Some Cautionary Notes on the Use of Principal Components Regression

Ali S. Hadi; Robert F. Ling

The American Statistician, Vol. 52, No. 1 (Feb., 1998), 15-19.

Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28199802%2952%3A1%3C15%3ASCNOTU%3E2.0.CO%3B2-K

The American Statistician is currently published by American Statistical Association.

Some Cautionary Notes on the Use of Principal Components Regression

Ali S. HADI and Robert F. LING

Many textbooks on regression analysis include the methodology of principal components regression (PCR) as a way of treating multicollinearity problems. Although we have not encountered any strong justification of the methodology, we have encountered, through carrying out the methodology in well-known data sets with severe multicollinearity, serious actual and potential pitfalls in the methodology. We address these pitfalls as cautionary notes, numerical examples that use well-known data sets. We also illustrate by theory and example that it is *possible* for the PCR to fail miserably in the sense that when the response variable is regressed on all of the p principal components (PCs), the first $(p-1)$ PCs contribute nothing toward the reduction of the residual sum of squares, yet the last PC alone (the one that is always discarded according to PCR methodology) contributes everything. We then give conditions under which the PCR totally fails in the above sense.

KEY WORDS: Hald's data; Longley data; Multicollinearity.

1. INTRODUCTION

Many textbooks on regression analysis include the methodology of principal components regression (PCR) as an alternative to ordinary least squares when computational or statistical problems arise in the presence of severe multicollinearity in the set of independent variables $\mathbf{X} = \mathbf{X}_1, \ldots, \mathbf{X}_p$. By keeping the first few principal components (PCs) of $\mathbf{X}$, the regression of the response variable $\mathbf{Y}$ on these PCs (which are orthogonal) will remove any computational problem that may arise as a result of multicollineary/ill-conditioning while keeping all of the original variables.

In our opinion, in such cases of severe multicollinearity, it is simply a sign of some statistical redundancy in $\mathbf{X}$, and the removal of one or more of the *superfluous* variables in the $\mathbf{X}$-space will have removed both the numerical and statistical difficulties in the problem. Moreover, as was pointed out by Beaton, Rubin, and Barone (1976), the insistence on a high degree of numerical accuracy in the estimated regression coefficients (when all original variables are kept, even the statistically redundant ones) may not lead to a set of *statistically correct* estimates anyway.

We have not encountered in the statistical literature any strong justification for using the PCR for multicollinearity problems. We are not here to discuss or argue *how* such multicollinearity problem should be handled or addressed. Instead, we wish to draw attention to some very serious potential pitfalls in the application of the PCR methodology to warn against such pitfalls by numerical examples.

As a methodological and theoretical characterization of the PCR defect we observed in several real-life data sets, we begin by presenting an extreme-case scenario to illustrate by theory and example that it is *possible* for the PCR to fail miserably in the sense that when the response variable is regressed on all of the p PCs, the first $(p-1)$ PCs contribute nothing toward the reduction of the sum of squares, yet the last PC alone (the one that is always discarded according to PCR methodology) contributes everything. We give conditions under which the PCR totally fails in that sense.

We then use two real-life data sets to illustrate the cautionary notes presented in this article. The first data set is the Longley (1967) data set, well-known for its multicollinearity. The second is the Hald's data set in (Draper and Smith 1981) which drew our attention to the pitfalls in the first place, though such pitfalls seemed to have been overlooked by those authors. Otherwise, we felt they might have included some cautionary notes as they had done on some other methods alternative to the use of ordinary least squares (OLS). We show that the same type of defects that appeared in our analysis of the Longley data by PCR occur in the analysis of the Hald's data by PCR also.

Section 2 summarizes the rationalization commonly given for using PCR. Section ?? gives the main cautionary remark and illustrates it both by an example and by theoretical arguments. Section ?? gives other cautionary remarks and illustrates them by examples. Section ?? discusses the relationships bewteen our cautionary notes and existing related literature. Section 6 gives concluding remarks.

