Ridge Regression: Applications to Nonorthogonal Problems

ARTHUR E. HOERL AND ROBERT W. KENNARD

University of Delaware and E. I. du Pont de Nemours & Co.

This paper is an exposition of the use of ridge regression methods. Two examples from the literature are used as a base. Attention is focused on the RIDGE TRACE which is a two-dimensional graphical procedure for portraying the complex relationships in multifactor data. Recommendations are made for obtaining a better regression equation than that given by ordinary least squares estimation.

1. INTRODUCTION

Multiple linear regression is one of the most widely used of all statistical methods. It is used by data analysts in nearly every field of science and technology as well as the social sciences, economics, and finance. Today it is a rare computer center that does not have a general purpose program of some kind to perform the standard calculations. But, as has been shown in [3], the estimation of regression coefficients can present problems when the data vectors for the predictors are not orthogonal. In particular, the coefficients tend to become too large in absolute value and it is possible that some will even have the wrong sign. And the probability of such difficulties increases the more the prediction vectors deviate from orthogonality.

Consider the standard model for multiple linear regression

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{1.0}$$

where $E[\varepsilon] = \mathbf{0}$, $E[\varepsilon\varepsilon'] = \sigma^2\mathbf{I}_n$ and $\mathbf{X}$ is $(n \times p)$ and full rank. Let

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{1.1}$$

be the least squares estimate of β. The difficulties in this standard estimation are a direct consequence of the average distance between $\hat{\beta}$ and β. In particular, if L_1^2 is the squared distance between $\hat{\beta}$ and β, then the following hold:

$$L_1^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta) \tag{1.2}$$

$$E[L_1^2] = \sigma^2 \text{ Trace } (\mathbf{X}'\mathbf{X})^{-1} \tag{1.3}$$

$$E[\hat{\beta}'\hat{\beta}] = \beta'\beta + \sigma^2 \text{ Trace } (\mathbf{X}'\mathbf{X})^{-1} \tag{1.4}$$

In terms of the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ of $\mathbf{X}'\mathbf{X}$, (1.3) can be written as:

$$E[L_1^2] = \sigma^2 \sum_1^p (1/\lambda_i) > \sigma^2/\lambda_p. \tag{1.5}$$

Received Aug. 1968; revised June 1969.

As the vectors of $\mathbf{X}$ deviate further from orthogonality, λ_p becomes smaller and $\hat{\beta}$ can be expected to be farther from β.

Ridge regression, as has been detailed in [3], is an estimation procedure based upon

$$\hat{\beta}^* = [\mathbf{X}'\mathbf{X} + \mathbf{K}]^{-1}\mathbf{X}'\mathbf{Y} \tag{1.6}$$

where $\mathbf{K}$ is a diagonal matrix of non-negative constants. A useful procedure uses $\mathbf{K} = k\mathbf{I}_p$, $k \geq 0$. Ridge regression has two aspects. The first is the RIDGE TRACE which is a two-dimensional plot of the $\hat{\beta}_i^*(k)$ and the residual sum of squares, $\phi^*(k)$, for a number of values of k in the interval [0, 1]. The trace serves to portray the complex interrelationships that exist between nonorthogonal prediction vectors and the effect of these interrelationships on the estimation of β. The second aspect is the determination of a value of k that gives a better estimate of β by dampening the effect of (1.5).

How Ridge Regression can be used by the data analyst is shown in the sections to follow by applying it to two problems that have already appeared in the literature.

2. Ten-Factor Example

Gorman and Toman [1] have published a summary of the data for a 10-factor regression problem that has nonorthogonal factors. The summary, $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$ in correlation form, is reproduced in Table I. There is a large number of significant interfactor correlations. This is reflected in the eigenvalues of $\mathbf{X}'\mathbf{X}$ which are:

$$\lambda_1 = 3.692 \qquad \lambda_6 = 0.659$$
$$\lambda_2 = 1.542 \qquad \lambda_7 = 0.357$$
$$\lambda_3 = 1.293 \qquad \lambda_8 = 0.220$$
$$\lambda_4 = 1.046 \qquad \lambda_9 = 0.152$$
$$\lambda_5 = 0.972 \qquad \lambda_{10} = 0.068$$

The sum of the reciprocals of the eigenvalues is $\sum (1/\lambda_i) = 33.825$. Thus, (1.5) shows that the expected squared distance of the coefficient estimate, $\hat{\beta}$, from β is $33.825\ \sigma^2$, which is more than three times what it would be for an orthogonal system.

Since the smallest eigenvalue λ_{10} is not zero, the factors do define a 10-dimensional space in the mathematical sense. However, the first six eigenvalues sum to 9.184 so that "most of the variation" can probably be accounted for in about six dimensions. And, in fact, Gorman and Toman use this problem as an example to portray a short-cut method for finding a "best" subset of factors of a specified size less than ten without having to compute all regressions of the specified size.

