A Simulation Study of Alternatives to Ordinary Least Squares

A. P. DEMPSTER, MARTIN SCHATZOFF, and NANNY WERMUTH*

Estimated regression coefficients and errors in these estimates are computed for 160 artificial data sets drawn from 160 normal linear models structured according to factorial designs. Ordinary multiple regression (OREG) is compared with 56 alternatives which pull some or all estimated regression coefficients some or all the way to zero. Substantial improvements over OREG are exhibited when collinearity effects are present, noncentrality in the original model is small, and selected true regression coefficients are small. Ridge regression emerges as an important tool, while a Bayesian extension of variable selection proves valuable when the true regression coefficients vary widely in importance.

KEY WORDS: Least squares; Multiple regression; Ridge regression; Simulation; Variable selection.

1. INTRODUCTION

We report here summary results of a numerical study undertaken to compare the properties of a collection of alternatives to ordinary least squares for multiple linear regression analysis. The approach is a broad-brush exploration of the relative performance, from the standpoints of estimation and prediction, of different techniques over a range of conditions which are systematically varied according to factorial designs. The variable factor levels include different patterns of true regression coefficients, different amounts of noncentrality, and different degrees of collinearity or multicollinearity among independent variables. The substantive conclusions from the study are indications of possible drastic improvements over least squares, especially through the technique of ridge regression, and especially when a high degree of correlation exists among the independent variables.

We have not attempted to study alternatives to standard regression analysis which are designed to be robust against failures of the normal error model. Instead we have focused on the recently prominent difficulties with least squares under the normal model. From a frequentist standpoint, it has long been recognized that good mean squared error properties do not necessarily follow from the celebrated minimum variance unbiasedness properties of least squares, since in certain regions of the parameter space the loss from increasing the squared bias can be overcompensated by reducing variance. Important work by Charles Stein and his colleagues, (e.g., [1, 13, 18]), has served to draw attention to the potential weakness of straightforward maximum likelihood estimation when more than a very few parameters must be estimated. Efron and Morris [4, 5, 6, 7, 8, 9] have recently extended and advocated the Stein approach at great length. Also taking a mainly frequentist point of view, Hoerl and Kennard [11, 12] introduced and defended ridge regression as having good mean squared error properties in practically relevant regions of the parameter space, at least when the independent variables multicorrelate strongly. (See [2, 10, 17] for various views of ridge regression.) From the standpoint of a subjectivist Bayesian theory, posterior mean squared error is substantially reduced if a flat prior distribution can be replaced by a prior distribution which clusters about some prior mean, taken here to be zero. (See [15, 21] for recent Bayesian discussions of Stein-type and ridge-type estimates.)

Alongside the alternative regression methods just cited, the present study includes forward and backward selection methods and Bayesian selection procedures proposed in [3]. Thus, we hope to draw attention to the substantial improvements over ordinary least-squares methods which are afforded by a wide variety of alternative methods.

Our results are empirical results derived from numerical experiments. Specifically, we created two series of artificial data sets. The first series, referred to as Experiment 1, contains 32 data sets in the form of a 2^5 factorial design, while the second series, or Experiment 2, contains 128 data sets in the form of a quarter replicate of a 2^9 design. Each data set was drawn from a normal linear model with 6 regression coefficients to be estimated and 14 degrees of freedom for error. Each of 57 estimation procedures was applied to each of the 160 data sets, yielding a set of 6 estimated regression coefficients, which were compared to the true regression coefficients in the simulated models, using mainly the two end-point criteria SEB (sum of squared errors of betas) and SPE (sum of squared prediction errors).

Basic notation, including precise definitions of SEB and SPE, appears in Appendix A. Further details of the 57 estimation procedures appear in Section 2 and Appendix

*A.P. Dempster is Professor, Theoretical Statistics, Harvard University, Cambridge, MA 02138. Martin Schatzoff is Manager of Operations Research, IBM Cambridge Scientific Center, Cambridge, MA 02139. Nanny Wermuth is with the Institute für Medizinische Statistik, Universität Mainz, Federal Republic of Germany. The first author's research was supported in part by National Science Foundation Grant GP-31003X. The second author's research was supported by the IBM Cambridge Scientific Center. The third author's research was supported by the Cambridge Project. Computing facilities were provided by IBM.

© Journal of the American Statistical Association
March 1977, Volume 72, Number 357
Theory and Methods Section

B, while the design of the study is described in Section 3 and Appendix C. The analysis of experimental data is presented in Section 4. Further detailed analyses may be found in Wermuth [19].

The study is broad in some ways, e.g., in its range of estimation procedures and range of underlying models. In other ways, the study is narrowly focused, e.g., in its restriction to 6 and 14 degrees of freedom and its limitation to just one replication of each model. In view of the latter restriction, it is clear that we are not attempting detailed analysis of the frequency properties of estimators under each specified model, and so we are not meeting the objectives of the mathematical statistician who wishes to calculate such frequency properties. We would prefer to shed light on the conceptually more difficult task of the data analyst, who knows only his data and not the underlying parameters. Our data base enables us to study the actual errors which a data analyst using specified rules of estimation would encounter under a simulated range of data sets which we believe could typify certain types of real world experience.

Any particular user of regression techniques may legitimately criticize us for not including the specific variant procedures which interest him in the context of a specific class of real world situations. For such a reader, we believe that we have provided a concrete illustration of a type of study which can produce interesting or even startling results. Given adequate computational facility, whose availability continues to develop rapidly, the study could be repeated holding the design matrix $\mathbf{X}$ fixed at the values for a given data set and varying the factors and procedures of greatest concern to a particular data analyst, including of course the shape of the error distribution and appropriate robust procedures, as may be indicated either by prior understanding of circumstances or by the properties of the given data set. We hope, therefore, that we may be contributing to the development of a methodology which will ultimately be of broader use than the specific results of this paper.

2. ALTERNATIVES TO LEAST SQUARES

2.1 Overview

The 57 estimation procedures under study can be grouped into several major classes or families, each containing a number of variants. Both the classes and the variants within each class are denoted by capitalized abbreviations. For example, RIDGE denotes a family of ridge regression techniques, while within the family we study five specific procedures labelled SRIDG, RIDGM, CRIDG, 1CRIDG, and 2CRIDG.

The classes are distinguished by different technical approaches adopted in the attempt to reduce error of estimation, but all approaches produce estimates which shrink or pull back the least-squares estimates toward the origin. Shrinking can be justified either by the frequentist yardstick of improvement in the sum of variance and squared bias, or by the Bayesian device of a prior distribution more or less clustered about the origin. The two extremes in our list of procedures are OREG, or ordinary least squares, which does not shrink, and ZERO which achieves total shrinkage by setting all estimated regression coefficients to zero. Between these extremes the procedures differ in the pattern and degree of shrinking. In Section 2.2 we discuss procedures in the classes STEIN and RIDGE which shrink each estimate according to a continuous formula, while in Section 2.3 we describe the families FSL, BSL, CP, REGF, RREG, and PRI which shrink discretely in the sense that selected coefficients are pulled back to zero, or nearly to zero. The concepts used to define specific variants within the families are described in Section 2.4.

2.2 Continuous Shrinking Methods

Following the notation established in Appendix A, the standard least-squares estimator $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ can be generalized to

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + k\mathbf{Q})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.1}$$

where $\mathbf{Q}$ is a positive definite symmetric matrix and k is a nonnegative scalar. In practice, $\mathbf{Q}$ is allowed to depend only on the design matrix $\mathbf{X}$, while k is generally allowed to depend on $\mathbf{Y}$ as well.

The choices $\mathbf{Q} = \mathbf{X}^T\mathbf{X}$ and $\mathbf{Q} = \mathbf{I}$ in (2.1) define the classes of estimators which we call STEIN and RIDGE.

The principal components transformation $\mathbf{C}$ defined in Appendix A simultaneously diagonalizes both $\mathbf{X}^T\mathbf{X}$ and $\mathbf{I}$, and therefore provides a simple representation for RIDGE estimators. After transforming we deal with $\hat{\alpha}_R = \mathbf{C}\hat{\beta}_R$, $\mathbf{a} = \mathbf{CB}$, and $\alpha = \mathbf{C}\beta$ where $\mathbf{C}$ is defined by (A.5) and (A.6). By transforming (2.1) we see that the components of $\hat{\alpha}_R$ and $\mathbf{a}$ are related by

$$\hat{\alpha}_{iR} = f_i a_i \ , \quad i = 1, 2, \ldots, p \ , \tag{2.2}$$

where

$$f_i = \lambda_i/(\lambda_i + k) \ , \tag{2.3}$$

and $\lambda_1, \lambda_2, \ldots, \lambda_p$ denote as in (A.6) the eigenvalues of $\mathbf{X}^T\mathbf{X}$. The analogs of (2.2) and (2.3) for the STEIN estimator $\hat{\alpha}_S = \mathbf{C}\hat{\beta}_S$ are likewise seen to be

$$\hat{\alpha}_{iS} = f a_i \ , \quad i = 1, 2, \ldots, p \ , \tag{2.4}$$

where

$$f = 1/(1 + k) \ . \tag{2.5}$$

Note that the STEIN estimator $\hat{\beta}_S = f\beta$, and so shrinks all components, whatever the coordinate system, by the same factor f.

As remarked in Appendix A, the a_i have independent $N(\alpha_i, \sigma^2/\lambda_i)$ sampling distributions, whence from (2.2) and (2.4) the $\hat{\alpha}_{iR}$ have independent $N(f_i\alpha_i, f_i^2\sigma^2/\lambda_i)$ sampling distributions and the $\hat{\alpha}_{iS}$ have independent $N(f\alpha_i, f^2\sigma^2/\lambda_i)$ sampling distributions. For a given value of k, the f_i are smaller when the λ_i are smaller, and at the same time the error variance σ^2/λ_i of the least squares a_i is larger. Thus the RIDGE approach applies more drastic shrinking where it has greater effect in reducing mean squared error. The STEIN approach by contrast shrinks

all components equally. The advantage of RIDGE is potentially large, therefore, when certain of the λ_i are close to zero, and when the loss function weights components equally, as does SEB defined in (A.10) or (A.12). We may anticipate less advantage of RIDGE over STEIN when the loss function weights the components by λ_i, as does SPE defined in (A.11) or (A.13).

The general estimator (2.1) has a simple Bayesian interpretation. If β has the multivariate normal prior distribution $N(\mathbf{0}, \omega^2\mathbf{Q}^{-1})$, then the posterior distribution of β is $N(\hat{\beta}, (\mathbf{X}^T\mathbf{X}/\sigma^2 + \mathbf{Q}/\omega^2)^{-1})$, where $\hat{\beta}$ is determined by (2.1) with k given by

$$k = \sigma^2/\omega^2 \ . \tag{2.6}$$

Thus $\hat{\beta}_R$ is a posterior mean corresponding to a prior $N(\mathbf{0}, \omega^2\mathbf{I})$ distribution for β, and $\hat{\beta}_S$ is a posterior mean corresponding to a prior $N(\mathbf{0}, \omega^2(\mathbf{X}^T\mathbf{X})^{-1})$ distribution for β. In principal component terms, if the α_i are *a priori* independently $N(0, \omega^2)$ distributed, then they are *a posteriori* independently $N(\hat{\alpha}_{iR}, f_i\sigma^2/\lambda_i)$ distributed for $i = 1, 2, \ldots, p$. The corresponding result for STEIN estimators is that if the α_i are *a priori* independently $N(0, \omega^2/\lambda_i)$ distributed, then the α_i are *a posteriori* independently $N(\hat{\alpha}_{iS}, f\sigma^2/\lambda_i)$ distributed for $i = 1, 2, \ldots, p$. For a Bayesian, therefore, the choice between RIDGE or STEIN hinges on whether he regards the prior variances of the α_i to be roughly equal or roughly inversely proportional to the λ_i. To assert that RIDGE is better in practice, is equivalent to asserting that its prior assumptions are more nearly correct over the range of the statistician's experience. Note especially that if the RIDGE prior is correct then the RIDGE estimator is optimum for any quadratic loss function, including both SEB and SPE. Corresponding remarks can of course be made about the Bayesian view of STEIN.

The precise realization of a RIDGE or STEIN estimator requires a rule for determining k from the sample data. These rules are discussed in Section 2.4.

2.3 Discrete Shrinking Methods

We consider here six classes of estimation procedures which may be classified into three groups: Group 1 includes FSL, BSL, and MCP; Group 2 includes REGF and RREG; and Group 3 includes PRI.

Group 1 consists of methods which partition the components of β into two subsets. The components in one subset are estimated by least squares under the constraint that the components in the remaining subset are zero. The FSL or forward selection methods proceed by introducing independent variables into the least-squares procedure one at a time, choosing at each step the variable which produces the largest reduction in sum of squares at that step. An FSL method chooses among a set of $p + 1$ partitions of the p independent variables, i.e., one partition which fits r variables for each of $r = 0, 1, 2, \ldots, p$. The BSL or backward selection methods proceed by dropping variables one at a time from the complete least-squares fit in such a way that the increase in residual sum of squares is minimized at each step. A BSL method chooses among $p + 1$ partitions of the independent variables, one for each number r of variables selected, as do the FSL methods, but the set of partitions may differ between BSL and FSL, depending on the data set. The MCP methods consider all 2^p possible partitions of the p independent variables and select one for fitting. The abbreviation MCP is an oblique reference to the C_p statistic which Mallows [16] uses as a criterion for selecting among all possible regressions. The specific variants of FSL, BSL, and MCP used in our study are described in Section 2.4.

From the standpoint of a Bayesian whose prior distribution for β is centered about the zero vector, the FSL, BSL, and CP methods have the weakness that zero posterior estimates for a subset of β components follow in general only from a prior judgment that those components are precisely zero. In practice it may be plausible to judge *a priori* that some subset of the β components are close to zero, but it would rarely be possible to match prior judgments with a subset chosen from the data by a somewhat *ad hoc* selection criterion. The REGF procedures proposed by Dempster [3] attempt to soften the difficulty by supposing that *some* subset of $p - r$ independent variables have zero β components, and supposing for fixed r that all $\binom{p}{p-r}$ possible subsets are *a priori* equally likely. The equiprobability assumption might well be altered in real world practice, but is the most plausible assumption to make in a study of automatic data analysis procedures. The precise definition of a REGF procedure requires a device for choosing r, and a specific rule for computing a posterior probability that each of the $\binom{p}{p-r}$ subsets is the true subset with zero regression coefficients. The least-squares estimates for each postulated set of r nonzero regression coefficients are then averaged over the posterior distribution of subsets to yield the REGF estimator $\hat{\beta}$.

The second class of procedures in Group 2, namely RREG, is a forward selection analog of REGF which avoids the necessity of computing least-squares estimates for all $\binom{p}{p-r}$ subsets of size r, a task which becomes increasingly onerous as r and p increase. An RREG procedure carries out the REGF procedure for $r = 1$, and then repeats the REGF $r = 1$ procedure on the residuals from the first pass, and so on until the residual sum of squares is judged to be small enough. The RREG estimate is formed by summing over the $\hat{\beta}$ produced by each application of REGF. We do not recommend that RREG should necessarily be pursued as a practical tool, in part because of its *ad hoc* nature and in part because the iterations were observed to converge very slowly in the presence of even a few regressors which are highly collinear.

Finally, the PRI procedures of Group 3 are selection procedures based on the principal components representation of the model. Assuming that the principal components, as defined in Appendix A, are ordered so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, a commonly used procedure is to modify the least-squres estimate $\mathbf{a}$ of $\boldsymbol{\alpha}$ so that the last

$p - r$ components are set to zero, yielding

$$\hat{\alpha} = (a_1, a_2, \ldots, a_r, 0, 0, \ldots, 0) \ . \tag{2.7}$$

The corresponding $\hat{\beta}$ is computed as $\mathbf{C}^T\hat{\alpha}$. A specific variant of PRI is defined by a rule for determining r.

2.4 Fifty-Seven Varieties

OREG and ZERO are single procedures, but each of the families STEIN, RIDGE, FSL, BSL, MCP, REGF, RREG, and PRI have several specific variants. Variants of three kinds are used. First, in the case of STEIN and RIDGE, there are methods which aim directly at reducing squared error, whether through frequency concepts or through empirical Bayes concepts. Second, in the case of the remaining families, there are methods related to F tests for the significance of variables not yet included in the fit. Third, there are three methods associated with each family, indicated by the prefixes C, $1C$, and $2C$, which control the maximum permissible deviation from the least-squares estimator $\mathbf{b}$, where such deviation is measured in terms of standard confidence contours about $\mathbf{b}$.

2.4.1. Continuous Shrinking Methods. In the RIDGE and STEIN classes, the specific variants are SRIDG, RIDGM, CRIDG, 1CRIDG, 2CRIDG, and STEINM, CSTEIN, 1CSTEIN, 2CSTEIN. The precise definitions of the last three procedures in each class are given in Section 2.4.5. The SRIDG procedure seeks to minimize the frequentist expectation of the criterion SEB defined by (A.12). It is easily shown that this expectation is minimum when

$$\sum_{i=1}^{p} \frac{\lambda_i(k\alpha_i^2 - \sigma^2)}{(\lambda_i + k)^3} = 0 \ . \tag{2.8}$$

The SRIDG method is defined by choosing k to satisfy (2.8), after s^2 from (A.3) and $\hat{\alpha}_{iR}$ from (2.2) are substituted for σ^2 and α_i in (2.8). The RIDGM and STEINM methods are motivated by the Bayesian interpretation of RIDGE and STEIN discussed in Section 2.2. The prior distribution for the α_i which leads to the posterior mean interpretation of RIDGE also implies that the observable least-squares estimators a_i are marginally independently $N(0, \omega^2 + \sigma^2/\lambda_i)$ distributed. It follows that the prior expectation of $\sum a_i^2/(\omega^2 + \sigma^2/\lambda_i)$ is p. RIDGM chooses k to make this quantity equal to its prior expectation, when s^2 is substituted for σ^2 and $k = \sigma^2/\omega^2$. Similarly, the STEIN procedure is associated with a marginal $N(0, (\omega^2 + \sigma^2)/\lambda_i)$ distribution for the a_i, and STEINM is defined by the choice of k such that $\sum \lambda_i a_i^2/(\omega^2 + \sigma^2)$ equals its marginal expectation p, where again s^2 is substituted for σ^2 and $k = \sigma^2/\omega^2$. It is perhaps unfortunate that we did not adopt the specific choice of k recommended by Stein. In retrospect, however, we can see that Stein's method would have performed worse on our data sets than STEINM, the reason being that improved estimates on our data sets require STEIN to shrink more than STEINM provides, while Stein's recommendation shrinks considerably less.

2.4.2. Subset Regression. The FSL variants group naturally into the three subclasses: FSLA, 0FSL, 1FSL; FSLN, 1FSLN; and CFSL, 1CFSL, 2CFSL. The first two subclasses are defined using different F statistics. Suppose that r variables have been included in the fit, and we are considering whether to include the $(r + 1)$st forward selected variable. Define

$$F_1 = \frac{\text{RSS}_r - \text{RSS}_{r+1}}{\text{RSS}_{r+1}/(n - r - 1)} \ , \tag{2.9}$$

and

$$F_2 = \frac{\text{RSS}_r - \text{RSS}_p}{\text{RSS}_p/(n - p)} \ , \tag{2.10}$$

where RSS_t denotes the residual sum of squares after fitting the best forward selected t variables. F_1 and F_2 are F statistics with nominal degrees of freedom $(1, n - r - 1)$ and $(p - r, n - p)$, respectively. The specific procedures FSLA, 0FSL, and 1FSL are based on the statistics F_1 computed at each stage of selection. FSLA selects a further variable if the F_1 test rejects at level $.05/(p - r)$, 0FSL selects if F_1 rejects at level .05, and 1FSL selects if $F_1 > 1$. FSLA thus sets a fairly rigorous standard of significance for a variable to be included, 0FSL uses a mild standard, and 1FSL includes a variable if there is any indication at all of positive effect with no requirement of a small tail area. Similarly, FSLN selects a further variable if F_2 exceeds its nominal .05 critical value, while 1FSLN selects if $F_2 > 1$. The C, $1C$, and $2C$ variants will be defined in Section 2.4.5.

The corresponding BSL variants are: BSLA, 0BSL, 1BSL; BSLN, 1BSLN; and CBSL, 1CBSL, 2CBSL. The first two subclasses are again determined by the statistics F_1 and F_2, still defined from (2.9) and (2.10) except that RSS_t refers to residuals from the backward selected fit of t independent variables. At the stage of deciding whether to retain the $(r + 1)$st variable or drop it from the fit, a value of F_1 less than its nominal $.05/(p - r)$ critical value indicates dropping the variable under procedure BSLA. Level .05 is used similarly for 0BSL, and the criterion $F_1 < 1$ is used for 1BSL. BSLN drops the $(r + 1)$st variable if F_2 is less than its nominal .05 level, while 1BSLN drops a variable if $F_2 < 1$.

In the case of CP procedures, the F_1 criterion is not always sensible because the best variable set of size $r + 1$ need not contain the best variable set of size r. We therefore consider only two subclasses: 0MCP, 1MCP; and CMCP, 1CMCP, 2CMCP. The procedures 0MCP and 1MCP use F_2 at nominal .05 level and $F_2 = 1$, respectively, as criteria for passing from r to $r + 1$.

2.4.3. REGF *Methods.* There are two parallel series of REGF methods, typically labelled REGF and DRGF, which are defined by alternative prior distributions of β. The two series each appear in three subclasses analogous to the three subclasses of FSL or BSL methods: FREGF, 1FREGF; 0REGF, 1REGF; CREGF, 1CREGF, 2CREGF; and FDRGF, 1FDRGF; 0DRGF, 1DRGF; CDRGF, 1CDRGF, 2CDRGF.

