and there is a growing body of empirical evidence that they perform well in many situations. The most popular regression method in chemometrics is partial least squares (PLS) (H. Wold 1975) and, to a somewhat lesser extent, principal components regression (PCR) (Massy 1965). Although PLS is heavily promoted (and used) by chemometricians, it is largely unknown to statisticians. PCR is known to, but seldom recommended by, statisticians. [The *Journal of Chemometrics* (John Wiley) and *Chemometrics and Intelligent Laboratory Systems* (Elsevier) contain many articles on regression applications to chemical problems using PCR and PLS. See also Martens and Naes (1989).]

The original ideas motivating PLS and PCR were entirely heuristic, and their statistical properties remain largely a mystery. There has been some recent progress with respect to PLS (Helland 1988; Lorber, Wangen, and Kowalski 1987; Phatak, Reilly, and Penlidis 1991; Stone and Brooks 1990). The purpose of this article is to view these procedures from a statistical perspective, attempting to gain some insight as to when and why they can be expected to work well. In situations for which they do perform well, they are compared to standard statistical methodology intended for those situations. These include ordinary least squares (OLS) regression, variable subset selection (VSS) methods, and ridge regression (RR) (Hoerl and Kennard 1970). The goal is to bring all of these methods together into a common framework to attempt to shed some light on their similarities and differences. The characteristics of PLS in particular have so far eluded theoretical understanding. This has led to unsubstantiated claims concerning its performance relative to other regression procedures, such as that it makes fewer assumptions concerning the nature of the data. Simply not understanding the nature of the assumptions being made does not mean that they do not exist.

Space limitations force us to limit our discussion here to methods that so far have seen the most use in practice. There are many other suggested approaches [e.g., latent root regression (Hawkins 1973; Webster, Gunst, and Mason 1974), intermediate least squares (Frank 1987), James–Stein shrinkage (James and Stein 1961), and various Bayes and empirical Bayes methods] that, although potentially promising, have not yet seen wide applications.

1.1 Summary Conclusions

RR, PCR, and PLS are seen in Section 3 to operate in a similar fashion. Their principal goal is to shrink the solution coefficient vector away from the OLS solution toward directions in the predictor-variable space of larger sample spread. Section 3.1 provides a Bayesian motivation for this under a prior distribution that provides no information concerning the direction of the true coefficient vector—all directions are equally likely to be encountered. Shrinkage away from low spread directions is seen to control the variance of the estimate. Section 3.2 examines the relative shrinkage structure of these three methods in detail. PCR and PLS are seen to shrink more heavily away from the low spread directions than RR, which provides the optimal shrinkage (among linear estimators) for an equidirection prior. Thus PCR and PLS make the assumption that the truth is likely to have particular preferential alignments with the high spread directions of the predictor-variable (sample) distribution. A somewhat surprising result is that PLS (in addition) places increased probability mass on the true coefficient vector aligning with the Kth principal component direction, where K is the number of PLS components used, in fact expanding the OLS solution in that direction. The solutions and hence the performance of RR, PCR, and PLS tend to be quite similar in most situations, largely because they are applied to problems involving high collinearity in which variance tends to dominate the bias, especially in the directions of small predictor spread, causing all three methods to shrink heavily along those directions. In the presence of more symmetric designs, larger differences between them might well emerge.

The most popular method of regression regularization used in statistics, VSS, is seen in Section 4 to make quite different assumptions. It is shown to correspond to a limiting case of a Bayesian procedure in which the prior probability distribution places all mass on the original predictor variable (coordinate) axes. This leads to the assumption that the response is likely to be influenced by a few of the predictor variables but leaves unspecified which ones. It will therefore tend to work best in situations characterized by true coefficient vectors with components consisting of a very few (relatively) large (absolute) values.

Section 5 presents a simulation study comparing the performance of OLS, RR, PCR, PLS, and VSS in a variety of situations. In all of the situations studied, RR dominated the other methods, closely followed by PLS and PCR, in that order. VSS provided distinctly inferior performance to these but still considerably better than OLS, which usually performed quite badly.

Section 6 examines multiple-response regression, investigating the circumstances under which considering all of the responses together as a group might lead to better performance than a sequence of separate regressions of each response individually on the predictors. Two-block multiresponse PLS is analyzed. It is seen to bias the solution coefficient vectors away from low spread directions in the predictor variable space (as would a sequence of separate PLS regressions) but also toward directions in the predictor space that preferentially predict the high spread directions in the response-variable space. An (empirical) Bayesian motivation for this behavior is developed by considering a joint prior on all of the (true) coefficient vectors that provides information on the degree of similarity of the dependence of the responses on the predictors (through the response correlation structure) but no information as to the particular nature of those dependences. This leads to a multiple-response analog of RR that exhibits similar behavior to that of two-block PLS. The two procedures are compared in a small simulation study in which multiresponse ridge slightly outperformed two-block PLS. Surprisingly, however, neither did dramatically better than the corresponding uniresponse procedures applied separately to the individual responses, even though the situations were designed to be most favorable to the multiresponse methods.

Section 7 discusses the invariance properties of these regression procedures. Only OLS is equivariant under all nonsingular affine (linear—rotation and/or scaling) transformations of the variable axes. RR, PCR, and PLS are equivariant under rotation but not scaling. VSS is equivariant under scaling but not rotation. These properties are seen to follow from the nature of the (informal) priors and loss structures associated with the respective procedures.

Finally, Section 8 provides a short discussion of interpretability issues.

2. REGRESSION

Regression analysis on observational data forms a major part of chemometric studies. As in statistics, the goal is to model the predictive relationships of a set of q response variables $\mathbf{y} = \{y_1 \ldots y_q\}$ on a set of p predictor variables $\mathbf{x} = \{x_1 \ldots x_p\}$ given a set of N (training) observations

$$\{y_i, x_i\}_1^N = \{y_{1i} \ldots y_{qi}, x_{1i} \ldots x_{pi}\}_1^N \tag{1}$$

on which all of the variables have been measured. This model is then used both as a descriptive statistic for interpreting the data and as a prediction rule for estimating likely values of the response variables when only values of the predictor variables are available. The structural form of the predictive relationship is taken to be linear:

$$y_j = a_{j0} + \sum_{k=1}^{p} a_{jk}x_k, \qquad j = 1, q. \tag{2}$$

The problem then is to use the training data (1) to estimate the values of the coefficients $\{a_{jk}\}_{j=1}^{q}{}_{k=0}^{p}$ appearing in Model (2).

In nearly all chemometric analyses, the variables are standardized ("autoscaled"):

$$\begin{aligned} y_j &\leftarrow (y_j - \bar{y}_j)/[\text{ave}(y_j - \bar{y}_j)^2]^{1/2} \\ x_k &\leftarrow (x_k - \bar{x}_k)/[\text{ave}(x_k - \bar{x}_k)^2]^{1/2}, \end{aligned} \tag{3}$$

with

$$\begin{aligned} \bar{y}_j &= \text{ave}(y_j) \\ \bar{x}_k &= \text{ave}(x_k), \end{aligned} \tag{4}$$

where the averages are taken over the training data (1); that is,

$$\text{ave}(\eta) = \frac{1}{N}\sum_{i=1}^{N} \eta_i,$$

where η is the quantity being averaged. (This notational convention will be used throughout the article.) The analysis is then applied to the standardized variables and the resulting solutions transformed back to reference the original locations and scales of the variables. The regression methods discussed later are always assumed to include constant terms (2), thus making them invariant with respect to the variable locations so that translating them to all have zero means is simply a matter of convenience (or numerics). Most of these methods are not, however, invariant to the relative scaling of the variables so that choosing them to all have the same scale is a deliberate choice on the part of the user. A different choice would give rise to different estimated models. This is discussed further in Section 7.

After autoscaling the training data, the regression models (2) (on the training data) can be expressed as

$$y_j = \mathbf{a}_j^T\mathbf{x}, \qquad j = 1, q, \tag{5}$$

with the jth coefficient vector being $\mathbf{a}_j^T = (a_{j1} \ldots a_{jp})$ or in matrix notation

$$\mathbf{y} = A\mathbf{x} \tag{6}$$

with the $q \times p$ matrix of regression coefficients being

$$A = [a_{jk}]. \tag{7}$$

The dominant regression methods used in chemometrics are PCR and PLS. The corresponding methods most used by statisticians (in practice) are OLS, RR, and VSS. The goal of this article is to compare and contrast these methods in an attempt to identify their similarities and differences. The next section starts with brief descriptions of PCR, PLS, and RR. (It is assumed that the reader is familiar with OLS and the various implementations of VSS.) We consider first the case of only one response variable ($q = 1$), since most of their similarities and differences emerge in this simplified setting. Multivariate regression ($q > 1$) is discussed in Section 6.

2.1 Principal Components Regression

PCR (Massy 1965) has been in the statistical literature for some time, although it has seen relatively little use compared to OLS and VSS. It begins with the training-sample covariance matrix of the predictor variables

$$V = \text{ave}(\mathbf{x}\mathbf{x}^T) \tag{8}$$

and its eigenvector decomposition

$$V = \sum_{k=1}^{p} e_k^2 \mathbf{v}_k \mathbf{v}_k^T. \tag{9}$$

Here $\{e_k^2\}_1^p$ are the eigenvalues of V arranged in descending order ($e_1 \ge e_2 \ge \cdots \ge e_p$) and $\{\mathbf{v}_k\}_1^p$ their corresponding eigenvectors. PCR produces a sequence of regression models $\{\hat{y}_0 \ldots \hat{y}_R\}$ with

$$\hat{y}_K = \sum_{k=0}^{K} [\text{ave}(y\mathbf{v}_k^T\mathbf{x})/e_k^2]\mathbf{v}_k^T\mathbf{x}, \qquad K = 1, R, \tag{10}$$

with R being the rank of V (number of nonzero e_k^2). The Kth model (10) is just the OLS regression of y on the "variables" $\{z_k = \mathbf{v}_k^T\mathbf{x}\}_0^K$ with the convention that for $K = 0$ the model is just the response mean, $\hat{y}_0 = 0$ (3). The goal of PCR is to choose the particular model $\hat{y}_K$ with the lowest prediction mean squared error

$$K^* = \underset{0 \le K \le R}{\text{argmin}}\ \widetilde{\text{ave}}(y - \hat{y}_K)^2, \tag{11}$$

where $\widetilde{\text{ave}}$ is the average over future data, not part of the training sample. The quantity K can thus be considered a *meta parameter* of the procedure whose value is to be estimated from the training data through some model-selection procedure. In chemometrics, model selection is nearly always done through ordinary cross-validation (CV) (Stone 1974),

$$\hat{K} = \underset{0 \le K \le R}{\text{argmin}} \sum_{i=1}^{N} (y_i - \hat{y}_{K\backslash i})^2, \qquad (12)$$

where $\hat{y}_{K\backslash i}$ is the Kth model (10) computed on the training sample with the ith observation removed. There are many other model selection criteria in the statistics literature [e.g., generalized cross-validation (Craven and Wahba 1979), minimum descriptive length (Rissiden 1983), Bayesian information criterion (Schwartz 1978), Mallows's Cp (Mallows 1973), etc.] that can also be used. (A discussion of their relative merits is outside the scope of this article.)

2.2 Partial Least Squares Regression

PLS was introduced by Wold (H. Wold 1975) and has been heavily promoted in the chemometrics literature as an alternative to OLS in the poorly conditioned or ill-conditioned problems encountered there. It was presented in algorithmic form as a modification of the NIPALS algorithm (H. Wold 1966) for computing principal components. Like PCR, PLS produces a sequence of models $\{\hat{y}_K\}_1^R$ [R = rank V (8)] and estimates which one is best through CV (12). The particular set of models constituting the (ordered) sequence are, however, different from those produced by PCR. Wold's PLS algorithm is presented in Table 1. [To simplify the description, random-variable notation is adopted; that is, a single symbol is used to represent the collection of values (scalar or vector) of the corresponding quantity over the data, and the observation index is omitted. This convention is used throughout the article.]

At each step, K (For loop pass, lines 2–10) y residuals from the previous step (y_{K-1}) are partially regressed on $\mathbf{x}$ residuals from the previous step ($\mathbf{x}_{K-1}$). In the beginning (line 1) these residuals are initialized to the original (standardized) data. The partial regression consists of computing the covariance vector $\mathbf{w}_K$ (line 3) and then using it to form a linear combination z_K of the $\mathbf{x}$ residuals (line 4). The y residuals are then regressed on this linear combination (line 5), and the result is added to the model (line 6) and subtracted from the current y residuals to form the new y residuals y_K (line 7) for the next step. New $\mathbf{x}$ residuals ($\mathbf{x}_K$) are then computed (line 8) by subtracting from $\mathbf{x}_{K-1}$ its projection on z_k. The test (line 9) will cause the algorithm to terminate after R steps, where R is the rank of V (8).

Table 1. Wold's PLS Algorithm

```
(1)  Initialize: $y_0 \leftarrow y$; $\mathbf{x}_0 \leftarrow \mathbf{x}$; $\hat{y}_0 \leftarrow 0$
(2)  For $K = 1$ to $p$ do:
(3)      $\mathbf{w}_K = \text{ave}(y_{K-1}\mathbf{x}_{K-1})$
(4)      $z_K = \mathbf{w}_K^T\mathbf{x}_{K-1}$
(5)      $r_K = [\text{ave}(y_{K-1}z_K)/\text{ave}(z_K^2)]z_K$
(6)      $\hat{y}_K = \hat{y}_{K-1} + r_K$
(7)      $y_K = y_{K-1} - r_K$
(8)      $\mathbf{x}_K = \mathbf{x}_{K-1} - [\text{ave}(z_K\mathbf{x}_{K-1})/\text{ave}(z_K^2)]z_K$
(9)      if $\text{ave}(\mathbf{x}_K^T\mathbf{x}_K) = 0$ then Exit
(10) end For
```

This PLS algorithm produces a sequence of models $\{\hat{y}_K\}_0^R$ (line 1 and line 6) on successive passes through the For loop. The one ($\hat{y}_{\hat{K}}$) that minimizes the CV score (12) is selected as the PLS solution. Note that straightforward application of many of the competing model-selection criteria is not appropriate here since, unlike PCR and RR, PLS is not a linear modeling procedure; that is, the response values $\{y_i\}_1^N$ enter nonlinearly into the model estimates $\{\hat{y}_i\}_1^N$.

The algorithm in Table 1 is the one first proposed by Wold that defined PLS regression. Since its introduction, several different algorithms have been proposed that lead to the same sequence of models $\{\hat{y}_K\}_1^R$ (e.g., see Naes and Martens 1985; Wold, Ruhe, Wold, and Dunn 1984). Perhaps the most elegant formulation (Helland 1988) is shown in Table 2.

Table 2 shows that the Kth PLS model $\hat{y}_K$ can be obtained by an OLS regression (OLS – line 5) of the response y on the K linear combinations $\{z_k = (V^{k-1}\mathbf{s})^T\mathbf{x}\}_1^K$.

2.3 Ridge Regression

RR (Hoerl and Kennard 1970) was introduced as a method for stabilizing regression estimates in the presence of extreme collinearity, V (8) being singular or nearly so. The coefficients of the linear model (5) are taken to be the solution of a penalized least squares criterion with the penalty being proportional to the squared norm of the coefficient vector $\mathbf{a}$:

$$\hat{\mathbf{a}}_\lambda = \underset{\mathbf{a}}{\text{argmin}}[\text{ave}(y - \mathbf{a}^T\mathbf{x})^2 + \lambda\mathbf{a}^T\mathbf{a}]. \qquad (13)$$

The solution is

$$\hat{\mathbf{a}}_\lambda = (V + \lambda I)^{-1}\mathbf{s}, \qquad (14)$$

Table 2. Helland's PLS Algorithm

```
(1) $V = \text{ave}(\mathbf{x}\mathbf{x}^T)$
(2) $\mathbf{s} = \text{ave}(y\mathbf{x})$
(3) For $K = 1$ to $R$ do:
(4)     $\mathbf{s}_K = V^{K-1}\mathbf{s}$
(5)     $\hat{y}_K = \text{OLS}[y \text{ on } \{\mathbf{s}_k^T\mathbf{x}\}_1^K]$
(6) end For
```

with

$$\mathbf{s} = \text{ave}(y\mathbf{x}) \tag{15}$$

and I being the $p \times p$ identity matrix. The inverse of the (possibly) ill-conditioned predictor-variable covariance matrix V is thus stabilized by adding to V a multiple of I. The degree of stabilization is regulated by the value of the "ridge" parameter $\lambda > 0$. A value of $\lambda = \infty$ results in the model being the response mean $\hat{y} = 0$, whereas $\lambda = 0$ gives rise to the unregularized OLS estimates. A value for λ in any particular situation is generally obtained by considering it to be a meta parameter of the procedure and estimating it through some model-selection procedure such as CV. Since here the response values $\{y_i\}_1^N$ do enter linearly in the model estimates $\{\hat{y}_i\}_1^N$, any of the competing model-selection criteria can also be straightforwardly applied (see Golub, Heath, and Wahba 1979).

