KM 12(1983)
pag 32-41

SOME NOTES ON RIDGE REGRESSION

John P. Van de Geer*

Summary

Solutions for ridge regression are compared to other solutions for relating a criterion variable to a set of predictor variables. The unifying frame is provided by a generalization of the determinantal equations for eigenvalues. This approach makes it possible to relate ridge regression to solutions based on eigenvalues both of the correlation matrix of the predictors and of the combined matrix of criterion and predictors. Some other possible solutions are briefly indicated.

The paper concludes with an illustration of the vector geometry of ridge regression.

KEY WORDS: Geometry of regression, Multiple correlation, Regression, Ridge regression.

1. Ridge regression

Ridge regression has been introduced by Hoerl and Kennard [1970a,b] as an alternative to classical multiple regression in cases where there is near-collinearity in the set of predictor variables. For a detailed discussion one could also consult Marquardt [1970] or Winer [1978]; Swindle [1981] shows some of the geometrical aspects.

Let y be an n x 1 vector of observed values on a criterion variable, and let X be the n x m matrix of observations on m predictor variables. It is assumed here that y and X are in deviations from column means, and also that y and X are scaled in such a way that columns have unit norm. The latter assumption is for convenience only. It ensures that X'X is a matrix of correlations, and that y'X is a row vector of correlations between criterion and predictors.

Classical multiple regression solves for regression weights

$b=(X'X)^{-1}X'y$

with the effect that the sum of squares of the difference vector y-Xb is

* Department of Data Theory, University of Leiden, Middelstegracht 4, 2312 TW Leiden . Tel. 071 - 148333 -tst 2254

minimized. However, results for b can be extremely unstable if there is near collinearity among the columns of X. Small changes in X then may produce dramatic changes in the values of the elements of b. Non-stability also will be revealed by the fact that the value of b'b (the sum of the squared weights) is suspiciously large.

Ridge regression replaces the solution for weights by

$\bar{b}=(X'X+\gamma I)^{-1}X'y$

where γ is a suitably chosen positive constant. The effect will be that $\bar{b}'\bar{b}<b'b$. But the effect is also that $\bar{b}$ no longer is an unbiased estimate. Ridge regression trades off some bias against a gain in stability.

2. Singular value decomposition

A re-formulation is obtained as follows. Let

$X=P\Phi Q'$

be the singular value decomposition of X. I.e., assuming that X has rank k, this decomposition requires that P is an n x k matrix with $P'P=I$, that Q is an m x k matrix with $Q'Q=I$, and that Φ is a diagonal k x k matrix with positive diagonal elements in descending order.

Mandel [1982] gives a detailed discussion of the relation between the singular value decomposition and multiple correlation. For the present we need only the following results. The classical solution Xb for multiple regression now can be written as $Xb=P\Phi t$, with $b=Qt$. This implies that b and t have the same norm: $b'b=t'Q'Qt=t't$. The solution for t becomes

$t=\Phi^{-1}P'y$

This equation explains why t (and therefore b) becomes unstable when there are elements in Φ close to zero. Corresponding elements in Φ^{-1} then become disproportionately large.

Table 1 gives a small example, with only two predictors x_1 and x_2, highly correlated. The solution for regression weights is $b_1=2.374$ and $b_2=-1.869$, with sum of squares 9.127. The solution for t becomes: $t_1=(1.96)^{-1}(.70)=.357$ and $t_2=(.04)^{-1}(.12)=3.000$, again with sum of squares 9.127. Clearly, t_2 becomes so large because the second singular value is so small.

Ridge regression replaces the solution for t by

$\bar{t}=(\Phi^2+\gamma I)^{-1}\Phi P'y$

with γ a positive constant. The effect of near-zero elements in Φ now will be damped, so to speak. In the example above, taking $\gamma=.088$, we obtain $\bar{t}_1=(2.048)^{-1}(.70)=.342$ and $\bar{t}_2=(.128)^{-1}(.12)=.940$, with sum of squares equal to 1.000. This shows that even a rather small value of γ may produce

Table 1

Correlations between y and X (left part) and sums of squares and cross products for y and PΦ (right part).

y	1	.5798	.4101	y	1	.70	.12
x_1	.5798	1	.96	$p_1\varphi_1$	.70	1.96	0
x_2	.4101	.96	1	$p_2\varphi_2$	.12	0	.04

a substantial decrease of the sum of squares of the weights. But the price one has to pay is a decrease of the correlation: whereas the squared multiple correlation is .610, the squared correlation between y and $P\Phi\bar{t}$ is .469.

