COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 402

A CRITICAL ANALYSIS OF RIDGE REGRESSION

Frank Campbell and Gary Smith

August 22, 1975

A CRITICAL ANALYSIS OF RIDGE REGRESSION*

by

Frank Campbell and Gary Smith

Researchers often find that their data does not contain enough information to decisively answer the questions that they have posed. If they are searching for a model, they may find that the variation in a particular dependent variable can seemingly be explained equally well by an annoyingly wide variety of theoretically motivated and even randomly selected explanatory variables. And researchers with a well-defined specification often obtain confidence regions which are so large that the point estimates and forecasts are of little interest.

One response to informational inadequacy in the data is shoulder shrugging, as with Johnston's widely quoted remark that one cannot "make bricks without straw." In this approach, one simply estimates the a priori preferred model and notes that more precise estimates will require more information.

More commonly, the researcher tries to improve his reported estimates by changing the model. He may limit himself to a small number of explanatory variables which are chosen in part for their high variances and relative orthogonality, or he may begin with a more complete model and then drop those variables with coefficients which are found to be incorrectly signed or statistically insignificant. At the extreme, there are those who toil endlessly for that elusive combination of variables

*The research described in this paper was undertaken by grants from the National Science Foundation and the Ford Foundation.

which will yield statistically significant and plausibly signed parameter estimates.

These specification searches can be viewed as the imposition of exact parameter restrictions on a more general model. If these implicit restrictions have no a priori weight behind them, then the final reported estimates and statistics will have little meaning. By emphasizing variance reduction (which can be measured) and neglecting the biases which have been introduced (but cannot be measured), the researcher will have done little more than disguise the imprecision of his estimates. Techniques such as stepwise regression, generalized inverses, and principal components analysis which impose wholly ad hoc parameter restrictions are without merit, in that they can only be successful by fortuitous accident.

While we are more sympathetic to concerns for correctly signed parameter estimates, this objective in practice usually motivates only the introduction of exact (typically exclusion) restrictions on an individual parameter-by-parameter basis. It is difficult to be comfortable with exact (or, even worse, exclusion) restrictions since they force one into the position of having to choose between acting as if one had either no knowledge or perfect knowledge. In addition, a myopic parameter-by-parameter procedure neglects the critical possibilities that the errors in one's restrictions may multiply or may cancel one another out. For example, parameter constraints which are individually more accurate than the corresponding estimates may collectively worsen the estimates of all of the remaining coefficients. In particular, setting two "incorrectly signed" coefficients equal to zero may worsen rather than improve one's model even when one is right that the two estimates do in fact have the wrong signs.

Recently there has been some interest in ridge regression as an alternative method for coping with inadequately informative data. This approach is similar to the popular practices which we have disparaged, in that it is sometimes motivated by priors which it does not accurately describe and is more often wholly ad hoc. It differs from the usual procedures in that it imposes stochastic rather than exact restrictions on the parameters. Thus, while ridge regression is considerably more flexible than common practices, it almost always represents an incorrect characterization of one's prior beliefs and, to the extent that it relies upon capricious information, it can only be accidentally beneficial.

To focus our comments, we have organized this paper as a discussion of a recent lead article by Marquardt and Snee (MS) in The American Statistician. This article provides both a mathematical and verbal justification of their procedure and several detailed examples. For simplicity, we will only discuss their first example here, which deals with acetylene data.

I. Variance Inflation Factors and Standardized Data:

Marquardt and Snee begin by arguing that the explanatory variables should always be standardized:

> In standardizing the predictor variables, the mean is subtracted from each variable ("centering") and then the centered variable is divided by its standard deviation ("scaling"). Centering removes the nonessential ill-conditioning, thus reducing the variance inflation in the coefficient estimates. In a linear model centering removes the correlation between the constant term and all linear terms. In addition, in a quadratic model centering reduces and in certain situations completely removes, the correlation between the linear and quadratic terms. Scaling expresses the equation in a form that lends itself to more straightforward interpretation and use.

This standardization applies a unique nonsingular linear transformation to the variables and consequently has no substantive effect on the model, in that it does not change the least squares forecasts or the least squares estimates of any estimatable coefficients. Looking specifically at their acetylene data example, the unstandardized model

$$(1) \qquad Y = b_0 + \sum_{i=1}^{3} b_i X_i + \sum_{1 \le i \le j}^{3} b_{ij} X_i X_j + \epsilon$$

can be rewritten in the entirely equivalent standardized form

$$(2) \qquad Y = \left[b_0 + \sum_{i=1}^{3} b_i \bar{X}_i + \sum_{1 \le i \le j}^{3} b_{ij} \bar{X}_i \bar{X}_j\right] + \sum_{i=1}^{3} \left[b_i + b_{ii} \bar{X}_i + \sum_{i=1}^{3} b_{ij} \bar{X}_j\right] S_i \left[\frac{X_i - \bar{X}_i}{S_i}\right] + \sum_{1 \le i \le j}^{3} b_{ij} S_i S_j \left(\frac{X_i - \bar{X}_i}{S_i}\right)\left(\frac{X_i - \bar{X}_i}{S_j}\right) + \epsilon$$

$$= \beta_0 + \sum_{i=1}^{3} \beta_i Z_i + \sum_{1 \le i \le j}^{3} \beta_{ij} Z_i Z_j + \epsilon$$

(where $Z_i = (X_i - \bar{X}_i)/S_i$ and the scaling has been by $S_i = \sqrt{\Sigma(X_i - \bar{X}_i)^2/n-1}$.)