In Figure 1 is the Ridge Trace for this problem. This trace was constructed by computing a total of 15 regressions using $\hat{\beta}^* = [\mathbf{X}'\mathbf{X} + k\mathbf{I}]^{-1}\mathbf{X}'\mathbf{Y}$ and 15 values of k in the interval (0, 1) as indicated by the dots. The Ridge Trace gives a two-dimensional portrayal of the effects of the factor correlations and makes

TABLE I

Correlation Coefficients for 10-Factor Example N = 36*

X_1	1									
X_2	-0.04	1								
X_3	0.51	-0.00	1							
X_4	0.12	-0.16	0.00	1						
X_5	-0.71	0.06	-0.59	-0.07	1					
X_6	-0.87	0.09	-0.65	-0.09	0.84	1				
X_7	-0.09	0.24	-0.02	0.03	0.38	0.13	1			
X_8	-0.00	0.01	0.34	0.08	-0.36	-0.20	-0.48	1		
X_9	-0.09	0.09	-0.08	0.02	-0.14	0.04	0.07	-0.18	1	
X_{10}	-0.36	-0.30	-0.44	-0.09	0.54	0.45	0.40	-0.46	0.05	1
Y	-0.81	-0.10	-0.63	-0.10	0.56	0.81	0.04	0.06	0.16	0.45 1

*The two-digit correlations were used in all computations. Gorman and Toman made the raw data available. No practical differences are found using all significant digits.

possible assessments that cannot be made even if "all" regressions are computed. The following are examples:

(i) The coefficients from the ordinary least squares are undoubtedly overestimated. At least, they are collectively not stable. It is unlikely that another set of y's would give $\hat{\beta}_i$ like these. Moving a short distance from the least squares point $k = 0$ shows a rapid decrease in absolute value of at least two of them, namely, those for factors 5 and 6. Figure 2 shows the

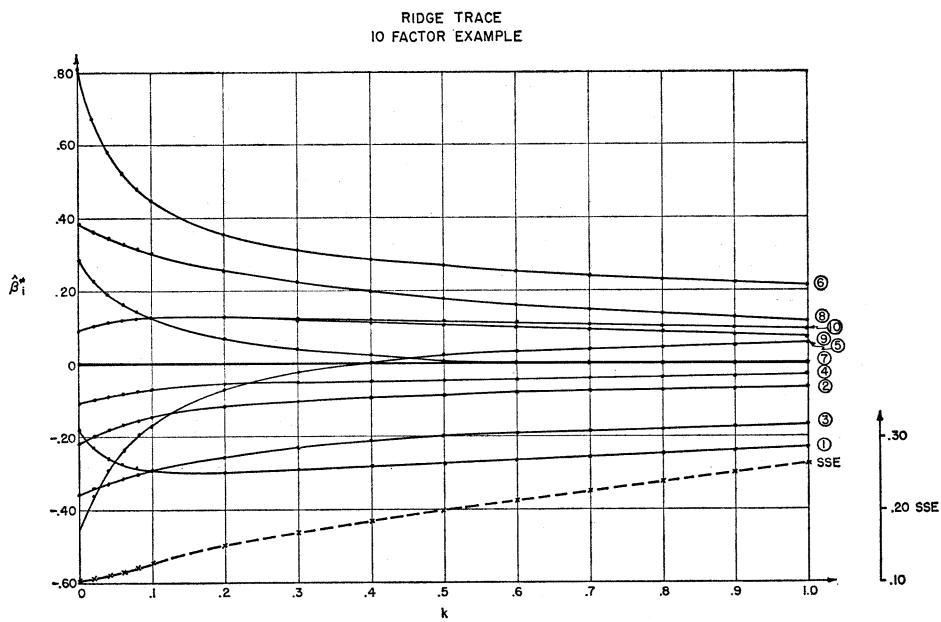

FIGURE 1—Ridge Trace. Ten Factor Example

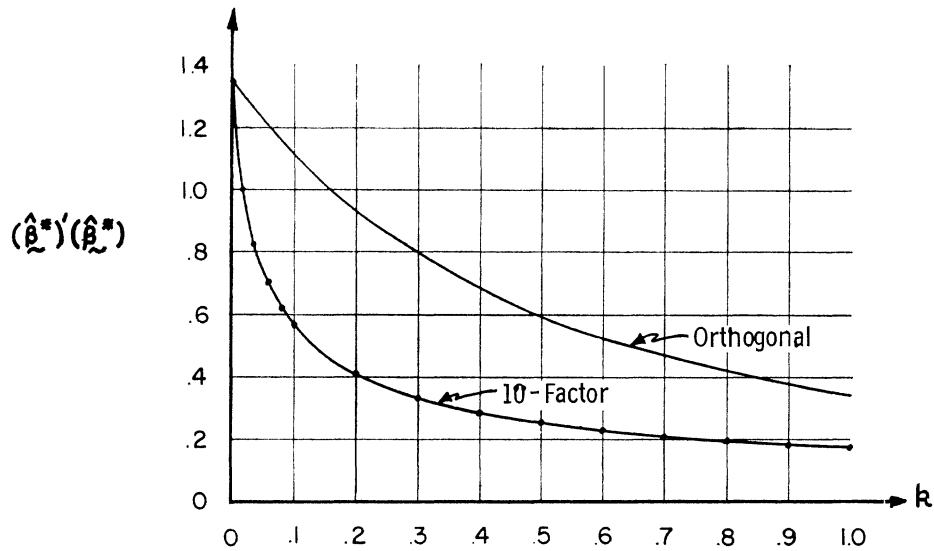

FIGURE 2—Ten Factor Example. Squared Length of Coefficient Vector

decrease in the squared length of the coefficient vector with k. When $k = .1$, it is 41.3% of its original value; for an orthogonal system it would be 83%.

(ii) Factor 5 has the negative coefficient with the largest absolute value. But the addition of $k > 0$ quickly drives it toward zero and it then becomes positive. Such action should not be surprising, especially when it is compared with the action of factor 6. Factor 6 also decreases rapidly but stabilizes and does not go to zero. Factors 5 and 6 have a simple correlation coefficient of 0.84 which says that to a first approximation, they are the same factor but with different names. It would be surprising if their true effects were opposite in sign. (Without a knowledge of the underlying technology, no definitive statement can be made.) The covariance of -4.33 is driving them apart so that they are opposite in sign. The phenomenon observed here is not atypical. Positive coefficients for highly correlated factors can be stable as a sum, especially when they are correlated to various degrees with other factors.