3. Determinantal equations

Solutions for multiple regression and ridge regression can be interpreted as special cases of a more general problem. This general problem is defined by the stationary equations

$$\begin{pmatrix} 1 & y'P\Phi \\ \Phi P'y & \Phi^2 \end{pmatrix} \begin{pmatrix} c \\ t \end{pmatrix} = \begin{pmatrix} c\mu_1 \\ t\mu_2 \end{pmatrix}$$

Special cases are the following.

(i) Multiple regression. The solution for multiple regression is obtained by taking $\mu_2=0$. The second stationary equation then implies

$\Phi P'yc=-\Phi^2 t$

or

$-t/c=\Phi^{-1}P'y$

so that $-t/c$ becomes the solution for regression weights for the regression of y on PΦ. Moreover, a little algebra shows that in this case $\mu_1=1-R^2_{y.X}$, where $R_{y.X}$ is the multiple correlation coefficient.

(ii) Ridge regression. Ridge regression corresponds to a solution where μ_2 has negative value. The second stationary equation then implies

$-t/c=(\Phi^2-\mu_2 I)^{-1}\Phi P'y$

which agrees with the solution given in section 2, with $\mu_2=-\gamma$.

(iii) Eigenvalues. The solution for eigenvalues and eigenvectors of the combined matrix of correlations between y and X is obtained by setting $\mu_1=\mu_2$, so that the stationary equations become classical equations for eigenvalues.

The stationary equations above imply the determinantal equation

$$g(\mu_1,\mu_2) = \begin{vmatrix} 1-\mu_1 & y'P\Phi \\ \Phi P'y & \Phi^2-\mu_2 I \end{vmatrix} = 0$$

It follows that one can make a graph of $g(\mu_1,\mu_2)$ in which feasible solutions for (μ_1,μ_2) are plotted. For the numerical example of Table 1 this graph is shown in Figure 1. Table 2 identifies a number of special points of the graph.

4. Properties of the graph

The graph in Figure 1 has a number of interesting properties some of which are listed below.

(i) Typically, the graph shows a curve consisting of (k+1) "branches" (k is the column rank of X). These branches have asymptotes at $\mu_1=y'y=1$, and at $\mu_2=\varphi_i^2$ $(i=1,..,k)$. For the numerical example there is a horizontal asymptote at $\mu_1=1$, and there are two vertical asymptotes at $\mu_2=1.96$ and $\mu_2=.04$.

(ii) For any point on the curve the tangential to the curve has slope $-(t't)/c^2$.

(iii) Eigenvalues of the combined correlation matrix for y and X are shown at the points where $\mu_1=\mu_2$; i.e., where the curve is intersected by the line $\mu_1=\mu_2$.

(iv) For any set of (k+1) solutions located on a straight line through the origin, corresponding solutions for the vectors $yc+P\Phi t$ are orthogonal to each other. The eigenvector solutions are an example.

(v) For any set of (k+1) solutions located on a line parallel to the line $\mu_1=\mu_2$, corresponding solutions for the vectors of weights $\begin{pmatrix} c \\ t \end{pmatrix}$ are orthogonal.

(vi) For any set of k solutions with identical value of μ_1, corresponding solutions for t are orthogonal to each other.

(vii) The classical solution for multiple regression is given by the point of the graph where $\mu_2=0$ and $\mu_1=1-R^2_{y.X}$. Figure 1 shows that the tangential at this point is rather steep, for this example. This indicates that c^2 is relatively small compared to t't, and therefore that the sum of the squared regression weights, equal to $t't/c^2$, is large. The figure makes it clear why this is likely to happen when there is a smallest φ_k^2 close to zero.