While the implicit estimates of all estimatable parameters are unchanged, the parameters that are actually being directly estimated are somewhat different, which means that one must be careful in making comparisons. This is true of such obvious things as the comparison of, say, the estimated coefficients of X_1 in equation (1) and $(X_1 - \bar{X}_1)/S_1$ in equation (2), and also of more subtle interpretations such as the strength or weakness of the data.

Since rewriting the model in the form (2) does not affect any of the implicit estimates, it clearly has no effect on the amount of information contained in the data. That information has only been packaged in

a different form. One can, of course, use simplistic measures of the informational content of the data which do depend upon how the variables are arranged. Such a dependency does not imply that there is a preferred linear arrangement of the variables, but only that the information measure is inadequate. For example, Marquardt and Snee offer the variance inflation factor (VIF) as a measure of ill-conditioning and as an indicator of whether a biased estimator should be used.

In the general model

$$y = \sum_i \alpha_i X_i + \epsilon_i$$

the variance of the ordinary least squares estimate of α_k is given by

$$\text{Var}(\hat{\alpha}_k) = \frac{\sigma_\epsilon^2}{(n-1)S_k^2}\left(\frac{1}{1 - R_k^2}\right)$$

where R_k^2 is the squared multiple correlation coefficient between the k^{th} variable and the remaining explanatory variables. The variance inflation factor is

$$\text{VIF}(\hat{\alpha}_k) = \frac{1}{1 - R_k^2}$$

which can be interpreted as the ratio of the variance of $\hat{\alpha}_k$ to what that variance would be if X_k were uncorrelated with the remaining X_i.

However, the VIF is of little interest since this measure is not invariant to linear transformations of the variables. At the cosmetic extreme, data can always be orthogonalized so that all of the VIF's are equal to one without affecting any of the parameter estimates. If the estimates are imprecise, this would then be attributable to the low variation

of the associated variables rather than to their high intercorrelation. In general, any collinearity problem can be equally well described as a problem of low variation. For example, one could attribute the imprecision in estimated interest rate elasticities to the high correlation between two interest rates or to the stability of the rate differential; but we surely do not want to use a measure of informational content which depends upon whether we use two rates as explanatory variables or instead use one rate and the rate differential.

Yet this is precisely what Marquardt and Snee have done. When their acetylene model is written in form (1), there are some very high intercorrelations among the variables, due in part to the fact that they have only 16 observations on 9 variables, with 6 of the variables constructed as products of the other 3. One of the more dramatic and yet easily understood facets of this example is the term $b_{11}X_1^2$. The only observations on X_1 are 6 at 1300, 6 at 1200 and 4 at 1100, which gives a simple correlation between X_1 and X_1^2 of .99967. Overall, the squared correlation between X_1^2 and the remaining variables is .9999996, which gives b_{11} a VIF of 2.5 million. This means that the variance of b_{11} will be large unless the variance of X_1^2 is large or the variance of the disturbance term is small. In this example, the former is 3.77×10^{10} and the latter is estimated to be .81258, so that

$$\widehat{\text{Var}}(\hat{b}_{11}) = \frac{.81258}{15(3.77 \times 10^{10})}(2.5 \times 10^{6}) = .36 \times 10^{-5} .$$

Is this large or small? Obviously one cannot say without knowing what b_{11} is and what the estimate will be used for. Nonetheless, Marquardt and Snee state that a VIF of two million "is unthinkable and unnecessary"

since the model can be written in the standardized form (2), in which the coefficient of $(X_1 - \bar{X}_1)^2/S_1^2$ has a VIF of less than two thousand, as the squared correlation of this variable with the remaining variables in (2) is "only" .99943. Since the coefficient of this variable is $b_{11}S_1^2$, we can obtain the implicit estimate $\hat{b}_{11} = (\widehat{b_{11}S_1^2})/S_1^2$ which will be identical to the estimate that is directly obtained from (1). As they note, scaling does not affect the VIF (since it doesn't affect the correlation coefficients), so that in form (2),

$$\text{VIF}(\hat{b}_{11}) = \frac{1}{1 - .99943} = 1762.58$$

while

$$\widehat{\text{VAR}}(\hat{b}_{11}) = \left[\frac{\sigma_\epsilon^2}{(n-1)\sigma^2_{(X_1 - \bar{X}_1)^2}}\right]\text{VIF}(\widehat{b_{11}S_1^2})$$

$$= \frac{.81258}{(15)(2.656 \times 10^7)}1762.58 = .36 \times 10^{-5} .$$

Thus, the use of form (2) has no effect on the estimate of b_{11} or on the precision of this estimate. The VIF has been reduced by a factor of 1000 but so has the variance of the associated variable, so that the imprecision has simply been relabeled a problem of low variation rather than one of high covariation. A similar analysis could be carried out for any of the estimatable coefficients.