3. A COMPARISON OF PCR, PLS, AND RR

From their preceding algorithmic descriptions, it might appear that PCR, PLS, and RR are very different procedures leading to quite different model estimates. In this section we provide a heuristic comparison that suggests that they are, in fact, quite similar, in that they are all attempting to achieve the same operational goal in slightly different ways. That goal is to bias the solution coefficient vector $\mathbf{a}$ (5) away from directions for which the projected sample predictor variables have small spread; that is,

$$\text{var}(\mathbf{a}^T\mathbf{x}/|\mathbf{a}|) = \text{ave}(\mathbf{a}^T\mathbf{x}/|\mathbf{a}|)^2 = \text{small}, \tag{16}$$

where the average is over the training sample.

This comparison consists of regarding the regression procedure as a two-step process as in VSS (Stone and Brooks 1990); first a K-dimensional subspace of p-dimensional Euclidean space is defined, and then the regression is performed under the restriction that the coefficient vector $\mathbf{a}$ lies in that subspace:

$$\mathbf{a} = \sum_{k=1}^{K} a_k \mathbf{c}_k, \tag{17}$$

where the unit vectors $\{\mathbf{c}_k\}_1^K$ span the prescribed subspace with $\mathbf{c}_k^T\mathbf{c}_k = 1$. The regression procedures can be compared by the way in which they define the subspace $\{\mathbf{c}_k\}_1^K$ and the manner in which the (constrained) regression is performed.

First, consider OLS in this setup. Here the subspace is defined by the (single) unit vector that maximizes the sample correlation (squared) between the response and the corresponding linear combination of the predictor variables

$$\mathbf{c}_{\text{OLS}} = \underset{\mathbf{c}^T\mathbf{c}=1}{\text{argmax}}\ \text{corr}^2(y, \mathbf{c}^T\mathbf{x}); \tag{18}$$

the OLS solution is then a simple least squares regression of y on $\mathbf{c}_{\text{OLS}}^T\mathbf{x}$,

$$\hat{y}_{\text{OLS}} = [\text{ave}(y\mathbf{c}_{\text{OLS}}^T\mathbf{x})/\text{ave}(\mathbf{c}_{\text{OLS}}^T\mathbf{x})^2]\mathbf{c}_{\text{OLS}}^T\mathbf{x}. \tag{19}$$

RR can also be cast into this framework. As in OLS, the subspace is defined by a single unit vector, but the criterion that defines that vector is somewhat different:

$$\mathbf{c}_{\text{RR}} = \underset{\mathbf{c}^T\mathbf{c}=1}{\text{argmax}}\ \text{corr}^2(y, \mathbf{c}^T\mathbf{x}) \frac{\text{var}(\mathbf{c}^T\mathbf{x})}{\text{var}(\mathbf{c}^T\mathbf{x}) + \lambda}, \tag{20}$$

where λ is the ridge parameter [(13)–(14)]. The ridge solution is then taken to be a (shrinking) ridge regression of y on $\mathbf{c}_{\text{RR}}^T\mathbf{x}$ with the same value for the ridge parameter

$$\hat{y}_{\text{RR}} = \left[\frac{\text{ave}(y\mathbf{c}_{\text{RR}}^T\mathbf{x})}{\text{ave}(\mathbf{c}_{\text{RR}}^T\mathbf{x})^2 + \lambda}\right]\mathbf{c}_{\text{RR}}^T\mathbf{x}. \tag{21}$$

(See Appendix.)

PCR defines a sequence of K-dimensional subspaces each spanned by the first K eigenvectors (9) of V (8). Thus each $\mathbf{c}_k$ $(1 \le k \le R)$ is the solution to
c

$$\mathbf{c}_k(\text{PCR}) = \underset{\substack{\{\mathbf{c}^TV\mathbf{c}_l = 0\}_1^{k-1} \\ \mathbf{c}^T\mathbf{c}=1}}{\text{argmax}}\ \text{var}(\mathbf{c}^T\mathbf{x}). \tag{22}$$

The first constraint in (22) (V orthogonality) ensures that the linear combinations associated with the different solution vectors are uncorrelated over the training sample

$$\text{corr}(\mathbf{c}_k^T\mathbf{x}, \mathbf{c}_l^T\mathbf{x}) = 0, \qquad k \neq l. \tag{23}$$

As a consequence of this and the criterion (22), they also turn out to be orthogonal $\mathbf{c}_k^T\mathbf{c}_l = 0$, $k \neq l$. The Kth PCR model is given by a least squares regression of the response on the K linear combinations $\{\mathbf{c}_k^T\mathbf{x}\}_1^K$. Since they are uncorrelated (23), this reduces to the sum of univariate regressions on each one (10).

PLS regression also produces a sequence of K-dimensional subspaces spanned by successive unit vectors, and then the Kth PLS solution is obtained by a least squares fit of the response onto the corresponding K-linear combinations in a strategy similar to PCR. The only difference from PCR is in the criterion used to define the vectors that span the K-dimensional subspace and hence the corresponding linear combinations. The criterion that gives rise to PLS (Stone and Brooks 1990) is

$$\mathbf{c}_k(\text{PLS}) = \underset{\substack{\{\mathbf{c}^TV\mathbf{c}_l = 0\}_1^{k-1} \\ \mathbf{c}^T\mathbf{c}=1}}{\text{argman}}\ \text{corr}^2(y, \mathbf{c}^Tx)\text{var}(\mathbf{c}^T\mathbf{x}). \tag{24}$$

As with PCR the vectors $\mathbf{c}_k$(PLS) are constrained to be mutually V orthogonal so that the corresponding linear combinations are uncorrelated over the training sample (23). This causes the K-dimensional least squares fit to be equivalent to the sum of K univariate regressions on each linear combination separately, as with PCR. Unlike PCR, however, the $\{\mathbf{c}_k(\text{PLS})\}_1^K$ are not orthogonal owing to the different criterion (24) used to obtain them.

The OLS criterion (18) is invariant to the scale of the linear combination of $\mathbf{c}^T\mathbf{x}$ and gives an unbiased estimate of the coefficient vector and hence the regression model [(18)–(19)]. The criteria associated with RR (20), PCR (22), and PLS (24) all involve the scale of $\mathbf{c}^T\mathbf{x}$ through its sample variance, thereby producing biased estimates. The effect of this bias is to pull the solution coefficient vector away from the OLS solution toward directions in which the projected data (predictors) have larger spread. The degree of this bias is regulated by the value of the model-selection parameter.

For RR, setting $\lambda = 0$ [(20)–(21)] yields the unbiased OLS solution, whereas $\lambda > 0$ introduces increasing bias toward larger values of $\text{var}(\mathbf{c}^T\mathbf{x})$ (20) and increased shrinkage of the length of the solution coefficient vector (21). For small values of λ, the former effect is the most pronounced; for example, for $\lambda > 0$ the RR solution will have no projection in any subspace for which $\text{var}(\mathbf{c}^T\mathbf{x}) = 0$, and very little projection on subspaces for which it is small.

In PCR, the degree of bias is controlled by the value of K, the dimension of the constraining subspace spanned by $\{\mathbf{c}_k(\text{PCR})\}_1^K$ (22)—that is, the number of components K used (10). If $K = R$ [rank of V (8)], one obtains an unbiased OLS solution. For $K < R$, bias is introduced. The smaller the value of K, the larger the bias. As with RR, the effect of this bias is to draw the solution toward larger values of $\text{var}(\mathbf{c}^T\mathbf{x})$, where $\mathbf{c}$ is a unit vector in the direction of the solution coefficient vector $\mathbf{a}$ (5) ($\mathbf{c} = \mathbf{a}/|\mathbf{a}|$). This is because constraining $\mathbf{c}$ to lie in the subspace spanned by the first K eigenvectors of V [(8)–(9)] places a lower bound on the sample variance of $\mathbf{c}^T\mathbf{x}$,

$$\text{var}(\mathbf{c}^T\mathbf{x}) \geq e_K^2. \qquad (25)$$

Since the eigenvectors (and hence the subspaces) are ordered on decreasing values of e_K^2, increasing K has the effect of easing this restriction, thereby reducing the bias.

For PLS, the situation is similar to that of PCR. The degree of bias is regulated by K, the number of components used. For $K = R$, an unbiased OLS solution is produced. Decreasing K generally increases the degree of bias. An exception to this occurs when $V = I$ (totally uncorrelated predictor variables), in which case an unbiased OLS solution is reached for $K = 1$ and remains the same for all K (though for $K \geq 2$ the regressions are singular, all of the regressors being identical). This can be seen from the PLS criterion (24). In this case, $\text{var}(\mathbf{c}^T\mathbf{x}) = 1$ for all $\mathbf{c}$, and the PLS criterion reduces to that for OLS (18). With this exception, the effect of decreasing K is to attract the solution coefficient vector toward larger values of $\text{var}(\mathbf{c}^T\mathbf{x})$ as in PCR. For a given K, however, the degree of this attraction depends jointly on the covariance structure of the predictor variables and the OLS solution, which in turn depends on the sample response values. This fact is often presented as an argument in favor of PLS over PCR. Unlike PCR, there is no sharp lower bound on $\text{var}(\mathbf{c}^T\mathbf{x})$ for a given K. The behavior of PLS compared to PCR for changing K is examined in more detail in Section 3.2.

3.1 Bayesian Motivation

Inspection of the criteria used by RR (20), PCR (22), and PLS (24) shows that they all can be viewed as applying a penalty to the OLS criterion (18), where the penalty increases as $\text{var}(\mathbf{c}^T\mathbf{x})$ decreases. A natural question to ask is: Under what circumstances should this lead to improved performance over OLS? It is well known (James and Stein 1961) that OLS is inadmissible in that one can always achieve a lower mean squared estimation error with biased estimates. The important question is: When can these estimators substantially improve performance and which one can do it best?

Some insight into these questions can be provided by considering a (highly) idealized situation. Suppose that in reality

$$y = \boldsymbol{\alpha}^T\mathbf{x} + \varepsilon \qquad (26)$$

for some (true) coefficient vector $\boldsymbol{\alpha}$ and ε is an additive (iid) homoscedastic error, with zero expectation and variance σ^2,

$$E(\varepsilon) = 0, \qquad E(\varepsilon^2) = \sigma^2. \qquad (27)$$

Since all of the estimators being considered here are equivariant with respect to rotations in the predictor variable space (after standardization), we will consider (for convenience) the coordinate system in which the predictor variables are uncorrelated; that is,

$$V = \text{diag}(e_1^2 \ldots e_p^2). \qquad (28)$$

Let $\mathbf{a}$ be an estimate of $\boldsymbol{\alpha}$ (26); that is,

$$\hat{y}(\mathbf{x}) = \mathbf{a}^T\mathbf{x} \qquad (29)$$

for a given point $\mathbf{x}$ in the predictor space (not necessarily one of the training-sample points). Consider training samples for which the (sample) predictor covariance matrix V has the eigenvalues (28).

The mean squared error (MSE) of prediction at $\mathbf{x}$ is

$$\text{MSE}[\hat{y}(\mathbf{x})] = E_\varepsilon[\boldsymbol{\alpha}^T\mathbf{x} - \mathbf{a}^T\mathbf{x}]^2, \qquad (30)$$

with the expected value over the distribution of the errors ε (26). Since $\boldsymbol{\alpha}$ (the truth) is unknown, the MSE (at $\mathbf{x}$) for any particular estimator is also unknown. One can, however, consider various (prior) probability distributions $\pi(\boldsymbol{\alpha})$ on $\boldsymbol{\alpha}$ and compare the properties of different estimators when the relative probabilities of encountering situations for which a particular $\boldsymbol{\alpha}$ (26) occurs is given by that distribution. For a given $\pi(\boldsymbol{\alpha})$, the mean squared prediction error averaged over the situations it represents is

$$E_{\boldsymbol{\alpha}} E_{\varepsilon}[\boldsymbol{\alpha}^T\mathbf{x} - \mathbf{a}^T\mathbf{x}]^2. \tag{31}$$

A simple and relatively unrestrictive prior probability distribution is one that considers all coefficient vector directions $\boldsymbol{\alpha}/|\boldsymbol{\alpha}|$ equally likely; that is, the prior distribution depends only in its norm $|\boldsymbol{\alpha}|^2 = \boldsymbol{\alpha}^T\boldsymbol{\alpha}$,

$$\pi(\boldsymbol{\alpha}) = \pi(\boldsymbol{\alpha}^T\boldsymbol{\alpha}). \tag{32}$$

For this exercise, we will consider simple linear shrinkage estimates of the form

$$a_j = f_j\hat{\alpha}_j, \qquad j = 1, p, \tag{33}$$

where $\hat{\boldsymbol{\alpha}}$ is the OLS estimate and the $\{f_j\}_1^p$ are shrinkage factors taken to be independent of the sample response values. In this case, the mean squared prediction error becomes [(33) and (31)]

$$\text{MSE}[\hat{y}(\mathbf{x})] = E_{\boldsymbol{\alpha}} E_{\varepsilon}\left[\sum_{j=1}^{p} (\alpha_j - f_j\hat{\alpha}_j)x_j\right]^2. \tag{34}$$

Averaging over $\boldsymbol{\alpha}$ using the probability distribution given by (32) [taking advantage of the fact that $E_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}\boldsymbol{\alpha}^T) = I \cdot E_{\boldsymbol{\alpha}}|\boldsymbol{\alpha}|^2$, with I being the identity matrix] yields

$$\text{MSE}[\hat{y}(\mathbf{x})] = \sum_{j=1}^{p} [(1 - f_j)^2 E_{\boldsymbol{\alpha}}|\boldsymbol{\alpha}|^2/p + f_j^2\sigma^2/(Ne_j^2)]x_j^2. \tag{35}$$

Here (35) $E_{\boldsymbol{\alpha}}|\boldsymbol{\alpha}|^2$ is the expected value of the length of the coefficient vector $\boldsymbol{\alpha}$ under the prior (32), p is the number of predictor variables, σ^2 is the variance of the error term [(26)–(27)], N is the training-sample size, and $\{e_j^2\}_1^p$ are the eigenvalues of the (sample) predictor-variable covariance matrix (28), which in this case are the sample variances of the predictor variables due to our choice of coordinate system (28).

The two terms within the brackets (35) that contribute to the MSE at $\mathbf{x}$ have separate interpretations. The first term depends on (the distribution of the) truth ($\boldsymbol{\alpha}$) and is independent of the error variance or the predictor-variable distribution. It represents the bias (squared) of the estimate. The second term is independent of the nature of the true coefficient vector $\boldsymbol{\alpha}$ and depends only on the experimental situation—error variance and predictor-design sample. It is the variance of the estimate. Setting $\{f_j = 1\}_1^p$ (33) yields the least squares estimates, which are unbiased but have variance given by the second term in (35). Reducing any (or all) of the $\{f_j\}_1^p$ to a value less than 1 causes an increase in bias [first term (35)] but decreases the variance [second term (35)]. This is the usual bias variance trade-off encountered in nearly all estimation settings. [Setting any (or all) of the $\{f_j\}_1^p$ to a value greater than 1 increases *both* the bias squared *and* the variance.]

This expression (35) for the MSE (in a simplified setting) illustrates the important fact that justifies the qualitative behavior of RR, PCR, and PLS discussed previously, namely, the shrinking of the solution coefficient vector away from directions of low (sample) variance in the predictor-variable space. One sees from the second term in (35) that the contribution to the variance of the model estimate from a given (eigen) direction (x_j) is inversely proportional to the sample predictor variance e_j^2 associated with that direction. Directions with small spread in the predictor variables give rise to high variance in the model estimate.

The values of $\{f_j\}_1^p$that minimize the MSE (35) are

$$f_j^* = e_j^2/(e_j^2 + \lambda), \qquad j = 1, p \tag{36}$$

with

$$\lambda = p(\sigma^2/E_{\boldsymbol{\alpha}}|\boldsymbol{\alpha}|^2)/N. \tag{37}$$

The quantity λ [(36)–(37)] is the number of (predictor) variables times the square of the noise-to-signal ratio, divided by the training-sample size. Combining (33), (36), and (37) gives the optimal (minimal MSE) linear shrinkage estimates

$$a_j = \hat{\alpha}_j \cdot \frac{e_j^2}{e_j^2 + \lambda}, \qquad j = 1, p. \tag{38}$$

One sees that the unbiased OLS estimates $\{\hat{\alpha}_j\}_1^p$ are differentially shrunk with the relative amount of shrinkage increasing with decreasing predictor variable spread e_j. The amount of differential shrinkage is controlled by the quantity λ (37): The larger the value of λ, the more differential shrinkage, as well as more overall global shrinkage. The value of λ in turn is given by the inverse product of the signal/noise squared and the training-sample size.