(iii) The correlations with other factors causes factor 1 to be underestimated. At $k = 0$ factor 1 is the second least important negative factor. But with the addition of $k > 0$ it increases in absolute value. The other negative factors are slightly overestimated and when sufficient $k > 0$ has been added to stabilize the system, factor 1 becomes the most important negative factor.

(iv) Factor 7 is overestimated and is driven toward zero.

(v) At a value of k in the interval (0.2, .3) the system has stabilized and coefficients chosen from a k in this range will undoubtedly be closer to β and more stable for prediction than the least squares coefficients or some subset of them.

It is interesting to compare how one would handle the subset problem with the RIDGE TRACE and what is obtained from a criterion such as that used by Gorman and Toman [1] and by Hocking and Leslie [2]. Take a best subset of size six, that is, the "best" six predictors. The four that are eliminated using the C_p statistic are factors 9, 10, 4, 1 and in that order for larger subsets.

First, it will be noted that eliminating 9, 10, 4, and 1 and using minimum sum of squares for estimation does not eliminate the tendency to overestimate and to have an unstable solution. With 9, 10, 4, and 1 omitted, the eigenvalues of the resulting $\mathbf{X}'\mathbf{X}$ are:

$$\lambda_1 = 2.703 \qquad \lambda_4 = .591$$
$$\lambda_2 = 1.300 \qquad \lambda_5 = .291$$
$$\lambda_3 = .999 \qquad \lambda_6 = .115$$

In Figure 3 is shown the RIDGE TRACE for the reduced system. Except for factor 1, all the instabilities and overestimation are still there. In fact, now that they are uncoupled from factor 1, factors 5 and 6 have even larger coefficients than for the full system. The squared length of the coefficients vector is now 1.91 compared to 1.35, so that the solution is even more unstable than it was prior to the deletion. The results for the reduced system should not be surprising. Factors 9, 10, and 4 do contribute the least and their estimation is not affected

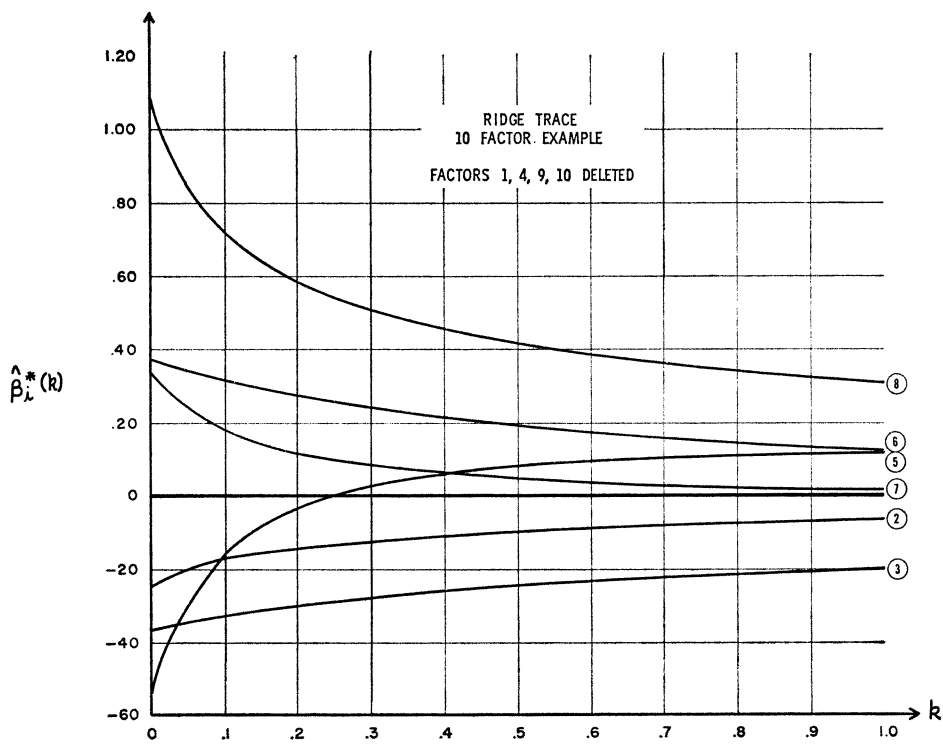

FIGURE 3

by their associations with the other factors (see the flat traces). Factor 1, at $k = 0$, is the next smallest contributor and would be the logical contender for elimination after 9, 10, and 4.

However, what would be logical to do from an examination of the RIDGE TRACE? As was pointed out in the earlier discussion, factors 5 and 7 are clearly unstable and are quickly driven toward zero with the addition of $k > 0$. Since these factors cannot hold their predicting power, it seems evident that they should be the factors to be eliminated. Figure 4 shows the RIDGE TRACE with factors 5 and 7 deleted. The wild overestimations and the instabilities associated with the full system are damped considerably. In fact, the system does not act unlike an orthogonal system. For example, Figure 5 shows the decrease in the length of the coefficient vector for this reduced system and that which would be observed for an orthogonal one. For the range of k of primary interest (< 0.5) they are practically coincident.