Thus, while some may find the VIF helpful in describing the sources of imprecision, it does not measure the amount of imprecision and cannot be used to justify the reliance on weakly held supplementary information; nor can it be used to motivate linear transformations of the variables.

Similarly, the examination of the eigenvalues of the sample moment matrix (which MS mention briefly) is insufficient as a measure of informational content since these also are not invariant to linear transformations of the data. For example, the data can again always be orthogonalized and rescaled so that all of the eigenvalues are equal to one.

The essential problem with these techniques is that they ignore the parameters while trying to assess the informational content of the data. Clearly, an evaluation of the strength of the data depends upon the scale and nature of the parameters. One cannot label a variance or a confidence interval (or, even worse, a part of the variance) as large or small without knowing what the parameter is and how much precision is required in the estimate of that parameter. In particular, a seemingly large variance may be quite satisfactory if the parameter is very large, if one has strong a priori information about the parameter, or if the parameter is uninteresting (perhaps because the associated variable will be constant during the forecast period). A meaningful assessment will require a well-defined loss function which must necessarily depend upon the particular problem being examined.

II. Ridge Regression

For estimation purposes, MS prefer to rewrite the model in correlation form:

$$(3) \qquad \frac{Y - \bar{Y}}{\sqrt{n-1}\,S_Y} = \sum_{i=1}^{3}\left(\frac{\beta_i}{S_Y}\right)\left[\frac{Z_i}{\sqrt{n-1}}\right] + \sum_{1\le i\le j}^{3}\left(\beta_{ij}\frac{S_{ij}}{S_Y}\right)\left[\frac{Z_iZ_j - \bar{Z}_i\bar{Z}_j}{\sqrt{n-1}\;S_{ij}}\right] + \frac{\epsilon - \bar{\epsilon}}{\sqrt{n-1}\;S_Y}$$

(which is again a transformation that does not alter the model). They then argue that "the 'fly in the ointment' with least squares is its

requirement of unbiasedness....Thus, it is meaningful to focus on the achievement of small mean square error as the relevant criterion, if a major reduction in variance can be obtained as a result of allowing a little bias. This is precisely what the ridge and generalized inverse solutions accomplish."

However, a MSE comparison for these suggested estimators is always ambiguous since the size of the bias depends upon the unknown population values of the parameters. Indeed, a major advantage of unbiased estimators is that the MSE's do not depend upon the actual values of the parameters.

Alternatively, least squares can be justified on likelihood grounds or as the mode of the posterior with improper uninformative priors. In response to these justifications, MS argue that a reasonable person would have bounded priors and that in correlation form "it is exceedingly rare for the population value of any regression coefficient to be larger than three in a real problem."* Now if one is as well informed as MS about the population values of parameters, then one would certainly want to use that information to choose among biased estimators or to more directly

---

*To formally analyze this strange statement, consider the simple model

$$Y = \beta + \beta_1 X_1 + \beta_2 X_2 + \epsilon \ .$$

In correlation form, the (squared) parameters are

$$\frac{\beta_i S_i}{S_Y}^2 = \frac{\beta_i^2 S_i^2}{\beta_1^2 S_1^2 + \beta_2^2 S_2^2 + 2\beta_1\beta_2 S_1 S_2 r + S_\epsilon^2}$$

where r is the sample correlation coefficient between X_1 and X_2, and ϵ is assumed for simplicity to be uncorrelated in sample with X_1 and X_2. Now, if $\beta_1 S_1 = -\beta_2 S_2$ (to take an extreme example), then $(\beta_i S_i / S_Y)^2 = 1/[2(1-r) + (S_\epsilon^2/\beta_i^2 S_i^2)]$ which will be arbitrarily large as $r \to 1$ and $S_\epsilon^2/\beta_i^2 S_i^2 \to 0$.

apply an optimal estimator based on the available information. Unfortunately, a comparison with such procedures only serves to reveal the sterility of ridge estimators.

More concretely, consider the model

$$Y = X\beta + \epsilon$$

$$\epsilon \sim N[0, \sigma_\epsilon^2 I]$$

where one has the supplementary information

$$\beta = b + u$$

$$u \sim N(0, \Sigma) \ .$$

Theil-Goldberger's mixed estimation is a classical approach which views β as fixed and b and Y as random and applies generalized least squares to the two sets of data to obtain

$$\beta^* = [X'X + \sigma_\epsilon^2 \Sigma^{-1}]^{-1} [X'Y + \sigma_\epsilon^2 \Sigma^{-1} b] \ .$$

Chipman's partially Bayesian analysis takes b and Y as fixed and β as random and obtains β^* as the linear minimum mean squared error estimator. In a fully Bayesian approach, β^* is found to be the posterior mean.

Since the ridge estimator is

$$\hat{\beta}^R = [X'X + kI]^{-1} X'Y$$

it is possible to motivate ridge regression from a wide variety of viewpoints when one actually has <u>a priori</u> information of the special form

$$b = 0 \quad \text{and} \quad \Sigma = \frac{\sigma_\epsilon^2}{K} I$$

which is to say that one has orthogonal priors with common variances centered at the origin. Conversely, the theoretical inadequacy of ridge regression is that it does not even attempt to assess the appropriateness of these implicit priors.