It is important to note that this high relative shrinkage in directions of small spread in the (sample) predictor-design distribution enters only to control the variance and not because of any prior belief that the true coefficient vector $\boldsymbol{\alpha}$ (26) is likely to align with the high spread directions of predictor design. The prior distribution on $\boldsymbol{\alpha}$, $\pi(\boldsymbol{\alpha})$ (32), that leads to this result (38) places equal mass on all directions

$\boldsymbol{\alpha}/|\boldsymbol{\alpha}|$ and by definition has no preferred directions for the truth. Therefore, one can at least qualitatively conclude that the common property of RR, PCR, and PLS of shrinking their solutions away from low spread directions mainly serves to reduce the variance of their estimates, and this is what gives them generally superior performance to OLS. The results given by (35), (37), and (38) indicate that their degree of improvement (over OLS) will increase with decreasing signal-to-noise ratio and training-sample size and increasing collinearity as reflected by the disparity in the eigenvalues (28) of the predictor-variable covariance matrix [(8)–(9)].

It is well known that (38) is just RR as expressed in the coordinate system defined by the eigenvectors of the sample predictor-variable covariance matrix [(8)–(9)]. Thus these results show (again well known) that RR is a linear shrinkage estimator that is optimal (in the sense of MSE) among all linear shrinkage estimators for the prior $\pi(\boldsymbol{\alpha})$ assumed here (32) and λ (37) known. PCR is also a linear shrinkage estimator

$$a_j(\text{PCR}) = \hat{\alpha}_j \cdot I(e_j^2 - e_K^2), \qquad (39)$$

where K is the number of components used and the second factor $I(\cdot)$ takes the value 1 for nonnegative argument values and 0 otherwise. Thus RR dominates PCR for an equidirection prior (32). PLS is not a linear shrinkage estimator, so RR cannot be shown to dominate PLS through this argument.

3.2 Shrinking Structure

One way to attempt to gain some insight into the relative properties of RR, PCR, and PLS is to examine their respective shrinkage structures in various situations. This can be done by expanding their solutions in terms of the eigenvectors of the predictor-sample covariance matrix [(8)–(9)] and the OLS estimate $\hat{\boldsymbol{\alpha}}$:

$$\mathbf{a}(\text{RR}:\text{PCR}:\text{PLS}) = \sum_{j=1}^{p} f_j(\text{RR}:\text{PCR}:\text{PLS})\hat{\alpha}_j v_j. \qquad (40)$$

Here $\hat{\alpha}_j$ is the projection of the OLS solution on $\mathbf{v}_j$ (the jth eigenvector of V),

$$\hat{\alpha}_j = \text{ave}(y\mathbf{v}_j^T\mathbf{x})/e_j^2, \qquad (41)$$

and $\{f_j(\cdot)\}_1^p$ can be regarded as a set of factors along each of these eigendirections that scale the OLS solution for each of the respective methods. As shown in (36) and (39), $f_j(\text{RR}) = e_j^2/(e_j^2 + \lambda)$ and

$$\begin{aligned} f_j(\text{PCR}) &= 1 \quad e_j^2 \geq e_K^2 \\ &= 0 \quad e_j^2 < e_K^2, \end{aligned} \qquad (42)$$

both of which are linear in that they do not involve the sample response values $\{y_i\}_1^N$.

The corresponding scale factors for PLS are not linear in the response values. For a K-component solution, they can be expressed as

$$f_{jK}(\text{PLS}) = \sum_{k=1}^{K} \beta_k e_j^{2k}, \qquad (43)$$

where the vector $\boldsymbol{\beta} = \{\beta_k\}_1^K$ is given by $\boldsymbol{\beta} = W^{-1}\mathbf{w}$, with the K components of the vector $\mathbf{w}$ being

$$w_k = \sum_{j=1}^{p} \hat{\alpha}_j^2 e_j^{2(k+1)},$$

and the elements of the $K \times K$ matrix W are given by

$$W_{kl} = \sum_{j=1}^{p} \hat{\alpha}_j^2 e_j^{2(k+l+1)}.$$

They depend on the number of components K used and the eigenstructure $\{e_j^2\}_1^p$ (as do the factors for RR and PCR), but not in a simple way. They also depend on the OLS solution $\{\hat{\alpha}_j\}_1^p$, which in turn depends on the response values $\{y_i\}_1^N$. The PLS scale factors are seen to be independent of the length of the OLS solution $|\hat{\alpha}|^2$, depending only on the relative values of $\{\hat{\alpha}_j\}_1^p$. Note that for all of the methods studied here the estimates (for a given value of the meta parameter) depend on the data only through the vector of OLS estimates $\{\hat{\alpha}_j\}_1^p$ and the eigenvalues of the predictor-covariance matrix $\{e_j^2\}_1^p$.

Although the scale factors for PLS (43) cannot be expressed by a simple formula (as can those for RR and PCR), they can be computed for given values of K, $\{e_j^2\}_1^p$, and $\{\hat{\alpha}_j\}_1^p$ and compared to those of RR and PCR [(36) and (42)] for corresponding situations. This is done in Figures 1–4, for $p = 10$. In each figure, the scale factors f_1 (PLS) – f_{10} (PLS) are plotted (in order—solid line) for the first six ($K = 1, 6$) component PLS models. Each of the four figures represents a different situation in terms of the relative values of $\{e_j^2\}_1^p$ and $\{\hat{\alpha}_j\}_1^p$. Also plotted in each frame for comparison are the corresponding shrinkage factors for RR (dashed line) and PCR (dotted line) for that situation, normalized so that they give the same overall shrinkage ($sh = |\mathbf{a}|/|\hat{\boldsymbol{\alpha}}|$); that is, for RR the ridge parameter λ (36) is chosen so that the length of the RR solution vector is the same as that for PLS ($|\mathbf{a}_{\text{RR}}| = |\mathbf{a}_{\text{PLS}}|$). In the case of PCR, the number of components was chosen so that the respective solution lengths were as close as possible ($|\mathbf{a}_{\text{PCR}}| \simeq |\mathbf{a}_{\text{PLS}}|$). The three numbers in each frame give the number of PLS components, the corresponding shrinkage factor ($sh = |\mathbf{a}|/|\hat{\boldsymbol{\alpha}}|$), and the ridge parameter (λ) that provides that overall shrink-

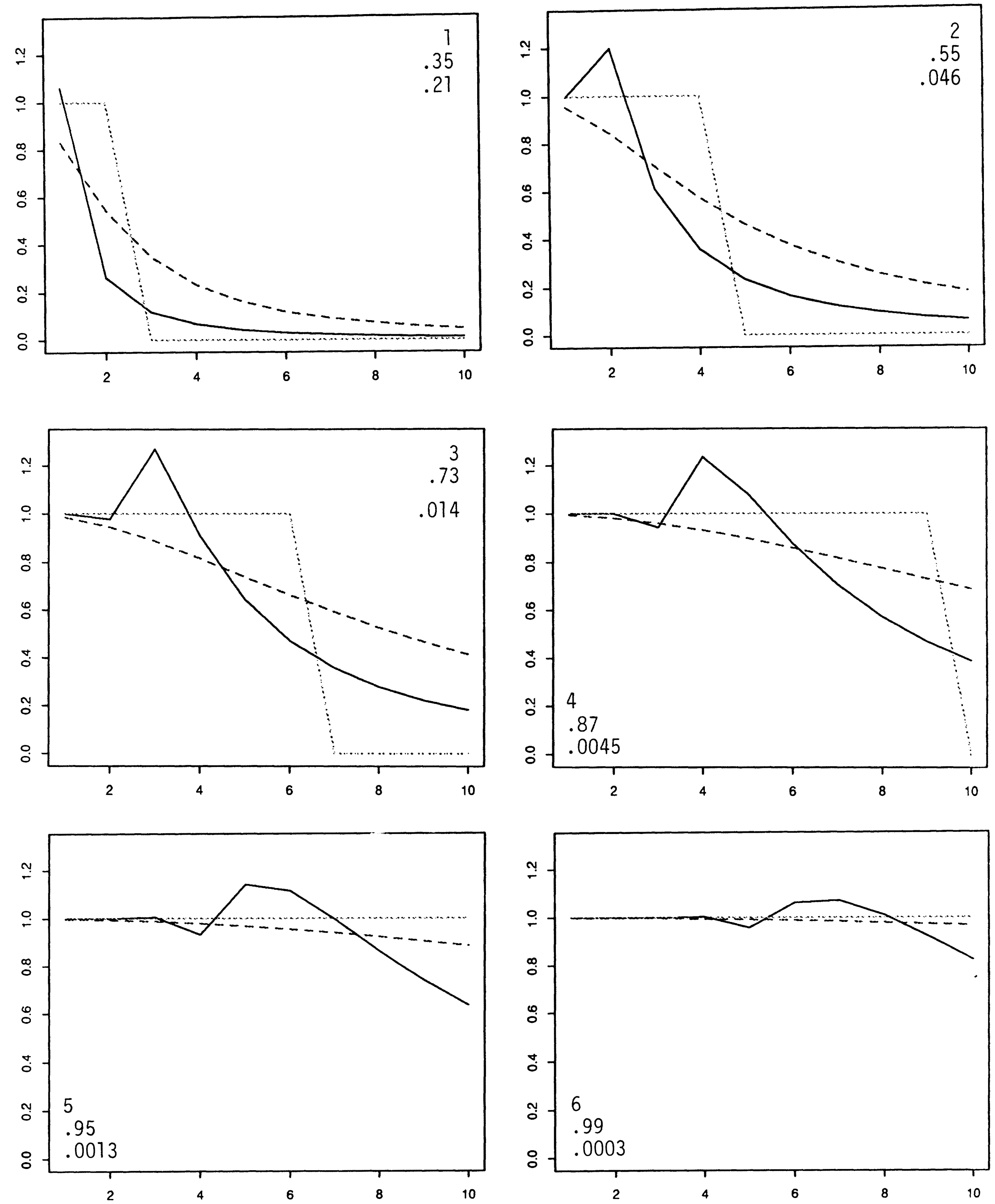

Figure 1. Scale Factors for PLS (solid), RR (dashed), and PCR (dotted) for Neutral Least Squares Solution and High Collinearity. Shown in each frame are the number of PLS components (upper entry), overall shrinkage (middle entry), and corresponding ridge parameter (lower entry).

age. The four situations represented in Figures 1–4 are as follows: $\{\hat{\alpha}_j = 1\}_1^p$ $\{e_j^2 \sim 1/j^2\}_1^p$ (neutral $\hat{\alpha}$'s, high collinearity), $\{\hat{\alpha}_j = 1\}_1^p$ $\{e_j^2 \sim 1/j\}$ (neutral $\hat{\alpha}$'s, moderate collinearity), $\{\hat{\alpha}_j = 1/j\}$ $\{e_j^2 \sim 1/j^2\}_1^p$ (favorable $\hat{\alpha}$'s, high collinearity), and $\{\hat{\alpha}_j = j\}_1^p$ $\{e_j^2 \sim 1/j^2\}_1^p$ (unfavorable $\hat{\alpha}$'s, high collinearity).

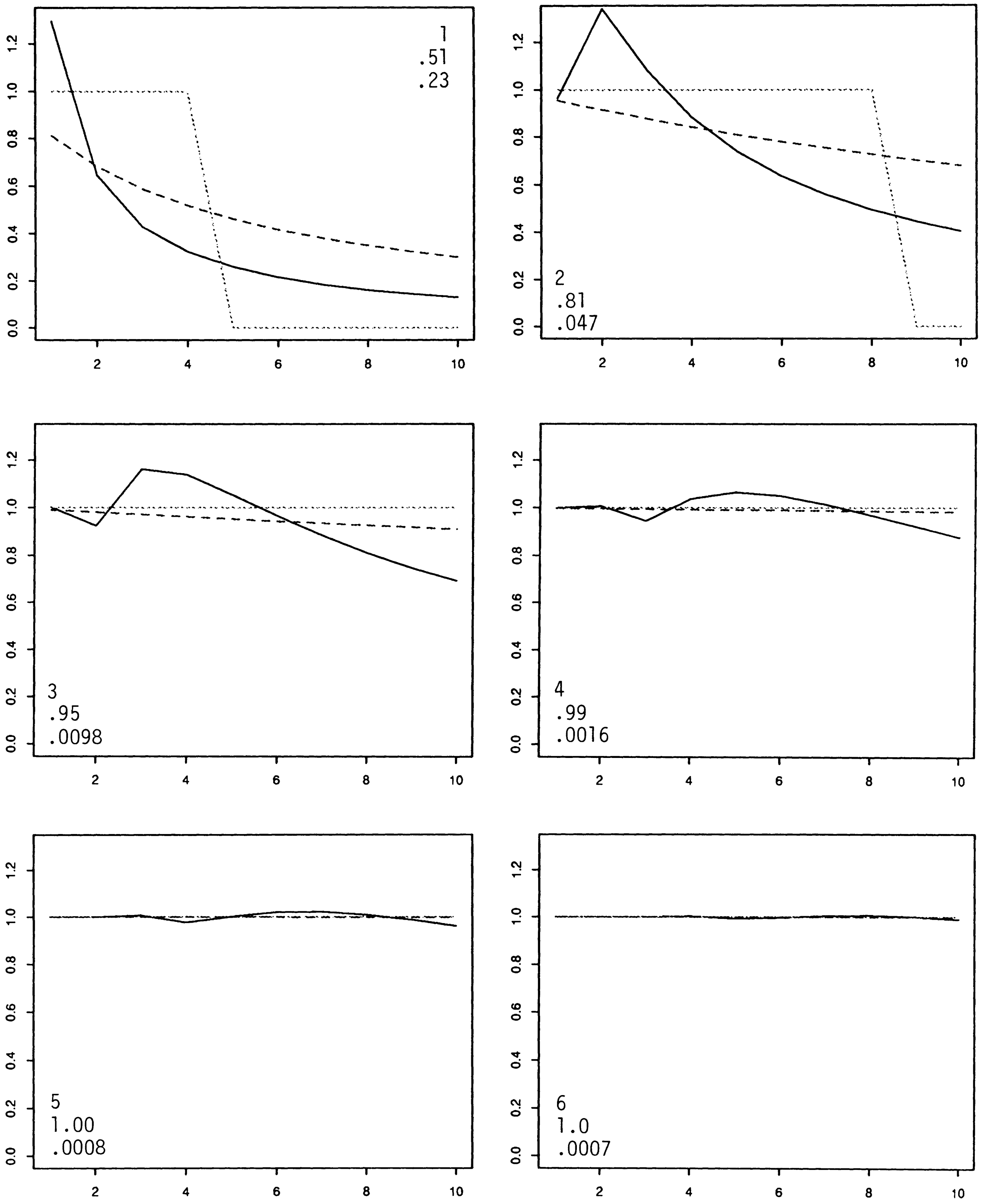

Figure 2. Scale Factors for PLS (solid), RR (dashed), and PCR (dotted) for Neutral Least Squares Solution and Moderate Collinearity. The entries in each frame correspond to those in Figure 1.