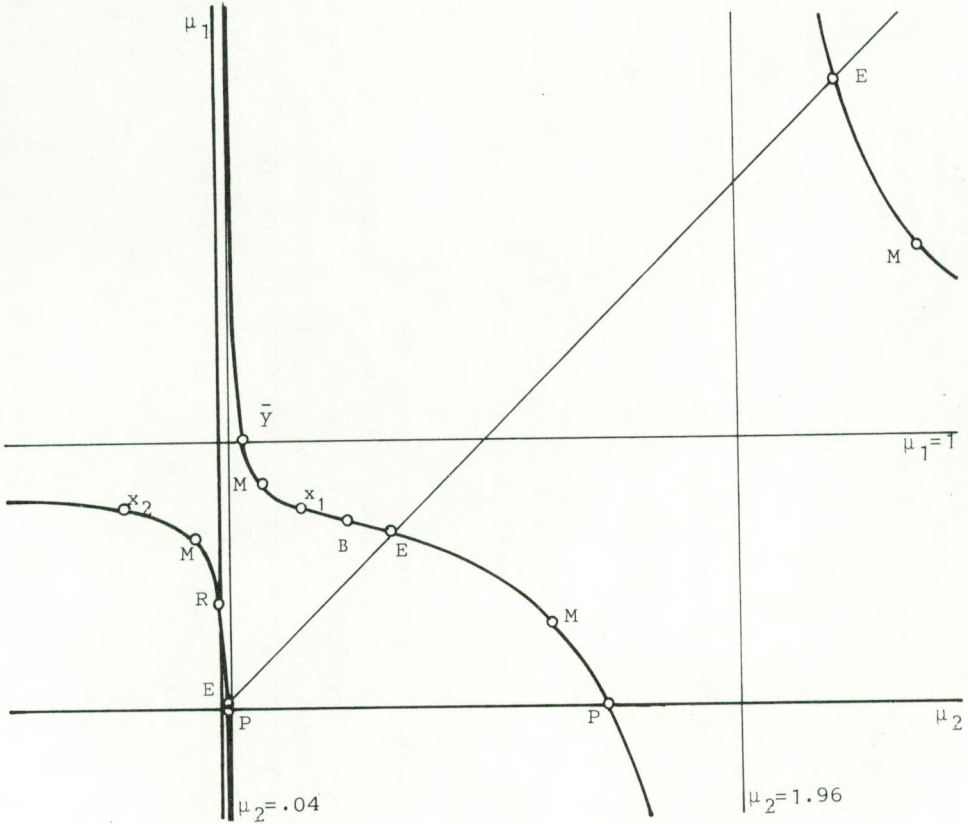

Figure 1.

Graph of determinantal equation.

Inhoud **Kwantitatieve Methoden** 12 (1983)

3	O.J.W.F. Kardaun	Over betrouwbaarheidsbanden en aanpassingstoetsen bij gecensureerde waarnemingen
20	A.J.M. Hagenaars & B.M.S. van Praag	Meetfout gemeten
32	J.P. van de Geer	Some notes on ridge regression
42	V. Fidler & A. de Jonge	Testing for independence of binary responses
49	C.B. Tilanus	Failures and successes of quantitative methods in management
65	M. Alderliesten & R. van Splunter	Cudif charts for detecting systematic differences between duplicate determinations
79	A. Nieuwenhuis	Twee bekende stelsels vraagvergelijkingen en hun mengvorm
107	D.A. Kodde	Comparatieve statica en Kuhn-Tucker voorwaarden
120	R.J. Noordman	KWANTITATIEVE METHODEN BIJ ... Hoogovens
124		VRAGEN
126-133	*Reacties op van Leeuwen & Mijlpaal KM 11(1983) p 5-11*	
126	W.K. Klein Haneveld	Ingezonden brief
127	P. van Beek	Enkele kritische kanttekeningen
129	T. Schandpaal & B. de Stier	Wiskunde en Operations Research
133	K. van Leeuwen & B.S. Mijlpaal	Naschrift
134		Inhoud van jaargang 1983: KM 9 t/m KM 12
136		Mededeling van de redactie
137-139		KM X LEZERS INTERACTIE

K W A N T I T A T I E V E M E T H O D E N

NUMMER 12 DECEMBER 1983 JAARGANG 4

TABLE 2

Solutions for (μ_1,μ_2) with corresponding solutions for c, t, and -t/c, based on the numerical example of Table 1. The normalization $c^2+t't=2$ is chosen. What makes the solution special is underlined.