It is seen then that with a computation of only 10 to 15 regressions a comprehensive picture of the nonorthogonality effects can be obtained. In the preceding paragraphs considerable attention has been given to a comparison of RIDGE with factor screening methods. However, it is the opinion of the authors that the best strategy is the choice of a good value of k. In this example, the

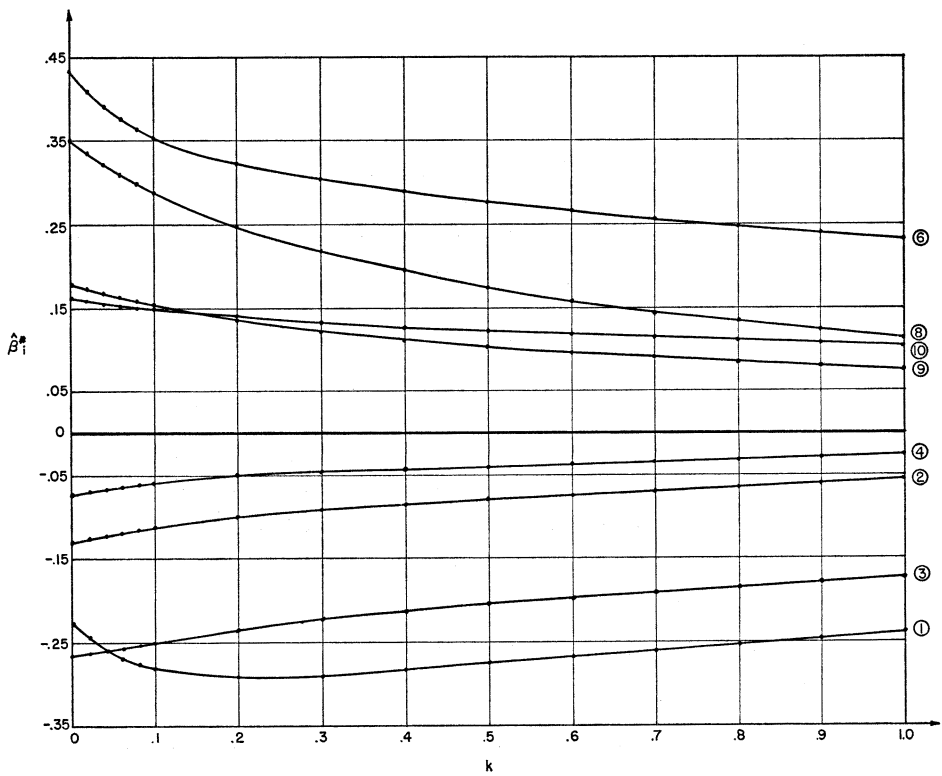

FIGURE 4—Ridge Trace. Ten Factor Example. Factors 5 and 7 deleted.

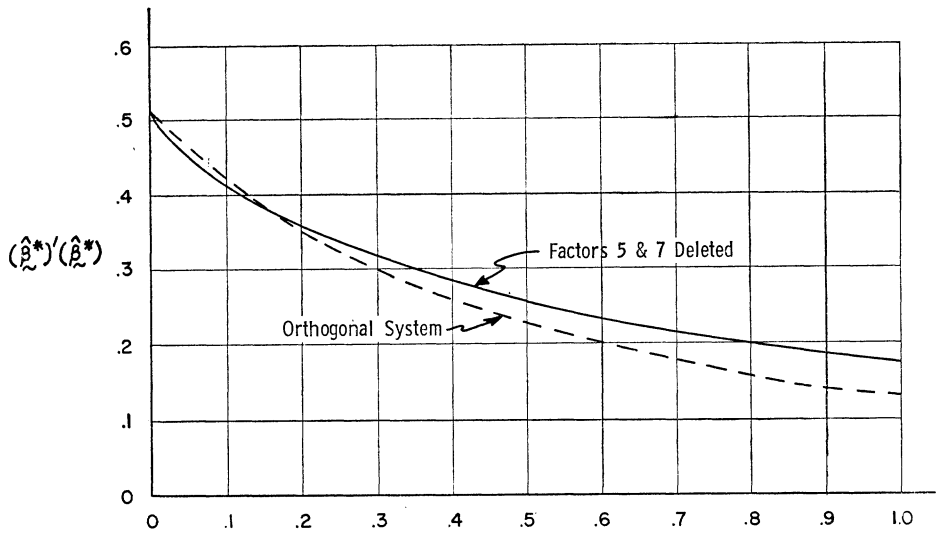

FIGURE 5—Ten Factor Example. Squared Length of Coefficient Vector. Factors 5 and 7 deleted.

system stabilizes in the region $k = 0.2$ to $k = 0.3$; therefore, choose $k = 0.25$ as a stable point solution. This gives (from the graph):

$$\hat{\beta}_1^* = -0.295 \qquad \hat{\beta}_6^* = 0.325$$
$$\hat{\beta}_2^* = -0.110 \qquad \hat{\beta}_7^* = 0.050$$
$$\hat{\beta}_3^* = -0.245 \qquad \hat{\beta}_8^* = 0.240$$
$$\hat{\beta}_4^* = -0.050 \qquad \hat{\beta}_9^* = 0.125$$
$$\hat{\beta}_5^* = -0.040 \qquad \hat{\beta}_{10}^* = 0.125$$

Then view the system as one of ten controlled factors with these coefficients as the "best" estimates. Factors with small effects have small coefficients. To "discard" a factor, set it at its average value for all predictions, which is the equivalent of setting the coefficient equal to zero. But do not delete and re-estimate; the possible effects of this strategy have already been demonstrated.

3. THIRTEEN-FACTOR EXAMPLE

In [4] Jeffers studies the maximum compressive strength of pitprops as a function of 13 physical factors that can be measured on the props. The sampling basis for the data is given in [4] and it can be assumed that a valid situation exists for applying some form of regression analysis. The relevant data in the form of correlation coefficients is reproduced in Table II.