In Figure 1, the OLS solution is taken to project equally on all eigendirections (neutral) and the eigenvalue structure is taken to be highly peaked toward the larger values (high collinearity). The one-component PLS model ($K = 1$, upper left frame) is seen to dramatically shrink the OLS coefficients for

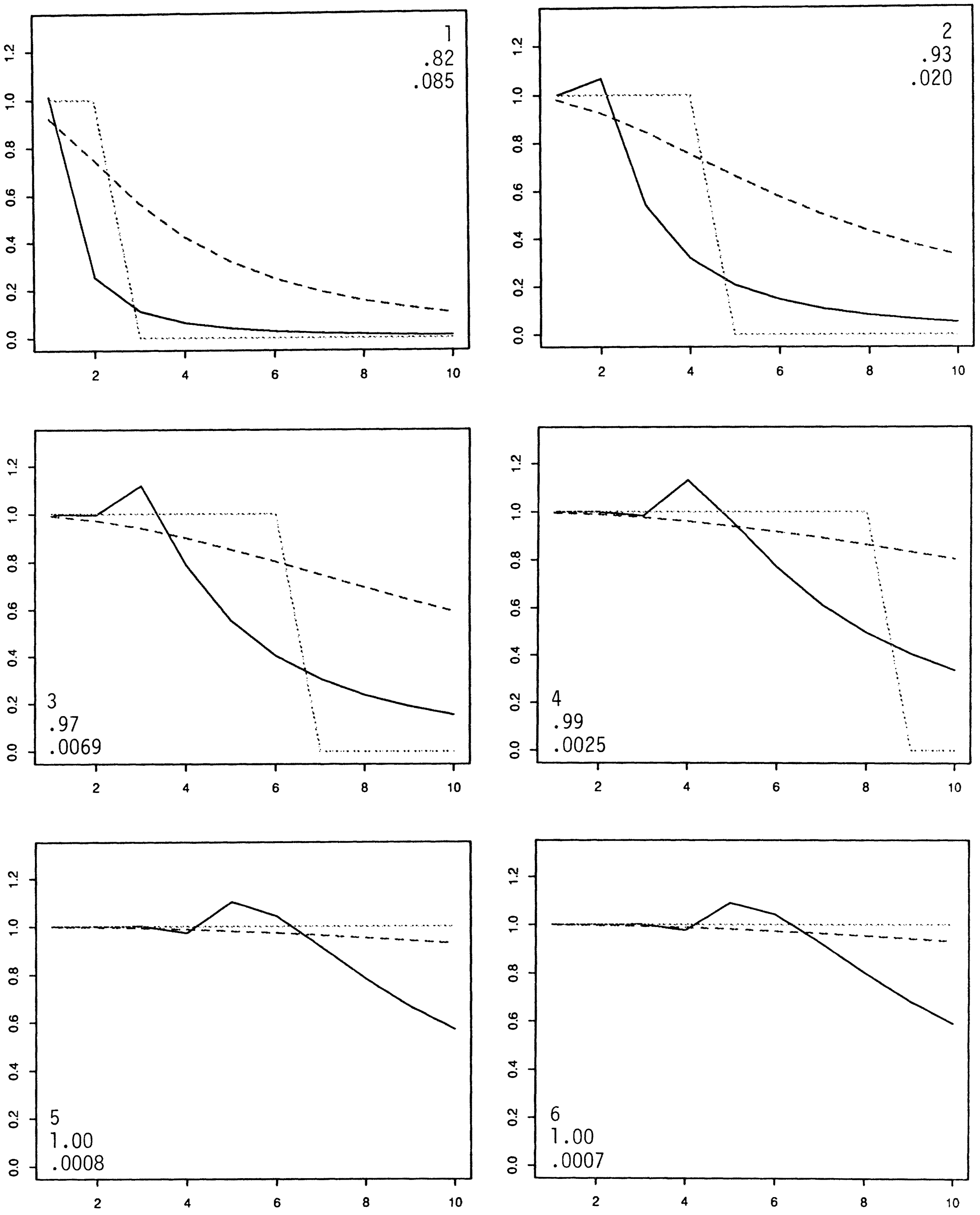

Figure 3. Scale Factors for PLS (solid), RR (dashed), and PCR (dotted) for Favorable Least Squares Solution and High Collinearity. The entries in each frame correspond to those in Figure 1.

the smallest eigendirections. It slightly "expands" the OLS coefficient for the largest (first) eigendirection, f_1(PLS) $>$ 1. The overall shrinkage is substantial; the length of the $K = 1$ PLS solution coefficient vector is about 35% of that for the OLS solution. For the same overall shrinkage, the relative shrink-

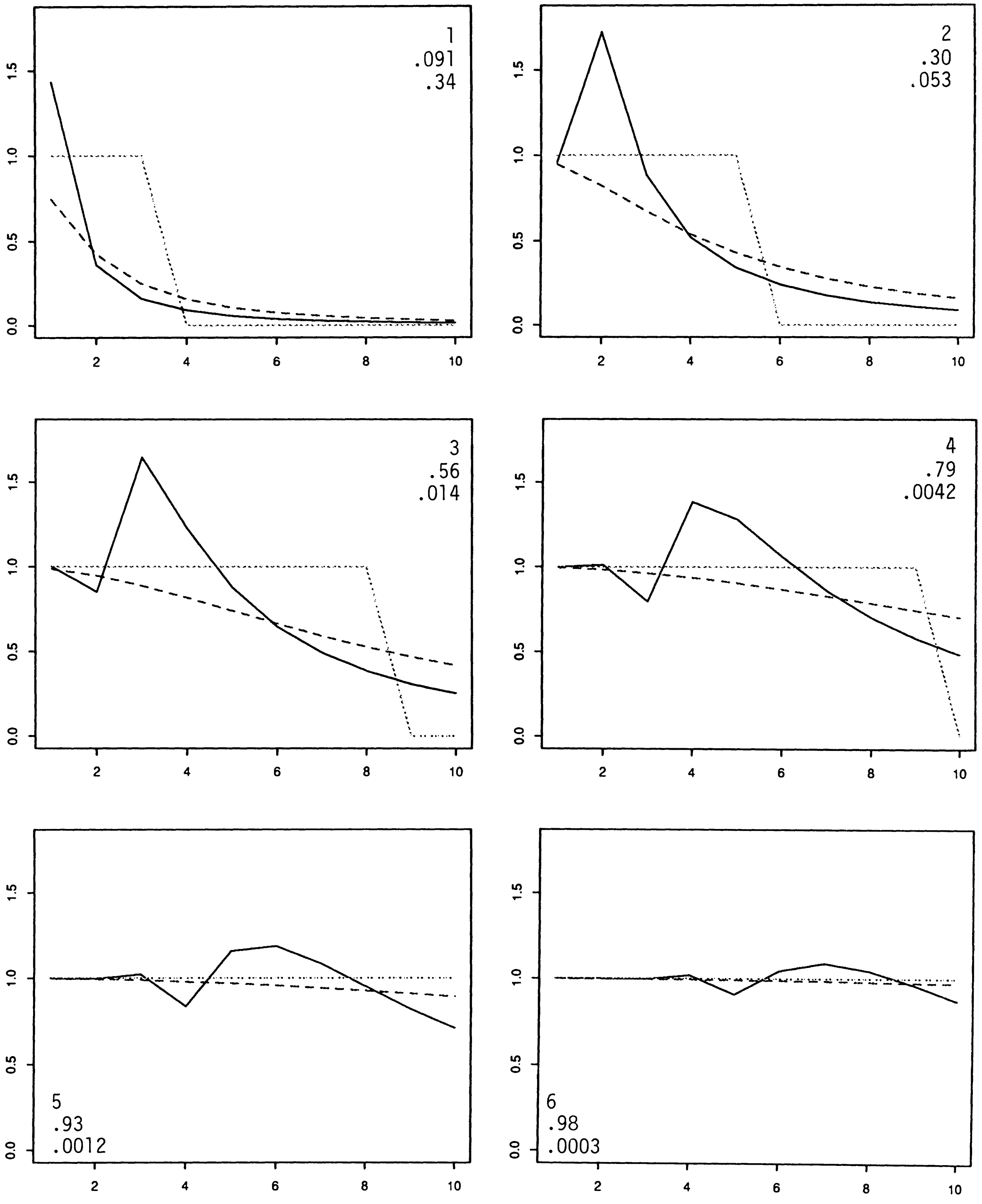

Figure 4. Scale Factors for PLS (solid), RR (dashed), and PCR (dotted) for Unfavorable Least Squares Solution and High Collinearity. The entries in each frame correspond to those in Figure 1.

age of RR tracks that of PLS but is somewhat more moderate. This is a consistent trend throughout all situations (Figs. 1–4). For PCR, a two-component model ($K = 2$) gives roughly the same overall shrinkage as the $K = 1$ PLS solution. Again this is a trend throughout all situations in that one gets roughly the same overall shrinkage for $K_{\text{PCR}} \simeq 2K_{\text{PLS}}$. As the number of PLS components is increased (left to right, top to bottom frames) the overall shrinkage applied to the OLS solution is reduced and the relative shrinkage applied to each eigendirection becomes more moderate. For $K = 6$, the PLS solution is very nearly the same as the OLS solution $\{f_j(\text{PLS}) \simeq 1\}_1^p$ even though it only becomes exactly so for $K = 10$. Again this feature is present throughout all situations (Figs. 1–4).

An interesting aspect of the PLS solution is that (unlike RR and PCR) it not only shrinks the OLS solution in some eigendirections ($f_j \le 1$) but expands it in others ($f_j > 1$). For a K-component PLS solution, the OLS solution is expanded in the subspace defined by the eigendirections associated with the eigenvalues closest to the Kth eigenvalue. Directions associated with somewhat larger eigenvalues tend to be slightly shrunk, and those with smaller eigenvalues are substantially shrunk. Again this behavior is exhibited throughout all of the situations studied here. The expression for the mean squared prediction error (35) suggests that, at least for linear estimators, using any $f_j > 1$ can be highly detrimental because it increases *both* the bias squared and the variance of the model estimate. This suggests that the performance of PLS might be improved by using modified scale factors $\{\tilde{f}_j(\text{PLS})\}_1^p$, where $\tilde{f}_j(\text{PLS}) \leftarrow \min(f_j(\text{PLS}), 1)$, although this is not certain since PLS is not linear and (35) was derived assuming linear estimates. It would, in any case, largely remove the preference of PLS for (true) coefficient vectors that align with the eigendirections whose eigenvalues are close to the Kth eigenvalue.

The situation represented in Figure 2 has the same (neutral) OLS solution but less collinearity. The qualitative behavior of the PLS, RR, and PCR scale factors are seen to be the same as that depicted in Figure 1. The principal difference is that PLS applies less shrinkage for the same number of components and (nearly) reaches the OLS solution for $K = 4$. Note that for no collinearity (all eigenvalues equal) PLS produces the OLS solution with the first component ($K = 1$).

Figures 3 and 4 examine the high collinearity situation for different OLS solutions. In Figure 3, the OLS solution is taken to be aligned with the major axes of the predictor design. The relative PLS shrinkage for different eigendirections for this favorable case is seen to be similar to that for the neutral case depicted in Figure 1. The overall shrinkage is much less, however, owing to the favorable orientation of the OLS solution. Figure 4 represents the contrasting situation in which the OLS solution is (unfavorably) aligned in orthogonal directions to the major axes of the predictor design. Here one sees qualitatively similar relative behavior as before, with a bit more exaggeration. Due to the unfavorable alignment of the OLS solution, the overall shrinkage here is quite considerable. Still the OLS solution is nearly reached by the $K = 6$-component PLS solution.

3.2.1. Discussion. Although the study represented by Figures 1–4 is hardly exhaustive, some tentative conclusions can be drawn. The qualitative behavior of RR, PCR, and PLS as deduced from (20), (22), and (24) is confirmed. They all penalize the solution coefficient vector **a** for projecting onto the low-variance subspace of the predictor design [i.e., $\text{ave}(\mathbf{a}^T\mathbf{x})^2 = \text{small}$). For PLS and PCR, the strength of the penalty decreases as the number of components K increases. For RR, the strength of the penalty increases for increasing values of the ridge parameter λ. For RR, the strength of this penalty is monotonically increasing for directions of decreasing sample variance. For PCR, it is a sharp threshold function, whereas for PLS it is relatively smooth but not monotonic. All three methods are shrinkage estimators in that the length of their solution coefficient vector is less than that of the OLS solution. RR and PCR are strictly shrinking estimators in that in any projection the length of their solution is less than (or equal to) that of the OLS solution. This is not the case for PLS. It has preferred directions in which it increases the projected length of the OLS solution. For a K-component PLS solution, the projected length is expanded in the subspace of eigendirections associated with eigenvalues close to the Kth eigenvalue.

In all situations depicted in Figures 1–4, PLS used fewer components to achieve the same overall shrinkage as PCR, generally about half as many components. PLS closely reached the OLS solution with about five to six components, whereas PCR requires all ten components. This property has been empirically observed for some time and is often stated as an argument in favor of the superiority of PLS over PCR; one can fit the data at hand to the same degree of closeness with fewer components, thereby producing more *parsimonious* models. The issue of parsimony is a bit nebulous here, since the result of any method that fits linear models (29) is a single component (direction)—namely, that associated with the solution coefficient vector **a**. One can decompose **a** arbitrarily into sums of any number of (up to p) other vectors and thus change its parsimony at will. For the same number of components, PCR applies more shrinkage than PLS and thus attempts to fit the data at hand less closely, thereby using fewer degrees of freedom to obtain the fit. In the situations studied here (Figs. 1–4) it appears that PLS is using twice the number of degrees of freedom per component as PCR, but this will depend on the structure of the predictor-sample covariance matrix. (For all eigenvalues equal, PLS uses p df for a one-component model.) Thus fitting the data with fewer (or more) components (in and of itself) has no bearing on the quality (future prediction error) of an estimator.

Another argument often made in favor of PLS over PCR is that PCR only uses the predictor sample to choose its components, whereas PLS uses the response values as well. This argument is not unrelated to the one discussed previously. By using the response values to help determine its components, PLS uses more degrees of freedom per component and thus can fit the training data to a higher degree of accuracy than PCR with the same number of components. As a consequence, a K-component PLS solution will have less bias than the corresponding K-component PCR solution. It will, however, have greater variance, and since the mean squared prediction error is the sum of the two (bias squared plus variance) it is not clear which solution would be better in any given situation. In any case, either method is free to choose its own number of components (bias-variance trade-off) through model selection (CV). Both PLS and PCR span a full (but not the same) spectrum of models from the most biased (sample mean) to the least biased (OLS solution). The fact that PLS tends to balance this trade-off with fewer components is (in general) neither an advantage nor disadvantage.

For all of the situations considered in Figures 1–4, PLS and PCR are seen to more strongly penalize for small $\text{ave}(\mathbf{a}^T\mathbf{x})^2$ than RR for the same degree of overall shrinkage $|\mathbf{a}|/|\hat{\boldsymbol{\alpha}}|$. The RR penalty (36) was derived to be optimal under the assumption that the (true) coefficient vector $\boldsymbol{\alpha}$ (26) has no preferred alignment with respect to the predictor-variable distribution; all directions are equally likely (32). Thus the set of situations that favor PLS and PCR would involve $\boldsymbol{\alpha}$'s that have small projections on the subspace spanned by the eigenvectors corresponding to the smallest eigenvalues. For example, an (improper) prior for a K-component PCR would place zero mass on any coefficient vector $\boldsymbol{\alpha}$ for which

$$\sum_{j=K+1}^{p} (\boldsymbol{\alpha}^T\mathbf{v}_j)^2 > 0 \qquad (44)$$

and equal mass on all others. Here $\{\mathbf{v}_j\}_{K+1}^{p}$ are the eigenvectors of the sample predictor-variable covariance matrix [(8)–(9)] associated with the smallest $N - K$ eigenvalues.

Judging from Figures 1–4, a corresponding prior distribution for PLS (if it could be cast in a Bayesian framework) would be more complicated. As with PCR a *prior* for a K-component PLS solution would put low (but nonzero) mass on coefficient vectors that heavily project onto the smallest eigendirections. It would, however, put highest mass on those that project heavily onto the space spanned by the eigenvectors associated with eigenvalues close to e_K^2 and moderate to high mass on the larger eigendirections.

In Figures 1–4, the scale factors for RR, PCR, and PLS were compared for the same amount of overall shrinkage ($|\mathbf{a}|/|\hat{\boldsymbol{\alpha}}|$). In any particular problem, there is no reason that application of these three methods would result in exactly the same overall shrinkage of the OLS solution, although they are not likely to be dramatically different. The respective scale factors were normalized in this way so that insight could be gained through the relative shape of their scale-factor spectra.

3.3 Power Ridge Regression

If one actually had a prior belief that the true coefficient vector $\boldsymbol{\alpha}$ (26) is likely to be aligned with the larger eigendirections of the predictor-sample covariance matrix V (8), PCR or PLS might be preferred over RR. Another approach would be to directly reflect such a belief in the choice of a prior distribution $\pi(\boldsymbol{\alpha})$ for the true coefficient vector $\boldsymbol{\alpha}$ (26). This prior would not be spherically symmetric (32) but would involve a more general quadratic form in $\boldsymbol{\alpha}$,

$$\pi(\boldsymbol{\alpha}) = \pi(\boldsymbol{\alpha}^T\Delta^{-1}\boldsymbol{\alpha}). \qquad (45)$$

The (positive definite) matrix Δ would be chosen to emphasize directions for $\boldsymbol{\alpha}/|\boldsymbol{\alpha}|$ that align with the larger eigendirections of V (8). One such possibility is to choose Δ to be proportional to V^{δ},

$$\Delta = \beta^2 V^{\delta}, \qquad (46)$$

where the proportionality constant

$$\beta^2 = E_{\boldsymbol{\alpha}}|\boldsymbol{\alpha}|^2/\text{tr}(V^{\delta}) \qquad (47)$$

is chosen to explicitly involve the expected value of $|\boldsymbol{\alpha}|^2$ [numerator (47)] under $\pi(\boldsymbol{\alpha})$ (45) and the denominator (47) is the trace of the matrix V^{δ}. The optimal linear shrinkage estimator (33) under this prior [(45)–(47)] is

$$\mathbf{a} = (V + \lambda V^{-\delta})^{-1}\,\text{ave}(y\mathbf{x}) \qquad (48)$$

with

$$\lambda = \sigma^2/(N\beta^2). \qquad (49)$$

Here σ^2 is the variance of the noise [(26)–(27)] and N is the training-sample size. This procedure [(48)–

(49)] is known as *power* ridge regression (Hoerl and Kennard 1975; Sommers 1964). The corresponding (solution) shrinkage factors (33) in the principal component representation are

$$f_j^* = \frac{e_j^{2(\delta+1)}}{e_j^{2(\delta+1)} + \lambda}. \qquad (50)$$

The prior parameter δ [(46)–(48)] regulates the degree to which the true coefficient vector $\boldsymbol{\alpha}$ (26) is supposed to align with the major axes of the predictor-variable distribution. The value $\delta = 0$ gives rise to RR (36) and corresponds to no preferred alignment. Setting $\delta > 0$ expresses a preference for alignment with the larger eigendirections corresponding (approximately) to PCR and PLS, whereas $\delta < 0$ places increased probability on the smaller eigendirections. The value $\delta = -1$ gives rise to James–Stein (James and Stein 1961) shrinkage in which the least squares solution coefficients are each shrunk by the same (overall) factor. If a value for δ were unspecified, one could regard it as an additional meta parameter of the procedure (along with λ) and choose both values (jointly) to minimize a model-selection criterion such as CV (12). Whether this will lead to better performance than one of the existing competing methods (RR, PCR, PLS) is an open question that is the topic of current research.