	μ_1	μ_2	c	t_1	t_2	$-t_1/c$	$-t_2/c$
x_2	.752	-.357	1.300	-.393	-.393	.302	.302
M (slope -1)	.648	-.088	1.000	-.342	-.940	.342	.940
R (mult.regression)	.390	0	.444	-.159	-1.333	.357	3.000
E (eigenvalue)	.020	.020	.230	-.083	-1.393	.361	6.060
P (partials of X)	0	.021	.224	-.081	-1.394	.361	6.227
$\bar{y}$	1	.095	.581	-.218	1.271	.375	-2.190
M (slope -1)	.837	.170	1.000	-.391	.920	.391	-.920
x_1	.752	.321	1.211	-.517	.517	.427	-.427
B (bending point)	.698	.493	1.241	-.592	.329	.477	-.265
E (eigenvalue)	.650	.650	1.229	-.657	.242	.534	-.197
M (slope -1)	.315	1.257	1.000	-.995	.099	.995	-.099
P (partials of X)	0	1.475	.805	-1.161	.067	1.443	-.084
E (eigenvalue)	2.330	2.330	.661	1.250	.035	-1.891	-.052
M (slope -1)	1.705	2.661	1.000	.999	.046	-.999	-.046

(viii) Solutions where $\mu_1=0$ define t as an eigenvector of the matrix $\Phi(I-P'yy'P)\Phi$. This is the matrix of sums of squares and cross-products of $P\Phi$ with y partialled out. Such solutions for t are orthogonal (property vi). Moreover, the solution with smallest value of μ_2 then corresponds to an estimate for linear relation between y and Xt if it is assumed that there is measurement error in X but not in y. This "model" in fact says that the sum of squares of $yc+P\Phi t$ must be minimized under the restriction that this sumvector is uncorrelated with y. Compare with the classical solution for multiple regression, where it is assumed that y has measurement error whereas X has not: we then minimize the sum of squares of $yc+P\Phi t$ under the restriction that this vector is uncorrelated with X.

(ix) Solutions with negative value of μ_2 are ridge regression solutions. Obviously, the tangential slope becomes flatter to the extent μ_2 becomes more negative, thereby decreasing the value of the sum of squares of the weights -t/c.

(x) In Figure 1 there is one particular ridge regression solution where the tangential has slope -1, which implies $t't=c^2=1$. It has the simple interpretation that the sum of squares of the difference vector y-Xb is minimized under the restriction that b'b=1. It follows that Xb can be interpreted in terms of projections of the row points of X on a vector with direction cosines b. Such a solution also is "fair" in the sense that it gives equal shares to y and X. Figure 1 shows other such "fair" solutions. E.g., the solution with largest eigenvalue is not fair: it depends mainly on the internal structure of X (X'X has large eigenvalue) and it more or less ignores y. The solution therefore "explains" much of the variance of X, but little of the variance of y. The solution where the tangential slope equals -1 corrects this bias.

(xi) The solution for smallest eigenvalue of the combined correlation matrix of y and X is on the left lower branch of Figure 1. It gives the best estimate of a linear relation between y and Xt under the assumption that both y and X are subject to measurement error of the same order of magnitude. This solution is probably the oldest example of application of eigenvector theory to data analysis (Pearson [1901]).The example of Figure 1 also shows that at this particular eigenvalue solution the tangential has steep slope, indicating that the solution depends more on the internal structure of X than on the relation between X and y.

(xii) Figure 1 and Table 2 also identify a solution on the left lower branch labeled x_2. It is a ridge regression solution where yc+PΦt is uncorrelated with x_2. Such a solution would be appropriate if it could be assumed that y and x_1 are subject to measurement error whereas x_2 is not. Its counterpart is the solution labeled x_1 on the middle branch, where yc+PΦt is uncorrelated with x_1.

(xiii) The solution labeled $\bar{y}$, and where the horizontal asymptote intersects the middle branch, has the characteristic that PΦt is uncorrelated with y.

(xiv) Finally, Table 2 also identifies the bending point of the middle branch. This point does not seem to have any particular meaning for data analysis. It explains, however, why it can happen that there are more than (k+1) solutions where the tangential slope is equal to -1 (labeled M in Figure 1 and Table 2).

5. Geometry of ridge regression

Figure 2 shows the geometry of ridge regression. First of all, this figure shows the two vectors $p_1\varphi_1$ and $p_2\varphi_2$ as principal axes of an ellipse. All vectors PΦt with t't=1 (and therefore all vectors Xb with b'b=1) will be located on this ellipse. Figure 2 also shows the projection of vector y on the plane spanned by X. Clearly, to obtain a weighted sum PΦt which coincides with the projection of y (this solution for PΦt would be the solution for classical multiple regression) requires that this vector PΦt is far outside the ellipse. In other word, this solution requires that t't is much larger than 1 (in fact, section 2 showed that t't=9.127, so that an ellipse passing through the projection of y should be a factor $\sqrt{9.127}$ larger than the ellipse shown in Figure 2.)