The 13 eigenvalues of $\mathbf{X}'\mathbf{X}$ are as follows:

$$\lambda_1 = 4.219 \qquad \lambda_6 = .815 \qquad \lambda_{11} = .051$$
$$\lambda_2 = 2.378 \qquad \lambda_7 = .576 \qquad \lambda_{12} = .041$$
$$\lambda_3 = 1.878 \qquad \lambda_8 = .440 \qquad \lambda_{13} = .039$$
$$\lambda_4 = 1.109 \qquad \lambda_9 = .353$$
$$\lambda_5 = .910 \qquad \lambda_{10} = .191$$

TABLE II

Correlation Coefficients for Pitprop Data $N = 180$

TOPDIAM	X_1	1															
LENGTH	X_2	0.945**	1														
MOIST	X_3	0.364**	0.297**	1													
TESTSG	X_4	0.342**	0.284**	0.882**	1												
OVENSG	X_5	-0.129	-0.118**	-0.148*	0.220**	1											
RINGTOP	X_6	0.313**	0.291**	0.153*	0.381**	0.364**	1										
RINGBUT	X_7	0.496**	0.503**	-0.029	0.174*	0.296**	0.813**	1									
BOWMAX	X_8	0.424**	0.419**	-0.054	-0.059	0.004	0.090	0.372**	1								
BOWDIST	X_9	0.592**	0.648**	0.125	0.137	-0.039	0.211**	0.465**	0.482**	1							
WHORLS	X_{10}	0.545**	0.569**	-0.081	-0.014	0.037	0.274**	0.679**	0.557**	0.526**	1						
CLEAR	X_{11}	0.084	0.076	0.162*	0.097	-0.091	-0.036	-0.113	0.061	0.085	-0.319**	1					
KNOTS	X_{12}	0.019	-0.036	0.220**	0.169*	-0.145*	0.024	-0.232**	-0.357**	-0.127	-0.368**	0.029	1				
DIAKNOT	X_{13}	0.134	0.144	0.126	0.015	-0.208**	-0.329	-0.424**	-0.202**	-0.076	0.291**	0.007	0.184*	1			
MAXCOMSTR	Y	-0.419**	-0.338**	-0.728**	-0.543**	0.247**	0.117	0.110	-0.253**	-0.235**	-0.101	-0.055	-0.117	-0.153*			

*Significance at 0.05 probability level.

**Significance at 0.01 probability level.

And $\sum (1/\lambda_i) = 86.128$ which from (1.5) indicates that if a linear predictor of $\mathbf{Y}$ is obtained from the 13 factors then $\hat{\beta}$ will probably be far from β.

In Figure 6 is shown a RIDGE TRACE for this problem. The effects of the

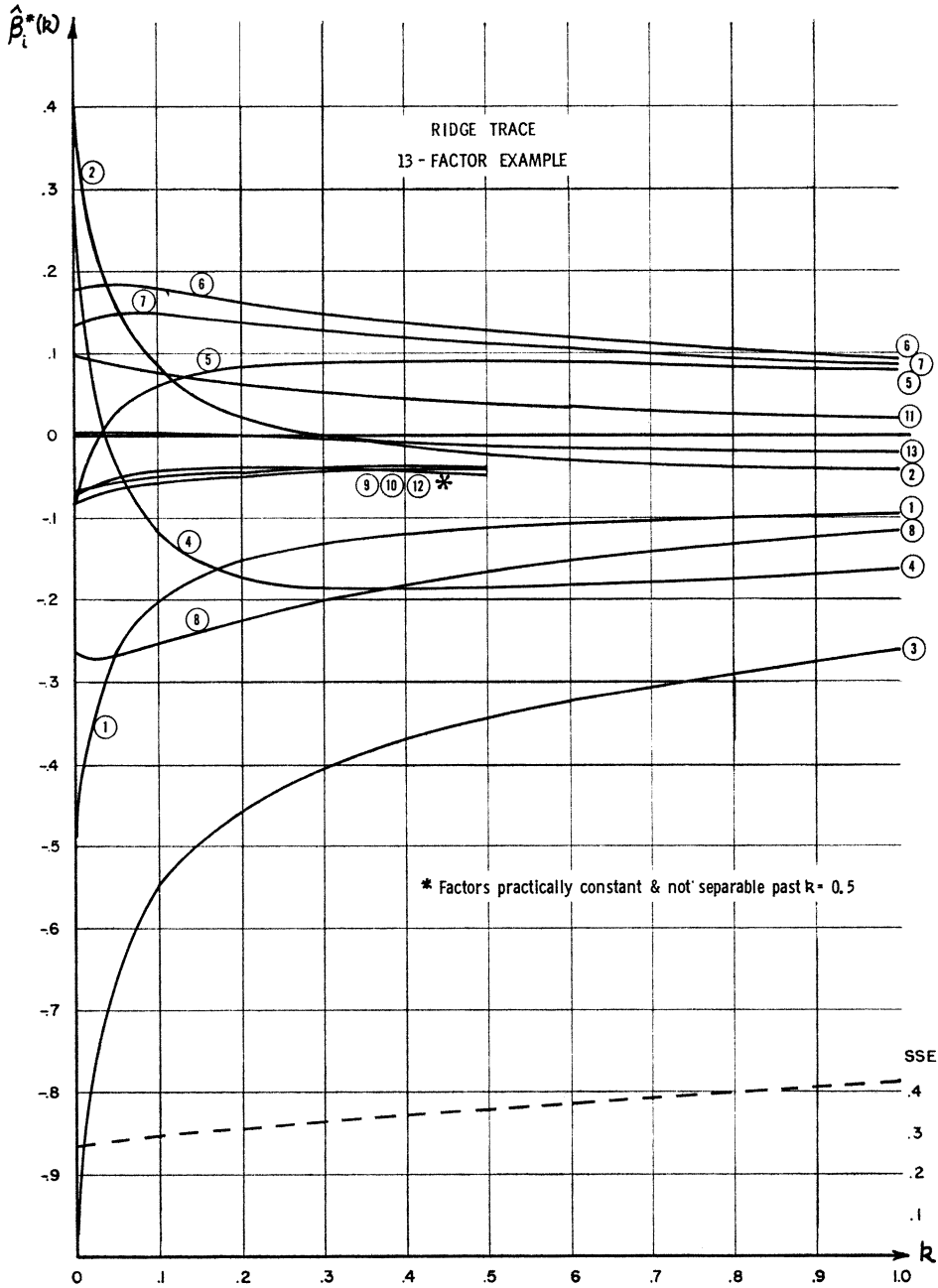

FIGURE 6

many correlations among the predicting factors (and, hence, the small eigenvalues) are readily apparent in the ridge trace of the coefficients.