One important issue is robustness of the procedure to the choice of a value for δ. Suppose that the true coefficient vector $\boldsymbol{\alpha}$ (26) occurred with relative probability $\pi(\boldsymbol{\alpha}|\delta = \delta^*)$ [(45)–(47)] but a different value, $\delta = \delta'$, was chosen for power ridge regression [(48)–(50)]. A natural question is: How much accuracy is sacrificed in such a situation for different (joint) values of (δ^*, δ')? This is examined in Table 3 for a situation characterized by $p = 20$ predictor variables, $N = 40$ training observations, signal $E_{\boldsymbol{\alpha}}|\boldsymbol{\alpha}|^2 = 1$, noise $\sigma = .3$, and predictor-variable covariance matrix eigenvalues $\{e_j^2 = j^2\}_1^{20}$. Shown in Table 3 are the ratios of actual to optimal expected squared error loss when $\delta = \delta'$ (vertical) is assumed and $\delta = \delta^*$ (horizontal) is the true parameter characterizing $\pi(\boldsymbol{\alpha})$ [(45)–(47)].

One sees from Table 3 that choosing $\delta' = 0$ (RR) is the most robust choice (over these situations). James–Stein shrinkage ($\delta' = -1$) is exceedingly dangerous except when $\delta^* = -1$, causing preferential alignment with the smaller eigendirections. For all entries in which δ^* and δ' are nonnegative, choosing $\delta' < \delta^*$ is better than vice versa. The evidence presented in Figures 1–4 indicates that PCR and PLS more strongly penalize the smaller eigendirections than RR, thereby more closely corresponding to $\delta' > 0$. The results presented in Table 3 then suggest that RR ($\delta' = 0$) might be the most robust choice if the nature of the alignment of the true coefficient vector $\boldsymbol{\alpha}$ (26) with respect to the predictor-variable distribution is unknown.

Table 3. Ratio of Actual to Optimal Expected Squared Error Loss When the Parameter $\delta = \delta'$ Is Used With Power Ridge Regression and the True Value Characterizing the Prior Distribution $\pi(\boldsymbol{\alpha})$ Is $\delta = \delta^*$

	δ^*			
δ'	−1	0	1	2
−1	1.00	3.57	6.87	9.43
0	1.37	1.00	1.10	1.27
1	1.58	1.20	1.00	1.08
2	1.76	1.81	1.15	1.00

4. VARIABLE SUBSET SELECTION

VSS is the most popular method of regression regularization used in statistics. The basic goal is to choose a small subset of the predictor variables that yields the most accurate model when the regression is restricted to that subset. A sequence of subsets, indexed by the number of variables K constituting each one, is considered. For a given K the subset of that cardinality giving rise to the best OLS fit to the data is selected ("all subsets regression"). Sometimes forward/backward stepwise procedures are employed to approximate this strategy with less computation. The subset cardinality K is considered to be a meta parameter of the procedure whose value is chosen through some model-selection scheme, such as CV (12). Other model-selection methods (intended for linear modeling) are also often employed, but their use is not strictly correct since VSS is not a linear modeling method for a given value of its meta parameter K; the particular variables constituting each selected subset are heavily influenced by the response values $\{y_i\}_1^N$ so that they enter into the estimates $\{\hat{y}_i\}_1^N$ in a highly nonlinear fashion (see Breiman 1989).

To try to gain some insight into the relationship between VSS and the procedures considered previously (RR, PCR, and PLS), we again consider the (highly) idealized situation [(26)–(27)] in a Bayesian framework:

$$\Pr(\text{model}|\text{data}) = \Pr(\text{data}|\text{model})\Pr(\text{model})/\Pr(\text{data}), \qquad (51)$$

where the left side ("posterior") is the quantity to be maximized, the first factor on the right side is the likelihood $\mathscr{L}$, the second factor is the prior $\pi(\boldsymbol{\alpha})$, and the denominator is a constant (given the data). If we further assume Gaussian errors $\varepsilon \sim N(0, \sigma^2)$, the likelihood becomes $\mathscr{L}(\mathbf{a}) \sim \exp[-(N/2\sigma^2)\text{ave}(y -$

$\mathbf{a}^T\mathbf{x})^2]$, and maximizing (51) is equivalent to minimizing the (negative) log-posterior

$$\text{ave}(y - \mathbf{a}^T\mathbf{x})^2 - 2 \log \pi(\mathbf{a}), \tag{52}$$

where $\pi(\boldsymbol{\alpha})$ is the (prior) relative probability of encountering a (true) coefficient vector $\boldsymbol{\alpha}$ (26). This is a penalized least squares problem with penalty $-2 \log \pi(\mathbf{a})$.

In Section 3, we saw that choosing an *equidirection* prior (32) leads to procedures that shrink the coefficient vector estimate $\mathbf{a}$ away from directions in the predictor-variable space for which $\text{ave}(\mathbf{a}^T\mathbf{x}/|\mathbf{a}|)^2$ is small to control the variance of the estimate. The prior that leads to RR is

$$\begin{aligned} -2 \log \pi_{\text{RR}}(\boldsymbol{\alpha}) &= \boldsymbol{\lambda}\boldsymbol{\alpha}^T\boldsymbol{\alpha} \\ &= \lambda \sum_{j=1}^{p} \alpha_j^2. \end{aligned} \tag{53}$$

Informal "priors" leading to PCR and PLS were seen (Figs. 1–4) to involve some preferential alignment of $\boldsymbol{\alpha}$ with respect to the eigendirections $\{\mathbf{v}_j\}_1^p$ (9) of the predictor covariance matrix (8).

To study VSS, consider a generalization of (53) to

$$-2 \log \pi(\boldsymbol{\alpha}) = \lambda \sum_{j=1}^{p} |\alpha_j|^\gamma, \tag{54}$$

where $\lambda > 0$ (as before) regulates the strength of the penalty and $\gamma > 0$ is an additional meta parameter that controls the degree of preference for the true coefficient vector $\boldsymbol{\alpha}$ (26) to align with the original variable $\{x_j\}_1^p$ axis directions in the predictor space. A value $\gamma = 2$ yields a rotationally invariant penalty expressing no preference for any particular direction—leading to RR. For $\gamma \neq 2$, (54) is not rotationally invariant, leading to a prior that places excess mass on particular orientations of $\boldsymbol{\alpha}$ with respect to the (original variable) coordinate axes.

Figure 5 shows contours of equal value for (54) [and thus for $\pi(\boldsymbol{\alpha})$] for several values of γ ($p = 2$). One sees that $\gamma > 2$ results in a prior that supposes that the true coefficient vector is more likely to be aligned in directions oblique to the variable axes, whereas for $\gamma < 2$ it is more likely to be aligned with the axes. The parameter γ can be viewed as the degree to which the prior probability is concentrated along the favored directions. A value $\gamma = \infty$ places maximum concentration along the diagonals, which is in fact not very strong. On the other hand, $\gamma \rightarrow 0$ places the entire prior mass in the directions of the coordinate axes.

The situation $\gamma \rightarrow 0$ corresponds to (all subsets) VSS. In this case, the sum in (54) simply counts the number of nonzero coefficients (variables that enter), and the strength parameter λ can be viewed as a penalty or cost for each one, controlling the number that do enter. Since the penalty term expresses no preference for particular variables, the "best" subset will be chosen through the minimization of the least squares term, $\text{ave}(y - \mathbf{a}^T\mathbf{x})^2$, of the combined criterion (52).

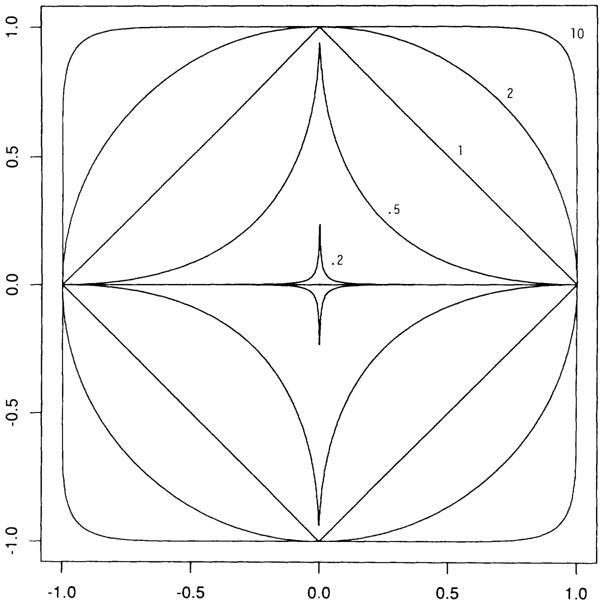

Figure 5. Contours of Equal Value for the Generalized Ridge Penalty for Different Values of γ.

This discussion reveals that a prior that leads to VSS being optimal is very different from the ones that lead to RR, PCR, and PLS. It places the entire prior probability mass on the original variable axes, expressing the (prior) belief that only a few of the predictor variables are likely to have high relative influence on the response, but provides no information as to which ones. It will therefore work best to the extent that this tends to be the case. On the other hand, RR, PCR, and PLS are controlled by a prior belief that many variables together collectively effect the response with no small subset of them standing out.

Expressions (52) and (54) reveal that VSS and RR can be viewed as two points ($\gamma = 0$ and $\gamma = 2$, respectively) on a continuum of possible regression-modeling procedures (indexed by γ). Choosing either procedure corresponds to selecting from one of these two points. For a given situation (data set), there is no a priori reason to suspect that the best value of γ might be restricted to only these two choices. It is possible that an optimal value for γ may be located at another point in the continuum ($0 < \gamma \leq \infty$). An alternative might be to use a model-selection criterion (say CV) to jointly estimate optimal values of λ and γ to be used in the regression, thereby greatly expanding the class of modeling procedures. It is an open question as to whether such an approach will actually lead to improved performance; this is the subject of our current research (with Leo Breiman). Note that this approach is different from those that use Bayesian methods to directly compute model-selection criteria for different variable subsets (e.g., see Lindley 1968; Mitchell and Beauchamp 1988).

5. A COMPARATIVE MONTE CARLO STUDY OF OLS, RR, PCR, PLS, AND VSS

This section presents a summary of results from a set of Monte Carlo experiments comparing the relative performance of OLS, RR, PCR, PLS, and VSS that were described in more detail by Frank (1989). The five methods were compared for 36 different situations. In all situations, the training-sample size was $N = 50$. The situations were differentiated by the number of predictor variables ($p = 5, 40, 100$), structure of the (population) predictor-variable correlation matrix (independent—all off-diagonal elements 0; highly collinear—all off-diagonal elements .9), true regression coefficient vector $\boldsymbol{\alpha}$ (26) (equal—$\{\alpha_j = 1\}_1^p$; unequal—$\{\alpha_j = j^2\}_1^p$), and signal-to-noise ratio [(26)–(27)] $(\sigma/[\mathrm{var}(\boldsymbol{\alpha}^T\mathbf{x})]^{1/2} = 7, 3, 1)$. A full $3 \times 2 \times 2 \times 3$ factorial design on the chosen levels for these four factors yields the 36 situations studied here.

For each situation, 100 repetitions of the following procedure were performed:

1. Randomly generate $N = 50$ training observations with a joint Gaussian distribution (with specified population correlation matrix) for the predictors and using (26) for the response, with ε drawn from a Gaussian with the specified σ^2 (27).
2. Apply OLS, RR, PCR, PLS, and VSS (forward stepwise) to the training sample using CV (12) for model selection.
3. Generate $N_t = 100$ independent "test" observations from the same prescription as in 1.
4. Compute the average squared prediction error (PSE) for the model selected for each method over these test observations:

$$\mathrm{PSE} = \frac{1}{N_t}\sum_{i=1}^{N_t}[y_i - a_0 - \mathbf{a}^T\mathbf{x}_i]^2, \qquad (55)$$

where $(a_0, \mathbf{a})$ is the solution transformed back to the original (unstandardized) representation.

The computed PSE values for each method were averaged over the 100 replications of this procedure.

Figures 6–10 present a graphical summary of selected results from this simulation study. [Complete results in both graphical and tabular form are in the work of Frank (1989).] The summaries are in the form of distances in a 36-dimensional Euclidean space. Average PSE (55) in each of the 36 situations are the axes for this space. There are six points in the space, each defined by the 36 simultaneous values of average PSE for OLS, RR, PCR, PLS, VSS, and the true (known) coefficient vector $\hat{y}_{\mathrm{true}} = \boldsymbol{\alpha}^T\mathbf{x}$ (26). The quantities plotted in Figures 6–10 are the Euclidean distances (bar height) of each of the first five points (OLS, RR, PCR, PLS, and VSS) from the sixth point, which represents the performance using the "true" underlying coefficient vector as the regression model in each situation. Thus smaller values indicate better performance.

Figure 6 shows these distances in the full 36-dimensional space, which characterizes average performance over all 36 situations. Figures 7–10 show the distances in various subspaces characterized by slicing (conditioning) on specific values of some of the design variables. These represent respective average performances conditioned on these particular values.

One sees from Figure 6 that (not surprisingly) OLS gives the worst performance overall. RR is seen to provide the best average overall performance, closely followed by PLS and PCR. Stepwise VSS gives distinctly inferior overall performance to the other biased procedures but still considerably better than OLS. Figure 7 shows that the biased methods improve very little on OLS in the well-conditioned ($p = 5$, $N = 50$) case, but as the conditioning of the problem becomes increasingly worse ($p = 40, 100$), their performance degrades substantially less than OLS, thereby providing increasing improvement over it. Figure 8 shows that the biased methods provide dramatic improvement (over OLS) in the highly collinear situations.

The results shown in Figure 9 represent something of a surprise. From the discussion in Section 4, one

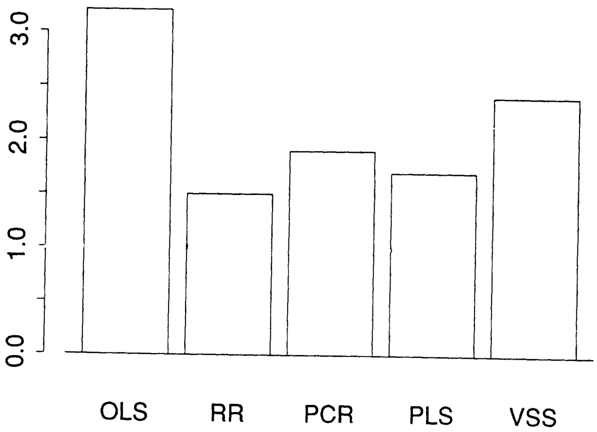

Figure 6. Distances of OLS, RR, PCR, PLS, and VSS From the Performance of the True Coefficient Vector, Averaged Over all 36 Situations.

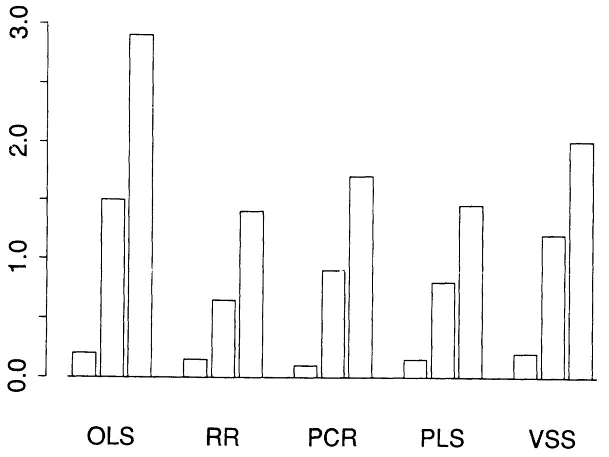

Figure 7. Performance Comparisons Conditioned on the p = 5, 40, and 100 Variable Situations.