Suppose now that we require a ridge regression solution with t't=1. I.e., the solution for PΦt should be located on the ellipse of Figure 2. At the same time we want PΦt to be as close to the projection of y as possible. This implies that PΦt should be selected at the point where y has smallest distance to the ellipse; in other words, where a circle around the point y is tangential to the ellipse. Figure 2 illustrates this for the four solutions labeled M (also in Figure 1 and Table 2). The ridge regression solution is, of course, the solution M which is closest to y.

Figure 2 also explains why there is a bending point in the middle branch of Figure 1. If we would require a solution for PΦt with t't much smaller than 1, the corresponding ellipse becomes also a very small one. Circles

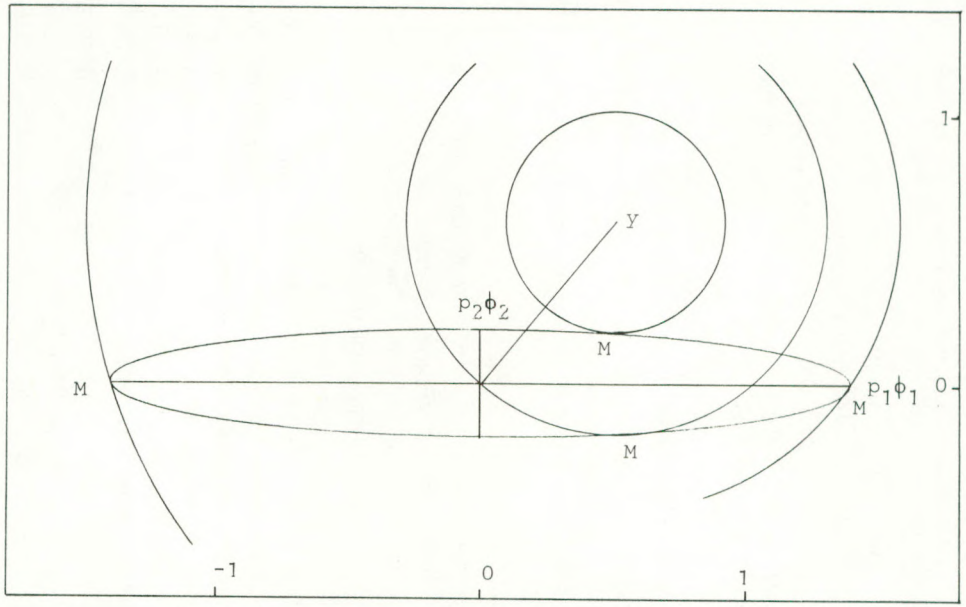

Figure 2. Geometry of ridge regression for Pφt with t't=1.

around y then are tangential to this small ellipse at only two points (instead of four). The bending point in Figure 1 indicates where the transition from four to two such solutions occurs.

6. Conclusions

Purpose of this paper was to relate ridge regression to a number of other least squares solutions which express relations between y and X. Such solutions are illustrated in the graph of the determinantal equation (Figure 1). Moreover, solutions on the left-lower branch are all of the regression type, with relatively small sum of squares for the vector yc+PΦt. Solutions on the right-upper branch, on the other hand, are characterized by the fact that the vector yc+PΦt explains much of the variance of y and X. Solutions on the intermediate branch are in-between these two "ideals".

A similar approach can be developed for the situation where there are more than one criterion variables. The classical solution then is the canonical solution. But it then becomes rather easy to develop "canonical ridge regression" solutions and other solutions for relating two sets of variables. This is further discussed in Van de Geer [1984, in press].

References

Hoerl, A.E., and Kennard, R.W. [1970a]. Ridge regression. Biased Estimation for Nonorthogonal Problems. Technometrics, 12, 55-67.

Hoerl, A.E., and Kennard, R.W. [1970b]. Ridge regression: Applications to nonorthogonal Problems. Technometrics, 12, 69-82.

Mandel, J. [1982]. Use of the Singular Value Decomposition in Regression Analysis. The American Statistician, 36, 15-24.

Marquardt, D.W. [1970]. Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. Technometrics, 12, 591-612.

Pearson, K. [1901]. On Lines and Planes of closest Fit to Points in Space. Phil. Magazine, 2, 559-572.

Swindle, B.F. [1981]. Geometry of Ridge Regression illustrated. The American Statistician, 35, 12-15.

Van de Geer, J.P. [1984, in press]. Relations among k sets of Variables. Psychometrika.

Ontvangen: 2-5-1983