(i) The coefficients tend to be unusually large. This is further demonstrated in Figure 7. The squared length of the coefficient vector drops from 1.6 at $k = 0$ to 0.5 at $k = 0.1$; at 0.1 it is only 31% of its ordinary least squares length. An orthogonal system would show a drop to only 83%.
(ii) Factors 4 and 5, Test SG and Oven SG, have incorrect signs at the least squares solution, $k = 0$. In fact, factor 4 would be assessed at this point as being the second largest positive contributor to compressive strength, whereas it is in reality the second largest detractor. With the addition of $k > 0$ to the diagonal of $\mathbf{X}'\mathbf{X}$, $\hat{\beta}_4^*$ decreases rapidly to zero, becomes negative, and then stabilizes at a value between -0.15 and -0.20.
(iii) At $k = 0$, factor 2 is indicated as being the largest positive contributor. However, it decreases quite quickly in value, and there is a question as to whether it predicts at all. Its sign is also open to question.
(iv) At about $k = 0.2$ the system stabilizes and an assessment of factors based on the coefficients in the interval (0.2, 0.4) should give a better prediction equation.

Jeffers uses this problem as an example to portray the use of principal components in regression analysis. It is not the intent here to critique principal component analysis. Jeffers does, by this method, get the coefficients down to reasonable sizes. The price is 63.97% of the variation in Y accounted for versus the 73.09% that it would be for the least squares solution using all 13 factors.

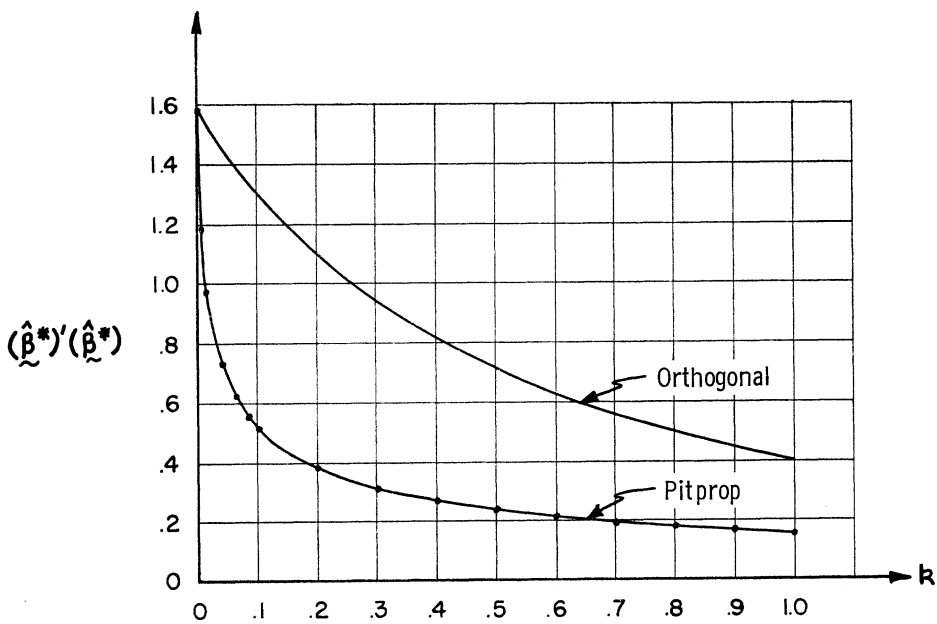

FIGURE 7—Pitprop Data. Squared Length of Coefficient Vector

The following observations can be made:

(i) The squared length of the regression vector from the principal component solution is about 0.4 compared to a value of 1.589 for the standard solution for 13 factors. If an analyst is willing to consider solutions with a squared length of the regression vector equal to 0.4, then in [3] it is shown that the ridge solution gives a minimum residual sum of squares. Referring to Figure 7, $(\hat{\beta}^*)'(\hat{\beta}^*) = 0.4$ gives a value of $k \simeq 0.2$. The percent variation accounted for at this point is 68.67 compared to the 63.97 for the principal components solution of Jeffers.

(ii) Another comparison that can be made is to fix the residual sum of squares (the percent variation accounted for). Again in [3] it is shown that with the residual sum of squares fixed, the Ridge solution gives a value of k that minimizes the squared length of the regression vector. For this example, this gives $k \simeq 0.5$.

However, the point to be made for the RIDGE TRACE is that it shows where the sensitivities of the system are. Using principal components, by comparison, is like flying blind. An examination of the trace makes it clear where the system stabilizes and where a good set of coefficients can be found. The stability occurs in the region $k = 0.2$ to $k = 0.4$, and a good stable point is at $k = 0.3$ which gives the solution:

$$\begin{array}{lll}
\hat{\beta}_1^* = -0.1347 & \hat{\beta}_6^* = 0.1496 & \hat{\beta}_{11}^* = 0.0547 \\
\hat{\beta}_2^* = 0.0000 & \hat{\beta}_7^* = 0.1312 & \hat{\beta}_{12}^* = -0.0419 \\
\hat{\beta}_3^* = -0.4065 & \hat{\beta}_8^* = -0.2034 & \hat{\beta}_{13}^* = -0.0062 \\
\hat{\beta}_4^* = -0.1876 & \hat{\beta}_9^* = -0.0430 & \\
\hat{\beta}_5^* = 0.0890 & \hat{\beta}_{10}^* = -0.0464 &
\end{array}$$

The same strategy can then be followed as outlined in the previous example.