2. RATIONALIZATION FOR USING PCR

The usual motivation and rationalization given for using PCR is as follows. Let $\mathbf{Y}$ denote the response variable and $\mathbf{X}$ denote the design matrix or the matrix containing p explanatory variables. Suppose that the columns of $\mathbf{X}$ are highly multicollinear, but the researcher wants to keep all the variables in $\mathbf{X}$. Based on this premise, the PCR-advocates would then:

1. Compute the standardized version of $\mathbf{X}$ and denote it by $\mathbf{Z}$.

Ali S. Hadi is Professor, Department of Social Statistics, Cornell University, 358 Ives Hall, Ithaca, NY 14853-3901 (E-mail: ali-hadi@cornell.edu). Robert F. Ling is Professor, Department of Mathematical Sciences, College of Engineering and Science, Martin Hall, Box 3431907, Clemson University, Clemson, SC 29634-1907 (E-mail: rfling@clemson.edu).

Table 1. Hald's Data, the Corresponding PCs $\mathbf{W}_1, \ldots, \mathbf{W}_4$, and a Constructed Variable $\mathbf{U}$

X_1	X_2	X_3	X_4	Y	W_1	W_2	W_3	W_4	U
7	26	6	60	78.5	1.467	1.903	−.530	.039	.077
1	29	15	52	74.3	2.136	.238	−.290	−.030	−.060
11	56	8	20	104.3	−1.130	.184	−.010	−.094	−.187
11	31	8	47	87.6	.660	1.577	.179	−.033	−.066
7	52	6	33	95.9	−.359	.484	−.740	.019	.038
11	55	9	22	109.2	−.967	.170	.086	−.012	−.024
3	71	17	6	102.7	−.931	−2.135	−.173	.008	.017
1	31	22	44	72.5	2.232	−.692	.460	.023	.045
2	54	18	22	93.1	.352	−1.432	−.032	−.045	−.090
21	47	4	26	115.9	−1.663	1.828	.851	.020	.040
1	40	23	34	83.8	1.641	−1.295	.494	.031	.063
11	66	9	12	113.3	−1.693	−.392	−.020	.037	.074
10	68	8	12	109.4	−1.746	−.438	−.275	.037	.074

2. Compute the principal components: Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the eigenvalues of $\mathbf{Z}^T\mathbf{Z}$ (or the correlation matrix $\mathbf{R}$) and $\mathbf{V}$ be the corresponding eigenvectors. Let $\mathbf{W} = \mathbf{ZV}$. The columns in $\mathbf{W}$ are the PCs of $\mathbf{Z}$. The jth column of $\mathbf{W}$ is called the jth PC, $j = 1, \ldots, p$.

3. Regress $\mathbf{Y}$ on the first m PCs, $\mathbf{W_1}, \ldots, \mathbf{W_m}$, where $m \leq p$.

For a more detailed description of the methodology, for example, see Draper and Smith (1981), and Chatterjee and Price (1991).

Advocates of PCR give the following reasons for the use of the methodology:

1. Because the PCs, $\mathbf{W}_1, \ldots, \mathbf{W}_m$, are orthogonal, the problem of multicollinearity disappears completely, and no matter how many PCs are actually used, the regression equation will always contain all of the variables in $\mathbf{X}$ (because each PC is a linear combination of the variables in $\mathbf{X}$ formed by an eigenvector of $\mathbf{Z}^T\mathbf{Z}$).

2. PCR presumably improves the numerical accuracy of the regression estimates because of the use of orthogonal PCs.

3. THE MAIN CAUTIONARY NOTE

3.1 Cautionary Note 1: The First m Principal Components can Totally Fail in Accounting for the Variation in the Response Variable

To illustrate this cautionary note we use the Hald's data set, which is taken from Draper and Smith (1981, p. 630) who used it to illustrate the PCR methodology (pp. 327–331). Hald's data consist of one response and four explanatory variables. The data set appears in the first five columns of Table **??**. The next four columns, $\mathbf{W}_1, \ldots, \mathbf{W}_4$, in Table 1 are the four PCs of $\mathbf{X}$. (Note that the PCs at the bottom of page 330 in Draper and Smith (1981) should be the same as the ones given here.) The last column $\mathbf{U}$ is a constructed response variable that we use to illustrate this cautionary note.