The structure of each of these methods runs as follows. A non-Bayesian scheme is used to select an r on $r = 0, 1, 2, \ldots, p$, whereupon a Bayesian analysis takes over. The Bayesian analysis assumes that exactly r of the p

components of β are nonzero, but assumes that all $\binom{p}{r}$ possible subsets are equally probable *a priori*. Suppose that $\mathcal{I}_r$ denotes the class of $\binom{p}{r}$ subsets consisting of r of the p independent variables. For each $I \in \mathcal{I}_r$, a posterior probability $\omega(I)$ is computed for the event that I is the true subset. Also, for each $I \in \mathcal{I}_r$ we compute an estimate $\hat{\beta}(I)$ whose components in the I positions consist of least-squares estimates of the corresponding β components, assuming that the remaining coefficients are all set to zero. These remaining coefficients are of course all estimated at zero in $\hat{\beta}(I)$. The final estimator $\hat{\beta}$ is defined to be

$$\hat{\beta} = \sum_{\mathcal{I} \in \mathcal{I}_r} \omega(I)\hat{\beta}(I) \ . \tag{2.11}$$

The details of how to compute $\omega(I)$ for each of the REGF and DRGF series are given in Appendix B.

In Appendix B we also give reformulated definitions appropriate for REGF of the F_1 and F_2 criteria defined in (2.9) and (2.10). FREGF and RDRGF use a nominal .05 critical level for the F_1 criterion, and 1FREGF and 1FDRGF use $F_1 = 1$ as the cutoff point, where in both cases larger values of F_1 force an increase from r to $r + 1$. The pairs 0REGF, 0DRGF and 1REGF, 1DRGF operate similarly in relation to the F_2 criterion.

The RREG variants are: RREG1; and CRREG, 1CRREG, 2CRREG. RREG1 makes repeated use of the FREGF technique with $r = 1$ and stops when the F_1 criterion associated with FREGF fails to indicate proceeding from $r = 1$ to $r = 2$.

2.4.4. Regression on Principal Components. The PRI variants are: PRIF, 1PRIF; PRIB, 1PRIB; and CPRI, 1CPRI, 2CPRI. The methods PRIF and PRIB both use the F_2 criterion (2.11) at its nominal .05 level, while 1PRIF and 1PRIB use the critical value $F_2 = 1$. The difference is that the F methods proceed through the estimators (2.7) in the order $r = 0, 1, \ldots, p$ while the B methods use the order $r = p - 1, \ldots, 0$.

2.4.5. Confidence Contour Constraints. Finally, we describe the C, $1C$, and $2C$ variants which are associated with each family. As is well known, an ellipsoidal $1 - \alpha$ confidence region for β centered at the least-squares estimate $\mathbf{b}$ is defined by

$$(\beta - \mathbf{b})^T\mathbf{X}^T\mathbf{X}(\beta - \mathbf{b})/ps^2 \leq F_{p,n-p,1-\alpha} \tag{2.12}$$

where $F_{p,n-p,1-\alpha}$ denotes the level α critical value for F on p and $n - p$ degrees of freedom. The idea of C, $1C$, and $2C$ methods is to limit the deviation of β from $\mathbf{b}$ by requiring that $\hat{\beta}$ lie within an ellipsoid of the form (2.12). This idea is similar to the limited risk proposal of Efron and Morris [4, 5]. The C method uses $\alpha = .05$ and the $1C$ method uses $F_{p,n-p,1-\alpha} = 1$. The criterion in the $2C$ method is that the residual sum of squares is allowed to rise by at most 20 percent. When $p = 6$ and $n = 20$, the $2C$ criterion is equivalent to the choice $F_{p,n-p,1-\alpha} = .46$.

Given any value of $F_{p,n-p,1-\alpha}$ we can adjust the k in RIDGE or STEIN methods, or the r in the selection methods, in such a way that the resulting estimator $\hat{\beta}$ is shrunk as much as possible subject to the constraint (2.12). This is the guiding principle of all of the C, $1C$, and $2C$ methods. For example, the 1CMCP method chooses the best subset of independent variables of size r, when r is as small as possible consistent with (2.12), choosing the right side in (2.12) to be unity.

3. DESIGN AND EXECUTION

The plan of our study required drawing from the model (A.1) with $p = 6$ and $n = 20$. Conceptually, this meant fixing $\mathbf{X}$, β, and σ^2, and then drawing a random vector $\mathbf{e}$ using a standard normal random number generator. In practice, since all of our methods depend only on the sufficient statistics (A.2) and (A.3), we did not actually generate $\mathbf{X}$, $\mathbf{e}$, or $\mathbf{Y}$, but instead generated first $\mathbf{X}^T\mathbf{X}$, and then $\mathbf{b}$ and $(n - p)s^2$, where the latter required random number generators to simulate the 6-variate $N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ and χ_{14}^2 distributions. The 6-variate normal was found by linear transformation of 6 standard normal deviates and the χ_{14}^2 was found by summing the squares of 14 standard normal deviates. Uniformly distributed pseudorandom numbers were generated by the algorithm described in [14], and were transformed to normal random deviates by a table look-up procedure applied to the cumulative normal distribution. Further details may be found in [19, 20].

We made one drawing from each of 160 different models. The factors and factor levels of the 2^5 structure of Experiment 1 are described. After creating and partially analyzing these data, we decided not simply to replicate Experiment 1, but instead to create a somewhat larger data set with more levels of certain factors and generally less correlation among the independent variables. The result was the 128 simulated models of Experiment 2, also described. At this point we decided to analyze and report the results from Experiments 1 and 2 without creating another series of models or drawing replicated data sets from the two available series.

The five factors in Experiment 1 are labelled EIG, ROT, COL, CEN, and BET. The first three of these define the $\mathbf{X}^T\mathbf{X}$ matrix of a model. The two levels of factor EIG were determined by

$$\begin{aligned} \mathbf{T} &= \text{diag } (32, 25, 16, 9, 4, 1) \\ &= \text{diag } (64, 16, 4, 1, .25, .0625) \end{aligned} \tag{3.1}$$

where diagonal vectors of the diagonal matrices $\mathbf{T}$ should be regarded as preliminary eigenvalues of $\mathbf{X}^T\mathbf{X}$. The two levels of EIG specified by (3.1) constitute one device for putting into the experiment variation in the amount of correlation among the independent variables.

The two levels of ROT correspond to two replications of matrices with eigenvalues fixed by a given level of EIG. From a pair of 6×6 arrays of simulated standard normal deviates, a pair of random 6×6 orthogonal matrices $\mathbf{G}$ was created by Gram-Schmidt orthogonalization followed by scaling to unit length. We then formed the four inner product matrices $\mathbf{G}^T\mathbf{T}\mathbf{G}$ corresponding to the four levels of EIG $\times$ ROT. Finally, these four inner product matrices

were reduced to four correlation matrices via scalar division of each row and column by the square root of its diagonal member. These correlation matrices specify the four choices of $\mathbf{X}^T\mathbf{X}$ actually used at one level of COL. Note that the eigenvalues of $\mathbf{G}^T\mathbf{T}\mathbf{G}$ are given by $\mathbf{T}$, but that these are no longer the eigenvalues of the final correlation matrices $\mathbf{X}^T\mathbf{X}$.

The second level of COL is defined by modifying the corresponding $\mathbf{X}^T\mathbf{X}$ at the first level so that it has .99 in the (1, 2) position, thus introducing substantial collinearity between the first two independent variables. The modification was not simply a replacement of the (1, 2) element by .99, which could have destroyed positive definiteness, but a linear transformation scheme described in Appendix C. We have now described, modulo $\mathbf{G}$, the eight matrices $\mathbf{X}^T\mathbf{X}$ used in Experiment 1, corresponding to the eight levels of EIG × ROT × COL.

The factors CEN and BET jointly define β and σ^2. CEN refers to two levels of the noncentrality parameter, specifically

$$\begin{aligned} \beta^T(\mathbf{X}^T\mathbf{X})\beta/\sigma^2 &= 100 \\ &= 200 \ . \end{aligned} \tag{3.2}$$

The factor BET refers to two vectors of regression coefficients, specifically

$$\begin{aligned} \beta &= (32, 25, 16, 9, 4, 1) \\ &= (64, 16, 4, 1, .25, .0625) \ . \end{aligned} \tag{3.3}$$

In practice, two values of σ^2 were determined from (3.2) for each of the two vectors β in (3.3). Since our endpoint criteria SEB (A.10) and SPE (A.11) are unaffected by scale changes $\beta \to c\beta$ and $\sigma \to c\sigma$, we could equally well have set σ arbitrarily and computed scalar multipliers from (3.2) for the β vectors in (3.3).

In Experiment 2, the factors are EIG at two levels, ROT at four levels, COL at two levels, MCL at two levels, CEN at four levels, and BET at four levels. The actual design is a quarter replicate of the $2^3 \times 4^3$ complete design. The preliminary eigenvalue levels of EIG in Experiment 2 were changed to

$$\begin{aligned} \mathbf{T} &= \text{diag } (30, 30, 30, 20, 20, 20) \\ &= \text{diag } (64, 16, 4, 2, 1, .5) \ . \end{aligned} \tag{3.1}'$$

For ROT we created four new random orthogonal matrices $\mathbf{G}$ by the same algorithm used in Experiment 1. The EIG and ROT levels were crossed as in Experiment 1, yielding now $2 \times 4 = 8$ correlation matrices $\mathbf{X}^T\mathbf{X}$, which were further crossed with the factors COL and MCL to obtain $8 \times 2 \times 2 = 32$ correlation matrices $\mathbf{X}^T\mathbf{X}$ altogether. COL means the presence or absence of a deliberately introduced correlation .95 between X_1 and X_2, while MCL means the presence or absence of a deliberately introduced correlation .92 between $X_1 - X_2$ and X_3, thus providing a partially hidden substantial correlation among the first three independent variables X_1, X_2, and X_3. The COL and MCL algorithms are described in Appendix C.

Given $\mathbf{X}^T\mathbf{X}$, the procedure for fixing β and σ^2 is the same as in Experiment 1, changing (3.2) to

$$\begin{aligned} \beta^T(\mathbf{X}^T\mathbf{X})\beta/\sigma^2 &= 100 \\ &= 500 \\ &= 10 \\ &= 50 \end{aligned} \tag{3.2}'$$

and changing (3.3) to

$$\begin{aligned} \beta &= (1, 1, 1, 1, 1, 1) \\ &= (32, 16, 8, 8, 8, 8) \\ &= (1, 1, 1, 0, 0, 0) \\ &= (32, 16, 8, 0, 0, 0) \ . \end{aligned} \tag{3.3}'$$

The experimental data sets were created and analyzed using the APL computer language as implemented under the CP-67 system at IBM Cambridge Scientific Center. An advantage of APL is that a large number of small but mathematically complex program units can be written and put together with relative ease. A disadvantage is that a large number of routine repetitions of the programs, as would be required for standard large sample Monte Carlo, becomes prohibitively expensive due to the interpretive nature of APL. The programs described in [20] make it feasible to reproduce much of the data generation and analysis or to replicate the experiments if so desired.

4. NUMERICAL RESULTS

4.1 Overall Comparisons of 57 Methods

Many different analyses were carried out for purposes of comparing the properties of the various estimators, and relating them to the design factors. Overall comparisons of the methods under study are provided in Table 1, which shows the mean values and medians of the two criteria, SEB and SPE, together with their ranks on the 57 methods, for each of the experiments. The methods are arranged so as to put together different versions of the same method, as indicated by the extra space separating the ten different groups.

Examination of Table 1 leads to a number of interesting observations.

1. Ordinary regression (OREG) is inferior to all nontrivial methods of estimation with respect to observed SEB averaged over each series of data sets. In the first experiment, it is even worse than the trivial method of estimating all coefficients to be equal to zero.
2. The reductions in SEB, on average over the observed data sets, achieved by some of the methods under study are as large as 90 percent.
3. Average reductions in SPE are at most 20–30 percent, and ordinary least squares performs better than a number of its competitors on this criterion. These results corroborate our observation in Section 2.3 to the effect that the advantage of RIDGE over STEIN would be less on SPE than on SEB, since SPE weights individual components by λ_i.
4. The methods which produced the best overall results were not the customary ones such as ordinary least squares, selection of variables, or regression on principal components, but rather versions of RIDGE and REGF. In particular, it is interesting to note that RIDGM was best with respect to mean

1. Means and Ranks and Medians and Ranks of 57 Methods

a. Means and ranks

Method	Experiment 1 Mean SEB	Experiment 1 Mean SPE	Experiment 1 Rank SEB	Experiment 1 Rank SPE	Experiment 2 Mean SEB	Experiment 2 Mean SPE	Experiment 2 Rank SEB	Experiment 2 Rank SPE
OREG	542.86	5.70	57	27	78.37	6.26	56	21
ZERO	143.92	150.00	35	57	134.19	165.00	57	57
FSLA	93.59	8.26	29	47	68.42	16.37	52	55
CFSL	84.35	8.12	27	46	61.63	10.38	39	50
1CFSL	93.13	4.56	28	10	51.53	6.79	22	29
2CFSL	466.92	4.82	52	17	63.42	6.04	42	16
0FSL	78.61	4.78	22	16	56.27	7.69	30	38
1FSL	215.22	4.89	36	20	60.09	6.08	35	17
FSLN	80.03	6.86	25	39	52.46	8.88	23	43
1FSLN	479.95	4.93	53	21	66.02	8.03	48	14
BSLA	76.22	5.96	19	29	64.34	9.94	43	48
CBSL	78.43	7.97	21	45	67.83	11.37	51	53
1CBSL	273.01	5.06	39	24	65.58	7.33	46	34
2CBSL	499.55	4.87	56	19	60.84	6.33	37	24
0BSL	78.88	4.77	23	13	58.39	7.23	32	33
1BSL	461.50	5.16	51	26	71.31	6.23	53	20
BSLN	77.50	7.38	20	41	60.86	9.82	38	47
1BSLN	421.26	4.97	49	23	74.20	6.36	54	25
CMCP	78.91	7.55	24	43	65.17	10.52	45	52
1CMCP	272.53	4.83	38	18	66.50	7.02	49	31
2CMCP	493.32	4.76	55	12	64.48	6.36	44	26
0MCP	81.47	6.74	26	37	56.46	9.18	31	44
1MCP	416.98	4.78	48	15	75.89	6.17	55	19
CPRI	73.75	11.46	18	53	28.12	8.58	7	42
1CPRI	66.10	6.82	16	38	27.72	6.02	6	13
2CPRI	131.44	6.01	33	30	51.09	5.94	21	12
PRIB	100.25	8.51	30	49	39.36	7.13	11	32
1PRIB	355.37	6.11	43	31	61.92	6.15	40	18
PRIF	102.37	9.21	31	52	27.34	7.50	5	37
1PRIF	363.24	6.35	44	33	55.57	6.03	20	15
STEINM	482.16	5.94	54	28	63.40	5.84	41	9
CSTEI	219.56	24.17	37	56	40.26	16.41	12	56
1CSTEI	325.61	11.76	40	54	46.86	8.54	19	41
2CSTEI	386.46	8.30	46	48	53.94	6.55	27	28
SRIDG	65.02	6.57	15	36	25.76	5.51	4	3
RIDGM	45.16	4.95	1	22	17.59	4.98	2	1
CRIDG	69.51	21.17	17	55	29.29	14.48	8	54
1CRIDG	53.85	9.00	6	51	18.71	6.84	3	30
2CRIDG	52.25	6.13	4	32	17.48	5.25	1	2
CREGF	57.67	7.54	11	42	52.97	9.73	24	46
1CREGF	60.76	4.01	13	2	46.53	6.40	18	27
2CREGF	137.73	3.90	34	1	38.60	5.54	10	4
0REGF	56.76	6.50	9	35	44.86	8.08	16	40
1REGF	355.31	4.39	42	8	58.76	5.62	33	7
FREGF	49.32	4.15	2	6	42.89	7.45	14	35
1FREGF	379.35	4.75	45	11	65.87	5.92	47	11
CDRGF	56.97	7.56	10	44	53.32	9.66	26	45
1CDRGF	126.37	4.12	32	4	48.88	6.30	20	22
2CDRGF	327.46	4.02	41	3	43.95	5.55	15	5
0DRGF	56.52	6.50	8	34	45.13	7.92	17	39
1DRGF	416.96	4.42	47	9	58.92	5.62	34	6
FDRGF	52.20	4.14	3	5	54.80	7.48	28	36
1FDRGF	457.61	4.78	50	14	67.59	5.90	50	10
RREG1	59.00	6.96	12	40	60.53	10.49	36	51
CRREG	62.15	8.97	14	50	53.18	10.22	25	49
1CRREG	52.61	5.11	5	25	42.64	6.31	13	23
2CRREG	54.30	4.35	7	7	38.23	5.73	9	8

1. Continued

b. Medians and ranks

Method	Experiment 1 Mean SEB	Experiment 1 Mean SPE	Experiment 1 Rank SEB	Experiment 1 Rank SPE	Experiment 2 Mean SEB	Experiment 2 Mean SPE	Experiment 2 Rank SEB	Experiment 2 Rank SPE
OREG	178.83	5.57	57	33	25.94	5.05	46	24
ZERO	128.87	150.00	54	57	55.42	75.00	57	57
FSLA	62.15	5.28	50	27	31.11	8.23	54	52
CFSL	39.55	6.01	30	40	30.36	7.86	51	49
1CFSL	33.23	3.02	24	3	23.24	5.19	39	26
2CFSL	58.77	4.36	49	19	20.22	4.84	30	16
0FSL	40.06	3.53	31	10	24.12	5.96	40	36
1FSL	56.71	4.79	44	25	19.47	4.83	27	15
FSLN	45.09	6.12	36	42	26.80	7.76	47	47
1FSLN	49.95	4.72	41	22	20.31	4.87	31	17
BSLA	25.20	4.86	8	26	33.76	7.74	56	46
CBSL	31.11	6.27	18	46	33.34	8.78	55	53
1CBSL	27.97	3.31	14	7	28.51	5.90	49	35
2CBSL	37.29	4.23	27	17	25.32	5.32	45	29
0BSL	31.87	3.92	21	13	28.99	5.62	50	33
1BSL	54.93	5.43	43	28	25.30	4.90	43	20
BSLN	27.15	6.69	13	47	31.11	7.91	53	51
1BSLN	37.29	4.76	28	23	25.31	5.43	44	30
CMCP	31.11	6.01	19	41	30.36	7.91	52	50
1CMCP	46.40	3.31	37	8	24.39	5.43	41	31
2CMCP	49.07	4.13	39	16	24.96	5.29	42	28
0MCP	31.11	5.97	20	39	27.70	7.78	48	48
1MCP	39.53	4.35	29	18	20.31	5.26	32	27
CPRI	57.82	11.63	47	54	13.05	7.04	10	42
1CPRI	43.36	5.46	35	30	9.41	5.03	4	22
2CPRI	47.27	5.69	38	35	12.93	5.02	8	21
PRIB	53.40	6.99	42	49	11.24	6.58	6	37
1PRIB	49.37	5.57	40	34	15.00	5.14	14	25
PRIF	57.75	8.62	46	52	11.92	6.72	7	39
1PRIF	58.36	5.71	48	36	12.93	5.03	9	23
STEINM	161.79	5.50	56	31	19.42	4.68	26	11
CSTEI	91.07	24.55	52	56	18.36	14.99	21	56
1CSTEI	113.48	11.39	53	53	15.87	7.51	17	43
2CSTEI	129.22	7.23	55	50	17.60	5.51	19	32
SRIDG	42.75	6.25	34	45	8.16	4.54	3	10
RIDGM	31.87	3.93	22	14	7.43	4.11	1	1
CRIDG	57.69	20.72	45	55	16.98	11.94	18	55
1CRIDG	42.59	8.29	33	51	10.13	5.71	5	34
2CRIDG	36.72	5.54	26	32	8.08	4.53	2	9
CREGF	25.43	6.18	9	43	22.80	7.67	36	45
1CREGF	21.21	3.09	3	4	18.58	4.45	23	8
2CREGF	24.62	3.27	7	6	15.47	4.11	16	2
0REGF	25.43	5.87	10	38	19.99	7.03	29	41
1REGF	28.68	4.47	15	20	14.38	4.35	11	7
FREGF	18.88	2.77	1	1	18.59	4.89	24	19
1FREGF	32.59	4.66	23	21	19.09	4.82	25	14
CDRGF	25.49	6.21	11	44	23.08	7.61	37	44
1CDRGF	23.21	3.17	5	5	19.92	4.35	28	6
2CDRGF	35.67	3.33	25	9	15.35	4.18	15	3
ODRGF	25.49	5.78	12	37	21.07	6.85	33	40
1DRGF	42.03	4.05	32	15	14.95	4.28	13	4
FDRGF	19.77	2.78	2	2	17.65	4.79	20	12
1FDRGF	77.45	4.76	51	24	21.11	4.81	34	13
RREG1	29.51	5.44	16	29	21.56	6.59	35	38
CRREG	30.69	6.94	17	48	23.17	9.24	38	54
1CRREG	23.57	3.90	6	12	18.53	4.87	22	18
2CRREG	22.30	3.57	4	11	14.91	4.29	12	5

SEB in the first experiment and second best in the second. In the latter instance, another version of RIDGE, namely 2CRIDG was slightly better than RIDGM. On the basis of median, rather than mean values, FREGF was best on both criteria in the first experiment, while RIDGM was the number one performer in the second experiment. Comparisons of RIDGM against OREG indicated improvements in average SEB of approximately 92 and 78 percent in the two experiments. RIDGM was also the best method on average SPE in Experiment 2, providing a reduction of 20 percent. In the first experiment it ranked only 22nd on mean SPE although it still provided an average reduction of about 14 percent.

5. The data do not indicate any consistent patterns of behavior with respect to variations in the confidence levels employed in the confidence contour procedures, or in the significance levels used in the various selection-type methods. In most instances, changes in level produced opposite effects on the two criteria. For example, referring to Table 1, for the PRI and STEIN methods, variations from C to $1C$ to $2C$ were generally accompanied by improved SPE, but degraded SEB. In other cases, such as BSL, REGF, and DRGF, the directions of change on SEB were opposite in the two experiments. Similar observations could be made with respect to choice of level for the various selection methods. Since the performance of these techniques is generally sensitive to the particular choice of level, their use can be risky.