If dimension reduction and factor screening are an aim, as it was implied they were in this case, then the RIDGE TRACE gives a number of clues on how to proceed. The following procedure would seem to be reasonable:

1–Examine the stable coefficients and eliminate the factors with the least predicting power.

In this case the TRACE clearly indicates 9, 10, 12, and 13 and 11 if one goes up a level. Hence, let's say that 9, 10, 11, 12, and 13 will be deleted.

2–Examine the unstable coefficients and eliminate those factors that cannot hold their predicting power.

Here the unstable ones are 1, 2, 3, 4, and 5. Of these the coefficient for factor 2 is quickly driven to zero. An examination of the simple correlations shows that 2 is correlated with all the others so there is a possibility of stabilizing the system by deleting it. (This occurred in the 10-factor example.) Thus, in addition to the factors above, also delete factor 2.

In Figure 8 is the RIDGE TRACE with factors 2, 9, 10, 11, 12, and 13 deleted.

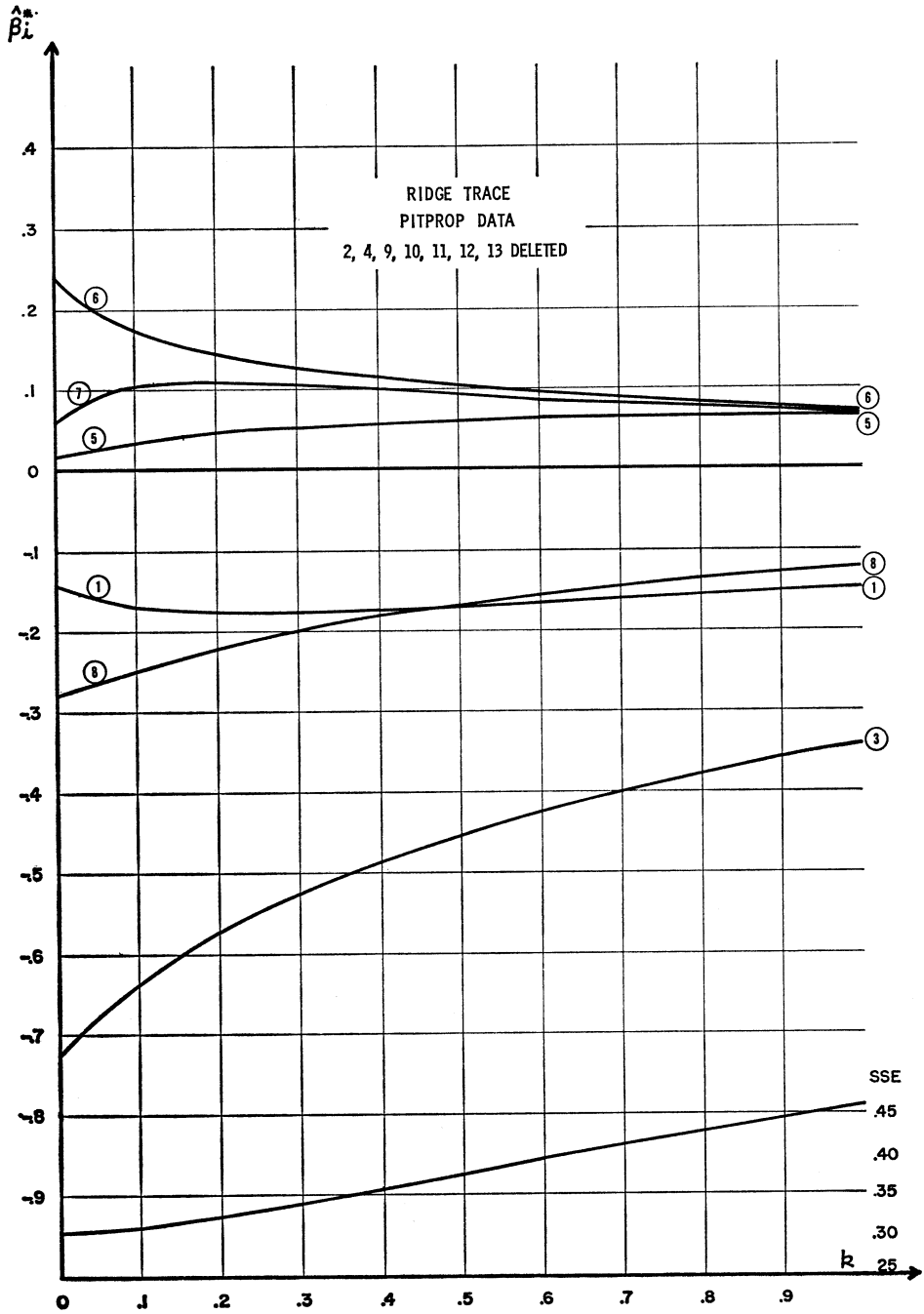

FIGURE 8

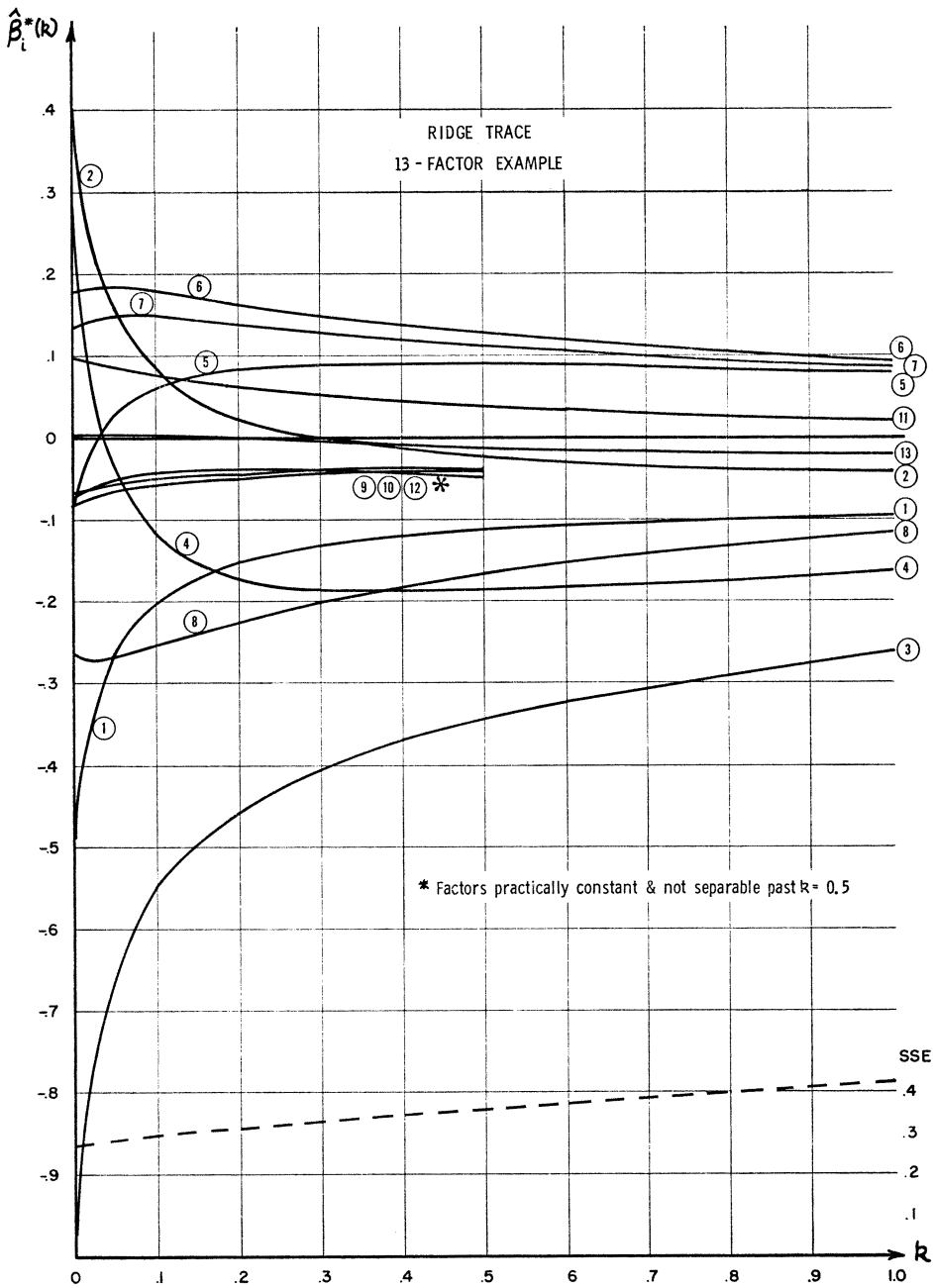

FIGURE 9

From this trace it is seen that the strength of the bonds of factor 2 with the other unstable factors was not such as to be able to stabilize the system completely. Factor 1 is no longer grossly overestimated and has stabilized. Factor 5 is still unstable. The tight bond between factors 3 and 4 has not been broken. Factor 4 is grossly overestimated with the wrong sign.

3–Delete one or more of the remaining unstable coefficients. Here it is factor 4 that shows the most instability.