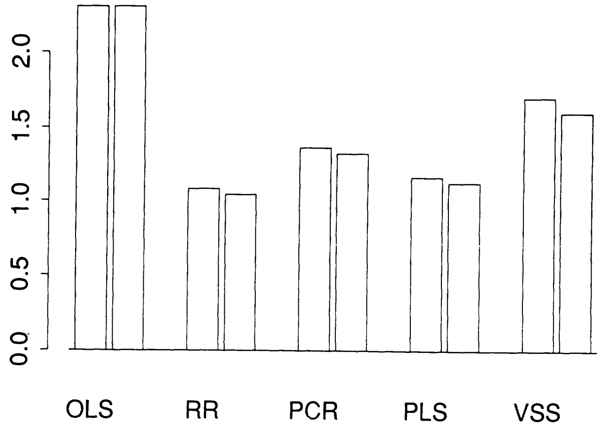

Figure 9. Performance Comparisons Conditioned on the Structure of the True-Coefficients Vector—Equal and Unequal Coefficients.

might have expected VSS to provide dramatically improved performance in the situations corresponding to (highly) unequal (true) coefficient values for the respective variables. For the situations studied here, $\{\alpha_j = j^2\}_1^p$, this did not turn out to be the case. All of the other biased methods dominated VSS for this case. Moreover, the performance of RR, PCR, and PLS did not seem to degrade for the unequal coefficient case. Since (stepwise) VSS must surely dominate the other methods if few enough variables only contribute to the response dependence, it would appear that the structure provided by $\{\alpha_j = j^2\}_1^p$ is not sharp enough to cause this phenomenon to set in.

Figure 10 contains few surprises. (Remember that bar height is proportional to distance from the performance of the true model, which itself degrades with decreasing signal-to-noise ratio.) Higher signal-to-noise ratio seems to help OLS and VSS more than the other biased methods. This may be because their performance degrades less than OLS and VSS as the noise increases.

For the situations covered by this simulation study, one can conclude that all of the biased methods (RR, PCR, PLS, and VSS) provide substantial improvement over OLS. In the well-determined case, the improvement was not significant. In all situations, RR dominated all of the other methods studied. PLS usually did almost as well as RR and usually outperformed PCR, but not by very much. Surprisingly, VSS provided distinctly inferior performance to the other biased methods except in the well-conditioned case in which all methods gave nearly the same performance. Although not discussed here, the performance ranking of these five methods was the same in terms of accuracy of estimation of the individual regression coefficients (see Frank 1989) as for the model prediction error shown here. Not surprisingly, the prediction error improves with increasing observation to variable ratio, increasing collinearity, and

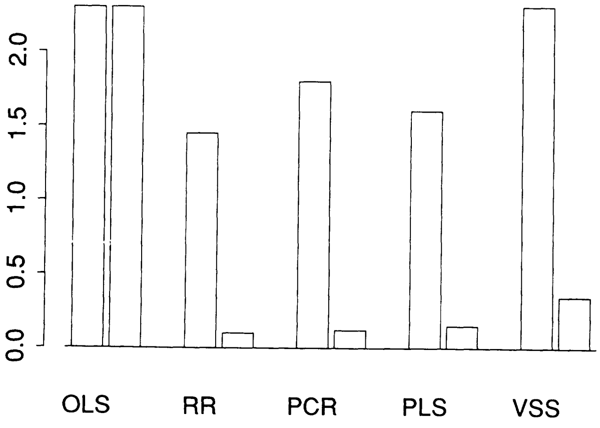

Figure 8. Performance Comparisons Conditioned on Low and High Collinearity Situations.

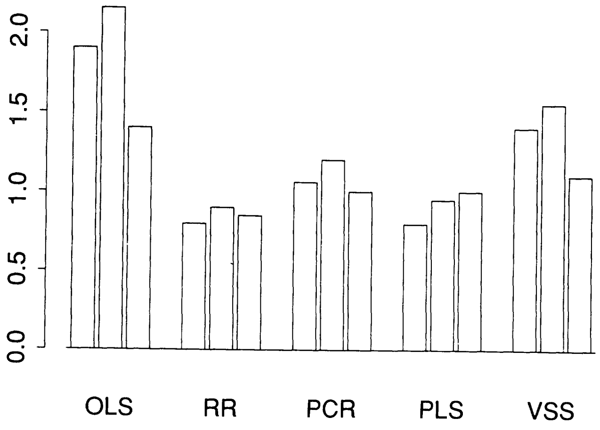

Figure 10. Performance Comparisons Conditioned on High, Medium, and Low Signal-to-Noise Ratio.

increasing signal-to-noise ratio. A bit surprising is the fact that performance seemed to be indifferent to the structure of the true coefficient values.

The results of this simulation study are in accord with the qualitative results derived from the discussion in Section 3.2.1—namely, that RR, PCR, and PLS have similar properties and give similar performance. (Although not shown here, the actual solutions given by the three methods on the same data are usually quite similar.) One can speculate on the reasons why the performance ranking RR > PLS > PCR came out as it did. PCR might be troubled by its use of a sharp threshold in defining its shrinkage factors (42), whereas RR and PLS more smoothly shrink along the respective eigendirections [(36) and Figs. 1–4]. This may be (somewhat) mitigated by linearly interpolating the PCR solution between adjacent components to produce a more continuous shrinkage (Marquardt 1970). PLS may give up some performance edge to RR because it is not strictly shrinking (some $f_j > 1$), which likely degrades its performance at least by a little bit.

The performance differential between RR, PCR, and PLS is seen here not to be great. One would not sacrifice much average accuracy over a lifetime of using one of them to the exclusion of the other two. Still one may see no reason to sacrifice any, in which case this study would indicate RR as the method of choice. The discussion in Section 3.2.1 and the simulation results presented here suggest that claims as to the distinct superiority of any one of these three techniques would require substantial verification.

The situation is different with regard to OLS and VSS. Although these are the oldest and most widely used techniques in the statistical community, the results presented here suggest that there might be much to be gained by considering one of the more modern methods (RR, PCR, or PLS) as well.

6. MULTIVARIATE REGRESSION

We now consider the general case in which more than one variable is regarded as a response ($q > 1$) [(1)–(7)] and a predictive relationship is to be modeled between each one $\{y_i\}_1^q$ and the complement set of variables, designated as predictors. The OLS solution to this (multivariate) problem is a separate ($q = 1$) uniresponse OLS regression of each y_i on the predictor variables $\mathbf{x}$, without regard to their commonality. The various biased regression methods (RR, PCR, PLS, VSS) could be applied to this problem by simply replacing each such uniresponse OLS regression with a corresponding biased ($q = 1$) regression, in accordance with this strategy. The discussion of the previous sections indicates that this would result in substantial performance gains in many situations.

Table 4. Wold's Two-Block PLS Algorithm

```
(1)  Initialize: $\mathbf{y}_0 \leftarrow \mathbf{y}$; $\mathbf{x}_0 \leftarrow \mathbf{x}$; $\hat{y}_0 \leftarrow 0$
(2)  For $K = 1$ to $p$ do:
(3)      $\mathbf{u}^T \leftarrow (1, 0, \ldots, 0)$
(4)      Loop (until convergence)
(5)          $\mathbf{w}_K = \text{ave}[(\mathbf{u}^T\mathbf{y}_{K-1})\mathbf{x}_{K-1}]$
(6)          $\mathbf{u} = \text{ave}[(\mathbf{w}_K^T\mathbf{x}_{K-1})\mathbf{y}_{K-1}]$
(7)      end Loop
(8)      $z_K = \mathbf{w}_K^T\mathbf{x}_{K-1}$
(9)      $\mathbf{r}_K = [\text{ave}(\mathbf{y}_{K-1}z_k)/\text{ave}(z_K^2)]z_K$
(10)     $\hat{\mathbf{y}}_K = \hat{\mathbf{y}}_{K-1} + \mathbf{r}_K$
(11)     $\mathbf{y}_K = \mathbf{y}_{K-1} - \mathbf{r}_K$
(12)     $\mathbf{x}_K = \mathbf{x}_{K-1} - [\text{ave}(z_K\mathbf{x}_{K-1})/\text{ave}(z_K^2)]z_K$
(13)     if $\text{ave}(\mathbf{x}_K^T\mathbf{x}_K) = 0$ then Exit
(14) end For
```

This approach is not the one advocated for PLS (H. Wold 1984). With PLS, the response variables $\mathbf{y} = \{y_i\}_1^q$ and the predictors $\mathbf{x} = \{x_k\}_1^p$ are separately collected together into groups ("blocks") which are then treated in a common manner more or less symmetrically. Table 4 shows Wold's two-block algorithm that defines multiple-response PLS regression.

If one were to develop a direct extension of Wold's ($q = 1$) PLS algorithm (Table 1) according to the strategy used by OLS (q-separate uniresponse regressions), line 3 of Table 1 would be replaced by the calculation of a separate covariance vector $\mathbf{w}_{Ki}$ for each separate response residual $y_{K-1,i}$ on each separate $\mathbf{x}$ residual $\mathbf{x}_{K-1,i}$, $\mathbf{w}_{Ki} = \text{ave}(y_{K-1,i}x_{K-1,i})$ ($i = 1, q$). These would then be used to update q-separate models $\hat{y}_{K,i}$ (line 6), as well as q-separate new y residuals, y_{Ki} (line 7), and $\mathbf{x}$ residuals, $\mathbf{x}_{Ki}$ (line 8).

Examination of Table 4 reveals a different strategy. A *single* covariance vector $\mathbf{w}_k$ is computed for all responses by the inner loop (lines 3–7), which is then used to update all of the models $\hat{\mathbf{y}}_K$ (line 10) and the response residuals to obtain $\mathbf{y}_K$ (line 11). A single set of $\mathbf{x}$ residuals $\mathbf{x}_K$ is maintained by this algorithm using the single covariance vector $\mathbf{w}_K$ (line 12) as in the uniresponse PLS algorithm (Table 1, line 8). The inner loop (lines 4–7) is an iterative algorithm for finding linear combinations of the response residuals $\mathbf{u}^T\mathbf{y}_{K-1}$ and the predictor residuals $\mathbf{w}_K^T\mathbf{x}_{K-1}$ that have maximal joint covariance. This algorithm starts with an arbitrary coefficient vector $\mathbf{u}$ (line 3). After convergence of the inner loop, the resulting $\mathbf{x}$ residual linear combination covariance vector $\mathbf{w}_K$ is then used for all updates.

This two-block multiple-response PLS algorithm produces R models [R = rank of V (8)] for each response $\{\hat{y}_{Kj}\}_{K=1}^R{}_{j=1}^q$ spanning a full spectrum of solutions from the sample means $\{\hat{y}_j = 0\}_1^q$ for $K = 0$ to the OLS solutions for $K = R$. The number of components K is considered a meta parameter of the procedure to be selected through CV,

$$\hat{K} = \underset{0 \le K \le R}{\operatorname{argmin}} \sum_{l=1}^{N} \sum_{j=1}^{q} [y_{jl} - \hat{y}_{Kj \setminus l}]^2, \tag{56}$$

where y_{jl} is the value of the jth response for the lth training observation and $\hat{y}_{Kj\setminus l}$ is the K-component model for the jth response computed with the lth observation deleted from the training sample. Note that the same number of components K is used for each of the response models.

As with the uniresponse PLS algorithm (Table 1), this two-block algorithm (Table 4) defining multiresponse PLS does not reveal a great deal of insight as to its goal. One can gain more insight by following the prescription outlined in the beginning of Section 3—that is, to consider the regression procedure as a two-step process. First, a K-dimensional subspace of p-dimensional Euclidean space is defined as being spanned by the unit vectors $\{\mathbf{c}_k\}_1^K$, and then q-OLS regressions are performed under the constraints that the solution coefficient vectors $\{\mathbf{a}_j\}_1^q$ (5) lie in that subspace,

$$\mathbf{a}_j = \sum_{k=1}^{K} a_{kj}\mathbf{c}_k. \tag{57}$$

A regression procedure is then prescribed by defining the ordered sequence of unit vectors $\{\mathbf{c}_k\}_1^R$ that span the successive subspaces $1 \le K \le R$. Defining each of these unit vectors to be the solution to

$$\mathbf{c}_k = \underset{\substack{\{\mathbf{c}^T V \mathbf{c}_l = 0\}_{l=1}^{k-1} \\ \mathbf{c}^T\mathbf{c} = 1}}{\operatorname{argmax}} \ \underset{u^T u = 1}{\operatorname{argmax}} \{\operatorname{var}(\mathbf{u}^T\mathbf{y})\operatorname{corr}^2[(\mathbf{u}^T\mathbf{y}), (\mathbf{c}^T\mathbf{x})] \cdot \operatorname{var}(\mathbf{c}^T\mathbf{x})\} \tag{58}$$

gives (in this framework) the same sequence of models $\{\hat{\mathbf{y}}_K\}_1^R$ as the algorithm in Table 4 defining two-block PLS regression. As with the uniresponse ($q = 1$) PLS criterion (24), the constraints on $\{\mathbf{c}_k\}_1^R$ require them to be unit vectors and to be V orthogonal so that the corresponding linear combinations are uncorrelated (23).

The multiresponse PLS criterion (58) bears some similarity to that for single-response PLS (24). It can be viewed as a penalized canonical correlation criterion. Using the middle factor $\operatorname{corr}^2[(\mathbf{u}^T\mathbf{y}), (\mathbf{c}^T\mathbf{x})]$ alone for the criterion would give rise to standard canonical correlation analysis, producing a sequence of uncorrelated linear combinations $\{\mathbf{c}_k^T\mathbf{x}\}_1^R$ that maximally predict (the corresponding optimal linear combinations $(\mathbf{u}^T\mathbf{y})$ of the responses. The (unbiased) canonical correlation criterion (middle factor) is invariant to the scales of the corresponding linear combinations $\mathbf{u}^T\mathbf{y}$ and $\mathbf{c}^T\mathbf{x}$. The complete PLS criterion (58) is seen to include two additional factors $[\operatorname{var}(\mathbf{u}^T\mathbf{y})$ and $\operatorname{var}(\mathbf{c}^T\mathbf{x})]$ that serve as penalties to bias the solutions away from low spread directions in both the $\mathbf{x}$ and $\mathbf{y}$ spaces. The penalty imposed on the predictor-variable linear combination coefficient vector $\mathbf{c}$ is the same as that used for single response PLS (24). The discussion in Section 3.1 indicates that this mainly serves to control the variance of the estimated model. The introduction of the $\mathbf{y}$-space penalty factor, along with optimizing with respect to its associated linear combination coefficient vector $\mathbf{u}$, serves to place an additional penalty on the $\mathbf{x}$-linear combination coefficient vectors $\{\mathbf{c}_k\}_1^R$ that define the sequence of PLS models $\{\hat{y}_{Kj}\}_{K=1}^{R}{}_{j=1}^{q}$; they are not only biased away from low (data) spread directions in the predictor-variable space but also toward $\mathbf{x}$ directions that preferentially predict the high spread directions in the response-variable space.

6.1 Bayesian Motivation

A natural question to ask is: To what extent (if any) should this multiresponse PLS strategy (58) improve performance over that of simply ignoring the response-space covariance structure and performing q-separate (single-response) regressions of each y_i on the predictors $\mathbf{x}$, using PLS (24) or one of the other competing biased regression techniques (RR, PCR)? One way to gain some insight into this is to adopt an (idealized empirical) Bayesian framework (as in Sec. 3.1) and see what (joint) prior on the (true) coefficient vectors $\{\boldsymbol{\alpha}_j\}_1^q$,

$$\pi(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q), \tag{59}$$

would lead to such a strategy being a good one. One can then judge the appropriateness of such a prior.