4.2 A Closer Look at Selected Methods

In this section, we focus our attention on six particular methods, with the objective of learning more about their detailed behavior patterns. The selected methods are representative of the total spectrum of 57 methods in that they include a member of each major class of pullback procedures, as follows:

Type of pullback	*Selected method*
None	OREG
Selection of variables	OFSL
Principal components	PRIF
STEIN	STEINM
RIDGE	RIDGM
REGF	FREGF

The specific choice of a member within each class was based on performance within the group, and similarity to procedures used in current practice. For example, RIDGM was chosen to represent the RIDGE class of estimators, because it appeared to be the best performer within that class. However, PRIF was chosen to represent the principal component class of estimators, even though 1CPRI seemed to be a better performer, because the type of selection procedure used by PRIF is more commonly used in practice. The OFSL, PRIF, and FREGF methods are comparable to one another in the sense that they all employ forward selection procedures, and use stopping rules based on significance at the .05 level.

The reason for simulating the performance of these methods is, of course, that desired distributions of estimation error can usually not be calculated analytically. In one instance, however, that of RIDGE regression, we can calculate stochastic lower bounds for frequentist expectations of our two criteria. It is of interest then to determine how closely RIDGM approximates these lower bounds, and to compare the performance of the other methods with these bounds as well. A description of the bounds follows.

In terms of the principal components transformation **C** defined in Appendix A, minimization of SEB (or SPE) in the direction of the ith principal axis, for estimators of the form $f_i a_i$, results in

$$f_i = \lambda_i/(\lambda_i + k_i) \ , \tag{4.1}$$

where $k_i = (\sigma/\alpha_i)^2$. One cannot use this estimator in practice, because the k_i are unknown. However, we can calculate the k_i for every sample in our simulation study, since we know the values of σ and $\boldsymbol{\alpha}$. The estimates based on (4.1) are referred to below as OPT.

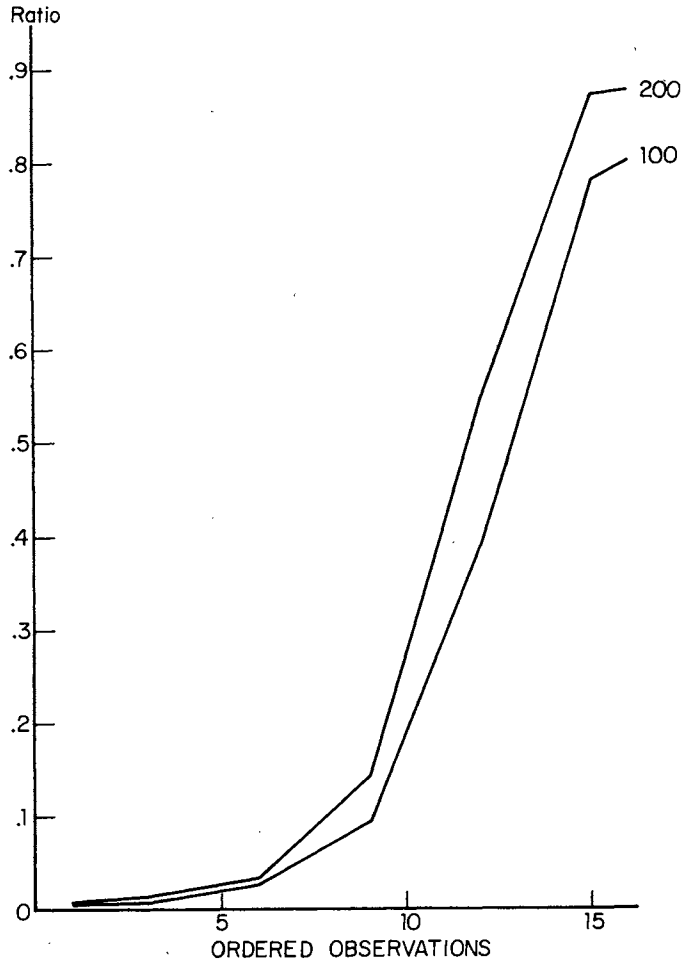
OPT is similar in form to RIDGE except that there is now a separate factor k_i along each principal axis. OPT provides a stochastic lower bound for RIDGE methods because it has additional degrees of freedom and also uses true, rather than estimated, values of the adjustment factors. The frequentist expectations of our two criteria for OPT are given by

$$E_{\text{OPT}}(\text{SEB}) = \sum_{1}^{p} f_i/\lambda_i \tag{4.2}$$

and

$$E_{\text{OPT}}(\text{SPE}) = \sum_{1}^{p} f_i \ , \tag{4.3}$$

A. Cumulative Distributions of $E_{\text{OPT}}(\text{SPE})/E_{\text{OREG}}(\text{SEB})$ by Noncentrality Level, Experiment 1



whereas the corresponding expectations for ordinary regressions are

$$E_{\text{OREG}}(\text{SEB}) = \sum_{1}^{p} 1/\lambda_i \qquad (4.4)$$

and

$$E_{\text{OREG}}(\text{SPE}) = p \, . \qquad (4.5)$$

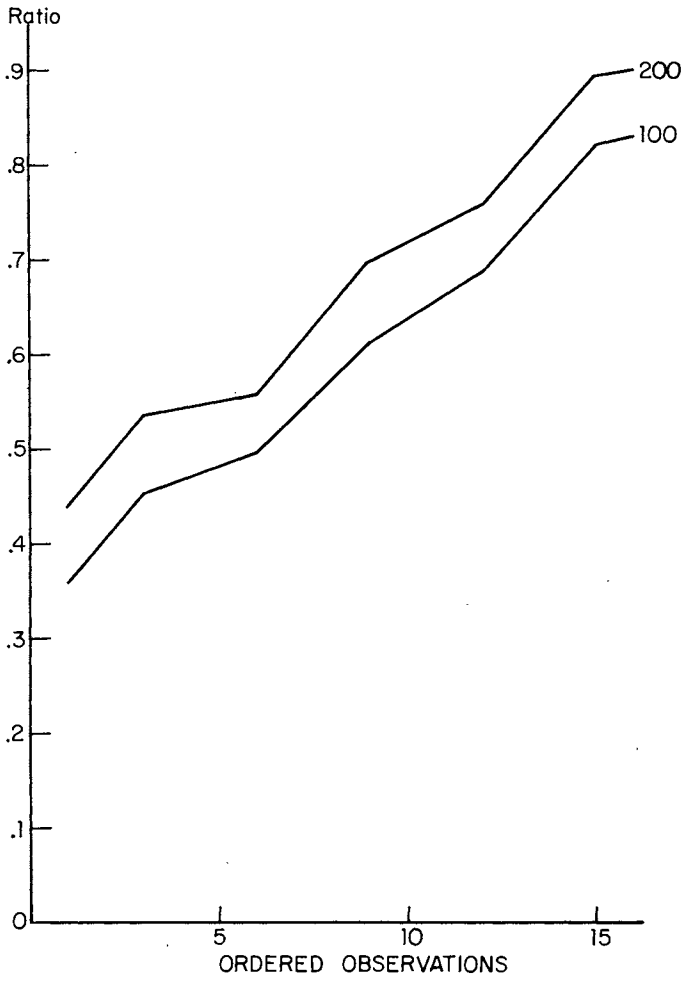
In evaluating the performance of different regression methods, we can use (4.1)–(4.5) in two ways to provide stochastic lower bound type comparisons. First, if we pretended that we knew the values of the f_i, we could apply these factors to the sample data, and evaluate the actual performance of OPT on such data. Second, the ratios of the corresponding expected values, (4.2)/(4.4) and (4.3)/(4.5), provide information as to the expected gains in SEB and SPE which could be realized by using the optimal pull-back procedure.

The shrinking factor f_i can be interpreted in terms of the noncentrality parameter, Δ_i^2, along the ith principal axis, as

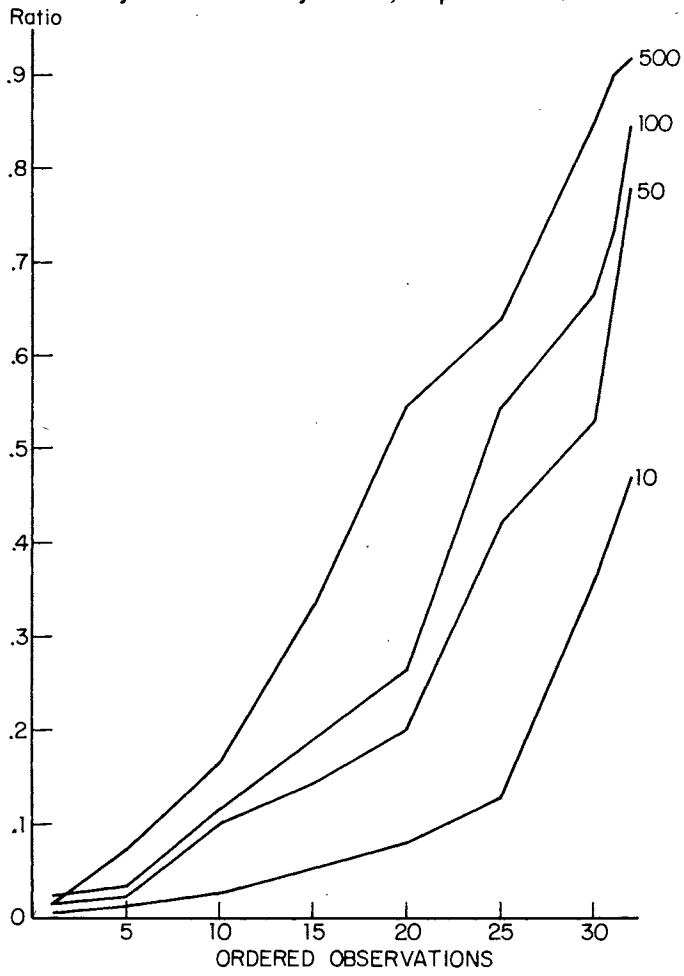
$$f_i = \Delta_i^2/(1 + \Delta_i^2) \, . \qquad (4.6)$$

Thus, when the noncentrality parameter is large, the shrinking factor is close to one, showing that least squares is asymptotically efficient.

B. Cumulative Distributions of $E_{\text{OPT}}(\text{SPE})/E_{\text{OREG}}(\text{SEB})$ by Noncentrality Level, Experiment 1



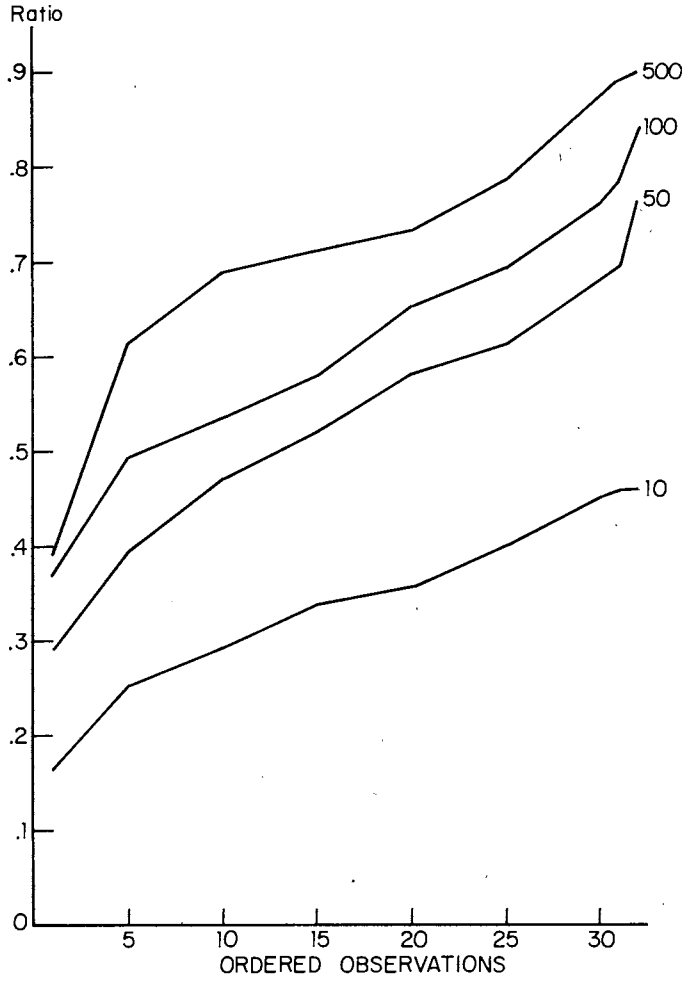
C. Cumulative Distributions of $E_{\text{OPT}}(\text{SEB})/E_{\text{OREG}}(\text{SEB})$ by Noncentrality Level, Experiment 2



One view of the particular parameter sets chosen for our two series of data sets is given in Figures A–D, which present cumulative sample distributions of the ratio of E_{OPT} to E_{OREG} for each of the criteria, by noncentrality level. It is readily apparent from these graphs that opportunities for large improvements via shrinking are greatest for small noncentralities. This property is borne out in our observed data, as illustrated in Table 2 which shows for the various methods the ratios of observed mean criterion values (SEB and SPE) to corresponding least-squares values, separately by noncentrality level. Generally, the largest relative gains have been achieved at the lowest noncentralities. This is particularly evident for SEB, which has larger expected gains that does SEP.

Further insight into the relative performance properties of the selected estimators is provided by Figures E–H, which show the cumulative sample distributions of the ratios of the observed values of the criteria for each of the methods, to their corresponding least-squares observed values. Examination of the SEB graphs (Figures E and G) shows that STEINM is the most conservative procedure of those being compared, in that it most closely approximates OREG. It provides improved performance in most cases (80 percent of the cases for Experiment 1, and 90 percent for Experiment 2), al-

D. Cumulative Distributions of E_{OPT}(SPE)/E_{OREG}(SPE) by Noncentrality Level, Experiment 2

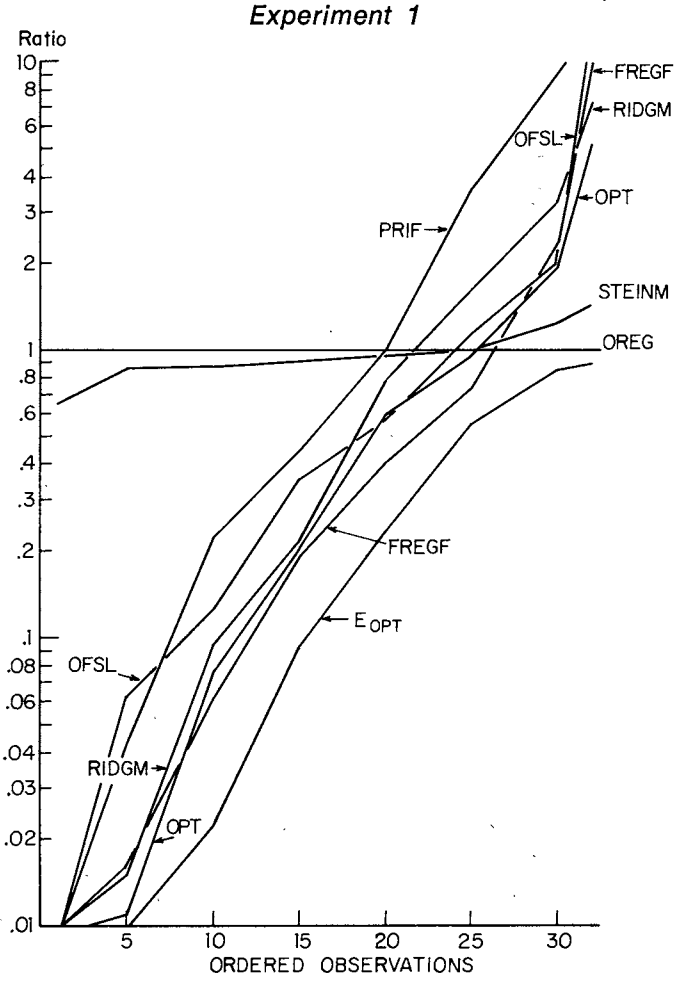


2. Ratios of Average Criteria Values for Individual Methods to Those for Least Squares, by Noncentrality Level

Method	Experiment 1		Experiment 2			
	NCP					
	100	200	10	50	100	500
SEB						
OFSL	.07	.46	.22	.49	1.03	.94
PRIF	.14	.38	.13	.31	.39	.46
STEINM	.88	.92	.38	.78	.90	.97
RIDGM	.05	.25	.09	.21	.27	.27
FREGF	.04	.31	.12	.37	.65	.84
EOPT	.03	.04	.03	.09	.12	.20
OPT	.03	.21	.06	.09	.17	.19
SPE						
OFSL	.65	1.06	.88	1.37	1.43	1.17
PRIF	1.70	1.54	1.14	1.26	1.13	1.22
STEINM	1.02	1.06	.75	1.03	.93	.98
RIDGM	.80	.94	.61	.88	.78	.88
FREGF	.57	.91	.73	1.50	1.30	1.10
EOPT	.59	.66	.34	.53	.61	.72
OPT	.54	.77	.35	.53	.61	.71

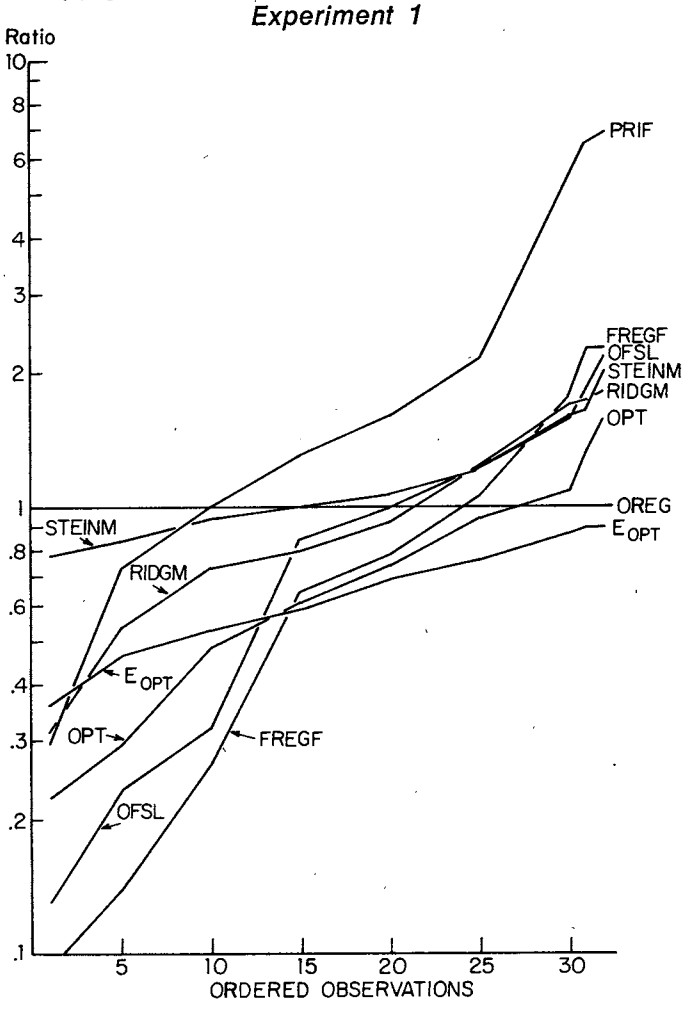
though the improvements are not as large on the average as those produced by the other methods. It also is less risky than the others in the sense that it seldom produces large degradations with respect to least squares.

E. Cumulative Distributions of SEB *Ratios, Experiment 1*



The FREGF procedure appears best on both criteria in Experiment 1, while RIDGM is the clear winner in Experiment 2. These conclusions are borne out also by Table 1, which shows that FREGF and RIDGM have the smallest median values on the two criteria for Experiments 1 and 2, respectively. The apparent reason for the reversal in performance from Experiment 1 to 2 is that FREGF is affected by the distribution of beta values, whereas RIDGM is not. In general, the performance of FREGF improves as the distribution of the beta values becomes more skewed, and as the degree of truncation (i.e., number of zero or near-zero values) increases. In Experiment 1, the mean SEB for RIDGM was approximately the same (46.2 and 44.1) at both levels of BETA, while the corresponding values for FREGF decreased from 75.1 to 23.5, for BET0 and BET1, respectively. Thus, the sharply improved performance of FREGF at BET1, corresponding to the extreme distribution (64, 16, 4, 1, .25, .0625), appears to be responsible for the better overall performance of FREGF in the first experiment.

F. Cumulative Distributions of SPE *Ratios, Experiment 1*



Comparisons of the distributions of SPE ratios (Figures F and H) yield similar results. Again, FREGF appears to be the best choice in the first experiment, while RIDGM looks best in the second.

4.3 Effects of Design Factors

In this section we assess further how the design factors affect the performance of the various methods. Logarithmic transformations were carried out on both SEB and SPE values because they have highly skewed chi-square distributions. Normal probability plots of the log-transformed sample values indicated reasonably normal behavior.