If one's priors are normally distributed, then they can always be centered, diagonalized and scaled. Thus, it is always possible to linearly transform a model so that a ridge estimator is appropriate. The problem with ridge estimation in practice is that the model is linearly transformed so as to center, scale, and partially diagonalize the variables rather than the parameters. And, indeed, one is never even asked to contemplate the reasonableness of the implicit assumption that the parameters have zero means, zero covariances and identical variances. Notice also that while linear transformations do not affect the model, they do change the parameters which are being explicitly estimated and consequently should change the variance-covariance matrix for one's priors on the explicitly estimated coefficients. In particular, if the priors implicit in a ridge approach are actually appropriate for the model in any one of the forms (1), (2), or (3), then ridge regression will be generally inappropriate for the remaining forms in that the variances will not all be equal and some of the covariances will be nonzero. Thus, "standardization" of the data does affect the ridge estimates and the appropriateness of a ridge technique. Unfortunately, the standardization is not based upon either of these considerations.

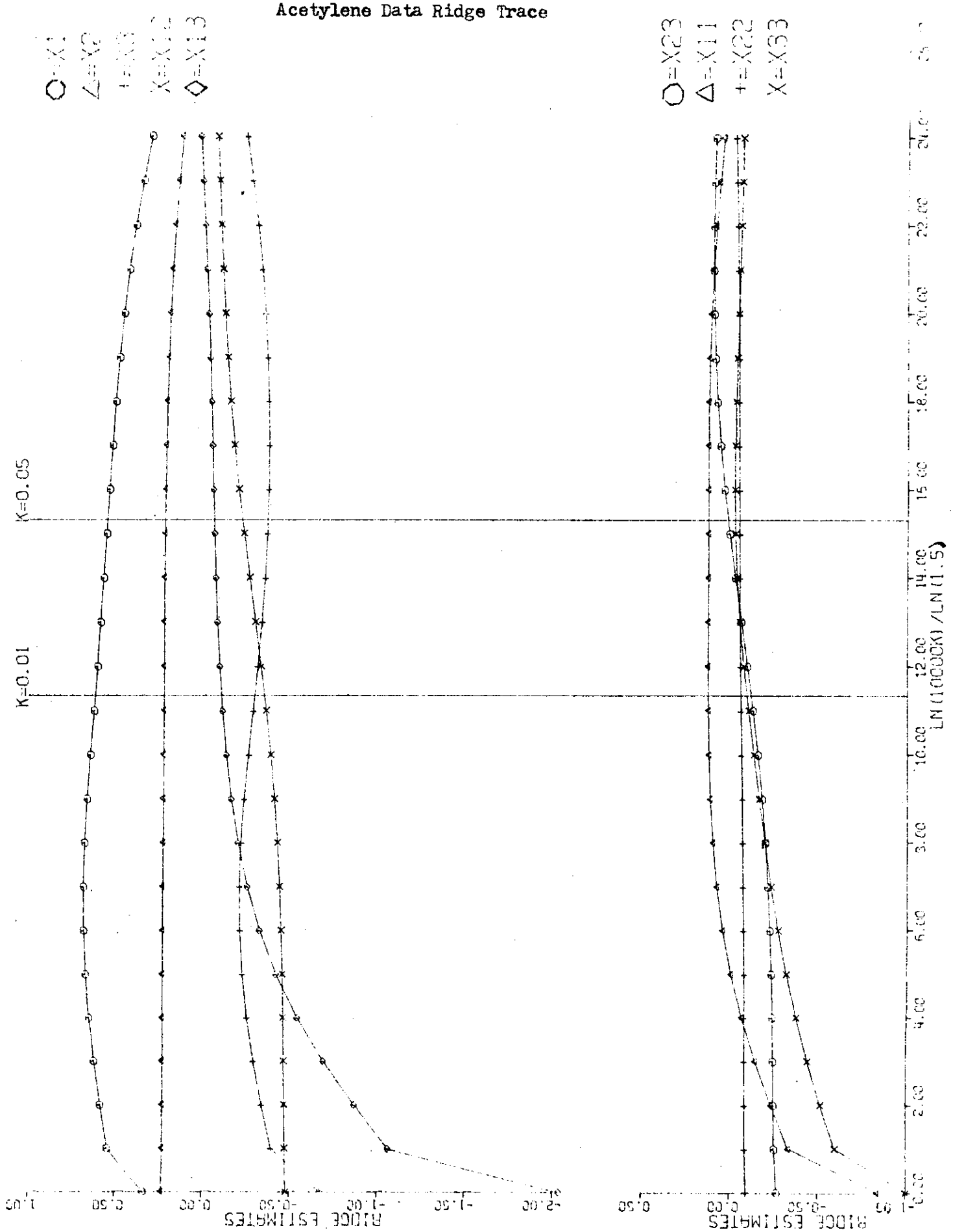
Similarly, in selecting k, the ridge user does not take into account the fact that the implicit variances on the priors are being set