Table 2. Hald's Data: Principal Components Decomposition

PC	Eigenvalues	% of Total	Cumulative %
$\mathbf{W}_1$	2.2357	55.893	55.893
$\mathbf{W}_2$	1.5761	39.402	95.294
$\mathbf{W}_3$	.18661	4.6652	99.959
$\mathbf{W}_4$	.0016237	.040594	100

The principal components decomposition (PCD) of Hald's data are given in Table 2. Accordingly, the number of PCs to keep is 2 or 3 (they account for 95.29% and 99.96% of the variation in $\mathbf{X}$, respectively). The sum of squares (SS) decomposition from the regression of the response variable $\mathbf{U}$ on the four PCs is given in Table 3. Note that $\mathbf{W}_1 - \mathbf{W}_3$, which constitute 99.96% of the variance in $\mathbf{X}$, contribute nothing to the fit while $\mathbf{W}_4$ alone contributes everything. This can also be seen from the scatterplots of $\mathbf{U}$ versus each of the PCs (see Fig. **??**). The scatter of points in the graphs of $\mathbf{U}$ versus each of the first three PCs are com-

Table 3. Hald's Data: Analysis of Variance Table for PCR Using $\mathbf{U}$ as the Response Variable

Source	SS	DF	MS	F value	P value
$\mathbf{W}_1$	.00000	1	.00000	.	1.0
$\mathbf{W}_2$	.00000	1	.00000	.	1.0
$\mathbf{W}_3$	.00000	1	.00000	.	1.0
$\mathbf{W}_4$	.07794	1	.07794	∞	.0
Residual	.00000	8	.00000	.	.

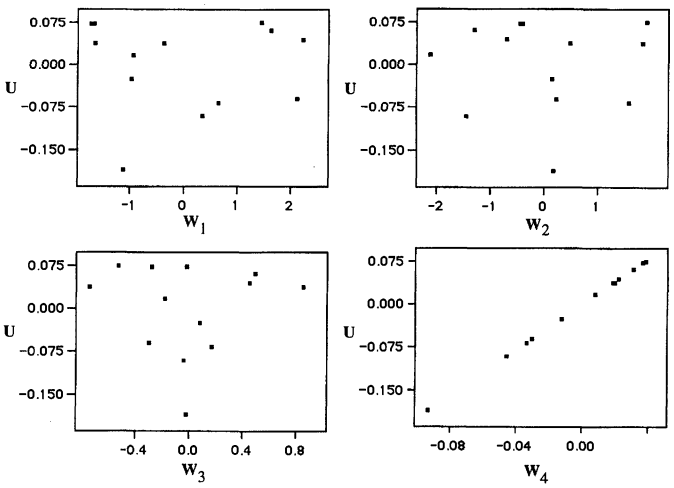

Figure 1. Hald's Data: Scatterplots of $\mathbf{U}$ Versus Each of the PCs $\mathbf{W}_1$–$\mathbf{W}_6$.

Table 4. Longley (1967) Data: Analysis of Variance Table for PCR

Source	SS	DF	MS	F value	P value
$\mathbf{W}_1$	169144914	1	169144914	1820.01	1.06603E−11
$\mathbf{W}_2$	2706733	1	2706733	29.1247	4.34863E−4
$\mathbf{W}_3$	10560944	1	10560944	113.637	2.09675E−6
$\mathbf{W}_4$	28586.8	1	28586.8	.307597	592672
$\mathbf{W}_5$	1457234	1	1457234	15.68	0033053
$\mathbf{W}_6$	273990	1	273990	2.94816	120105
Residual	836424	9	92936		

pletely random, whereas the relationship between $\mathbf{U}$ and the last PC $\mathbf{W}_4$ is perfectly linear.

This example raises a natural question: What are the conditions under which this phenomenon occurs? The answer is provided by the following theorem.