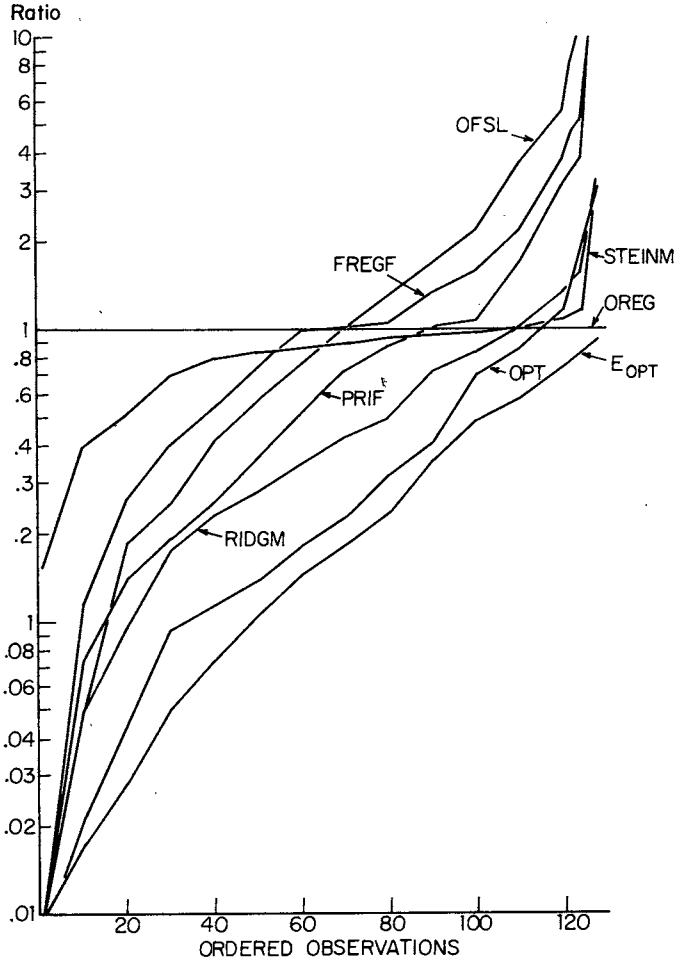
Initial analyses treated the data simply by computing analyses of variance and looking for significant main effects and first-order interactions. A typical analysis of this type for log SEB, Experiment 2, is presented in Table 3 for OREG, OFSL, STEINM, RIDGM, and FREGF. It shows F values for main effects and first-order interactions, together with signs of the effects, for all effects which are significant at the .05 level. The error sums of squares are based on pooling of second- and higher-order interactions. We note that there are relatively few significant interactions, and that they generally have much smaller F values than the significant main effects. In the case of OREG, there are no significant interactions at all. For most methods, eigenvalue structure, multicollinearity, and noncentrality are significant. Noncentrality is a significant factor in all methods except for OREG, and performance of each of these methods degrades, as previously noted in Section 4.2, as the degree of noncentrality increases.

The largest effect in OREG is the eigenvalue structure, as might be expected from (4.4). Collinearity and multicollinearity also have significant effects in OREG, displaying approximately the same F values and directions. The effects of rotation are not significant, which provides some reassurance as to the validity of the experimental results, since the random rotations can be regarded as replications in the case of ordinary regression.

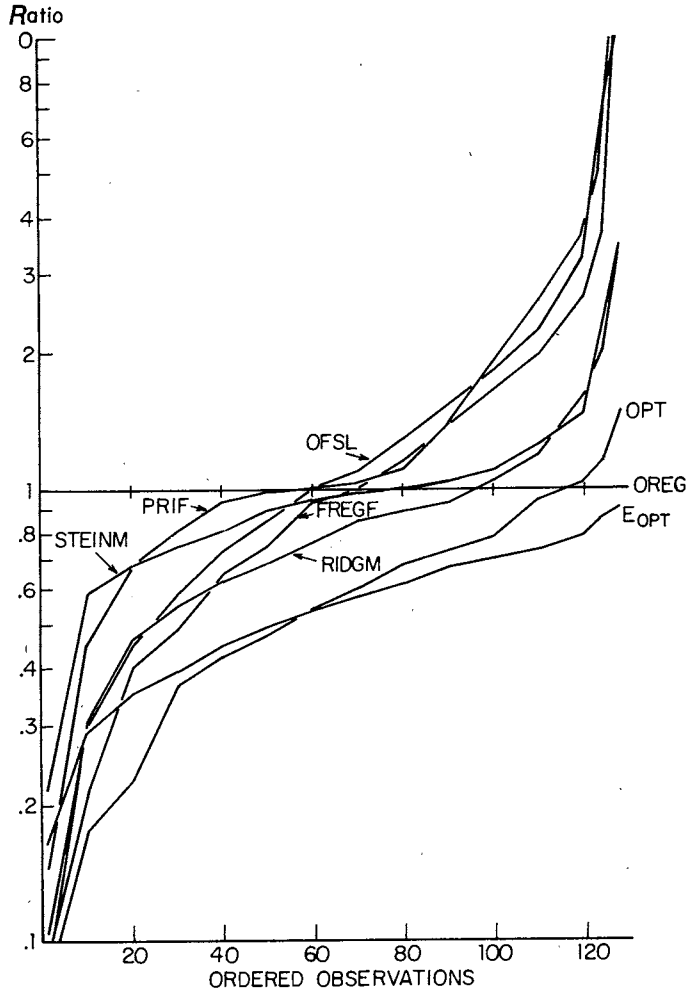
The effects of the regression coefficient structure are significant only for those methods which select variables (i.e., OFSL and FREGF). The coefficient structure has no effect on OREG, a result which should be expected since $E_{\text{OREG}}(\text{SEB})$ does not depend on the coefficient structure. Nor is it a significant factor in the smoother pull-back procedures RIDGM and STEINM.

A closer examination of the data is provided by Table

G. Cumulative Distributions of SEB *Ratios, Experiment 2*



H. Cumulative Distributions of SPE *Ratios, Experiment 2*



4 which presents the mean differences in log SEB and log SPE between each of five methods and either OREG, RIDGM, and FREGF. The results differ somewhat in the two experiments, but the overall picture is unchanged. Least-squares estimation (SEB) is strongly improved by FREGF, 1CPRI, and RIDGM in both experiments, and outstanding gains over least-squares prediction (SPE) are either observed for FREGF (in Experiment 1) or for RIDGM (in Experiment 2). From the paired comparisons with RIDGM it can be seen that FREGF is its only serious competitor. STEINM gives significantly larger estimation and prediction errors in both experiments, and generally, OFSL and 1CPRI give worse results as well. The classical selection-of-variables method OFSL turns out to be clearly inferior to the Bayes approach to variable selection: the significantly larger estimation errors of OFSL relative to FREGF in both experiments are not counterbalanced by smaller prediction errors.

3. F Values for .05 Level Significant Effects, Together with Signs of Effects, for Log SEB*, Experiment 2*

	OREG	STEINM	RIDGM	FREGF	OFSL
EIG	57.42	65.17	40.14	24.94	28.85
MCL	16.81	15.91	6.00	10.69	13.90
COL	15.43	13.68			
CEN1		12.92	28.80	27.16	13.74
CEN2		−8.00	−32.36	−59.46	−42.19
Beta shape				−5.16	
Beta truncation					−15.45
EIG × COL					−5.41
EIG × CEN2				−8.98	−7.65
MCL × CEN1					10.68
MCL × CEN2				−6.61	−13.97
COL × CEN2				−6.71	
CEN1 × CEN2		8.29	6.16	19.94	23.06
CEN2 × beta trunc			−4.28	−7.64	
MCL × ROT1					4.53

4. Mean Differences in Logs SEB *and* SPE

Standard method	Experiment	Compared methods: OREG	OFSL	FREGF	1CPRI	RIDGM	STEINM
			SEB				
OREG	1		−1.14[b]	−1.76[b]	−0.99[b]	−1.32[b]	−0.05
	2		−0.14	−0.49[b]	−0.84[b]	−1.12[b]	−0.22[b]
RIDGM	1	1.32[b]	0.12	−0.44	0.33		1.26[b]
	2	1.12[b]	0.98[b]	0.63[b]	0.27[b]		0.85[b]
FREGF	1	1.76[b]	0.61[b]		0.77	0.44	1.70[b]
	2	0.49[b]	0.36[b]		−0.35[b]	−0.63[b]	0.23
			SPE				
OREG	1		−0.41[a]	−0.66[b]	0.10	−0.15	0.06
	2		−0.04	−0.15	−0.10	−0.29	−0.08
RIDGM	1	0.15	−0.27	−0.51[b]	0.25[a]		0.21[b]
	2	0.29[b]	0.25[b]	0.14	0.19[b]		0.20[b]
FREGF	1	0.66[b]	0.25[b]		0.76[b]	0.51[b]	0.72[b]
	2	0.15	0.11		0.05	−0.14	0.06

[a] Indicates significance at the 0.05 level.
[b] Indicates significance at the 0.01 level.

The next level of analysis attempted to introduce the eigenvalue and noncentrality structures into the regression models in such a way as to reflect the ways in which they enter the expressions for the expected values of our two criteria, as given by (4.2)–(4.5). Specifically, using SEB as an example, we attempt to partition the total sums of squares of log SEB into components due to $\log E_{OPT}(SEB)$, $(\log E_{OREG}(SEB) - \log E_{OPT}(SEB))$, $(\log SEB_{OREG} - \log E_{OREG}(SEB))$, and components due to the various design factors.

The results of these analyses are presented in Table 5, which shows F values corresponding to terms in the models together with the squared multiple correlation coefficients. The rows of the tables are ordered according to the sequence of introduction of terms into the regression models. In all cases, error sums of squares are based on pooled interactions of all orders. In all of these analyses, the salient feature is that most of the explained variation is associated with the partitioning of the covariate (i.e., log SEB_{OREG} or log SPE_{OREG}) into its indicated components. Thus, after introducing these terms, which represent specific formulations of the ways in which the eigenvalue and noncentrality structures affect the re-

5. *F Values for Regression Models*

Regression model	Experiment 1						Experiment 2					
	OPT	PRIF	STEINM	RIDGM	FREGF	OFSL	OPT	PRIF	STEINM	RIDGM	FREGF	OFSL
					a. Log SEB							
Rotations	4.68[a]	.81	.93	9.24[b]	2.51	2.63	.26	1.05	5.10[b]	.84	.76	.16
Log $(\Sigma f_i/\lambda_i)$	64.43[b]	16.43[b]	19.13[b]	60.06[b]	2.83	5.29[a]	290.89[b]	227.11[b]	591.53[b]	173.73[b]	105.24[b]	76.32[b]
Log $(\Sigma 1/\lambda_i)/(\Sigma f_i/\lambda_i)$	.03	.18	3417.99[b]	1.60	5.70[a]	21.16[b]	1.20	1.93	762.22[b]	4.47[a]	3.93[a]	14.18[b]
Log (SEB OREG/$\Sigma 1/\lambda_i$)	.74	3.65	2329.73[b]	4.21	3.45	4.73[a]	8.38[b]	11.22[b]	1180.22[b]	46.04[b]	14.31[b]	3.92[a]
Log NCP	.00	2.07	.36	3.05	.07	.19	.04	1.77	43.91[b]	4.83[a]	11.20[b]	4.70[a]
EIG	.06	.05	.01	.46	.00	.66	1.33	1.30	.82	.85	.90	2.09
MCL							2.92	.59	.02	3.31	1.20	.08
COL	.82	2.15	.01	.78	3.20	3.48	.13	.39	.01	.00	.25	.02
Beta level	.11	3.48	.35	1.08	11.54[b]	12.80[b]						
Beta shape							1.85	.05	.31	.02	4.38[a]	2.06
Beta trunc							.18	.33	6.41[a]	2.76	3.30	14.15[b]
R^2	.75	.56	1.00	.78	.56	.69	.72	.68	.96	.67	.76	.50
					b. Log SPE							
Rotations	8.51[b]	4.63[a]	.55	.95	1.12	.05	.91	.42	2.05	1.49	.05	.23
Log SPE OREG	2.74	63.70[b]	.06	.18	1.03	.35	120.99[b]	22.77[b]	257.80[b]	136.61[b]	56.07[b]	32.52[b]
Log Σf_i	76.07[b]	.03	.98	1.81	22.56[b]	49.80[b]	70.26[b]	3.00	8.93[b]	20.21[b]	14.72[b]	13.48[b]
Log NCP	6.96[a]	.03	.98	.33	1.08	1.21	.03	.87	4.01[a]	.66	1.04	1.56
EIG	.64	.12	.01	.07	.33	.43	.56	.29	1.25	.47	1.33	1.51
MCL							3.01	2.24	7.44[a]	5.07[a]	.00	.30
COL	.03	.15	.27	.87	3.81	7.69[a]	.00	.10	.00	1.53	1.54	1.59
Beta level	2.49	.15	.00	.03	.89	.07						
Beta shape							.58	.81	.45	.17	19.37[b]	15.31[b]
Beta trunc							.05	1.18	.46	2.70	1.90	10.62[b]
R^2	.80	.74	.11	.15	.56	.71	.63	.22	.71	.60	.45	.40

[a] Indicates significance at the 0.05 level.
[b] Indicates significance at the 0.01 level.

spective criteria, the residual effects of the original design factors eigenvalue structure, collinearity, and multicollinearity, usually become insignificant. In most cases, noncentrality level also disappears as a significant factor, although it has a sizeable effect on log SEB for the STEINM and FREGF methods. The effects of beta structure are highly significant only for FREGF and OFSL, which select variables corresponding to significant beta values.

5. CONCLUDING REMARKS

We have illustrated the large improvements in SEB and SPE which are theoretically known to be possible through the substitution of either smooth or discontinuous pull-back procedures for ordinary least squares. The potential relative gains in accuracy for individual estimated regression coefficients are seen to be typically much larger than those for predicted **Y** values. The expected gains in SEB and SPE are fairly small for STEINM, while for RIDGM and PRIF they are larger and exhibit dependencies on design variables similar to those of the artificial method OPT, i.e., dependencies mainly on eigenvalues and noncentrality factors. Methods such as OFSL or FREGF, which select variables, are highly dependent additionally on the pattern of true regression coefficients. The relatively conservative STEINM procedure has one advantage, in that it less often does worse than OREG. The relative accuracy of traditional selection of variables methods based on significance testing behaves erratically in relation to significance levels adopted. While these results could have been predicted in a general way from either Bayesian or frequentist theoretical considerations, we believe numerical depictions make the results both concrete and vivid.

We would like to see the methodology of the study moved closer to helping the statistician with a specific body of data under analysis. For example, as remarked in Section 1, we could perform similar studies with the design matrix **X** fixed at the values of a particular data set. Another device would be to apply regression methods to $n - 1$ rows of $(\mathbf{Y}, \mathbf{X})$ and to evaluate predictions for the complementary row, and repeat with different selection of rows, much in the spirit of jackknifing.

Yet another, and more Bayesian way, would be to ask the statistician to specify 5 or 10 prior distributions covering a plausible range, and then to compare the closeness of various alternative regression procedures to the 5 or 10 posterior means. The Bayesian view offers some clarification of the real problem posed by a given set of data: if the correct analysis depends critically on the model and prior adopted, over some reasonable range, then the statistician should not expect any favorite procedure taken from his kit of tools to be automatically applicable. Our study suggests that conflicts will often appear among REGF, RIDGE, and STEIN estimates which should cause statisticians to reexamine both their data and their prior understanding for clues.

APPENDIX A

The normal linear model is written in the form

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e} \qquad \text{(A.1)}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of values on the response or dependent variable, $\mathbf{X}$ is an $n \times p$ matrix giving corresponding values on p independent variables, and $\mathbf{e}$ is an $n \times 1$ vector of independent $N(0, \sigma^2)$ disturbances. The least-squares estimator $\mathbf{b}$ of β is expressible in terms of $\mathbf{Y}$ and $\mathbf{X}$ according to

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y}) , \tag{A.2}$$

which together with the residual mean square

$$s^2 = (\mathbf{Y} - \mathbf{X}\mathbf{b})^T(\mathbf{Y} - \mathbf{X}\mathbf{b})/(n - p) , \tag{A.3}$$

defines sufficient statistics for the parameters β and σ^2 of (A.1). We can summarize the sampling distribution aspects of the model by asserting that $\mathbf{b}$ and $(n - p)s^2$ are independently distributed as $N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ and $\sigma^2\chi_{n-p}^2$, given fixed nonsingular $\mathbf{X}^T\mathbf{X}$.

The RIDGE and PRI methods are closely related to a principal components analysis of the p independent variables. Accordingly, it is convenient to have notation for replacing the given independent variables $\mathbf{X}$ by p linear combinations

$$\mathbf{X}^* = \mathbf{X}\mathbf{C}^T \tag{A.4}$$

where $\mathbf{C}$ is chosen such that

$$\mathbf{X}^{*T}\mathbf{X}^* = \mathbf{C}(\mathbf{X}^T\mathbf{X})\mathbf{C}^T = \text{diag}\ (\lambda_1, \lambda_2, \ldots, \lambda_p) \tag{A.5}$$

and

$$\mathbf{C}\mathbf{C}^T = \mathbf{I} . \tag{A.6}$$

The model (A.1) is correspondingly reexpressed as

$$\mathbf{Y} = \mathbf{X}^*\alpha + \mathbf{e} \tag{A.7}$$

where

$$\alpha = \mathbf{C}\beta . \tag{A.8}$$

The least-squares estimates

$$\mathbf{a} = \mathbf{C}\mathbf{b} \tag{A.9}$$

of α have a simple sampling distribution. Specifically, the components a_i are independently distributed with $N(\alpha_i, \sigma^2/\lambda_i)$ distribution for $i = 1, 2, \ldots, p$.

The principal components transformation $\mathbf{C}$ is not invariant in general under linear transformations of the p independent variables, and in particular is not invariant under linear changes of scale in each variable separately. In our work we have rescaled the variables so that the diagonal elements of $\mathbf{X}^T\mathbf{X}$ are all unity, so that our principal components are, apart from a single scale factor, the principal components of the correlation matrix among the independent variables. This convention is an essential part of the definition of the RIDGE and PRI methods which we study.

Two criteria are used throughout our study to measure the deviation of an estimated vector $\hat{\beta}$ from its true value β, namely

$$\text{SEB} = (\hat{\beta} - \beta)^T(\hat{\beta} - \beta)/\sigma^2 \tag{A.10}$$

and

$$\text{SPE} = (\hat{\beta} - \beta)^T\mathbf{X}^T\mathbf{X}(\hat{\beta} - \beta)/\sigma^2 . \tag{A.11}$$

SEB abbreviates sum of errors of betas, while SPE abbreviates sum of prediction errors. The connection between SPE and prediction follows. Suppose that a new set of responses $\mathbf{Y}^*$ is drawn from the linear model $\mathbf{Y}^* = \mathbf{X}\beta + \mathbf{e}^*$ using the same design matrix $\mathbf{X}$ appearing in the original data set which yielded $\hat{\beta}$, and using the same β, but using new $\mathbf{e}^*$ independent of the original $\mathbf{e}$. If $\mathbf{X}\hat{\beta}$ is used to predict $\mathbf{Y}^*$, then the sum of squares of prediction errors averaged over $\mathbf{e}^*$ for fixed $\hat{\beta}$ is $\sigma^2 + \sigma^2$ SPE.

The formulas expressing SEB and SPE in terms of principal components are

$$\text{SEB} = \sum_{i=1}^{p} (\hat{\alpha}_i - \alpha_i)^2/\sigma^2 \tag{A.12}$$

and

$$\text{SPE} = \sum_{i=1}^{p} \lambda_i(\hat{\alpha}_i - \alpha_i)^2/\sigma^2 . \tag{A.13}$$

APPENDIX B

Further details of the REGF methods are sketched here. In particular, we derive the posterior weights $\omega(I)$ used in (2.11), and we define the stopping criteria F_1 and F_2. The notation established in Section 2.4 represents by I a subset of r of the p independent variables, and by $\mathcal{I}_r$ the class of all such subsets. Given that only the components $\beta(I)$ of β are nonzero, (A.1) can be written in the form

$$\mathbf{Y} = \mathbf{X}(I)\beta(I) + \mathbf{e} \tag{B.1}$$

where $\mathbf{X}(I)$ and $\beta(I)$ denote the parts of $\mathbf{X}$ and β corresponding to I. The obvious generalizations of (A.2) and (A.3) define $\mathbf{b}(I)$ and $s(I)^2$.

For a given value of r, the REGF procedures assume that the $\binom{p}{r}$ members of $\mathcal{I}_r$ are *a priori* equally likely candidates to specify the r nonzero components of β. Given any particular $I \in \mathcal{I}_r$, it remains to specify a prior density for $\beta(I)$ and σ^2. The REGF and DRGF subfamilies are specified by flat prior density elements of the form $K(I)d\beta(I)$ where

$$K(I) = 1 \quad \text{and} \quad K(I) = [\det\ (\mathbf{X}(I)^T\mathbf{X}(I))]^{\frac{1}{2}} , \tag{B.2}$$

respectively. The alternative forms of $K(I)$ arise from considering $\beta(I)$ to have a multivariate normal distribution with alternative covariance matrices CI and $C(\mathbf{X}(I)^T\mathbf{X}(I))^{-1}$, respectively, and then letting $C \to \infty$. We also use the common form of flat gamma prior density for $h = 1/\sigma^2$, namely density element $h^{a/2-1}dh$ where $a = 0$, 1, 2 are typical choices. Thus the joint prior density element for I, $\beta(I)$ and h is

$$[K(I)d\beta(I)] \times [h^{a/2-1}dh] , \tag{B.3}$$

where a remains to be chosen.

The likelihood of I, $\beta(I)$, and h from (B.1) is proportional to

$$h^{n/2} \exp\ (-(h/2)(Q_1 + Q_2)) \tag{B.4}$$

where

$$Q_1 = (\mathbf{b}(I) - \beta(I))^T\mathbf{X}(I)^T\mathbf{X}(I)(\mathbf{b}(\mathbf{I}) - \beta(I)) \tag{B.5}$$

and

$$Q_2 = (n - r)s(I)^2 . \tag{B.6}$$

Multiplying (B.3) and (B.4) and integrating out $\beta(I)$ and h, we find that the posterior probabilities are proportional to

$$\omega^*(I) = K(I)[\det(\mathbf{X}(I)^T\mathbf{X}(I))]^{\frac{1}{2}}s(I)^{(r-a-n)/2} \tag{B.7}$$

whence the actual posterior weights used in (2.11) are

$$\omega(I) = \omega^*(I)/\sum_{J \in \mathcal{I}_r} \omega^*(J) . \tag{B.8}$$

The choices

$$a = r \quad \text{and} \quad a = r - 1 \tag{B.9}$$

were associated with the choices (B.2) for $K(I)$ to define REGF and DRGF, respectively.

The selection criterion F_1 used by FREGF, FDRGF, IFREGF, and IFDRGF is an analog of a 2 log-likelihood ratio statistic. Specifically, the statistic used to judge whether r should be increased from r to $r + 1$ is

$$F_1 = L_{r+1} - L_r \tag{B.10}$$

where

$$L_r = -\sum_{I \in \mathcal{I}_r} \omega(I) \log\ ((n - r)s(I)^2) . \tag{B.11}$$

We treated F_1 as nominally an F on $(1, \infty)$ degrees of freedom.