at σ_ϵ^2/k . Instead, k is chosen by an ad hoc procedure whose loose theoretical underpinnings are highlighted by the wholly arbitrary restriction that k be less than one, which compels one to assume that the variance of the disturbance term is less than the variance on one's priors. Inside this range, k is selected so as to yield "reasonable" variance inflation factors (which we have seen to be a meaningless objective) or so as to yield estimates which are relatively insensitive to small changes in k . Figure 1 gives a ridge trace for the acetylene data example. The intercepts are the least squares estimates (k = 0) ; the next unit corresponds to k = .0001, and each unit after that corresponds to a 50% increase in k over the preceding unit. Despite MS's assurances that it is easy to select k from the ridge trace, we do not view the estimates as stable for the k's of .01 or .05 which they selected and, indeed, do not see much stability for any values of k less than one. In addition, there is the problem that this sensitivity analysis is not invariant to linear transformations of the model. That is, ridge traces of linear combinations of the parameters will not generally show the same regions of relative stability. More fundamentally, we cannot understand why k should be chosen on the basis of local insensitivity. The fact that the estimates are not much different with k at .01 or .02 is not a convincing argument for using these estimates rather than the very different estimates which result from a k of .001 or 2.0.

Another puzzle for us is MS's statement that, "If the predictor variables are orthogonal, then the coefficients would change very little (i.e., the coefficients are already stable) indicating the least squares solution is a good set of coefficients." We have already pointed out that orthogonality does not imply strong data; therefore the ridge trace is

Figure 1

Acetylene Data Ridge Trace



misleading if it favors the least squares point in this situation. Similarly, since the data can always be orthogonalized, their statement illustrates our previous point that the selection of k depends upon which parameters are being examined. On the other hand, it is difficult for us to agree with their statement. It is true that in correlation form with orthogonal data the curve décolletage will be a straight line, so that the ridge estimates will move directly toward zero; but this does not imply that they will not move, or move slowly, or move only a small distance.

Seemingly, without prior information, the only theoretically defensible use of ridge regression would be with a value of k that was so small as to give, for all practical purposes, the least squares estimates. One could then argue that one was simply assuming proper locally almost uniform priors. MS do in fact argue that, "the ridge estimate is equivalent to placing mild boundedness requirements on the coefficient vector."

Were this so, ridge estimates would be only a formal curiosity. But of course it is not so in that in practice ridge regression is used to obtain estimates which are significantly different from the presumedly unsatisfactory least squares estimates. This is clear in the acetylene data example from a comparison of the estimates in MS's Table 2 and also from an examination of the values of k that were used, .01 and .05.

Using the least squares estimated value of σ_e^2, the implicit common variances on the priors are respectively .0328 and .0066, which most would feel are fairly tight priors given the size of the least squares estimates. In terms of confidence intervals, MS acted as if for each parameter they were 95% confident that the population value of the parameter is no further than .36 or .16 from zero. In contrast, 5 of the 9

least squares estimates are outside of the larger interval and 8 outside the smaller. Clearly, one is assuming more than "mild boundedness."

In fact, using Theil's suggested test, MS's implicit priors are actually incompatible with the data. For $k = .01$, the chi-square statistic here is 24.7 and for $k = .05$ it is 93.5, as compared to a critical point of 16.9 for a test at the 5% level. With priors of the ridge type, one would have to use a $k \leq .006$ to pass Theil's test and a much smaller k to be assuming only "mild boundedness."