In the case of single-response regression (Sec. 3.1), we saw that a prior distribution that placed no preference on any coefficient vector direction $\boldsymbol{\alpha}/|\boldsymbol{\alpha}|$ (32) gave rise to the preferential shrinkage of the corresponding estimate $\mathbf{a}/|\mathbf{a}|$ away from directions of low predictor spread (36) common to RR, PCR, and PLS. In particular, a Gaussian prior (53) (with Gaussian errors) leads to the optimality of RR. Consider a general (mean 0) joint Gaussian prior (59)

$$\pi(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q) \sim \exp\left(-\frac{1}{2}\sum \alpha_{ik}\alpha_{jl}\Gamma_{ijkl}^{-1}\right), \tag{60}$$

where the sum is over all indices ($1 \le i \le q$, $1 \le j \le q$, $1 \le k \le p$, $1 \le l \le p$). The covariance structure of such a prior distribution (60) is given by the ($q \times q \times p \times p$) array Γ with elements Γ_{ijkl}; namely,

$$E_{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q}(\alpha_{ik}\alpha_{jl}) = \Gamma_{ijkl}. \tag{61}$$

As in (32) and (53), we choose this covariance structure to have no preferred directions in the predictor-variable space—but not necessarily in the response-variable space. This corresponds to (with some abuse of notation)

$$\Gamma_{ijkl} = \Gamma_{ij}\delta_{kl} \qquad (62)$$

with $\delta_{kl} = 1$ if $k = l$ and $\delta_{kl} = 0$ otherwise. The corresponding resulting prior [(60) and (62)] provides information [through Γ_{ij} (62)] on the degree of similarity of the dependence of y_i and y_j on the predictors $\mathbf{x}$ but no information as to the nature of that $\mathbf{x}$ dependence. A relatively large positive value for Γ_{ij} suggests that y_i and y_j have highly similar dependencies on $\mathbf{x}$, whereas a large negative value indicates highly opposite dependencies. A relatively small value indicates dissimilar dependencies of y_i and y_j on the predictors. To further idealize the situation, suppose that

$$y_i = \boldsymbol{\alpha}_i^T\mathbf{x} + \varepsilon_i, \qquad i = 1, q, \qquad (63)$$

with the errors $\boldsymbol{\varepsilon} = \{\varepsilon_i\}_1^q$ having a joint Gaussian distribution

$$\boldsymbol{\varepsilon} \sim N(0, \Sigma), \qquad (64)$$

and, in addition, the error covariance is a multiple of the identity matrix

$$\Sigma = \sigma^2 I. \qquad (65)$$

If Σ were known, one could rotate and scale the $\mathbf{y}$-space coordinates so that (65) is obtained in the transformed coordinate system. Otherwise (65) remains a simplifying assumption. Under these assumptions [(60)–(65)], the following generalization of RR to multiple responses is optimal (smallest MSE):

$$A(\mathrm{RR}) = \underset{A}{\operatorname{argmin}}\left[\operatorname{ave}(\mathbf{y} - A\mathbf{x})^T\Gamma(\mathbf{y} - A\mathbf{x}) + \frac{\sigma^2}{N}\,||A||^2\right]. \qquad (66)$$

Here A is a $(q \times p)$ matrix of regression coefficients (7), Γ is the $(q \times q)$ "prior" matrix (62), σ^2 is the (common) error variance (65), and $||A||^2$ is the Frobenius norm

$$||A||^2 = \sum_{i=1}^{q}\sum_{j=1}^{p} A_{ij}^2. \qquad (67)$$

[For a different Bayesian approach to combining regression equations on the same predictor variables, see Lindley and Smith (1972).]

If the elements of the matrix Γ (62) are unknown one can take an "empirical" Bayesian approach and estimate them from the (training) data. Assuming (60)–(65), one has

$$\underset{\mathbf{x}}{\operatorname{ave}} E_{\varepsilon}E_{\boldsymbol{\alpha}_1 \,\ldots\, \boldsymbol{\alpha}_q}(y_iy_j) = \Gamma_{ij}\mathrm{tr}(V) + \sigma^2, \qquad (68)$$

where the left side is the expected value of (y_iy_j) over both the error [(64)–(65)] and the coefficient prior [(60) and (62)] distributions, then averaged over the predictor-training sample. The quantity $\mathrm{tr}(V)$ is the trace of the predictor-sample covariance matrix V (8). If the data are standardized [(3)–(4)], then

$$\mathrm{tr}(V) = p. \qquad (69)$$

Let W be the $(q \times q)$ sample covariance matrix of the response variables

$$W_{ij} = \operatorname{ave}(y_iy_j). \qquad (70)$$

Then from (68) an "estimate" for the elements of the matrix Γ would be

$$\hat{\Gamma} = (W - \sigma^2 I)/p, \qquad (71)$$

which could then be used in conjunction with Criterion (66) to obtain the resulting estimate $A(\mathrm{RR})$ (given σ^2). The common error variance σ^2 remains unknown and can be regarded as a meta parameter of the procedure to be estimated (from the training sample) through CV:

$$\hat{\sigma}^2 = \underset{\sigma^2}{\operatorname{argmin}} \sum_{k=1}^{N} ||\mathbf{y}_k - A_{\setminus k}(\mathrm{RR}|\sigma^2)\mathbf{x}_k||^2, \qquad (72)$$

where $A_{\setminus k}(\mathrm{RR}|\sigma^2)$ is the coefficient matrix $A(\mathrm{RR})$ estimated from (66) and (71) with the kth observation deleted from the training sample.

Insight into the nature of solutions provided by (66) and (71) can be enhanced by rotating in the $\mathbf{x}$ and $\mathbf{y}$ spaces to their respective principal component representations using orthonormal rotation matrices U_x and U_y such that

$$V = U_x^TE^2U_x$$
$$W = U_y^TH^2U_y \qquad (73)$$

with E^2 and H^2 being diagonal matrices constituting the respective (ordered) eigenvalues

$$E^2 = \operatorname{diag}(e_1^2 \ldots e_p^2)$$
$$H^2 = \operatorname{diag}(h_1^2 \ldots h_q^2). \qquad (74)$$

In this coordinate system, solutions to (66) and (71) simplify to

$$A_{ij}(\mathrm{RR}) = \hat{\alpha}_{ij} \cdot \frac{g_{ij}^2}{g_{ij}^2 + p\sigma^2/N}, \qquad i = 1, q; j = 1, p, \qquad (75)$$

with $\hat{\alpha}_{ij}$ being the OLS coefficient estimates (in the PP coordinate systems) and

$$g_{ij}^2 = e_j^2(h_i^2 - \sigma^2)_+. \qquad (76)$$

Here the subscript "+" indicates the positive part of the argument

$$\begin{aligned}(\eta)_+ &= \eta \quad \text{if } \eta > 0\\ &= 0 \quad \text{otherwise.}\end{aligned} \qquad (77)$$

This RR solution for multiple responses [(75)–(76)] bears considerable resemblance to that for single-response regression (38) in that each coefficient estimate is obtained by (differentially) shrinking the corresponding (unbiased) OLS estimates. Here (for a given value σ^2) the relative shrinkage is controlled both by e_j^2 (corresponding **x**-direction sample spread) and h_i^2 (corresponding **y**-direction sample spread) in a more or less symmetric way through their product (76). A smaller value for either results in more shrinkage. The overall result is to bias the coefficient vector estimates (7) simultaneously away from low sample spread directions in *both* spaces. The overall degree of this bias is controlled by the value of σ^2 [the variance of the noise (65)]. The larger its value the more bias is introduced.

The solution [(75)–(76)] can be recast as

$$A_{ij}(\text{RR}) = \hat{\alpha}_{ij} \cdot \frac{e_j^2}{e_j^2 + \lambda_i} \tag{78}$$

with

$$\lambda_i = p\sigma^2/N(h_i^2 - \sigma^2). \tag{79}$$

Comparing (78) to (38) shows that this multiresponse RR simply applies separate (uniresponse) RR's to each principal component linear combination of the responses $\{y_i(PP)\}_1^q$, with

$$\mathbf{y}(PP) = U_y\mathbf{y} \tag{80}$$

(73), using separate ridge parameters $\{\lambda_i\}_1^q$ for each one. As in single response RR (37), the ridge parameters (79) are related to the (inverse) signal-to-noise ratio.

Since the $\{y_i(PP)\}_1^q$ are uncorrelated, they represent a natural response set on which to perform separate regressions. The basic difference between this approach [(66), (71), (78), (79)] and one in which totally separate RR's are used is that the latter would separately estimate its own ridge parameter [for each $y_i(PP)$] through model selection (say CV) thereby giving rise to q-meta parameters $\{\lambda_i\}_1^q$ to be estimated for the entire procedure. The method previously developed [(66), (71), (78), (79)] attempts to estimate all $\{\lambda_i\}_1^q$ with a single meta parameter, σ^2, selected through CV. This is made possible through the assumption embodied in (65). To the extent that (65) represents a good approximation, this should give rise to better performance. If not, totally separate RR's on each $y_i(PP)$ may work better.

6.2 Discussion

The assumptions that lead to the $\{y_i(PP)\}_1^q$ (80) as being the natural coordinates for the single-response regressions are (60) and (62) through the results (68) and (71). Informally, these (quite reasonably) state that the degree of similarity of the dependence of a pair of responses (y_i, y_j) on the predictors is reflected in their correlation. A large positive (or negative) correlation between y_i and y_j means that the corresponding (true) coefficient vectors $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_j$ should be closely related; that is, $\boldsymbol{\alpha}_i \sim \boldsymbol{\alpha}_j$ (or $\boldsymbol{\alpha}_i \sim -\boldsymbol{\alpha}_j$). Small correlations imply no special relationship. This information is incorporated into the regression procedure by using the empirical response correlational structure to estimate the transformation to linear combinations of the responses $\{y_i(PP)\}_1^q$ that are uncorrelated (no relationship between any of the coefficient vectors) in which separate independent regressions are then performed.

These results suggest that, unless the original response variables happen to be uncorrelated, there is profit to be gained in considering them together rather than simply performing separate regressions on the original responses. This is accomplished by doing the separate regressions on their principal component linear combinations $\{y_i(PP)\}_1^q$ (80). For OLS, this, of course, has no effect, but for the shrinking procedures (RR, PCR, and PLS) this can make quite a difference.

The qualitative behavior of two-block multiresponse PLS (Table 4) as reflected in (57)–(58) is seen to be captured also in multiresponse RR [(66) and (71)] as reflected in (75)–(76)—namely, simultaneous shrinkage of the coefficient vector estimates away from low (sample) spread directions in both the **x** and **y** spaces. This fact serves then to justify this strategy on the part of the two-block PLS algorithm under the same assumptions that lead to multiresponse RR [(66) and (71)]. The principal assumption is that the respective response errors $\{\varepsilon_i\}_1^q$ (63) are independent between the responses and all have approximately the same variance $\{\sigma_i^2 \simeq \sigma^2\}_1^q$ (65). To the extent that this tends to be the case, the low spread directions in the **y** space will be dominated more by the noise than the high spread directions, and biasing the estimates away from these low spread directions will reduce the variance of the estimates. If the error covariance matrix Σ (64) is not well approximated by (65), then the two-block PLS strategy (Table 4) might be counterproductive and a series of uniresponse PLS regressions (Table 1) of each of the response principal component linear combinations $y_i(PP)$ (80) separately on the predictors could be (much) more effective. The same is, of course, also true for the respective versions of RR.

As noted previously, if Σ (64) were known, it could be used to derive a transformation (rotation and scaling) of the **y**-coordinate system so that (65) was obtained in the transformed coordinate system. The analysis (two-block PLS or multiresponse RR) would then be performed in the transformed system and the inverse transform applied to the resulting solutions. Such a transformation can be derived by decomposing Σ into the product

$$\Sigma = R^T R \qquad (81)$$

and taking $\mathbf{Z} = R\mathbf{y}$ as the new responses.

The case of Σ (64) unknown can be directly treated in the context of OLS (Box and Draper 1965). Here the residual covariance matrix is used as an estimate of Σ,

$$\hat{\Sigma}(A) = \text{ave}[(\mathbf{y} - A\mathbf{x})(\mathbf{y} - A\mathbf{x})^T]. \qquad (82)$$

Since this estimate depends on the estimated coefficient matrix A (which in turn depends on $\hat{\Sigma}$), an iterative algorithm is required. Using (82), the multiresponse (negative) log-likelihood [assuming Gaussian errors (64)] can be shown (see Bates and Watts 1988, p. 138) to reduce to $-L(A) = \log \det[\hat{\Sigma}(A)]$. This is minimized with respect to the coefficient matrix A, using (iterative) numerical optimization techniques, to obtain the estimate. It is an open question as to whether an analog of this approach can be developed for biased regression procedures such as RR, PCR, or PLS.

6.3 Monte Carlo Study

We end this section by presenting results of a small Monte Carlo study comparing multivariate RR [(66) and (71)] with two-block PLS (Table 4) in several situations. We also compare both multivariate methods to that of applying separate univariate ($q = 1$) regressions on each (original) response separately. The situations are characterized by the respective eigenstructures of the (population) predictor- and response-variable covariance matrices [(8) and (70)], signal-to-noise ratio, and alignment of the true coefficient vectors $\{\boldsymbol{\alpha}_i\}_1^q$ (63) with the eigenstructure of the (population) predictor covariance matrix.

For the first study, there are $p = 64$ predictor variables, $q = 4$ response variables, and $N = 40$ training observations. The study consisted of 100 replications of the following procedure. First, $N = 40$ training observations were generated with the $p = 64$ predictors having a joint (population) Gaussian distribution with the specified covariance matrix. The corresponding $q = 4$ response variables were obtained from (63) with the $\{\varepsilon_i\}_1^N$ generated from a Gaussian distribution with the (same) specified variance σ^2. The true coefficient vectors $\{\boldsymbol{\alpha}_i\}_1^q$ (63) were each independently generated from $\pi(\boldsymbol{\alpha})$ [(45)–(47)] under the constraint that the (population) response covariance matrix be the one specified. Several values of the prior parameter δ were used. After each of the models were obtained {using CV [(56) and (72)]}, 1,000 new observations were generated according to the same prescription and the average squared prediction error evaluated with them.

Table 5. Mean Squared Prediction Error of Multivariate RR (upper entry) and Two-Block PLS (lower entry) for Several Signal-to-Noise Ratios S/N (rows) and Different Prior Parameter Values δ (columns) for a Highly Collinear Situation

	δ		
S/N	*0*	*1*	*10*
10	.22	.15	.14
	.24	.14	.12
5	.35	.28	.26
	.38	.27	.24
1	.68	.61	.60
	.72	.63	.59

Table 5 compares (in terms of MSE) multivariate RR [(66) and (71)] (upper entry) with two-block PLS (lower entry) for (population) predictor covariance matrix eigenvalues $\{e_j^2 = 1/j^2\}_1^p$ and response covariance matrix eigenvalues $\{h_i^2 = 1/i^2\}_1^q$ (74). The rows correspond to different signal-to-noise ratios and the columns to different prior parameters δ, reflecting differing alignment of the true coefficient vectors $\{\boldsymbol{\alpha}_i\}_1^q$ (63) with the predictor (population) distribution eigendirections. One sees that for $\delta = 0$ (equidirection prior) RR does a bit better than PLS. For $\delta = 1$ (moderate alignment) performance is nearly identical, whereas for $\delta = 10$ (very heavy alignment) PLS has a slight advantage. These results hold for all signal-to-noise ratios.

Table 6 presents a similar set of results for the same situation except with less collinearity in both spaces: $\{e_j^2 = 1/j\}_1^p$ and $\{h_i^2 = 1/i\}_1^q$. Here overall performance is worse for both methods, but their respective relative performance is similar to that reflected in Table 5. These results lend further support to the conclusion that PLS assumes a prior distri-

Table 6. Mean Squared Prediction Error of Multivariate RR (upper entry) and Two-Block PLS (lower entry) for Several Signal-to-Noise Ratios S/N (rows) and Different Prior Parameter Values δ (columns) for Moderate Collinearity

	δ		
S/N	*0*	*1*	*10*
10	.44	.27	.18
	.47	.26	.15
5	.57	.41	.32
	.62	.39	.27
1	.84	.73	.67
	.92	.74	.62

Table 7. Mean Squared Prediction Error of Multivariate RR and Two-Block PLS Along With That of Their Corresponding (separate) Uniresponse Procedures for Several Signal-to-Noise Ratios

S/N	Multi-ridge	Uni-ridge	Two-block PLS	Uni-PLS
10	.23	.25	.25	.27
5	.36	.39	.39	.44
1	.68	.74	.73	.79

NOTE: *S/N* (rows), and prior parameter $\delta = 0$.

bution on the true coefficient vectors $\{\boldsymbol{\alpha}_i\}_1^q$ (63) that preferentially aligns them with the larger eigendirections of the predictor covariance matrix ($\delta > 0$).

Table 7 compares the multivariate RR [(66) and (71)] and two-block PLS (Table 4) procedures with the corresponding strategies of applying q-separate uniresponse ($q = 1$) regressions on the original responses. Here the situation is the same as that of Table 5 except that there are $q = 8$ responses and the comparison is made only for $\delta = 0$ [(45)–(47)]. The relative relationship between multivariate RR and two-block PLS is seen to be the same as that reflected in Table 5 (first column). Each multiresponse method outperforms its corresponding univariate method, but by a surprisingly small amount. In fact, separate RR's do as well as two-block PLS. These results are especially surprising since the situation represented here is set up to provide optimal advantage for the multivariate procedures. Thus even in this optimal setting separate regressions do almost as well as their multiresponse counterparts. This result seems to run counter to the preceding discussion in which it appeared that using the additional information provided by y-space correlational structure ought to help improve performance. This might well be the case if the population correlations were known. The simulation results indicate that having to estimate them from the data induces enough uncertainty to substantially mitigate this potential advantage, at least for the cases studied here.

Overall, the performance of multivariate RR [(66) and (71)] and two-block PLS (Table 4) are comparable. The RR procedure has the advantage of requiring about three times less computation, however.

7. VARIABLE SCALING

OLS is equivariant with respect to rotation and scaling of the variable axes; that is, if one were to apply any (nonsingular) affine (linear-rotation and/or scaling) transformation to the variable axes, perform the (OLS) analysis in the transformed system, and then apply the inverse transformation to the solution, the result would be the same as if the analysis were done in the original coordinate system. None of the biased regression procedures discussed here (RR, PCR, PLS, or VSS) enjoy this affine equivariance property. Applying such transformations on the variables can change the analysis and its result. RR, PCR, and PLS are equivariant under (rigid) rotations of the coordinates. This property allowed us to study them in the sample principal component representations in which the (transformed) covariance matrices were diagonal. They are not, however, equivariant to transformations that change the scales of the coordinates. VSS is equivariant under scaling of the variables but not under rotations. All of these procedures are equivariant under translation of (the origin of) the coordinate systems.