The F_2 criterion used by OREGF, ODRGF, IREGF, and IDRGF is also defined by posterior averaging. Specifically,

$$F_2 = \sum_{I \in \mathcal{I}_r} \omega(I)F_2(I) \tag{B.12}$$

where $F_2(I)$ is defined by applying (2.11) to the fitted subset I.

APPENDIX C

The devices used for introducing prespecified collinearity and multicollinearity into a given correlation matrix **R** consist of repeated applications of transformations of the type

$$\mathbf{R} \rightarrow \mathbf{CRC}^T \qquad \text{(C.1)}$$

where **C** is square and nonsingular. **C** may be naturally interpreted as a mapping from one basis of the linear space of six independent variables to another basis, where **R** and $\mathbf{CRC}^T$ are the initial and final covariance matrices. For example, in Experiment 1, to introduce .99 into the (1, 2) position of **R** we first used a **C** which replaced the standardized variables by their sum and difference, so that (C.1) took the form

$$\begin{bmatrix} 1 & r_{12} & \dots \\ r_{12} & 1 & \dots \\ \vdots & \vdots & \end{bmatrix} \rightarrow \begin{bmatrix} 1+r_{12} & 0 & \dots \\ 0 & 1-r_{12} & \dots \\ \vdots & & \end{bmatrix} . \qquad \text{(C.2)}$$

We then rescaled the new variables by $[1.99/(1 + r_{12})]^{\frac{1}{2}}$ and $[.01/(1 - r_{12})]^{\frac{1}{2}}$ so that the right side of (C.2) achieved a similar form with r_{12} replaced by .99. Finally, we reversed the transformation used in (C.2) to get the desired result. In Experiment 2, the first step (C.2) was the same. We then rescaled the difference variable to have unit variance, and in order to introduce a correlation .92 between this variable and the third variable we applied the same technique as in Experiment 1. Finally, we unstandardized the difference variable to get back the form on the right side of (C.2) and we repeated the technique of Experiment 1 to introduce correlation .95 between the first two variables.

[*Received May 1974. Revised September 1976.*]

REFERENCES

[1] Baranchik, Alvin J., "A Family of Minimax Estimators of the Mean of a Multivariate Normal Distribution," *Annals of Mathematical Statistics*, 41, 2 (1970), 642–5.

[2] Conniffe, D., and Stone, J., "A Critical View of Ridge Regression," *The Statistician*, 22 (1973), 181–7.

[3] Dempster, Arthur P., "Alternatives to Least Squares in Multiple Regression," in D. Kabe and R.P. Gupta, eds., *Multivariate Statistical Inference*, Amsterdam: North-Holland Publishing Co., 1973, 25–40.

[4] Efron, Bradley, and Morris, Carl, "Limiting the Risk of Bayes and Empirical Bayes Estimators—Part I: The Bayes Case," *Journal of the American Statistical Association*, 66 (December 1971), 807–15.

[5] ———, and Morris, Carl, "Limiting the Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case," *Journal of the American Statistical Association*, 67 (March 1972), 130–9.

[6] ———, and Morris, Carl, "Empirical Bayes on Vector Observations—an Extension of Stein's Method," *Biometrika*, 59, 2 (1972), 335–47.

[7] ———, and Morris, Carl, "Stein's Estimation Rule and Its Competitors—an Empirical Bayes Approach," *Journal of the American Statistical Association*, 68 (March 1973), 117–30.

[8] ———, and Morris, Carl, "Combining Possibly Related Estimation Problems," *Journal of the Royal Statistical Society, Ser. B*, 35, 3 (1973), 379–421.

(9) ———, and Morris, Carl, "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association*, 70 (June 1975), 311–9.

[10] Goldstein, M., and Smith, Adrian F.M., "Ridge-Type Estimations for Regression Analysis," *Journal of the Royal Statistical Society, Ser. B*, 36, 2 (1974), 284–91.

[11] Hoerl, Arthur, and Kennard, Robert, "Ridge Regression: Biased Estimation for Non-Orthogonal Problems," *Technometrics*, 12 (February 1970), 55–63.

[12] ———, and Kennard, Robert, "Ridge Regression: Applications to Non-Orthogonal Problems," *Technometrics*, 12 (February 1970), 69–82.

[13] James, W., and Stein, Charles, "Estimation with Quadratic Loss," in J. Neyman, ed., *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: The University of California Press, 1961, 361–79.

[14] Lewis, Peter A.W., Goodman, A.S., and Miller, J.M., "A Pseudo-Random Number Generator for the System/360," *IBM Systems Journal*, 8, 2 (1969), 136–46.

[15] Lindley, Dennis V., and Smith, Adrian F.M., "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society, Ser. B*, 34, 1 (1972), 1–41.

[16] Mallows, Colin L., "Some Comments on C_p," *Technometrics*, 15 (November 1973), 661–75.

[17] Rolph, John E., "Choosing Shrinkage Estimators for Regression Problems," *Communications in Statistics—Theory and Methods*, A5(9) (1976), 789–801.

[18] Sclove, Stanley, "Improved Estimators for Coefficients in Linear Regression," *Journal of the American Statistical Association*, 63 (June 1968), 596–606.

[19] Wermuth, Nanny, *An Empirical Comparison of Regression Methods*, unpublished Ph.D. thesis, Department of Statistics, Harvard University, Cambridge, Mass., 1972.

[20] ———, *APL-Functions for Data Simulation, Regression Methods and Data Analysis Techniques*, Research Report CP-15, Department of Statistics, Harvard University, Cambridge, Mass., 1972.

[21] Zellner, Arnold, and Vandaele, Walter, "Bayes-Stein Estimators for *k*-Means Regression and Simultaneous Equations," *H.G.B. Alexander Research Foundation*, Graduate School of Business, University of Chicago, 1972.

Comment

BRADLEY EFRON and CARL MORRIS*

Drs. Dempster, Schatzoff, and Wermuth deserve commendation for their substantial and novel efforts which have resulted in a comparison of 57 competing estimation methods over a variety of situations. We applaud their effort to shed light on the task of the data analyst, and share their hope that their methodology also will be of use in other problems.

* Bradley Efron is Professor and Chairman, Department of Statistics and Biostatistics, Stanford University, Stanford, CA 94305. Carl Morris is Senior Statistician, The RAND Corporation, Santa Monica, CA 90406.

In these comments, we shall be concerned with just two variants of ordinary regression, RIDGE and STEIN. Our interest is focused in this way because we have given considerable thought to estimators of these types in our own research over recent years on empirical Bayes and Stein-type estimation, because more analytical power can be brought to bear on these methods than suggested in the paper, and because for the loss functions used, the performance of these estimators depends on the experimental inputs only through the eigenvalues of the matrix $X'X$. This last reason simplifies matters greatly for understanding RIDGE and STEIN, although it does not apply to most of the other 57 varieties. For lack of space, and because we haven't studied them carefully, we will not discuss the confidence contour constraint variants of Section 2.4.5 of RIDGE and STEIN.

This leaves STEINM, RIDGM, and SRIDG as the only three rules under consideration. When the eigenvalues $\{\lambda_i\}$ of $X'X$ are equal, STEINM and RIDGM reduce to the same rule, but SRIDG behaves very badly, being highly nonminimax. By making use of arguments in [3], it easily can be shown to be substantially and uniformly improved upon by the James-Stein rule. We also expect SRIDG to perform poorly when the eigenvalues are unequal, which happened in the simulation. Since SRIDG is so complicated in general, and has not been recommended elsewhere, we shall take RIDGM to be the only interesting version of RIDGE considered in the study.

Certainly the most dramatic conclusion of the study is that a version of the ridge method, in the form of RIDGM, is the best method used in the study and dominates a Stein-type method, in the form of STEINM. While we agree with this conclusion, and have made similar statements in our own work [2, 3], we believe different language should be used; language based on an understanding of why this is so. In particular, we see justification for RIDGM arising out of the empirical Bayes literature in combination with the implicit assumptions of this experiment, and not from the ridge-trace graphical technique used for estimation of regression coefficients.

The reader should note that STEINM is not the James-Stein rule (which we shall term JSTEIN) [4], and overshrinks JSTEIN by a factor of $[(n - p + 2)/(n - p)]p/(p - 2)$, or 12/7 in this case. This sacrifices about 51 percent of the improvement in risk of JSTEIN over the ordinary regression estimator (OREG) for the loss function SPE. We make this statement for the data of these experiments, knowing that the authors claim the contrary at the end of Section 2.4.1. The risk function for SPE loss is a function only of CEN, the noncentrality parameter, which takes on the five values CEN = 10, 50, 100, 200, 500 in this experiment. The approximate value of the risk at these points (actually an upper bound) is 5.00, 5.74, 5.87, 5.93, and 5.972 for JSTEIN while STEINM has 5.51, 5.87, 5.93, 5.97, and 5.986. Ordinary regression has SPE risk of 6.00 for all values of CEN, and so even JSTEIN would not improve least squares substantially in these experiments. The James-Stein estimator should not be applied with the loss SEB, and is known not to be minimax in this case.

But RIDGM also is better than JSTEIN in these experiments. This is expected for experiments or problems where the regression parameters $\{\beta_i\}$ have an exchangeable prior distribution. In the Dempster, Schatzoff, and Wermuth experiments, the random rotation matrix G tends to symmetrize the prior distribution, and this fact combined with the dispersed set of eigenvalues insures that RIDGM will be better. Professor Thisted's comments are much more complete on this issue.

In relation to the preceding paragraph, one of our concerns about the widespread application of ridge regression is that, being a data-based Bayes rule against an exchangeable prior, it is necessary to feel confident in the exchangeability assumption, while in most real *regression* problems the statistician couldn't be. In certain nonregression problems, for instance in the examples of [2], exchangeability seems plausible *a priori*, but when this assumption is violated significantly, ridge rules can be much worse than the ordinary regression rule. That is, ridge rules are not minimax in general. We will put these issues in more mathematical form to make the argument clearer.

If we allow ourselves the luxury of thinking of n as large (instead of 20) so that σ^2 may be assumed known (this could be relaxed at the price of additional mathematical complexity), the Dempster, Schatzoff, and Wermuth estimation problem may be expressed in canonical form as the problem of observing

$$Y_i \sim N(\theta_i, V_i) \ , \quad i = 1(1)p \tag{1.1}$$

independently with $V_i \equiv \sigma^2/\lambda_i$ and σ^2 known, $Y_i \equiv a_i = (\mathbf{C}b)_i$, $\mathbf{C}$ being the principal components orthogonal transformation defined in Section 2.2. In this notation $\theta_i \equiv (\mathbf{C}\beta)_i$ is to be estimated with risk function

$$R = E \sum_{i=1}^{p} L_i(\hat{\theta}_i - \theta_i)^2 \tag{1.2}$$

where $L_i = 1$ for SEB and $L_i = 1/V_i$ for SPE.

Letting $A \equiv \sigma^2/k$, the independent prior distributions

$$\theta_i \sim N(0, A) \ , \quad i = 1(1)p \ , \tag{1.3}$$

lead to the Bayes estimator

$$E\theta_i | Y_i = (1 - B_i)Y_i \ , \quad B_i \equiv V_i/(V_i + A) \ . \tag{1.4}$$

We say that (1.4) is an "empirical Bayes estimator" if A is estimated from the independent marginal distributions (derived by integrating out the distributions of θ_i in (1.1) with respect to (1.3))

$$Y_i^2 \sim (A + V_i)\chi_1^2 \ , \quad i = 1(1)p \ , \tag{1.5}$$

and then the estimate $\hat{A}$ is used in (1.4) in place of the unknown value A.

Defining $\hat{A}_i \equiv Y_i^2 - V_i$, we have $E\hat{A}_i = A$, so these are p independent unbiased statistics which may be used to estimate the unknown value A. Obvious unbiased

estimates of A are of the form

$$\hat{A} \equiv \sum \hat{A}_i W_i \ , \quad \sum W_i = 1 \ . \tag{1.6}$$

The choice $W_i = W_i(A) \propto 1/\text{Var}\ (\hat{A}_i) = .5/(V_i + A)$ results in the RIDGM estimate if $W_i(A)$ is replaced by $W_i(\hat{A})$ in (1.6). The rule we proposed in [2] for the toxoplasmosis data, which we label EBMLE, derives from $W_i(A) \propto 1/\{\text{Var}\ (\hat{A}_i)\}^2 = .25/(V_i + A)^2$, $W_i(A)$ replaced by $W_i(\hat{A})$ in (1.6). This is the optimal linear estimator in (1.6), and is equivalent to the maximum likelihood estimator of A from (1.5). Hence the term EBMLE signifies the empirical Bayes model with maximum likelihood estimation of A.

Both of the estimators of the preceding paragraph take advantage of the exchangeability of (1.3) and therefore are superior to STEINM and JSTEIN for priors of the form (1.3). It is essential to note, however, that if (1.3) were replaced by $\theta_i \sim N(0,\ V_i A)$ then JSTEIN, which in this notation is $(1 - (p - 2)/\sum Y_j^2/V_j)Y_i$, and STEINM would outperform EBMLE and RIDGM. The experiment therefore has proved that the array of experimental regression coefficients is better represented by Var $(\theta_i) = A$ than Var $(\theta_i) = V_i A$.

Carter and Rolph used their own version of RIDGM quite successfully in an empirical Bayes application to spatial analysis [1]. (Actually both the Carter-Rolph rule and EBMLE [2] were modified slightly so that they reduce to the James-Stein rule when the V_i all are equal.) While we don't know whether EBMLE or RIDGM is better for small or moderate p, EBMLE must be better for large p if (1.3) holds because of its relation to the maximum likelihood estimator. It would be interesting to compare these rules on the 160 data sets of the experiment.

If the V_i are sufficiently unequal, then for certain configurations of the parameters $\theta_1, \ldots, \theta_p$, both EBMLE and RIDGM can be much worse than OREG for both losses SEB and SPE. (The problem rests precisely with the component having the large V_i or small λ_i, i.e., the component most ridge papers are concerned about.) JSTEIN, of course, is guaranteed to improve upon OREG for SPE, while it is not minimax for SEB.

It should be clear from the preceding discussion about SRIDG, RIDGM, and EBMLE that there are many ways to estimate the constant k from the data. Although almost every ridge paper published, including this one, has presented a different method, the expression "*the* ridge estimator" continues to be used. In fact, ridge estimators are a class of Bayes rules against normal priors indexed by k, and the effectiveness of a given rule depends upon how k is estimated. Some published ridge estimators are drastically different from others, and some are disastrously bad. We believe that the important problem now is to find estimators of k which have good risk properties in the class of all possible estimators.

Because most applications of Stein's rule require its generalization to the unequal variances situation, and because ridge regression formally reduces to this situation, we have given a great deal of thought to the problem framed by (1.1)–(1.5) over the past several years. This includes derivation of a wide class of minimax estimators which encompasses most estimators already proven to be minimax by other writers. We also have considered numerous empirical Bayes rules, partly in light of a necessary condition for minimaxity. In the equal variance situation of James and Stein [4], minimax rules with Bayesian properties against exchangeable priors exist, but when orthogonal invariance is sacrificed this happy result disappears. When the variances V_i are sufficiently unequal, our current understanding is as follows. A fundamental tension exists between minimax and empirical Bayes (or ridge) requirements, and no rules appear to exist which are satisfactory from both standpoints. One cannot approximate the Bayes rule against the prior (1.3) without risk of doing worse than the Gauss-Markov estimator. Improvement on the Gauss-Markov estimator in regression situations therefore can be guaranteed only with external information about the prior distribution of the regression coefficients. Unfortunately, such information is not available for many applications.

To summarize, the statistician has a choice of shrinkage rules to consider in applications to real data, and must be careful in exercising this choice because, while the rewards can be great, so also can be the penalties. No choice uniformly dominates the Gauss-Markov estimator for all loss functions. The statistician, therefore, must know enough about his data and about the properties of the alternative estimators available to him to make an intelligent choice of rule. For the Dempster, Schatzoff, and Wermuth experiments, the exchangeable prior, and therefore the use of RIDGM or EBMLE, seems to be justified in aggregate, although the statistician who looks at the 160 individual problems might choose not to use the same rule in all of these situations. Other experiments could give opposite conclusions, so the reader's faith in the results of this experiment ultimately depends on how much he believes the Dempster, Schatzoff, and Wermuth data sets typify real world experience.

REFERENCES

[1] Carter, Grace M., and Rolph, John E., "Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities," *Journal of the American Statistical Association*, 69 (December 1974), 880–5.

[2] Efron, Bradley, and Morris, Carl, "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association*, 70 (June 1975), 311–9.

[3] ———, and Morris, Carl, "Families of Minimax Estimators of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 4, 1 (1976), 11–21.

[4] James, W., and Stein, Charles, "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability, Vol. 1*, Berkeley: University of California Press, 1961, 361–79.

Comment

ARTHUR E. HOERL*

The authors are to be congratulated for so clearly presenting such a breadth of material. In addition, the theoretical Bayesian development which leads to the RIDGM algorithm is most ingenious and goes a long way in attaining a realistic share of the ridge potential in reducing mean square error.

In the joint papers [2, 3], where simulation was used to examine ridge algorithms, orthogonal **Z** matrices of various dimensions and conditionings were defined. Random vectors α with specified norm (designated as $a^2 = \alpha'\alpha$) and ϵ were generated, and the resultant least squares and ridge regression were characterized through a quadratic loss function. In these simulations our original algorithm relied on the minimum quadratic loss estimate $k_i = \sigma^2/\alpha_i^2$. This is, as the authors state, an interesting identity since it depends on α_i^2 (and σ^2) only and not on its corresponding eigenvalue λ_i. However, one can readily satisfy oneself that when using sample estimates the resultant loss is relatively large. Therefore, this motivated us to use a single k_a equal to the harmonic mean of the p values $k_1, \ldots, k_p$, i.e.,

$$k_a = ps^2/\hat{\alpha}'\hat{\alpha} \ .$$

Subsequent to this simulation publication [2], an iterative algorithm with an empirical stopping rule based on successive estimates of $\alpha'\alpha$ has been published [3]. Since the sample value $\hat{\alpha}'\hat{\alpha}$ overestimates $\alpha'\alpha$, k_a tends to underestimate its goal value. Therefore, since the first ridge cut, k_{ao}, results in an improved estimate, $\hat{\alpha}_{Ro}$, of α, then $\hat{\alpha}_{R1}$ based on the square length of $\hat{\alpha}_{Ro}$ would perhaps be a better estimate of α. The successive estimates k_{ai} can be repeated until the rate of change in k_{ai} has stabilized. This can be specified under an empirical stopping rule.

For the purposes of multiple comparisons of many estimation techniques, the authors were wise in formulating their comparisons by averaging their results over a range of signal-to-noise. However, for the purposes of evaluating a few algorithms it is perhaps more illustrative to display the relative effectiveness of each at specific values of signal-to-noise since it is not exclusively the frequentists who would argue that in formulating a simulation strategy the following criterion be used: if it can be shown that one estimator is superior to another for all specified values of $\alpha'\alpha$ over a wide spectrum of conditioning and a range of p, then regardless of the real world frequency distribution of $\alpha'\alpha$ (assuming a finite domain), that estimator is preferable to the other. If the two estimators vacillate in superiority as a function of $\alpha'\alpha$, some subjective judgment would then be required. As an example of this approach, the following is presented.

Using the Gorman-Toman data [1] with two spectrums of eigenvalue structure used in [2, 3], a series of 200 simulations was performed for each $\alpha'\alpha$ value. The uniform random number generator described in the Appendix served as the basis for the simulation. Unit normals were generated by summing 12 random uniforms on $(-5, .5)$. Least-squares estimates $\hat{\alpha}_i$ were generated by $\hat{\alpha}_i = \alpha_i + E_i$ with E_i defined by the sum of 12 uniforms divided by $\sqrt{\lambda_i}$. Sample values s^2, of $\sigma^2 = 1$, were defined by the sum of 25 squared unit normals divided by the df 25.

For the 10-factor basic (the same eigenvalue structure as originally published) the results in Section a of the table were obtained. As a guide, the expectation for the F ratio is included. The critical value of $F(10, 25)$ with $\alpha = .05$ is 2.24. For Section a, the expectation for the least-squares error is 32.58 and the maximum potential is defined to be the minimum possible square error for

Average Square Error

$\alpha'\alpha$	E(F Ratio)	Ordinary L.S.	Basic k_a	Iterative k_{ai}	RIDGM	Maximum potential
			a. 10-Factor Basic			
1	1.20	34.12	7.99	6.80	2.18	.88
5	1.63	31.85	8.83	8.57	5.15	3.47
10	2.17	30.03	9.66	9.89	7.53	5.52
15	2.72	31.55	11.06	11.15	9.10	7.01
25	3.80	30.70	12.75	12.81	12.05	9.51
50	6.52	34.28	18.23	18.23	17.83	13.82
100	12.0	31.84	20.76	20.76	20.56	16.53
200	22.8	30.12	24.69	24.69	24.92	19.55
500	55.4	32.99	30.33	30.33	30.45	23.84
1000	110.	30.08	30.41	30.41	30.51	24.67
2500	273.	32.09	30.64	30.64	30.64	23.99
10000	1088.	32.77	32.73	32.73	32.73	27.78
100000	10871.	35.44	35.38	35.38	35.38	28.29
			b. 10-Factor Wide			
1	1.20	530	50.1	.99	2.92	.87
5	1.63	584	64.4	4.93	6.71	3.52
10	2.17	602	55.3	9.64	7.68	5.87
15	2.72	544	54.9	14.1	10.1	7.88
25	3.80	589	57.7	21.4	12.9	10.7
50	6.52	559	62.8	26.2	22.1	17.6
100	12.0	572	71.6	32.0	31.6	27.5
200	22.8	615	97.5	49.8	54.0	43.2
500	55.4	561	120.	88.7	89.4	62.2
1000	110.	568	160.	133.	138.	85.3
2500	273.	555	253.	240.	243.	138.
10000	1088.	597	448.	448.	449.	218.
100000	10871.	565	546.	546.	545.	282.