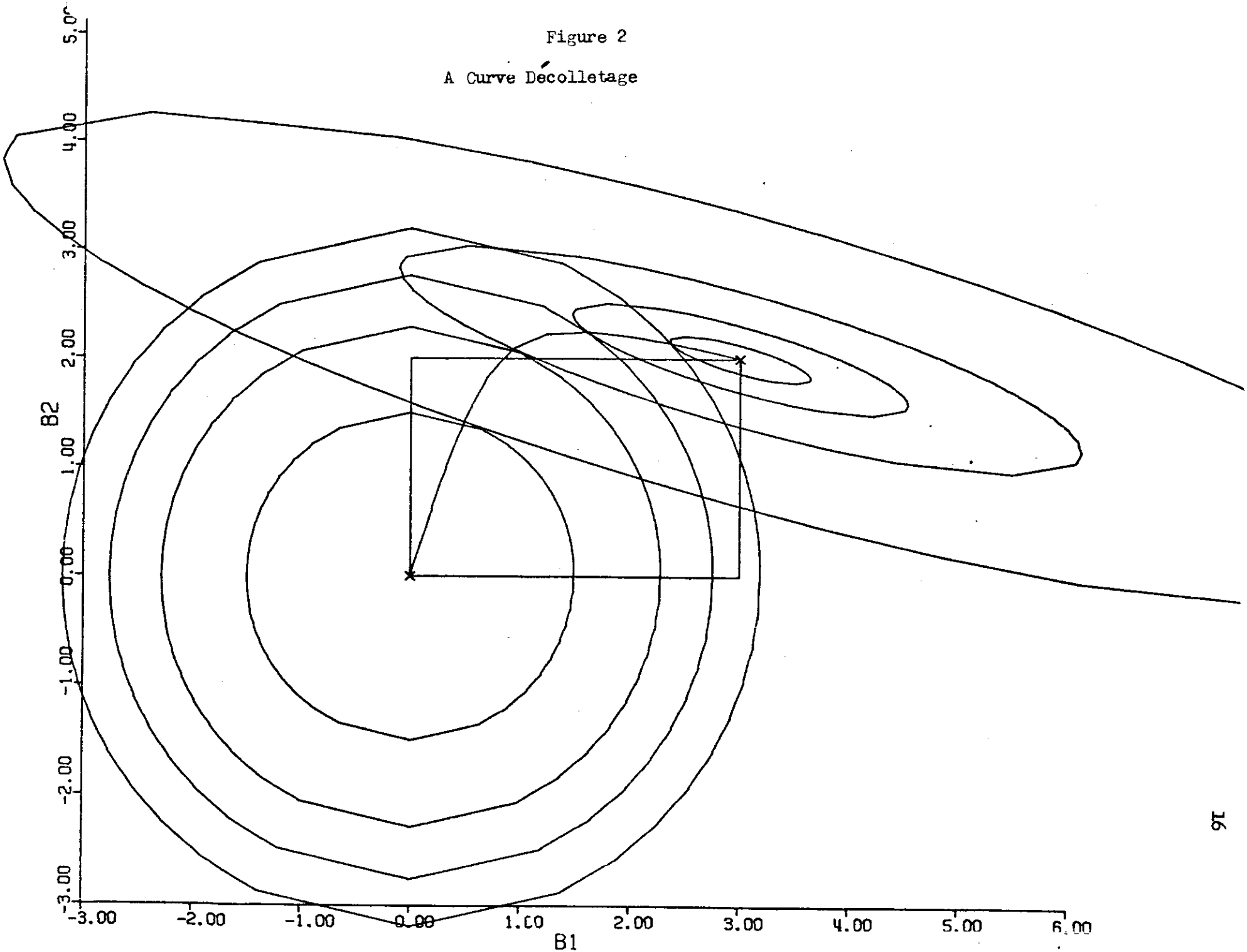
III. Ridge Regression, Principal Components and Generalized Inverses

In ridge regression, one works with orthogonal priors and typically very nonorthogonal data. In this situation, the estimates are only shrunk towards the origin in a matrix sense, in that individual parameter estimates may move away from the origin or may move past the origin and change signs. Both of these phenomena in fact occur with MS's acetylene data example. For a rough intuitive explanation, consider the 2 parameter ridge case where the explanatory variables are highly positively correlated; the data contains more information about β_2 than β_1 (due to the higher variance of X_2); and the data is more informative than the priors. In this case the priors will tend to pull $\hat{\beta}_1$ towards the origin while the data will move $\hat{\beta}_2$ in the opposite direction as illustrated in Figure 2.

One implication of this is that ridge regression cannot be used as an ad hoc way of shrinking all of the estimates. More generally, with nonorthogonal data and more than two parameters, it will be exceedingly difficult to foresee the consequences of using ridge regression. Thus, one's priors should be truly based upon one's priors and not upon the

Figure 2

A Curve Décolletage



hoped for posterior.

However, it is possible to make a simple parameter-by-parameter analysis by translating the data into its principal components. And in this framework we find a clear relationship between ridge regression, principal components analysis, and Marquardt's generalized inverse technique.

If the columns of A are the orthonormal eigenvectors of $X'X$, then

$$Y = X\beta + \epsilon = XAA'\beta + \epsilon$$
$$= P\gamma + \epsilon$$

where the columns of P are the principal components of $X'X$ and $P'P$ is a diagonal matrix with the eigenvalues (λ_i) of $X'X$ as its diagonal elements. Similarly we can transform the implicit ridge priors on β

$$\beta = 0 + U \ , \quad E(UU') = (\sigma_\epsilon^2/K)I$$

$$\gamma = A'\beta = 0 + A'U \ , \quad E(A'UU'A) = (\sigma_\epsilon^2/K)I \ .$$

Thus, one could equivalently carry our a ridge analysis using either the original data or the principal components. In the latter form, however, the orthogonality of both the data and the priors yields estimates which are simple weighted averages of the likelihood estimate and the prior mean

$$\hat{\gamma}_i^R = \left(\frac{\lambda_i}{\lambda_i + k}\right)\hat{\gamma}_i + \left(\frac{k}{\lambda_i + k}\right)0$$

with those estimates with the largest variances $(\sigma_\epsilon^2/\lambda_i)$ being shrunk

the most. Similarly, the larger is k, the closer are all of these estimates to zero.

There is an obvious similarity between this and the usual principal components analysis. In the latter, one typically uses

$$\hat{\gamma}_i^{P} = \begin{cases} \hat{\gamma}_i , & \lambda_i \geq h \\ 0 , & \lambda_i < h \end{cases}$$

where h is some selected cutoff point for retaining components. In this form of principal components analysis, then, one chooses between zero and the least squares point based upon the relative variance of the least squares estimate. In ridge regression, one chooses a point between zero and the least squares estimate based again upon the relative variance of the estimate.