In Figure 9 is the RIDGE TRACE with factors 2, 4, 9, 10, 11, 12, and 13 deleted. The gross instabilities have now been eliminated and there has been no significant inflation in the residual sum of squares. The percent variation accounted for is 69.55% versus that of 73.09% for all factors (and the 63.97% of Jeffers' principal component solution). There is some apparent overestimation still and the residual sum of squares is flat in the neighborhood of $k = 0$. Taking the values at k in (.1, .2) does not seem unreasonable.

4. Summary

In multiple linear regression the effect of nonorthogonality of the prediction vectors is to pull the least squares estimates of the regression coefficients away from the true coefficients that one is trying to estimate. The coefficients can be both too large in absolute value and incorrect with respect to sign. Furthermore, the least squares solution is unstable. A slight movement away from this point can give completely different estimates of the coefficients. Commonly used screening and dimension reduction procedures have two faults. First, they are done in the dark and do not give a picture of how the nonorthogonalities are causing the instabilities, overestimations, and wrong signs. And secondly, they can actually amplify the deficiencies of ordinary least squares for nonorthogonal data. As demonstrated in the examples, the RIDGE TRACE is a diagnostic tool that gives a readily interpreted picture of the effects of nonorthogonality and it can guide one to a better point estimate.

References

[1] Gorman, J. W. and Toman, R. J. (1966). Selection of variables for fitting equations to data *Technometrics 8*, 27–51.

[2] Hocking, R. R. and Leslie, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics 9*, 531–540.

[3] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression. Biased estimation for nonorthogonal problems. *Technometrics 12*.

[4] Jeffers, J. N. R. (1967). Two case studies in the application of principal component analysis. *Applied Statistics 3*, 225–236.