Theorem 1. For the usual regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, let $\mathbf{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_p)$ be the PCs of $\mathbf{X}$. If the true vector of regression coefficients β is in the direction of the j*th* eigenvector of $\mathbf{Z}^T\mathbf{Z}$, then when $\mathbf{Y}$ is regressed on $\mathbf{W}$, the jth PC $\mathbf{W}_j$ alone will contribute everything to the fit while the remaining PCs will contribute nothing.

Proof. Let $\mathbf{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_p)$ be the matrix containing the eigenvectors of $\mathbf{Z}^T\mathbf{Z}$. Then, we have

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \varepsilon \\ &= \mathbf{X}\mathbf{V}\mathbf{V}^T\beta + \varepsilon \quad (\text{because} \quad \mathbf{V}\mathbf{V}^T = \mathbf{I}) \\ &= \mathbf{W}\theta + \varepsilon, \end{aligned} \tag{1}$$

where $\theta = \mathbf{V}^T\beta$ is the regression coefficients of $\mathbf{W} = \mathbf{Z}\mathbf{V}$. If β is in the direction of the jth eigenvector $\mathbf{V}_j$, then $\mathbf{V}_j = \alpha\beta$, where α is a nonzero scalar. Consequently, $\theta_j = \mathbf{V}_j^T\beta = \alpha\beta^T\beta$ and $\theta_k = \mathbf{V}_k^T\beta = 0$, whenever $k \neq j$. Therefore, the regression coefficient θ_k corresponding to $\mathbf{W}_k$ is equal to zero, for $k \neq j$, hence (**??**) can be written as

$$\mathbf{Y} = \theta_j\mathbf{W}_j + \varepsilon. \tag{2}$$

Because, a variable $\mathbf{W}_k$ does not produce any reduction in the sum of squares (SS) iff its regression coefficient is zero, then $\mathbf{W}_j$ alone will contribute everything to the fit while the remaining PCs will contribute nothing. This completes the proof.

Two implications of Theorem 1 are:

Table 5. Longley (1967) Data: Principal Components Decomposition

PC	Eigenvalues	% of Total	Cumulative %
$\mathbf{W}_1$	4.60338	76.723	76.723
$\mathbf{W}_2$	1.17534	19.589	96.312
$\mathbf{W}_3$	.203425	3.39042	99.7024
$\mathbf{W}_4$	.0149283	.248804	99.9512
$\mathbf{W}_5$	.00255207	.0425344	99.9937
$\mathbf{W}_6$	3.76708E−4	.00627847	100

1. If β is substantially in the direction of the pth eigenvector, then the PCR will fail miserably.
2. Data sets that satisfy the conditions of Theorem 1 can easily be constructed as follows:
   a. Choose any data set $\mathbf{X}$ (collinear or not).
   b. Let V_p be the eigenvector corresponding to the smallest eigenvalue of $\mathbf{X}^T\mathbf{X}$.
   c. Generate $\mathbf{Y}$ using $\mathbf{Y} = \alpha\mathbf{X}\mathbf{V}_p + \varepsilon$, where α is a nonzero scalar and ε is a random error.

For example, the variable $\mathbf{U}$ in last column of Table **??** was generated using this procedure where $\mathbf{X}$ is replaced by its standardized version $\mathbf{Z}$, $\alpha = 2$, and $\varepsilon = 0$ (i.e., a degenerate normal random variable with mean 0 and standard deviation 0).

4. OTHER CAUTIONARY NOTES

4.1 Cautionary Note 2: When Using $m < p$ Principal Components, the Increase in the Resulting Sum of Squared Errors (SSE) may be Grossly Discrepant with the Magnitudes of the Eigenvalues in the PC Decomposition of the X Space

To illustrate this point we use Longley's (1967) classic example. This data set consists of six explanatory variables $\mathbf{X}_1, \ldots, \mathbf{X}_6$ and a response variable $\mathbf{Y}$. Table 4 shows the SS decomposition accounted for by each of the orthogonal PCs. The sum of squared errors, SSE, is 836424 for the OLS fit. For the full model, where $m = p$, the PCR gives the same SSE as that of the OLS. For $m < p$, the PCR methodology will never give smaller SSE than the SSE obtained by the OLS. This is quite obvious from standard results in regression analysis.