Marquardt's generalized inverse technique is a minor variant of this principal components analysis. He allows for the case of an intermediate eigenvalue that is neither obviously large nor small, for which one chooses a point between the least squares estimate and zero based upon the closeness of the eigenvalue to being large rather than small:

$$\hat{\hat{\gamma}}_i^{G} = \begin{cases} \hat{\gamma}_i , & \lambda_i \geq h_1 \\ r\hat{\gamma}_i , & h_2 < \lambda_i < h_1 \\ 0 , & \lambda_i \leq h_2 \end{cases} .$$

All three of these techniques consequently fit into the class of estimators which use simple weighted averages of the least squares estimate and some (for generality, possibly non-zero) point C_i

$$\hat{\hat{\gamma}}_i = \alpha_i \hat{\gamma}_i + (1 - \alpha_i) C_i$$

where the weights $0 \leq \alpha_i \leq 1$ depend upon the variances of the least squares estimates. These techniques consequently all suffer from the same inherent weaknesses.

First, the estimates depend upon the wholly arbitrary initial standardization of the model. Simple innocent linear transformations of the data will change the principal components and the characteristic roots and thereby alter the parameters (and the variances) which are explicitly estimated. Thus, different parameters will be set equal to or shrunk towards zero; equivalently, a particular set of parameters will be shrunk towards different points. Such linear transformations will also affect the degree of movement away from the least squares estimates through the number of parameters which are shrunk and/or the extent to which they are shrunk.

Second, these estimators incorrectly use the relative variance of an estimate as a measure of the strength of the data. One inadequacy in this is the neglect of the absolute size of the variances, which depends upon the size of σ^2_ε. More fundamentally, the variances alone cannot adequately describe the strength of the data relative to one's priors. For instance, a seemingly large variance should not motivate the abandonment of the least squares estimate if this point is far from one's priors (here the origin) or if one's priors are very weak. By ignoring these factors, these shrinkage techniques can yield estimates which would have little posterior weight and which would be rejected by classical hypothesis tests.

This is in fact the case with MS's acetylene data example. Figure 3 displays the least squares estimates of the coefficients of the principal components, the associated eigenvalues, MS's ridge estimates, and MS's generalized inverse estimates.

FIGURE 3

Estimates of the Coefficients of the Principal Components

Parameter	Least Squares Estimate (t-statistic)	λ_i	Ridge Estimates k = .01	Ridge Estimates k = .05	Generalized Inverse r = 3.8
γ_1	.352 (39.88)	4.205	.351	.348	.352
γ_2	-.005 (-.39)	2.163	-.005	-.005	-.005
γ_3	-.600 (-35.38)	1.138	-.595	-.575	-.600
γ_4	.238 (13.43)	1.041	.236	.227	.190
γ_5	.009 (.33)	.3845	.009	.008	.0
γ_6	.217 (2.67)	.0495	.181	.108	.0
γ_7	-.383 (-2.47)	.0136	-.221	-.082	.0
γ_8	.521 (2.06)	.0051	.176	.048	.0
γ_9	-2.401 (-1.31)	.0001	-.023	-.005	.0
SSR	.0023		.0039	.0067	.0110

Many of the coefficients with low eigenvalues are significantly different from zero because of the size of the estimates and the small value of σ_ε^2. As a consequence, four of the six restrictions imposed by MS's generalized inverse procedure would be rejected by individual classical hypothesis tests at the 5% level; and the 5 exact restrictions would even be rejected by a joint test. Due to the very limited degrees of freedom, a 95% nine-dimensional confidence region based solely on the least squares estimates is extremely large and does include each of the three vectors of estimates selected by MS. On the other hand, as we've stated before, if the ridge procedures are viewed as data augmenting techniques, then this supplementary information is incompatible with the data.

The inadequacy of retaining components on the basis of eigenvalues was recognized several years ago by Hotelling, who pointed out that components which are of little use in explaining the explanatory variables may be very powerful in explaining the dependent variable. This has led Massy and others to advocate the deletion of components whose coefficients are statistically insignificant. While this avoids the imposition of weakly held restrictions that are rejected by the data, it still mechanically overrules the likelihood point whenever an arbitrarily selected point (zero) is inside the confidence interval. Thus, this approach permits one to impose ad hoc constraints where the data is very informative (when zero is inside a tight band) and to refrain when constraints are badly needed (when zero is outside a large band).

The third inadequacy of the procedures considered here is that they do not even ask of the researcher what would be the most reasonable point to shrink the estimates towards or which estimates he feels most confident

about shrinking. Thus, one acts as if there were supplementary information available without ever confronting the validity of that presumed information. As a consequence, the desirability of the estimates depends upon unasked questions and, in addition, one loses the potential gains from introducing truly held a priori beliefs.