* Arthur E. Hoerl is Professor, Department of Statistics and Computer Science, University of Delaware, Newark, DE 19711.

each data set using a single k. This is defined by

$$\text{Max Pot} = \min_{k \geq 0} \left\{ \sum \left[\frac{\lambda_i}{\lambda_i + k} \hat{\alpha}_i - \alpha_i \right]^2 \right\} .$$

Similarily, for the second simulation, with an assumed eigenstructure (4.25, 1.50, 1.25, 1.00, .778, .6, .4, .2, .02, .002), the 200 trial averages are given in Section b of the table. Here, the expectation for the least-squares error is 563.15.

These results suggest, then, the resultant effectiveness similarity between the author's RIDGM and the current iterative algorithm. This prompts a need for a broad comparison, over a significant range of p and conditioning, of these ridge algorithms together with other ridge algorithms which are currently under study. This suggested study would also need to be concerned with the broader aspects of estimation including weighting and prediction. Means for disseminating this information at an early date would help to reduce unnecessary duplication and computational effort. Perhaps even a group of interested participants could pool their resources and talents in defining, formulating, and carrying out the detailed simulations.

With the increasing reliance on simulation in regression it would seem propitious to develop a standardized procedure that would be reasonably acceptable. Here it is suggested that sampling conditioned on an α norm is one approach. This has the major attribute of avoiding the question of what constitutes a typical regression problem.

Dempster, Schatzoff, and Wermuth suggest that another device to evaluate prediction might be a jackknifing type of technique. This has been extensively investigated by Hoerl and Kennard and found deficient. In fact, the basic idea was extended to include what we called duplex (splitting the data into two groups under a variety of criteria) and multiplex (defining all possible subsets $\binom{n}{p}$ and selecting all or some fractionated proportion of all nondegenerate sets). In every instance over a variety of simulations, the technique was found wanting.

APPENDIX: UNIFORM RANDOM NUMBER ON [−.5, .5]

Define an arbitrary irrational number A truncated to 12 significant digits with a normalized floating-point as

$$A = (0.xxx - - -x)10^a .$$

The generation of the uniform random numbers was obtained by the following steps.

1. Form the normalized floating-point number

   $$B = 1/A = (0.xxx - - -x)10^b$$

   with the remainder

   $$R = (0.xxx - - -x)10^r$$

   where $r \leq b - 12$ and $A \times B + \text{REMAINDER} = 1$ with a 24 digit product $A \times B$. The mantissas of the respective numbers R and A satisfy the condition

   $$0 < \text{MANT}\ (R) < \text{MANT}\ (A) .$$

   The assumption here is that the mantissa of the remainder R after 12 significant places is independent of the divisor.
2. Form $C = R/A$ as $(0.xxx - - -x)10^c$. Set c to zero and store as the new A.
3. The new C is assumed uniform on [.1, 1.].
4. Subtract 0.1 from C and divide by 0.9 for [0, 1].
5. Uniform numbers on [−.5, .5] can be defined by subtracting 0.5 from $C/0.9$.

No more than 100,000 consecutive numbers should be used with the same original seed to be well assured of a nonrepeated chain. An unlimited run (with no repeating chain) can be achieved by adding one to A on a fixed count of say every 50,000 numbers.

In no instance has the algorithm degenerated in over 10^8 uses.

REFERENCES

[1] Gorman, John W., and Toman, R.J., "Selection of Variables for Fitting Equations to Data," *Technometrics*, 8 (1966), 27–51.

[2] Hoerl, Arthur E., Kennard, Robert W., and Baldwin, Kent F., "Ridge Regression: Some Simulations," *Communications in Statistics*, 4 (1975), 105–23.

[3] ———, and Kennard, Robert W., "Ridge Regression: Iterative Estimation of the Biasing Parameter," *Communications in Statistics*, A5 (1976), 77–88.

Comment

DAVID M. ALLEN*

To establish some points for reference, I will begin by expressing some of my own thoughts regarding regression. We have a random vector $\mathbf{y}$ and a full-rank, nonstochastic matrix X. The matrix X is said to be ill-conditioned if there exists at least one vector ℓ such that $\|\ell\| = 1$ and $\|X\ell\|$ is "small." We denote $E(\mathbf{y})$ by $\boldsymbol{\mu}$, and we suppose there exists a vector $\boldsymbol{\beta}$ such that $\boldsymbol{\mu} = X\boldsymbol{\beta}$. I will make some harsh statements about $\boldsymbol{\beta}$ and then illustrate them using a simple example. The example depends on $X = (\mathbf{x}_1|\mathbf{x}_2|\mathbf{x}_3)$, $X^* = (\mathbf{x}_1|\mathbf{x}_2|\mathbf{x}_3^*)$, and $\boldsymbol{\mu}$ where the vectors are given in the table.

What is the interpretation of $\boldsymbol{\beta}$? Let μ_i and x_{ij} denote the ith elements of $\boldsymbol{\mu}$ and $\mathbf{x}_j$. Since $\mu_i = \sum_j x_{ij}\beta_j$ for all i, it is often appealing to suppress the subscript i and regard

* David M. Allen is Associate Professor, Department of Statistics, University of Kentucky, Lexington, KY 40506.

Vectors Used in Examples

x_1	x_2	x_3	x_3^*	μ
87.36	−87.36	−.07	.07	−84.
−43.17	−47.97	−65.08	−65.12	−69.
−51.68	−26.72	−55.99	−56.01	−68.
−54.32	−.56	−39.20	−39.20	−56.
−28.74	−59.46	−62.97	−63.03	−54.
−51.14	25.66	−18.21	−18.19	−38.
50.31	−12.09	27.24	27.36	−33.
−42.19	47.09	3.48	3.52	−19.
−8.34	−56.34	−46.16	−46.24	−18.
18.78	34.14	37.75	37.85	−6.
−27.52	58.88	22.37	22.43	−4.
−7.18	56.18	34.96	35.04	2.
18.08	−33.76	−11.15	−11.25	44.
50.57	13.13	45.56	45.44	137.
89.18	89.18	127.47	127.33	266.

μ as a linear function of continuous variables x_1, x_2, and x_3. A common interpretation is: β_j is the change in μ accompanying a unit change in x_j with all other x's constant. The problem with this interpretation is that an ill-conditioned X precludes all other x's constant. For the X of the example the relationship

$$|.5x_1 + .5x_2 - .7x_3| < .05 \qquad (1)$$

is always satisfied. This requires x_2 or x_3 to change if x_1 changes as much as .2. If x_1 increases by one and the sum of absolute changes in x_2 and x_3 is kept as small as possible subject to (1), then x_2 does not change and x_3 increases by .5714. Thus the change in μ accompanying a unit change in x_1 with minimum changes in x_2 and x_3 is $\beta_1 + .5714\beta_3$ and not β_1. If X is ill-conditioned then, β has little or no interpretation.

The vector $\boldsymbol{\mu}$ is unique. Its elements are expected values of observable random variables and are interpretable. If X has not been correctly specified (Who really knows X?), then there may not exist a β such that $\boldsymbol{\mu} = X\beta$. If X is correctly specified but ill-conditioned, then β is fickle with regard to small perturbations of the elements of X. In the example, both X and X^* are correct specifications in that their columns span the same vector space and $\boldsymbol{\mu}$ is in that space. The maximum absolute difference between corresponding elements of X and X^* is .14, yet $\beta = (-41950/49, -41950/49, 1200)'$ and $\beta^* = (42050/49, 42050/49, -1200)'$ are drastically different.

The estimation of a linear combination of the elements of β where the variance of the least-squares estimator is greater than σ^2 will be termed an extrapolation. That is, the precision of the estimator is less than the precision of a direct observation on that linear combination (if such observation were possible). If X is ill-conditioned, then estimation of an individual element of β is often an extreme form of extrapolation. For the X of our example the respective variances of estimators of β_1, β_2, and β_3 are $18.22\sigma^2$, $18.22\sigma^2$, and $35.716\sigma^2$.

Because of the high dimensionality of typical regression problems it is impossible to conduct a comprehensive simulation study. However, the authors have come closer than any other study I have seen. I believe interpretation of regression is most difficult when X is highly ill-conditioned and thus regard Experiment 1 as being more valuable than Experiment 2. In view of my harsh statements about β, I would rather have α (A.8) than β as a factor in the study design. This would systematically generate different $\boldsymbol{\mu}$ in the appropriate vector space. For similar reasons, I place more credence on the analysis by SPE (A.11) than on the analysis by SEB (A.10). I am impressed by the potential of REGF methods and look forward to studying them further.

The authors mention conflicts among different methods. If we do not extrapolate, these apparent conflicts may be of less consequence than they indicate. Evaluation of SPE cannot involve extrapolation, while evaluation of SEB often is extrapolation. This statement is supported by the fact that the coefficient variation of SPE is less than the coefficient of variation of SEB.

The authors caution against expecting any favorite procedure to be automatically applicable. I emphatically agree. In my example, the variance of the least-squares estimator of $(-51.14, 25.66, -18.21)\beta$ is $.097554\sigma^2$, which is quite good for 15 observations. However, for $(-50.74, 26.06, -18.77)\beta$ the variance is $46.62\sigma^2$. Except for unrealistically large values of σ^2, ridge is worse than least squares in mean square error for both cases. We should recognize the existence of situations where no estimation, by any method, is warranted. The second linear combination just mentioned is such a case. The data simply does not contain much information about that linear combination.

Comment

A. F. M. SMITH*

Interest in alternatives to ordinary least-squares procedures for the analysis of the normal linear model is now widespread among both Bayesian and frequentist statisticians, and the authors are to be congratulated on their timely and stimulating contribution.

I shall confine my comments to just two aspects of this

* A.F.M. Smith is Lecturer, Department of Statistics and Computer Science, University College London, Gower Street, London WC1E 6BT, England.

wide-ranging study: the first concerns the relationship between continuous and discrete shrinking methods; the second relates to the authors' formulation of the prediction problem.

A link between continuous and discrete shrinking methods is implicit in a result established by Leamer and Chamberlain [4, Theorem 1], which shows that the RIDGE estimator—given here by (2.1) with $\mathbf{Q}$ equal to the identity matrix—can be written as a weighted-average of the 2^p least-squares estimators corresponding to all possible ways of constraining subsets of the p regression coefficients to be zero. The weighted-average forms REGF and DRGF are not quite of this form, since only $\binom{p}{r}$ such least-squares estimators are combined (for some chosen r), but it would, perhaps, be worthwhile exploring the connection further. In particular, an examination of the weights given by Leamer and Chamberlain [4, Eq. (4)] and those discussed in Appendix B of this paper reveals great similarity, especially for DRGF. Indeed, apart from the fact that Leamer and Chamberlain assume the variance σ^2 to be known, it appears that DRGF could be derived by expressing RIDGE as a weighted-average and then forming a renormed weighted-average using only those terms corresponding to r nonzero-constrained components (see [4, Eq. (3) and (4)] and Appendix B, (B.2), (B.7), (B.8), (B.9) with $a = r$). A closer study of this connection might well give some insight into the comparative performances of these estimators. (A similar analysis could be made of the relationship between PRI and an alternative weighted-average representation of RIDGE given in [4, Theorem 2].)

I find the comparison of estimators using prediction mean square error rather difficult to interpret. Indeed, it seems to me that this part of the study is both misleading and misguided in so far as it identifies the prediction problem with that of predicting a set of future values at precisely the same design points as have been used in estimating the regression coefficients. This obscures many of the features of interest that are present in the more general prediction problem and leads the authors to conclude that there is less scope for improvement over least squares in the prediction context. Brown [1] has shown that this is not necessarily true, and his arguments have been extended to a more general shrinkage estimation setting by Goldstein [3].

The point at issue can be summarized as follows. Let us assume that $\mathbf{b}^*$ is an estimator of $\boldsymbol{\beta}$ from (A.1), and that we wish to predict m future values of $\mathbf{Y}$ corresponding to a design matrix $\mathbf{X}_o(m \times p)$, the latter scaled in such a way that $\mathbf{X}_o\mathbf{b}^*$ is the desired predictor. (Note that the authors consider only the special case $m = n$, $\mathbf{X}_o = \mathbf{X}$.) It can be shown that if $\mathbf{b}^*$ is taken to be the RIDGE estimator, then the derivative, with respect to k, of the predictive mean square error, evaluated at $k = 0$, is equal to $-2\sigma^2 \sum_i (B_{ii}/\lambda_i)$, where $B_{ii} = (\mathbf{C}\mathbf{X}_o{}^T\mathbf{X}_o\mathbf{C}^T)_{ii}$, and $\mathbf{C}$ and λ_i are defined by (A.5). In this more general setting, it is the relationship between the B_{ii} and the λ_i which determines the scope for saving in predictive mean square error. The case considered by the authors has the special form $B_{ii} = \lambda_i$ and gives no insight into the greater scope for improvement which occurs when the larger values of B_{ii} correspond to small values of λ_i; i.e., when the directions in which predictions are required turn out to be those which are poorly estimated on the basis of the original design matrix.

Finally, I should like to draw attention to a splendid example of RIDGE in action—that of "Election Night Forecasting" in the U.K. [2]—where a version of RIDGE triumphs over all-comers, including OREG ($k = 0$) and ZERO ($k = \infty$).

REFERENCES

[1] Brown, P.J., "Predicting by Ridge Regression," unpublished report, Department of Mathematics, Imperial College, London 1974.

[2] ———, and Payne, C.D., "Election Night Forecasting (with discussion)," *Journal of the Royal Statistical Society, Ser. A*, 138, 1975, 463–98.

[3] Goldstein, M., "Aspects of Linear Statistical Inference," unpublished Ph.D. thesis, Mathematical Institute, University of Oxford, 1974.

[4] Leamer, E.E., and Chamberlain, G., "A Bayesian Interpretation of Pretesting," *Journal of the Royal Statistical Society, Ser. B*, 38, 85–94.

Comment

CHRISTOPHER BINGHAM and KINLEY LARNTZ*

The methods of estimation compared in the paper fall into two classes: those which are best understood in terms of the coordinates defined by the given independent variables, and those which are best understood in terms of canonical variables defined by eigenvectors of the cross product or correlation matrix of the independent variables. What seems to be important in both these

*Christopher Bingham is Associate Professor and Kinley Larntz is Assistant Professor, both at Department of Applied Statistics, University of Minnesota, St. Paul, MN 55108.

The research of the second author was facilitated by a single quarter leave granted by the Regents of the University of Minnesota.

cases is the *orientation* of the *true* coefficient vector relative to the relevant coordinate system.

In the case of best-subset methods or Bayesian mixtures of them, it seems to us that the relevant coordinates are in terms of the orthogonal basis in variable space defined by the independent variables as given, either standardized or not. The coefficient β_j can be considered the length of the projection of $\boldsymbol{\beta}$ on the jth basis vector. Coefficient vectors oriented near the space spanned by a small set of the basis vectors will be well approximated by subset regression models, and methods assuming that they are so oriented should be an improvement over methods, such as least squares, that do not. In effect, subset methods *shrink* the coefficient vector in the direction of planes determined by a subset of the basis vectors. If that is appropriate, they do well.

The other class of estimators, ridge methods and their generalizations, can be defined as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{Q})^{-1}\mathbf{X}'\mathbf{Y}$$

where $\mathbf{Q}$ is a positive-semi-definite matrix, or as a limit of such estimators. The independent variables matrix $\mathbf{X}$ is almost always assumed to be corrected for the mean, and is usually assumed to be standardized so that $\mathbf{X}'\mathbf{X}$ is a correlation matrix. The properties of such an estimator are best understood in terms of the relative eigenvectors of $\mathbf{X}'\mathbf{X}$ and $\mathbf{Q}$. It is well known that $\mathbf{X}'\mathbf{X}$ and $\mathbf{Q}$ can be simultaneously diagonalized. That is, there is a nonsingular matrix $\mathbf{A}$ such that

$$\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A} = \boldsymbol{\Lambda} = \text{diag}\,[\lambda_1, \lambda_2, \ldots, \lambda_p]$$

and

$$\mathbf{A}'\mathbf{Q}\mathbf{A} = \mathbf{K} = \text{diag}\,[k_1, k_2, \ldots, k_p] \ .$$

The rows of $\mathbf{A}^{-1}$ are proportional to the eigenvectors of $\mathbf{X}'\mathbf{X}$ relative to $\mathbf{Q}$. For ordinary ridge regression, $\mathbf{Q} = \mathbf{K} = k\mathbf{I}_p$ and $\mathbf{A}$ is orthogonal. For generalized ridge regression, $\mathbf{A}$ is orthogonal and $\mathbf{Q}$ is the matrix having the same eigenvectors as $\mathbf{X}'\mathbf{X}$ and eigenvalues $k_1, \ldots, k_p$. For Marquardt's generalized inverse, $\mathbf{Q}$ can be taken as a limit of matrices of this form, with the k_i corresponding to the smallest eigenvalues λ_j of $\mathbf{X}'\mathbf{X}$ approaching infinity and those corresponding to larger eigenvalues being zero.

This diagonalization induces transformations of the parameters and independent variables:

$$\boldsymbol{\beta} \rightarrow \mathbf{A}^{-1}\boldsymbol{\beta} = \boldsymbol{\alpha} \quad \text{or} \quad \boldsymbol{\beta} = \mathbf{A}\boldsymbol{\alpha} \ ,$$

and

$$\mathbf{X} \rightarrow \tilde{X} = \mathbf{X}\mathbf{A} \ .$$

We can express the estimator $\hat{\boldsymbol{\beta}}$ as

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= [(\mathbf{A}')^{-1}\boldsymbol{\Lambda}\mathbf{A}^{-1} + (\mathbf{A}')^{-1}\mathbf{K}\mathbf{A}^{-1}]^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{A}(\boldsymbol{\Lambda} + \mathbf{K})^{-1}\mathbf{A}'\mathbf{X}'\mathbf{Y} = \mathbf{A}(\boldsymbol{\Lambda} + \mathbf{K})^{-1}\tilde{X}'\mathbf{Y} \\ &= \mathbf{A}\hat{\alpha} \ , \quad \text{where} \quad \hat{\alpha} = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}} = (\boldsymbol{\Lambda} + \mathbf{K})^{-1}\tilde{X}'\mathbf{Y} \ .\end{aligned}$$

Thus $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ can be expressed in terms of $\boldsymbol{\alpha}$ and $\hat{\alpha}$, which represent the coordinates relative to the rows of $\mathbf{A}^{-1}$ (the columns of $\mathbf{A}$, if $\mathbf{A}$ is orthogonal). These characterize the orientation of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ relative to these eigenvectors.

From the point of view of the paper, as well as of many other authors, the relevant property of an estimator is a generalized mean square error criterion

$$\tau^2 = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = E[(\hat{\alpha} - \boldsymbol{\alpha})'\tilde{W}(\hat{\alpha} - \boldsymbol{\alpha})] \ ,$$

where $\mathbf{W} = \mathbf{W}'$ and $\tilde{W} = \mathbf{A}\mathbf{W}\mathbf{A}'$. In the usual cases $\tilde{W} = \text{diag}\,[w_1, w_2, \ldots, w_p]$, i.e., $\mathbf{W}$ is also diagonalized by $\mathbf{A}$. Both SEB ($\mathbf{W} = \mathbf{I} = \tilde{W}$) and SPE ($\mathbf{W} = \mathbf{X}'\mathbf{X}$, $\tilde{W} = \boldsymbol{\Lambda}$) are of this form. Then

$$\tau^2 = \text{tr}\,(\tilde{W}\,\text{Cov}\,[\hat{\alpha}]) + \text{tr}\,\tilde{W}(E[\hat{\alpha} - \boldsymbol{\alpha}]E[\hat{\alpha} - \boldsymbol{\alpha}]') \ .$$

But $\text{Cov}\,[\hat{\alpha}] = \sigma^2(\boldsymbol{\Lambda} + \mathbf{K})^{-1}\boldsymbol{\Lambda}(\boldsymbol{\Lambda} + \mathbf{K})^{-1}$ and

$$E[\hat{\alpha} - \boldsymbol{\alpha}] = ((\boldsymbol{\Lambda} + \mathbf{K})^{-1}\boldsymbol{\Lambda} - \mathbf{I})\boldsymbol{\alpha} = -(\boldsymbol{\Lambda} + \mathbf{K})^{-1}\mathbf{K}\boldsymbol{\alpha} \ .$$

Thus

$$\tau^2 = \sigma^2 \sum [w_i\lambda_i/(\lambda_i + k_i)^2] + \sum [w_i\alpha_i^2k_i^2/(\lambda_i + k_i)^2] \ .$$

It is easy to check that τ^2 is minimized for any choice of $w_i > 0$ [1] by $k_i = \sigma^2/\alpha_i^2$, in accordance with Hoerl and Kennard's result for $\mathbf{W} = \mathbf{I}$ [3]. The minimized value, which in some sense represents the best that one could do using any estimator in this class of estimators, is

$$\begin{aligned}\tau_{\min}^2 &= \sum w_i(\sigma^2\lambda_i + \sigma^4/\alpha_i^2)/(\lambda_i + \sigma^2/\alpha_i^2)^2 \\ &= \sum w_i\alpha_i^2/(1 + \lambda_i\alpha_i^2/\sigma^2) \ .\end{aligned}$$

For least squares, ($\mathbf{K} = 0$), we have

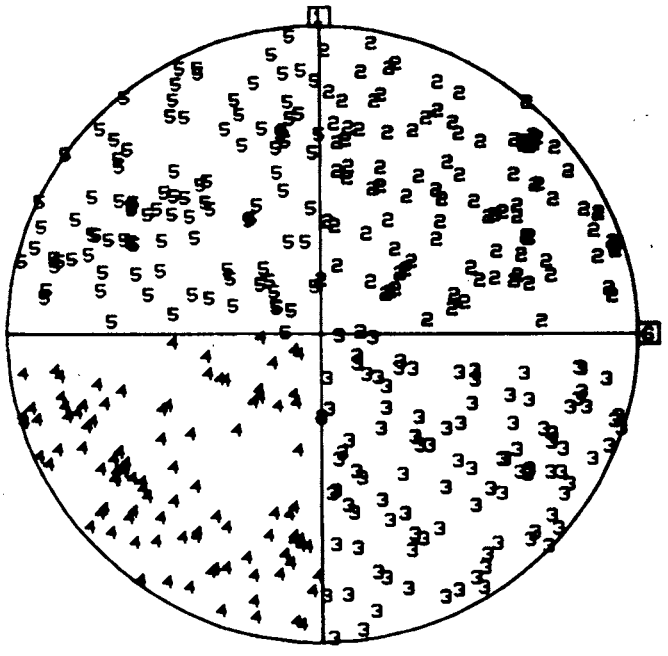
$$\tau_{\text{LS}}^2 = \sum w_i\sigma^2/\lambda_i \ .$$

Thus in each canonical direction the amount of possible improvement over least squares is $(\lambda_i\alpha_i^2/\sigma^2)/(1 + \lambda_i\alpha_i^2/\sigma^2)$. For fixed α_i/σ the improvement is greatest for smallest λ_i. This provides much of the motivation for Marquardt's generalized inverse estimator. However, for fixed λ_i, no matter how small, this ratio can be made as close to one as desired if the corresponding canonical coefficient α_i is large enough. In summary, we may conclude that the closer the coefficient vector is to the space spanned by the eigenvectors corresponding to the larger eigenvalues λ_i, the more improvement ought to be possible over least squares.