This cautionary note actually has two parts. First, if we follow the usual PCR methodology deciding the number of PCs to keep, even an extremely high "cumulative %" in the PCs kept can result in much larger SSE than the SSE obtained by OLS.

This phenomenon can be seen by the following PCR methodology applied to the Longley data. To determine the number of PCs m, we would first look at the PCD. These are given in Table 5. We would then probably have chosen to keep only 2 or 3 of the PCs which account for 96.3% and

Table 6. Longley (1967) Data: PCR Using the First Two PCs

Variable	Estimate	St. Error	T value	P value	SSE
$\mathbf{W}_1$	−1565.11	121.067	−12.9277	4.25643E−9	13157179
$\mathbf{W}_2$	391.828	239.597	1.63536	.0629717	
Constant	65317	251.507	259.703	0	

Table 7. Longley (1967) Data: PCR Using the First Three PCs

Variable	Estimate	St. Error	T value	P value	SSE
$\mathbf{W}_1$	−1565.11	55.9754	−27.9607	1.35515E−12	2596235
$\mathbf{W}_2$	391.828	110.778	3.53705	.00204649	
$\mathbf{W}_3$	−1860.39	266.277	−6.98667	7.30703E−6	
Constant	65317	116.284	561.701	0	

99.7% of the variation in $\mathbf{X}$, respectively. The PCR results for $m = 2$ and $m = 3$ are shown in Tables 6 and 7, respectively. The cumulative percent of total variance of $\mathbf{X}$ and the SSE for the various PCR models are given in Table **??**, from which we observe that the three PCs which account for 99.7 of the variation in $\mathbf{X}$ (or a seemingly negligible .3% loss) actually resulted in *threefold* increase in the SSE, from 836424 to 2596235. An even more dramatic increase in SSE is seen when only two PCs are kept (corresponding to a cumulative % of 96.3).

Second, the PCs selected according to the magnitudes of their eigenvalues *do not* contribute monotonically to the SSE. For example, from Table 4 one can see that $\mathbf{W}_3$ contributes more to the fit than $\mathbf{W}_2$ and that $\mathbf{W}_4$ is substantially less than $\mathbf{W}_5$ or $\mathbf{W}_6$.

4.2 Cautionary Note 3: There May Not Be Any Improvement on Numerical Accuracy Via the PCR Procedure

Many alternatives to OLS have been proposed because of the instability of parameter estimates (numerically as well as statistically) as a result of the ill-conditioning of the problem arising from multicollinearity in the predictors. PCR was at least in part so motivated to consider orthogonal PC coordinates free from any collinearity problems. Historically, the numerical accuracy of the Longley data posed challenge because of the *short precision* in the computation, but with long precision and improved basic algorithms for matrix computations, numerical accuracy is no longer a reason for using PCR. For example, using a mere double-precision to compute the OLS regression coefficients $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, we get the parameter estimate results accurate to *at least* nine significant figures in all of the coefficients.

If all of the PCs are used (no PCR-advocate does that to the best of our knowledge), then the PCR equation is (theoretically) identical to the OLS regression equation using all the variables in $\mathbf{X}$. Going through the PCR route, first computing the six PCs and then regressing $\mathbf{Y}$ on the six PCs and convert back to the estimated-beta solution in $\mathbf{X}$, we obtain the numerical results to be virtually the same—that is, the results are accurate to the same orders of magnitude of being correct to 9 to 11 significant digits.

Table 8. Longley (1967) Data: Cumulative Percent of Total Variance and SSE for Three PCR Models

Model	Cumulative %	SSE
Full	100.0	836424
3 PCs	99.7	2596235
2 PCs	96.3	13157179

Tables showing the detailed comparisons of the numerical accuracies of the two solutions (PC vs OLS), when all of the predictors are used, can be obtained from the authors by request.