These points are illustrated in more detail by the consideration of a specific loss function. MS follow Hoerl and Kennard in using average mean squared error, or expected squared distance to β ,*

$$L = E(\beta-\hat{\hat{\beta}})'(\beta-\hat{\hat{\beta}}) = \sum_i MSE(\hat{\hat{\beta}}_i)$$

Again, we can work with principal components as these preserve average mean square error

$$\begin{aligned} E(\beta-\hat{\hat{\beta}})'(\beta-\hat{\hat{\beta}}) &= E(\beta-\hat{\hat{\beta}})'AA'(\beta-\hat{\hat{\beta}}) \\ &= E(A'\beta - A'\hat{\hat{\beta}})'(A'\beta - A'\hat{\hat{\beta}}) = E(\gamma-\hat{\hat{\gamma}})'(\gamma-\hat{\hat{\gamma}}) \\ &= \sum_i MSE(\hat{\gamma}_i) \ . \end{aligned}$$

Now, for the procedures considered here

$$\hat{\hat{\gamma}}_i = \alpha_i\hat{\gamma}_i + (1 - \alpha_i)C_i$$

the MSE can be broken into two parts

$$\begin{aligned} E(\gamma_i - \hat{\hat{\gamma}}_i)^2 &= \alpha_i^2 MSE(\hat{\gamma}_i) + (1 - \alpha_i)^2(\gamma_i - C_i)^2 \\ &= Var(\hat{\hat{\gamma}}_i) + [Bias(\hat{\hat{\gamma}}_i)]^2 \ . \end{aligned}$$

---

*For nonorthonormal data, this loss function does not insure smaller mean square prediction error.

Thus, as compared to the least squares estimate $(\hat{\gamma}_i)$, shrinking unambiguously reduces the variance and increases the bias. If the least squares variance is not zero and $(\gamma_i - C_i)^2$ is bounded then there will always be some weights α_i which reduce the mean squared error. (If $(\gamma_i - C_i)^2 < \text{MSE}(\hat{\gamma}_i)$ then this will be true of all $\alpha_i < 1$.) If, however, γ_i is unknown then one will also not know whether or not the MSE has been reduced. Thus the gains from shrinking to wholly arbitrary points are necessarily accidental.

Notice also that the variance reduction is entirely independent of the shrinking target, C_i. That is, shrinking towards the origin is of no advantage for variance reduction. Where the choice of target does show up is in the squared bias, and here a more accurate target is unambiguously beneficial. Thus, the origin can only be justified as a shrinking target if it is favored over other potential targets on a priori grounds.* But if one has such a priori beliefs, then they should be incorporated in a straightforward fashion rather than being mangled by an inflexible procedure. And if one doesn't have such beliefs, then there is no justification for using procedures which mechanically impose arbitrary priors.

---

*Note that all of this discussion refers to procedures in which one acts as if one had diagonal priors. The analysis is considerably more complex with nondiagonal priors.

IV. Summary

Ridge estimation has several characteristics in common with other estimation practices which are designed to overcome weak data. Our indictment of these characteristics is consequently also a criticism of these other procedures.

One characteristic is the incorrect labelling of nonorthogonal data as uninformative data. Collinearity can be one source of weak data, but the strength of the data cannot be measured solely by the orthogonality of the data.

A second characteristic is the use of an estimator which is not invariant to nonsingular linear transformations of the model. Every model has an infinite number of implicitly estimatable parameters. Since the choice of parameters to explicitly estimate is wholly arbitrary, estimates which depend upon this arbitrary choice are necessarily capricious.

A third characteristic is the use of wholly ad hoc constraints on the parameters. Such pseudo information yields estimates of unknowable reliability which can only be accidentally successful.

A fourth characteristic is the abstention from the use of actual additional information about the parameters. Such prior information is a defensible supplement to weak data and is unambiguously superior to arbitrarily constructed pseudo information.

REFERENCES

Chipman, J. S. (1964): "On Least Squares with Insufficient Observations," Journal of the American Statistical Association, 59, 1078-1111.

Hoerl, A. E. and Kennard, R. W. (1970): "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, 12, 55-67.

_____ (1970): "Ridge Regression: Applications to Nonorthogonal Problems," Technometrics, 12, 69-82.

Hotelling, H. (1957): "Relation of the Newer Multivariate Statistical Methods to Factor Analysis," British Journal of Statistical Psychology, 10, 69-79.

Johnston, J. ( ): Econometric Methods, Second Edition, McGraw-Hill Book Company, New York.

Leamer, E. E. (1973): "Multicollinearity: A Bayesian Interpretation," The Review of Economics and Statistics, 55, 371-380.

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

_____ and Snee, R. D. (1975): "Ridge Regression in Practice," The American Statistician, 29, 3-20.

Massy, W. (1965): "Principal Components in Exploratory Statistical Research," Journal of the American Statistical Association, 60, 234-256.

Mayer, L. S. and Wilke, T. A. (1973): "On Biased Estimation in Linear Models," Technometrics, 15, 497-508.

Theil, H. (1971): Principles of Econometrics, John Wiley and Sons, New York.

_____ and Goldberger, A. S. (1961): "On Pure and Mixed Statistical Estimation in Economics," International Economic Review, 2, 65-78.

Theobald, C. M. (1974): "Generalizations of Mean Square Error Applied to Ridge Regression," Journal of the Royal Statistical Society, Series B, 36, 103-106.