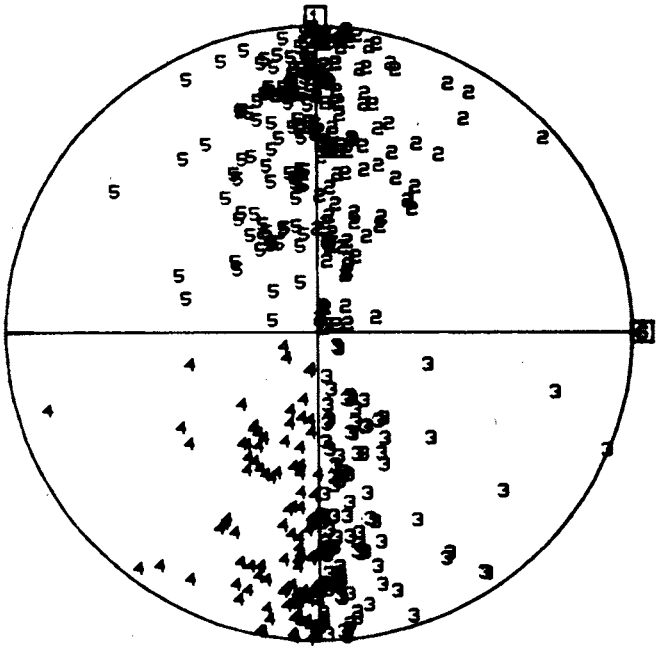
The above considerations indicate to us that any experiment designed to explore the merits of various adaptive ridge estimators, i.e., estimators with $\mathbf{K}$ chosen depending on the data (usually on $\hat{\alpha}_{LS}/s$), should have, as one of the primary factors, variation of $\boldsymbol{\beta}$ relative to the eigenvectors of $\mathbf{X}'\mathbf{X}$, i.e., variation of $\boldsymbol{\alpha}$. The second important factor is, of course, the pattern of eigenvalues of $\mathbf{X}'\mathbf{X}$. The direction of the eigenvectors is meaningful only with respect to their relationship with $\boldsymbol{\beta}$. This is why variation of $\boldsymbol{\alpha}$ should be a factor rather than the eigenvectors.

One difficulty in evaluating the results of the present experiment is that the eigenvectors, or more importantly, the $\boldsymbol{\alpha}$'s, are not given. The orientation is left to chance, without much indication of how the construction of patterned correlation matrices constrains $\boldsymbol{\alpha}$. Even when the eigenvector matrix of the nonstandardized form of $\mathbf{X}'\mathbf{X}$ is chosen randomly (uniformly over the orthogonal

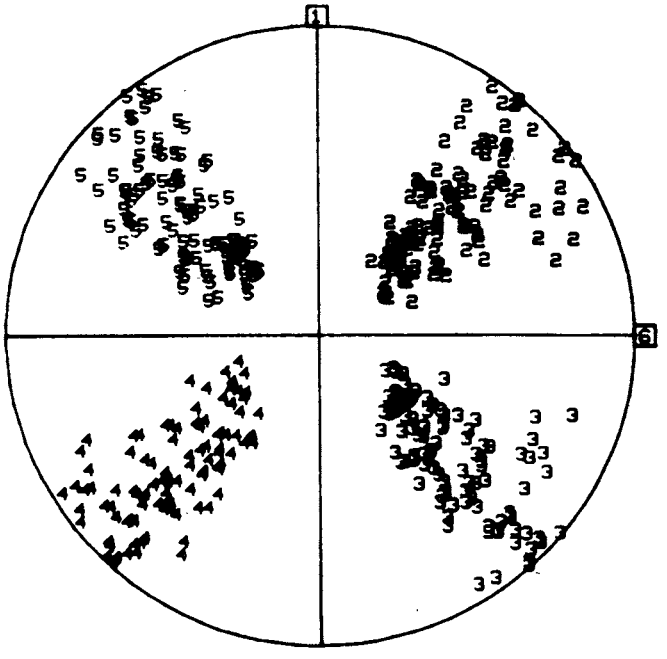
A. Orientation of Simulated Alpha's



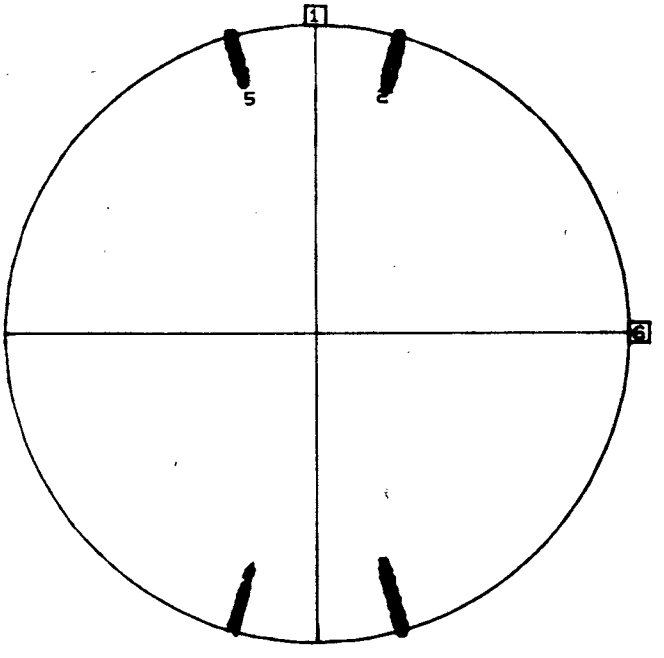
A1. EIG = *64.0 16.0 4.0 2.0 1.0 0.5* COR = *no*
BETA = *32.0 16.0 8.0 8.0 8.0 8.0* MCL = *no*



A2. EIG = *64.0 16.0 4.0 2.0 1.0 0.5* COR = *no*
BETA = *1.0 1.0 1.0 0.0 0.0 0.0* MCL = *yes*



A3. EIG = *30.0 30.0 30.0 20.0 20.0 20.0* COR = *yes*
BETA = *1.0 1.0 1.0 1.0 1.0 1.0* MCL = *no*



A4. EIG = *30.0 30.0 30.0 20.0 20.0 20.0* COR = *yes*
BETA = *32.0 16.0 8.0 0.0 0.0 0.0* MCL = *yes*

NOTE: EIG denotes eigenvalues of $X'X$ matrix, and BETA denotes the "true" regression coefficients.

group), the eigenvectors of **R**, the correlation matrix, are not random. Still less random are the eigenvectors of **R** after it is massaged to have high collinearity and/or multicollinearity. To get a clearer picture of what might have happened in the experiments described in the paper, we conducted a small simulation study to investigate the distribution of α, for fixed β, when the eigenvectors of $\mathbf{X'X}$ were chosen randomly and $\mathbf{X'X}$ was standardized and massaged exactly as described in the paper. There is considerable difficulty in presenting the results because the orientation of α is best expressed as a point on the unit sphere in 6-space. One simplification followed from the observation that we could always take α as being in the orthant defined by $\alpha_j \geq 0$, $j = 1, \ldots, 6$. To reduce the dimensionality, we looked at the orientation of α in some of the twenty three-dimensional subspaces defined by sets of three coordinate axes. An effective way of displaying such three-dimensional orientations is by means of an equiareal plot of a hemisphere (in this case, an octant) on a disk (quadrant). Figure A

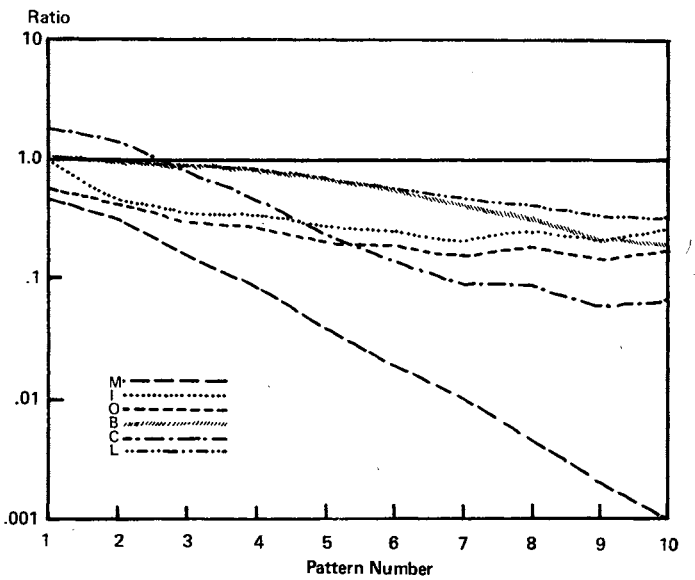
Eigenvalue Patterns for SEB *Simulation Study*

Eigenvalues	Pattern no.									
	1	2	3	4	5	6	7	8	9	10
					$p = 2$					
$\lambda_1 = \lambda_1/\lambda_2$	1.0	2.0	5.0	10.0	25.0	50.0	100.0	200.0	500.0	1000.0
λ_2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
					$p = 6$					
λ_1	1.0	1.23	2.05	2.00	3.40	4.63	1.96	1.994	3.55	3.14
λ_2	1.0	1.20	1.20	1.15	1.88	.733	1.22	1.010	1.50	1.66
λ_3	1.0	1.11	.958	1.02	.436	.333	1.06	1.005	.583	.877
λ_4	1.0	.847	.921	.962	.159	.164	.936	.995	.228	.257
λ_5	1.0	.827	.790	.816	.0738	.0927	.823	.990	.126	.0615
λ_6	1.0	.778	.0803	.0484	.0502	.0420	.00716	.00600	.00733	.000850
λ_1/λ_6	1.0	1.59	25.5	41.3	67.7	110	274	332	484	3694
EIG	—	30	30	30	64	64	30	—	64	64
MCL	—	no	yes	no	no	yes	yes	—	no	yes
COL	—	no	no	yes	no	no	yes	—	yes	yes

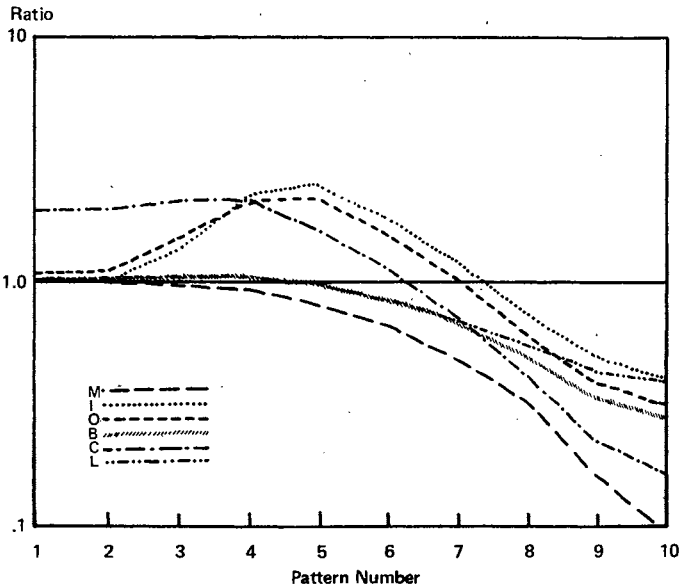
shows four typical plots. Each circle is actually four plots, one for each of the four three-dimensional subspaces containing the eigenvectors $\mathbf{a}_1$ and $\mathbf{a}_6$ corresponding to the largest eigenvalue and the smallest eigenvalue of the correlation matrix, respectively. Thus, starting at the upper right and proceeding clockwise, the four quadrants display the orientations of the simulated $\boldsymbol{\alpha}$'s in the spaces spanned by $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_6$; $\mathbf{a}_1$, $\mathbf{a}_3$, and $\mathbf{a}_6$; $\mathbf{a}_1$, $\mathbf{a}_4$, and $\mathbf{a}_6$; and $\mathbf{a}_1$, $\mathbf{a}_5$, and $\mathbf{a}_6$, respectively, where the $\mathbf{a}_j$'s are the eigenvectors corresponding to λ_j, the jth eigenvalue in decreasing order of magnitude.

Figure A1 corresponds to a situation in which there is no introduced collinearity or multicollinearity. In this case the distribution of the $\boldsymbol{\alpha}_j$'s is exchangeable, although probably not completely random (isotropic), and hence we would not expect to see any marked pattern. In fact the display shows a fairly uniform distribution of direc-

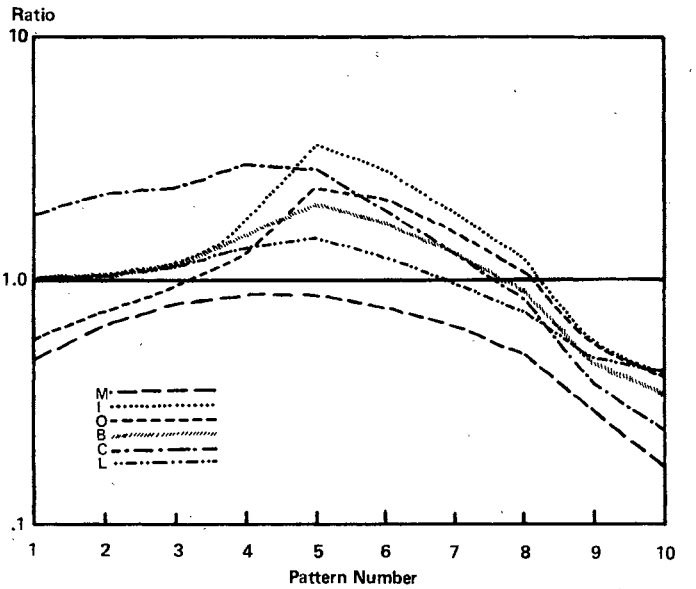
B. Ratio of SEB *for Ridge Methods to* SEB *for Least Squares*[a]



B1. Canonical Coefficients: $\alpha' = (10,0)$



B2. Canonical Coefficients: $\alpha' = (\sqrt{50}, \sqrt{50})$



B3. Canonical Coefficients: $\alpha' = (0,10)$

[a] Number of regressors: $p = 2$.

tions. In the other three cases with either collinearity, multicollinearity, or both, the distributions are clearly not random. Figures A2 and A4 display a marked tendency for α_1 to be large relative to α_6, exactly the situation for which we would expect ridge methods to be an improvement over least squares. The degree of consistency displayed by Figure A4 is, indeed, quite remarkable. The "inner" α_j's are, however, quite random. Figure A3 shows a case for which both α_1 and α_6 tend to be bounded away from zero, and to be rather highly correlated. Again the "inner" α_j's are relatively random. For emphasis, we would like to repeat that these α's were chosen as described in the paper, using 100 different random sets of eigenvectors. For the data sets discussed in the paper corresponding to Figure A4, for which α's are not given, it is clear we can say quite a lot about the orientation of α relative to $\mathbf{a}_1$ and $\mathbf{a}_6$, even though the original eigenvectors were chosen randomly.

To study the effect of varying the α's more explicitly than was done in the paper, we conducted another small simulation study. We restricted our investigation to a comparison of least squares with a few ridge-type estimators. No best-subset methods or their relatives were included. The particular procedures selected were (with the mnemonics used in our plots):

B: RIDGM in the paper, ridge with empirical Bayes k;

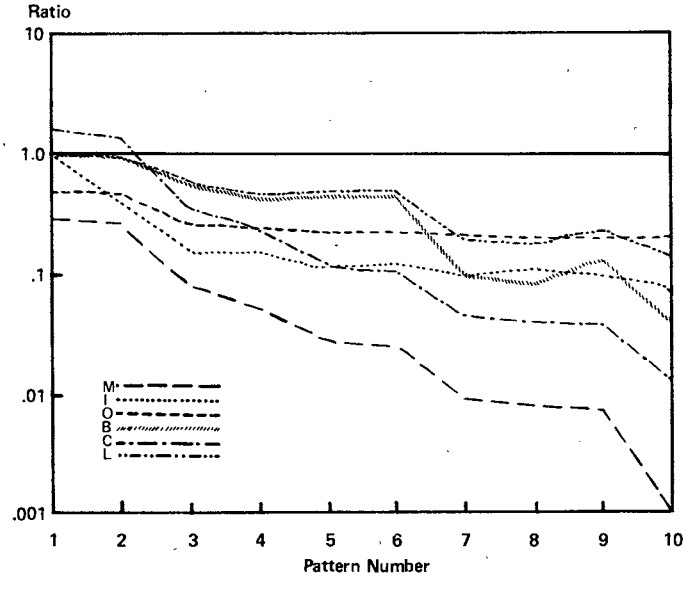
C: ICRIDG in the paper, ridge with shrinkage to the $F = 1$ contour;

I: PRIF in the paper, adaptive form of Marquardt's generalized inverse [5];

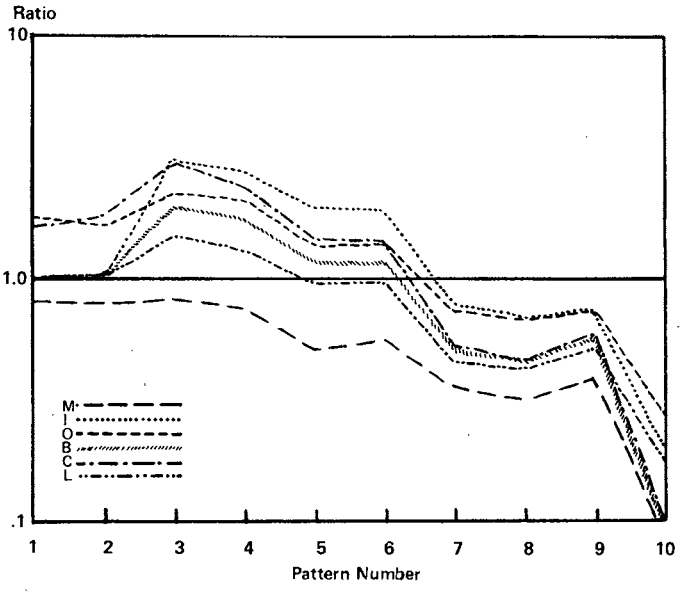
L: Ridge regression with k estimated as $ps^2/\hat{\beta}_{LS}'\hat{\beta}_{LS}$ as suggested by Hoerl, Kennard, and Baldwin [4];

O: Generalized ridge regression as proposed by Hoerl and Kennard [3] with $\mathbf{K} = \text{diag}[k_1, \ldots, k_p]$ computed using a method of Hemmerle [2];

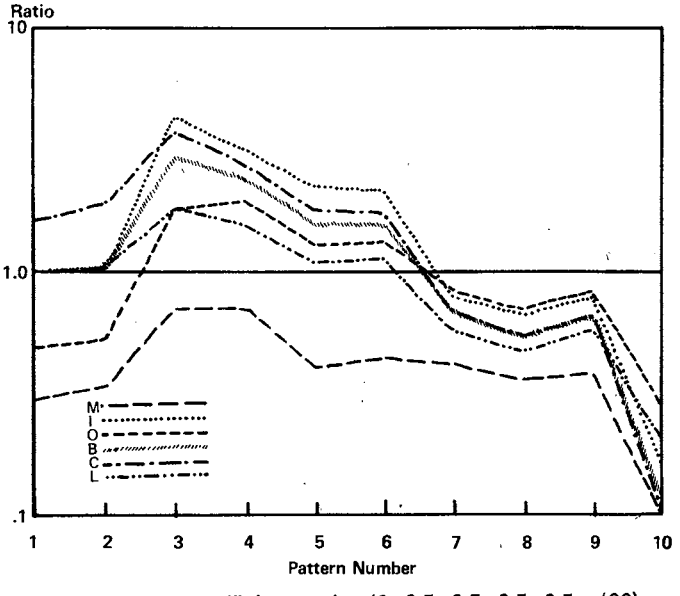
C. Ratio of SEB for Ridge Methods to SEB for Least Squares[b]



C1. Canonical Coefficients: $\alpha' = (\sqrt{99}, 0.5, 0.5, 0.5, 0.5, 0)$



C2. Canonical Coefficients: $\alpha' = (3, \sqrt{2.5}, \sqrt{2.5}, \sqrt{2.5}, \sqrt{2.5}, 9)$



C3. Canonical Coefficients: $\alpha' = (0, 0.5, 0.5, 0.5, 0.5, \sqrt{99})$

[b] Number of regressors: $p = 6$.

M: OPT in paper, generalized ridge with the correct (unrealizable) optimal k_i's, yields a lower bound for ridge type estimators.

Computations were carried out for $p = 2$ and $p = 6$ with variety of canonical regression coefficients α and eigenvalues Λ. The α's were standardized so that $\alpha'\alpha = 100$. The variance σ^2 was assumed to be 1. For each combination (α, Λ), 1000 regressions with $n = 20$ were simulated and the average SEB calculated. The eigenvalue patterns for $p = 2$ and $p = 6$ are given in the table with the patterns ordered by the ratio of the largest eigenvalue to the smallest eigenvalue (i.e., the condition number of the correlation matrix). For $p = 6$, eight of the eigenvalue combinations correspond to correlation matrices constructed according to the 2^3 combinations of the factors EIG, COL, and MCL in Experiment 2 of the paper.

Different randomly chosen rotations were used in constructing each of these matrices.