In Section 3 we saw that the basic regularization provided by RR, PCR, and PLS was to shrink their solutions away from directions of small spread in the predictor space. This is not an affine invariant concept. If an original predictor variable x_j has a (relatively) small scale compared to the other predictor variables, $\mathrm{var}(x_j) \ll \mathrm{var}(x_k)$ $(k \neq j)$, then the coordinate axis represented by this variable represents a direction of small spread in the predictor space and the solution will be biased away from involving this variable. Standardizing (autoscaling) the variables [(3)–(4)] to all have the same scale represents a deliberate choice on the part of the user to make all variables equally influential in the analysis. If it were known (a priori) that some variables ought to be more influential than others, this information could be incorporated by adjusting their relative scales to reflect that importance.

Lack of affine equivariance with respect to the predictor variables can be understood in the Bayesian framework adopted in Section 3.1. For RR, the prior (32) leading to its optimality is invariant under rotations; that is, if one were to apply a (rigid) rotation characterized by an orthonormal matrix U ($U^TU = UU^T = I$)

$$\boldsymbol{\alpha}' = U\boldsymbol{\alpha}, \tag{83}$$

then

$$\pi(\boldsymbol{\alpha}'^T\boldsymbol{\alpha}') = \pi(\boldsymbol{\alpha}^TU^TU\boldsymbol{\alpha}) = \pi(\boldsymbol{\alpha}^T\boldsymbol{\alpha}) \tag{84}$$

and the prior is unchanged, resulting in rotational equivariance. A more general prior would be

$$\pi(\boldsymbol{\alpha}) = \pi(\boldsymbol{\alpha}^T\Delta^{-1}\boldsymbol{\alpha}), \tag{85}$$

where Δ is a $p \times p$ positive definite matrix. All rigid rotations (84) involve taking $\Delta = I$, the identity matrix. This makes all directions for $\boldsymbol{\alpha}$ (the truth) equally likely using the original coordinate scales to define the metric. Taking Δ to represent a more general quadratic form in $\boldsymbol{\alpha}$ (85) imposes a specific prior belief on the relative importance of various directions in the predictor space (again using a metric defined by the original variable scales). In particular, choosing Δ to be diagonal,

$$\Delta = \operatorname{diag}(\delta_1^2 \ldots \delta_p^2), \qquad (86)$$

alters the prior belief of the relative importance of the original predictor variables (coordinates). The particular choice $\delta_j^2 = \operatorname{var}(x_j)$ $(j = 1, p)$ in (86) imposes the belief that all predictor variables have equal (a priori) importance, leading to the (data) scale invariant penalty

$$-2 \log \pi(\boldsymbol{\alpha}) = \lambda \sum_{j=1}^{p} \operatorname{var}(x_j) a_j^2$$

for RR (still using the original variable scales to form the metric). This is equivalent to changing the metric by standardizing the variables [(3)–(4)] and then using $\Delta = I$ with respect to one's new metric.

The similarity of PCR and PLS to RR extends to this property as well. Standardizing the variables so that all have the same scale imposes the prior belief that all of the predictor variables ought to be equally important. A different choice for the relative scales would reflect a different prior belief on their relative importance.

In Section 4 we saw that a prior leading to VSS places all of its mass on certain preferred directions in the predictor-variable space—namely, the coordinate axes [(54), $\gamma \to 0$]). Changing the definition of the coordinate axes (preferred directions) through a rotation clearly alters such a prior, causing VSS to not be equivariant under rotations. As $\gamma \to 0$, the (VSS prior) penalty (54) simply counts the number of nonzero coefficients and thus does not involve the variable scales. This causes VSS to be equivariant under predictor-variable scaling.

Since one would not expect (or want) a procedure to be invariant to the user's imposed prior beliefs as reflected in the chosen prior $\pi(\boldsymbol{\alpha})$, it is no surprise that the regularized regression procedures RR, PCR, PLS, and VSS are not affine equivariant in the predictor space. [See Smith and Campbell (1980), and associated comments, for a spirited discussion of this isssue.]

Changing the scales of the response variables in multiple-response regression (Sec. 6) has a similar effect but for a different reason. Changing their relative scales changes their relative influence on the solution. This change, however, is reflected through the loss criterion rather than prior belief. The squared-error loss criterion is

$$L = \sum_{i=1}^{q} E(y_i - \hat{y}_i)^2. \qquad (87)$$

A more general (squared-error) loss criterion would be

$$L_M = E(\mathbf{y} - \hat{\mathbf{y}})^T M^{-1} E(\mathbf{y} - \hat{\mathbf{y}}) \qquad (88)$$

with M some positive definite matrix chosen (by the user) to reflect the (relative) preference of accurately predicting certain linear combinations of the responses. Choosing M to be a diagonal matrix

$$M = \operatorname{diag}(m_1 \ldots m_q) \qquad (89)$$

chooses the response variables themselves to reference the preferred linear combinations (axis directions). In particular, the choice $M = I$ causes their relative importance to be proportional to their sample variance, whereas the choice

$$m_i = \operatorname{var}(y_i), \qquad i = 1, q, \qquad (90)$$

causes them to have equal influence on the loss criterion (88).

For OLS, a choice for M is irrelevant since this procedure chooses $\{\hat{y}_i\}_1^q$ such that each $E(y_i - \hat{y}_i)$ $(i = 1, q)$ is minimized separately, without regard for the other responses. Performing separate biased regressions on each of the individual original responses has a similar effect in that M is irrelevant; the result is the same regardless of a choice for M. This is not, however, the case for the biased procedures that operate collectively on the responses such as two-block PLS (Table 4) or multiresponse RR [(66) and (71)]. It is also not the case if the biased procedures are (separately) applied to the response principal component linear combinations $\{y_i(PP)\}_1^q$ (80) as suggested in Section 6.2. (An exception occurs when the chosen values of the regularization parameters turn out to give rise to unbiased OLS.)

Standardizing the response variables [(3)–(4)] and using $M = I$ (88) in the transformed system is equivalent to using (88), (89), and (90) in the original coordinate system, thereby making all original responses (but not their linear combinations) equally important (influential) in deriving the biased regression models for different levels of bias. If this is not what is wanted (i.e., it is important to accurately predict some responses more than others), then this desire can be incorporated into a choice for M (88) or equivalently a choice for the relative scales of each response (if M is diagonal), or their linear combinations (if M is not diagonal).

8. INTERPRETATION

In the preceding sections, we have compared the various regression methods from the point of view of prediction. This is because prediction error provides an objective criterion (once all definitions and assumptions have been stated) less subject to philosophical or emotional argument. As is well known, the goal of a regression analysis is often not solely prediction but also description; one uses the computed regression equation(s) as a descriptive statistic to attempt to interpret the predictive relationships derived from the data. The loss structure for this enterprise is difficult to specify and depends on the experience and skill of the user in relation to the method used.

It is common to interpret the solution coefficients on the (standardized) original variables as a measure of strength of the predictive relationship between the response(s) and the respective predictors. In this case accuracy of estimation of these coefficients is a relevant goal. As noted in Section 5, the relative ranking of the methods studied there on coefficient accuracy was the same as that for prediction (see Frank 1989). Interpretation is also often aided by the simplicity or parsimony of the representation of the result. This concept is somewhat subjective depending on the user's experience. In statistics, parsimony is often taken to refer to the number of (original) predictor variables that "enter" the regression equation—that is, the number with nonzero coefficients. The smaller this number, the more parsimonious and interpretable is the result. This leads to VSS as the method of choice, since it attempts to reduce mean squared (prediction) error by constraining coefficients to be 0. Moreover, it is often the original variables (as opposed to their linear combinations) that are most easily related to the system under study that produced the data.

It is well known that, in the presence of extreme collinearity, interpretation of individual regression coefficients as relating to the strength of the respective partial predictive relationships is dangerous. In chemometrics applications, the number of predictor variables often (greatly) exceeds the number of observations. Thus there are many exact (as well as possibly many approximate) collinearities among the predictors. This has led chemometricians to attempt to interpret the solution in terms of various linear combinations of the predictors rather than the individual predictor variables themselves. (This approach is somewhat similar to the use of factor-analytic methods in the social sciences.) The linear combinations associated with the principal component directions are a natural set to consider for this purpose, since they represent a set of uncorrelated "variables" that are mutually orthogonal (with respect to the standardized predictors) and satisfy a simple optimality criterion (22). Moreover, principal components analysis has long been in use and is a well-studied method for describing and condensing multivariate data.

The PLS procedure also produces a set of uncorrelated (but not orthogonal) linear combinations. It is often (subjectively) argued that these are a more "natural" set to interpret regression solutions because the criterion [(24) and (58)] by which they are defined involves the data response as well as predictor values. Linear combinations with low response correlation will tend to appear later in the PLS sequence unless their (data) variance is very large. One consequence of this is that a solution regression coefficient vector $\hat{\boldsymbol{\alpha}}$ can generally be approximated to the same degree of accuracy by its projection on the space spanned by fewer PLS components than principal components. As noted in Section 3.2.1, however, this parsimony argument is not compelling, since any vector $\hat{\boldsymbol{\alpha}}$ can be completely represented in a subspace of dimension 1—namely, that defined by a unit vector proportional to it.

The choice of a set of coordinates in which to interpret a regression solution is largely independent of the method by which the solution was obtained. One is not required to use a solution gotten through PCR or PLS to interpret it in terms of their respective components. One could interpret a regression equation(s) obtained by either OLS, VSS, RR, PCR, or PLS in terms of the original predictor variables, the principal components, or PLS linear combinations (or all three). Prediction and interpretation are separate issues, the former being amenable to (more or less) objective analysis but the latter always depending on subjective criteria associated with a particular analyst.

ACKNOWLEDGMENTS

This article was prepared in part while one of us (JHF) was visiting the Statistics Group at AT&T Bell Laboratories, Murray Hill, New Jersey. We acknowledge their generous support and especially thank Trevor Hastie and Colin Mallows for valuable discussions.

APPENDIX: PROOF OF (20) AND (21)

For convenience, center the data so that $E(y) = E(\mathbf{x}) = 0$. The RR solution $\hat{\alpha}_\lambda$ is given by (13). Let $\mathbf{a}^T\mathbf{a} = f^2$ so that $\mathbf{a} = f\mathbf{c}$, with $\mathbf{c}^T\mathbf{c} = 1$. Then, given $\mathbf{c}$, the solution to (13) for f, $f(\mathbf{c})$, is

$$\begin{aligned} f(\mathbf{c}) &= \underset{f}{\operatorname{argmin}}[\operatorname{ave}(y - f\mathbf{c}^T\mathbf{x})^2 + \lambda f^2] \\ &= \operatorname{ave}(y\mathbf{c}^T\mathbf{x})/[\operatorname{ave}(\mathbf{c}^T\mathbf{x})^2 + \lambda], \qquad \text{(A.1)} \end{aligned}$$

and the ridge solution is (21) with

$$\mathbf{c}_{\mathrm{RR}} = \underset{\mathbf{c}^T\mathbf{c}=1}{\operatorname{argmin}}\{\operatorname{ave}[y - f(\mathbf{c})\mathbf{c}^T\mathbf{x}]^2 + \lambda f^2(\mathbf{c})\}. \qquad \text{(A.2)}$$

Substituting (A.1) for $f(\mathbf{c})$ in (A.2) and simplifying gives

$$\mathbf{c}_{\mathrm{RR}} = \underset{\mathbf{c}^T\mathbf{c}=1}{\operatorname{argmin}}\left\{ \operatorname{ave}(y^2) - \frac{\operatorname{ave}^2(y\mathbf{c}^T\mathbf{x})}{\operatorname{ave}(\mathbf{c}^T\mathbf{x})^2 + \lambda} \right\},$$

or, equivalently,

$$\mathbf{c}_{\mathrm{RR}} = \underset{\mathbf{c}^T\mathbf{c}=1}{\operatorname{argmax}}\left\{ \frac{\operatorname{ave}^2(y\mathbf{c}^T\mathbf{x})}{\operatorname{ave}(y^2)[\operatorname{ave}(\mathbf{c}^T\mathbf{x})^2 + \lambda]} \right\}$$

$$= \underset{\mathbf{c}^T\mathbf{c}=1}{\operatorname{argmax}}\left\{ \frac{\operatorname{ave}^2(y\mathbf{c}^T\mathbf{x})}{\operatorname{ave}(y^2)\operatorname{ave}(\mathbf{c}^T\mathbf{x})^2} \, \frac{\operatorname{ave}(\mathbf{c}^T\mathbf{x})^2}{\operatorname{ave}(\mathbf{c}^T\mathbf{x})^2 + \lambda} \right\}.$$

If the data are uncentered then mean values would have to be subtracted from all quantities, giving (20).

[*Received December 1991. Revised September 1992.*]

REFERENCES

Bates, D. M., and Watts, D. G. (1988), *Nonlinear Regression Analysis*, New York: John Wiley.

Box, G. E. P., and Draper, N. R. (1965), "Bayesian Estimation of Common Parameters From Several Responses," *Biometrika*, 52, 355–365.

Breiman, L. (1989), "Submodel Selection and Evaluation in Regression I. The *x*-Fixed Case and Little Bootstrap," Technical Report 169, University of California, Berkeley, Dept. of Statistics.

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.

Frank, I. E. (1987), "Intermediate Least Squares Regression Method," *Chemometrics and Intelligent Laboratory Systems*, 1, 233–242.

——— (1989), "Comparative Monte Carlo Study of Biased Regression Techniques," Technical Report LCS 105, Stanford University, Dept. of Statistics.

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

Hawkins, D. M. (1973), "On the Investigation of Alternative Regressions by Principal Components Analysis," *Applied Statistics*, 22, 275–286.

Helland, I. S. (1988), "On the Structure of Partial Least Squares Regression," *Communications in Statistics—Simulation and Computation*, 17, 581–607.

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 8, 27–51.

——— (1975), "A Note on a Power Generalization of Ridge Regression," *Technometrics*, 17, 269.

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 California Press, pp. 361–379.

Lindley, D. V. (1968), "The Choice of Variables in Multiple Regression" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 30, 31–66.

Lindley, D. V., and Smith, A. F. M. (1972), "Bayes Estimates for the Linear Model" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 1–40.

Lorber, A., Wangen, L. E., and Kowalski, B. R. (1987), "A Theoretical Foundation for the PLS Algorithm," *Journal of Chemometrics*, 1, 19–31.

Mallows, C. L. (1973), "Some Comments on Cp," *Technometrics*, 15, 661–667.

Marquardt, D. W. (1970), "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," *Technometrics*, 12, 591–612.

Martens, H., and Naes, T. (1989), *Multivariate Calibration*, New York: John Wiley.

Massy, W. F. (1965), "Principal Components Regression in Exploratory Statistical Research," *Journal of the American Statistical Association*, 60, 234–246.

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression" (with discussion), *Journal of the American Statistical Association*, 83, 1023–1037.

Naes, T., and Martens, H. (1985), "Comparison of Prediction Methods for Multicollinear Data," *Communications in Statistics—Simulation and Computation*, 14, 545–576.

Phatak, A., Reilly, P. M., and Penlidis, A. (1991), "The Geometry of 2-block Partial Least Squares," Technical Report, University of Waterloo, Dept. of Chemical Engineering.

Rissiden, Y. (1983), "A Universal Prior for Integers and Estimation by Minimum Description Length," *The Annals of Statistics*, 11, 416–431.

Schwartz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

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

Sommers, R. W. (1964), "Sound Application of Regression Analysis in Chemical Engineering," unpublished paper presented at the American Institute of Chemical Engineers Symposium on Avoiding Pitfalls in Engineering Applications of Statistical Methods, Memphis, TN.

Stone, M. (1974), "Cross-validatory Choice and Assessment of Statistical Predictions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

Stone, M., and Brooks, R. J. (1990), "Continuum Regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 52, 237–269.

Webster, J. T., Gunst, R. F., and Mason, R. L. (1974), "Latent Root Regression Analysis," *Technometrics*, 16, 513–522.

Wold, H. (1966), "Estimation of Principal Components and Related Models by Iterative Least Squares," in *Multivariate Analysis*, ed. P. R. Krishnaiah, New York: Academic Press, pp. 391–420.

——— (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, London: Academic Press.

——— (1984), "PLS Regression," in *Encyclopaedia of Statistical Sciences* (Vol. 6), eds. N. L. Johnson and S. Kotz, New York: John Wiley, pp. 581–591.

Wold, S., Ruhe, A., Wold, H., and Dunn, W. J. (1984), "The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses," *SIAM Journal on Scientific and Statistical Computing*, 5, 735–743.