The results are too lengthy to give in full here. The general flavor is given in Figures B and C. Figure B ($p = 2$) and Figure C ($p = 6$) are semi-log plots of the ratio of the average SEB for each method to the average SEB for least squares. A point above the SEB = 1 line indicates the superiority of least squares. The abscissa is simply the eigenvalue pattern. Thus, the condition number of the correlation matrix increases from left to right. For both $p = 2$ and $p = 6$ there are clear gains for the ridge methods relative to least squares when α_1 (the canonical regression coefficient associated with the largest eigenvalue) is large. However, the gains become losses when there are substantial α_i's associated with the smaller eigenvalues, provided the condition number of the matrix is not too large. For extreme eigenvalue patterns, there appear to be guaranteed gains from the ridge methods, irrespective of the α's, at least within the range of α patterns we studied. The basic point is that for moderately ill-conditioned matrices (say corresponding to the degree of collinearity and multicollinearity studied in the paper) it is not at all clear that ridge methods offer a clear-cut improvement over least squares except for particular orientations of β relative to the eigenvectors of $\mathbf{X}'\mathbf{X}$.

Looking again more closely at Figure A (as well as other similar plots not given here), we see that there were cases where α_1 was in fact the dominant component, even though no explicit decision was made to make it so. This is a result of the choice of particular levels of factors BETA, MCL, and COR. Perhaps a more suitable procedure would have been to choose the orientations of the α's randomly, or even better, to choose combinations $(\alpha, \mathbf{A})$ in a systematic experimental design.

The Monte Carlo computations just discussed were performed using FORTRAN programs on a CDC 6400 computer. The random normal deviates used were generated using a library routine NORMAL based on a method proposed by Marsaglia and Bray [6]. The uniform random numbers used by NORMAL were produced by a multiplicative congruential generator using modulus 2^{48} and multiplier 5^{13}. Because the simulations were intended to be illustrative and preliminary, no attempt has been made to determine standard errors for the ratios of SEB in Figures B and C. All the curves in a plot were based on the same sets of randomly generated least-squares estimates $\hat{\alpha}_{LS}$. However, different plots were based on independent samples of random deviates.

REFERENCES

[1] Goldstein, M., and Smith, A., "Ridge-Type Estimations for Regression Analysis," *Journal of the Royal Statistical Society, Ser. B*, 36 (1974), 284–91.

[2] Hemmerle, W.J., "An Explicit Solution for Generalized Ridge Regression," *Technometrics*, 17 (1975), 309–14.

[3] Hoerl, A., and Kennard, R., "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12 (1970), 55–67.

[4] ———, Kennard, R., and Baldwin, K., "Ridge Regression: Some Simulations," *Communications in Statistics*, 4 (1975), 105–23.

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

[6] Marsaglia, G., and Bray, T.A., "A Convenient Method for Generating Normal Variables," *SIAM Review*, 6 (1964), 260–4.

Comment

RONALD A. THISTED*

Dempster, Schatzoff, and Wermuth have taken on the task of determining how best to achieve in practice the gains over least squares that are guaranteed to us in theory when the regression coefficients number three or more. They have given us a catalog of rules and, within the limitations of their study, have given us much insight into the behavior that these rules display and their performance relative to one another. Their conclusions are striking, particularly their assertion that ridge-regression rules are markedly superior to Stein-type estimators, and it is primarily toward this result that I shall direct my attention. Several remarks are in order which perhaps will clarify the scope and generality of their findings. These comments are primarily concerned with the relative merits of RIDGE estimators and Stein-type estimators.

The study attempts to separate the effects of collinearity, multicollinearity, and eigenvalue pattern by including separate factors for each of them in the experiments. However, both SPE and SEB for OREG, RIDGM, and STEINM depend upon $\mathbf{X}^T\mathbf{X}$ only through its eigenvalues. Furthermore, higher levels of COL (an MCL in Experiment 2) simply represent additional broadening of the eigenvalue spectrum. Consequently, it is not surprising to see significant main effects for each of these

* Ronald A. Thisted is Assistant Professor, Department of Statistics, University of Chicago, Chicago, IL 60637. This work was funded in part by a National Science Foundation graduate fellowship.

Different randomly chosen rotations were used in constructing each of these matrices.

The results are too lengthy to give in full here. The general flavor is given in Figures B and C. Figure B ($p = 2$) and Figure C ($p = 6$) are semi-log plots of the ratio of the average SEB for each method to the average SEB for least squares. A point above the SEB = 1 line indicates the superiority of least squares. The abscissa is simply the eigenvalue pattern. Thus, the condition number of the correlation matrix increases from left to right. For both $p = 2$ and $p = 6$ there are clear gains for the ridge methods relative to least squares when α_1 (the canonical regression coefficient associated with the largest eigenvalue) is large. However, the gains become losses when there are substantial α_i's associated with the smaller eigenvalues, provided the condition number of the matrix is not too large. For extreme eigenvalue patterns, there appear to be guaranteed gains from the ridge methods, irrespective of the α's, at least within the range of α patterns we studied. The basic point is that for moderately ill-conditioned matrices (say corresponding to the degree of collinearity and multicollinearity studied in the paper) it is not at all clear that ridge methods offer a clear-cut improvement over least squares except for particular orientations of β relative to the eigenvectors of $\mathbf{X}'\mathbf{X}$.

Looking again more closely at Figure A (as well as other similar plots not given here), we see that there were cases where α_1 was in fact the dominant component, even though no explicit decision was made to make it so. This is a result of the choice of particular levels of factors BETA, MCL, and COR. Perhaps a more suitable procedure would have been to choose the orientations of the α's randomly, or even better, to choose combinations $(\alpha, \mathbf{A})$ in a systematic experimental design.

The Monte Carlo computations just discussed were performed using FORTRAN programs on a CDC 6400 computer. The random normal deviates used were generated using a library routine NORMAL based on a method proposed by Marsaglia and Bray [6]. The uniform random numbers used by NORMAL were produced by a multiplicative congruential generator using modulus 2^{48} and multiplier 5^{13}. Because the simulations were intended to be illustrative and preliminary, no attempt has been made to determine standard errors for the ratios of SEB in Figures B and C. All the curves in a plot were based on the same sets of randomly generated least-squares estimates $\hat{\alpha}_{LS}$. However, different plots were based on independent samples of random deviates.

REFERENCES

[1] Goldstein, M., and Smith, A., "Ridge-Type Estimations for Regression Analysis," *Journal of the Royal Statistical Society, Ser. B*, 36 (1974), 284–91.

[2] Hemmerle, W.J., "An Explicit Solution for Generalized Ridge Regression," *Technometrics*, 17 (1975), 309–14.

[3] Hoerl, A., and Kennard, R., "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12 (1970), 55–67.

[4] ———, Kennard, R., and Baldwin, K., "Ridge Regression: Some Simulations," *Communications in Statistics*, 4 (1975), 105–23.

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

[6] Marsaglia, G., and Bray, T.A., "A Convenient Method for Generating Normal Variables," *SIAM Review*, 6 (1964), 260–4.

Comment

RONALD A. THISTED*

Dempster, Schatzoff, and Wermuth have taken on the task of determining how best to achieve in practice the gains over least squares that are guaranteed to us in theory when the regression coefficients number three or more. They have given us a catalog of rules and, within the limitations of their study, have given us much insight into the behavior that these rules display and their performance relative to one another. Their conclusions are striking, particularly their assertion that ridge-regression rules are markedly superior to Stein-type estimators, and it is primarily toward this result that I shall direct my attention. Several remarks are in order which perhaps will clarify the scope and generality of their findings. These comments are primarily concerned with the relative merits of RIDGE estimators and Stein-type estimators.

The study attempts to separate the effects of collinearity, multicollinearity, and eigenvalue pattern by including separate factors for each of them in the experiments. However, both SPE and SEB for OREG, RIDGM, and STEINM depend upon $\mathbf{X}^T\mathbf{X}$ only through its eigenvalues. Furthermore, higher levels of COL (an MCL in Experiment 2) simply represent additional broadening of the eigenvalue spectrum. Consequently, it is not surprising to see significant main effects for each of these

* Ronald A. Thisted is Assistant Professor, Department of Statistics, University of Chicago, Chicago, IL 60637. This work was funded in part by a National Science Foundation graduate fellowship.

factors and nonsignificant interactions in Table 3. It is important to recognize that these factors are not different effects but one and the same—the effect of highly unequal eigenvalues. As the reduction to principal components shows, multicollinearity and collinearity affect SEB and SPE only to the extent that they spread out the eigenvalues of $\mathbf{X}^T\mathbf{X}$.

After discussing the optimality of RIDGE and STEIN rules, each of which is Bayes for a particular prior distribution on $\boldsymbol{\alpha}$ and any quadratic loss, the authors proceed to require a "rule for determining k from the sample data." Of course this simply won't do for the subjectivist Bayesian, for whom k represents a judgment on the precisions of components of $\boldsymbol{\alpha}$. Further, it is important to note that we may forfeit the previously mentioned optimality if k is a function of the data.

It is curious that STEINM does so badly with respect to SPE. It is well known [2] that the problem of estimating regression coefficients with SPE loss is equivalent to estimating the mean of a multivariate normal distribution with equal variances and loss function $L(\boldsymbol{\delta}, \boldsymbol{\beta}) = \|\boldsymbol{\delta} - \boldsymbol{\beta}\|^2/\sigma^2$. Furthermore, Efron and Morris [1] show that in the latter problem the James-Stein positive-part rule cannot be substantially improved upon in very much of the parameter space. From the fact that STEINM is so badly beaten in SPE by RIDGM we must conclude that: STEINM is not equivalent to the James-Stein rule; the parameters chosen in the study are restricted to regions of the parameter space more favorable to RIDGE rules than to STEIN-type rules; or that in this particular trip to Monte Carlo the house has taken its cut, and that the results we see are not representative. This observation brings us to our final point.

Bayes rules are not optimal *only* when the statistician has quantified his prior beliefs about $\boldsymbol{\alpha}$ by specifying a probability distribution for it. They are also optimal, for instance, when the parameters in each experiment actually are generated by some random mechanism, the distributional properties of which are known to the statistician. In the latter case it makes sense to speak of a "correct" prior distribution for $\boldsymbol{\alpha}$. The authors are correct in their remark (p. 80) that,

> To assert that RIDGE is better [than STEIN] in practice is equivalent to asserting that its prior assumptions are more nearly correct over the range of the statistician's experience. Note especially that if the RIDGE prior is correct then the RIDGE estimator is optimum for any quadratic loss function, including SEB and SPE.

Consequently, when we observe RIDGM to be the big winner both in SEB and SPE, a rough application of Bayes theorem leads us to conclude with high posterior probability, that for these data, the RIDGE prior and not the STEIN prior is "more nearly correct."

Consider, then, the random mechanism by which $\boldsymbol{\alpha}$ is selected in this study. First of all, $\boldsymbol{\beta}$ is fixed, then a random orthonormal matrix $\mathbf{G}$ is generated. For any fixed vector $\mathbf{u}$, $\mathbf{Gu}$ is uniformly distributed on the p sphere of radius $\|\mathbf{u}\|$. The matrix $\mathbf{G}$ corresponds to $\mathbf{C}^T$ of Appendix A. Consequently, $\boldsymbol{\alpha} = \mathbf{G}^T\boldsymbol{\beta}$ has a uniform distribution on the p sphere of radius $\|\boldsymbol{\beta}\|$. It is easy to see that, since $\boldsymbol{\alpha}$, $-\boldsymbol{\alpha}$, and $(-\alpha_1, \alpha_2, \ldots, \alpha_p)^T$ have the same distribution,

$$\begin{aligned} E\alpha_i &= 0 \ , \\ \text{Var } (\alpha_i) &= p^{-1}\|\boldsymbol{\beta}\|^2 \ , \\ E\alpha_i\alpha_j &= 0 \ , \quad i \neq j \ . \end{aligned}$$

Hence, the method used to generate $\boldsymbol{\alpha}$ has mean zero and equal component variances. Thus the prior variances of the α_i, by which we mean the variances of the random mechanism generating the α_i *in this study*, are equal and not proportional to the inverse eigenvalues.

As the authors point out in the quoted passage, this setup is highly favorable to RIDGE, and it ought not to be surprising that RIDGM beats STEINM even on SPE, the loss function most favorable to STEIN-type estimators. Furthermore, the more disparate the eigenvalues, the worse STEIN-type rules will do in this experiment compared to RIDGE rules, since the STEIN prior is less like the "correct" prior. It is easy to predict on these grounds that STEINM will improve its performance in Experiment 2, since there are two vectors of eigenvalues added to those of the first experiment, each of which is less extreme than one of the vectors from the first experiment, so that the average spread in the eigenvalues is reduced. STEINM improves dramatically.

Let us return then to the data analyst, "who knows only his data and not the underlying parameters," and let us leave him with two words of caution. The conditions represented in the present experiment may not represent those likely to occur in practice. Further, it is perhaps still too early to recommend ridge regression for routine use in data analysis.

REFERENCES

[1] Efron, Bradley, and Morris, Carl, "Stein's Estimation Rule and Its Competitors—an Empirical Bayes Approach," *Journal of the American Statistical Association*, 68 (March 1973), 117–30.

[2] Sclove, Stanley L., "Improved Estimators for Coefficients in a Linear Regression," *Journal of the American Statistical Association*, 63 (June 1968), 596–606.

Rejoinder

A. P. DEMPSTER, MARTIN SCHATZOFF, and NANNY WERMUTH

1. INTRODUCTION

Since the discussants' comments fall into a few distinct categories, we shall organize our responses by topic rather than respond to each discussant separately. We preface our remarks with some comments of a general nature.

Our first observation on reading the six sets of comments is that all of the discussants are primarily interested in RIDGE methods, or in the comparison of RIDGE and STEIN procedures. Thus despite years of widespread use of techniques such as stepwise regression and regression on principal components, there is no mention of these methods by the discussants. We believe that the indicated direction of interest is due to a combination of the lack of theoretical understanding of the latter classes of procedures, and the analytical and philosophical attractiveness of the RIDGE and STEIN approaches. Possibly the performance comparisons produced by our study are such as to dampen interest in many common methods. Two of the discussants, Allen and Smith, expressed interest in the REGF methods, but did not offer substantive remarks.

A second observation is that there are no comments on the analysis of the results of the experiments. We were worried that someone might question our use of OREG in producing Table 5, while we clearly suggested in the paper that methods such as RIDGM and FREGF offered possibilities for improved estimation.

Third, we are impressed by the variety and seriousness of the commenters' views on the broad classes of methods covered by the labels RIDGE and STEIN. We believe that the state of the art in these areas is well reflected in the discussion.

In considering the specific points raised by the discussants, it appears that most of these may be appropriately classified as follows:

1. Design of the experiment,
2. Theoretical aspects of RIDGE and STEIN methods,
3. Estimation of the RIDGE parameter (k),
4. Criteria for evaluating alternate methods, and
5. What to do with real data.

We discuss in the next section what we consider to be the relevant aspects of various comments pertaining to these issues.

2. DISCUSSION OF SPECIFIC COMMENTS

2.1 Design of the Experiment

As with any Monte Carlo type of study, hard conclusions must usually be confined to the domain of investigation, with extrapolation to unexplored regions of the parameter space difficult at best, and often hazardous. Accordingly, we have not made sweeping claims as to the general applicability of our results, but rather have attempted to explore the effects of some parameters of interest on a large number of different estimation procedures.

Two of the discussants' papers (Allen; and Bingham and Larntz) argued for variation of α rather than β in the experimental design, while a third (Thisted) stressed that the design factors COL and MCL affect the risk functions based on SPE and SEB only through the λ's, for OREG, RIDGM, and STEINM. In both instances, our rationale was to provide comparative evaluation of these procedures with various types of stepwise selection of variables. We expected these comparisons to be sensitive to variation in the β's as well as to the pattern and degree of correlations in the independent variables. It should be pointed out that in the simulation examples presented by Hoerl, based on random selection of the α's with specified norm, RIDGM had very high efficiency relative to the maximum potential, over wide ranges of the norm.

A second comment on the design, made both in the Thisted and Efron and Morris papers, has to do with our use of a random rotation matrix. Their claim is that this type of randomization would tend to symmetrize the prior distribution of the β's, resulting in exchangeable prior distributions that would naturally favor RIDGM against other methods. This idea is intriguing, but is not made very precise in the comments. Perhaps it means that random rotation makes the coordinates α distribute in a way which appears exchangeable over the 160 data sets. Note that both RIDGE and REGF assume prior exchangeability among the components of β, but are very different methods which dominate each other in different stituations, so exchangeability is not a sufficient description of a prior distribution of β to guide the data analyst. In any case, the actual distribution of β over our 160 data sets is a very simple, known, discrete distribution, as opposed to the symmetrized distribution, whatever that is. It seems a small swindle to base interpretations on an artificially scrambled distribution of α's rather than the simple known distribution of β. An interesting question remains: how should we have systematically varied our factor β to produce fairer comparisons of the relative strengths and weaknesses of RIDGE and STEIN?

2.2 Theoretical Aspects of RIDGE and STEIN Methods

Efron and Morris state that STEINM is not the James-Stein rule, as we have been careful to note in Section 2.4.1. We wish to point out that JSTEIN and STEINM are both STEIN-type procedures in that they shrink uniformly on all principle axes, and that they differ only in the degree of shrinkage. We maintain our contention that JSTEIN would have performed worse on our data than STEINM. It thus appears that statistical theory is in conflict with our empirical results.

We believe that statisticians should no longer accept without question the assumptions of Efron, Morris, and Thisted that statistical techniques should be evaluated theoretically by means of frequentist risk functions depending on unknown parameter values. We were careful in our paper to define SPE and SEB in terms of actual errors of estimation, and not in terms of theoretical averages of such errors, whether frequentist or Bayesian. Our comment may be illustrated by Thisted's statement, "both SPE and SEB for OREG, RIDGM, and STEINM depend upon X^TX only through its eigenvalues," which is literally false according to our definitions. It is also false, in general, when a prior distribution of β is available, whether SEB and SPE are reinterpreted as posterior expectations given a data set, or are prior expectations of such Bayes risks. The statement is true for prior expectations of a game player who knows β, but the relevance and applicability of this game to data analysis is a matter of current dispute.

Having expressed serious reservations about the meaning of the Efron-Morris-Thisted theory, we do wish to express our admiration for their efforts, and our wish to understand the insights which they feel the theory gives. Their comment about the general incompatibility of minimax and empirical Bayes seems to us to capture a real dilemma of much of statistics: except in rare, mathematically nice, and overly taught, circumstances, there is no sure-thing principle to protect us against the need for hard prior judgments.

2.3 Estimation of the RIDGE Parameter

Efron and Morris's statement that "... ridge estimators are a class of Bayes rules against normal priors indexed by k, and the effectiveness of a given rule depends upon how k is estimated" summarizes the situation very concisely.

We believe that we have demonstrated remarkable empirical properties for the RIDGM rule for estimating k. We have received a letter from Professor Hoerl, written after his original commentary on our paper, in which he alludes to a recent comparative evaluation of a number of ridge estimators over a spectrum of signal-to-noise. He states, "Based on a broad comparison of all the algorithms, with $p = 10$, yours is the most effective. In fact, the degree to which your algorithm achieves near potential is startling." We are not sure whether he is referring to the study presented in his discussion of our paper, or to a further exploration not yet reported.

We agree with Efron and Morris that it would have been desirable to include EBMLE in our set of RIDGE methods. Our failure to do so was due to an error which led us to believe until too late in the study that RIDGM was equivalent to EBMLE. We would conjecture that EBMLE should be slightly better than RIDGM on our criteria.

Perhaps there are some Bayesian statisticians as Thisted states "for whom k represents a judgment on the precisions of components of α." A more usual contemporary Bayesian formulation would be to regard k as an unknown which needs a prior distribution just like other unknowns. The use of an estimated β associated with an estimated k is a crude approximation to the center of a posterior distribution, which is reasonably stable across a plausible range of smooth priors on k. We did not spell this out because our paper is not primarily Bayesian in outlook. We do feel, however, and Efron, Morris, and Thisted apparently agree, that the success of RIDGM must relate to some type of fit between the design of our study and the Bayesian assumptions which make RIDGM a near-optimum technique.

2.4 Criteria for Evaluation of Alternate Methods

Smith, and Efron and Morris have addressed themselves to the question of criteria for comparing different methods.

Specifically, Smith is concerned about our use of SPE as a measure of predictive error, because it is defined only at the same design points used in the experiment. He correctly points out that it "... gives no insight into the greater scope for improvement which occurs ... when the directions in which predictions are required turn out to be those which are poorly estimated on the basis of the original design matrix." It would have been interesting to expand the design to incorporate evaluation of predictive errors at points other than those included in the original design.

Efron and Morris state that JSTEIN should not be applied with the loss SEB, because it is not minimax in this case. We believe, however, that SEB is a very important criterion, since it often happens that the principal objective of a regression study is to estimate the values of the regression coefficients.

2.5 What to Do with Real Data

The problem of what to do with real data is not solved by our study. Efron and Morris, and Thisted correctly caution against the routine application of any shrinkage rule, and indeed we have adopted exactly the same posture in our paper. None of the discussants offered any concrete proposals, however, as to how one should proceed when analyzing real data. Nor were there any comments on our suggestions other than those by Hoerl, who indicated that he has experimented with a number of versions of our suggestion to divide data into subsets as a basis for comparing different estimation techniques.

Although he claims to have found such techniques to be deficient, we would be most interested in seeing the results. In terms of a predictive error criterion such as SPE, or the predictive mean square error advocated by Smith, it would seem that comparison of the predictive capabilities of various methods from one subset to another would provide a reasonable empirical basis for selecting a particular method in a given situation.

3. CONCLUSION

We feel that a number of interesting and useful points have emerged from the various discussions of our paper, and believe that the combined effect will be to stimulate further research, both theoretical and experimental. We view the problem of what to do with real data as being of paramount importance and we hope that some of the suggestions made in the concluding section of our paper will be followed up. This is not meant to preclude independent approaches, for there is certainly ample room for development and exploration of new ideas on many facets of the problem. The potential for large gains clearly exists. We need to develop tools for better exploiting this potential.