5. RELATED LITERATURE

Several authors in the statistical literature have hinted or mentioned *some* potential problems related to the pitfalls that we are presenting here explicitly as cautionary notes in the preceding sections. 1 Hotelling (1957) is the earliest reference to our knowledge that seems to be related to our cautionary notes 1 and 2, in a very vague and nonspecific way. Jolliffe (1982) gave four examples in which he considered the use of PCR to be unwise. Later, Jolliffe (1986) gave an example (more specifically related to our cautionary note 2, and a much less dramatic realization of the principle behind our cautionary note 1) in which two of the smaller principal components were among the better predictors. Jackson (1991, p. 276ff) gave an excellent exposition of the PCR methodology, together with some warnings, as well as mentioning a few works that tried to remedy the defects related to our cautionary note 2. These include Lott (1973) and Gunst and Mason (1973).

Traditional use of PCR selects some principal components but retains all of the original variables because each PC is a linear combination of the original predictor variables. A related line of development has been the use of PCR methodology in the selection of the predictor variables in the regression. Among these are Jeffers (1967), Jolliffe (1972, 1973), Hawkins (1973), Mansfield, Webster, and Gunst (1977), and a stepwise procedure proposed by Boneh and Mendieta (1994). Olaya (1997) questions the efficacy of the Boneh and Mendieta methodology, but these methods of selecting the predictor (or independent) variables in the regression are beyond the scope of our coverage in this article.

6. CONCLUDING REMARKS

The basic conclusion of this article is that, in general, the PCs may fail to account for the regression fit. As stated in Theorem 1, it is theoretically possible that the first $(p-1)$ PCs, which can have almost 100% of the variance, contribute nothing to the fit, while the response variable $\mathbf{Y}$ may fit perfectly the last PC which is always ignored by the PCR methodology.

The reason for the failure of the PCR in accounting for the variation of the response variable is that the PCs are chosen based on the PCD which depends only on $\mathbf{X}$. Thus, if PCR is to be used, it should be used with caution and the selection of the PCs to keep should be guided not only by

the variance decomposition but also by the contribution of each principal component to the regression sum of squares.

[Received November 1995. Revised March 1997.]

REFERENCES

Beaton, A. E., Rubin, D. B., and Barone, J. L. (1976), "The Acceptability of Regression Solutions: Another Look at Computational Accuracy," *Journal of the American Statistical Association*, 71, 158–168.

Boneh, S., and Mendieta, G.R. (1994), "Variable Selection in Regression Models Using Principal Components," *Communications in Statistics—Theory and Methods*, 23, 197–213.

Chatterjee, S., and Price, B. (1991), *Regression Analysis by Example*, 2nd ed., New York: Wiley.

Draper, N., and Smith, H. (1981), *Applied Regression Analysis*, 2nd ed.0, New York: Wiley.

Gunst, R.F., and Mason, R.L. (1973), "Biased Estimation in Regression: An Evaluation Using Mean Squared Error," *Journal of the American Statistical Association*, 72, 616–628.

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

Hotelling, H. (1957), "The Relations of Newer Multivariate Statistical Methods to Factor Analysis," *British Journal of Statistical Psychology*, 10, 69–79.

Jackson, E. (1991), *A User's Guide to Principal Components*, New York: Wiley.

Jeffers, J.N. (1967), "Two Case Studies in the Application of Principal Component Analysis," *Applied Statistics*, 16, 225–236.

Jolliffe, I.T. (1972), "Discarding Variables in a Principal Component Analysis, I: Artificial Data," *Applied Statistics*, 21, 160–173.

——— (1973), "Discarding Variables in a Principal Component Analysis, II: Real Data," *Applied Statistics*, 22, 21–31.

——— (1982), "A Note on the Use of Principal Components in Regression," *Applied Statistics*, 31, 300–303.

——— (1986), *Principal Components Analysis*, New York: Springer-Verlag.

Longley, J.W. (1967), "An Appraisal of Least-Squares Programs for the Electronic Computer from the Point of View of the User," *Journal of the American Statistical Association*, 62, 819–831.

Lott, W. F. (1973), "The Optimal Set of Regression Components Restrictions on a Least Squares Regression," *Communications in Statistics—Theory and Methods*, 2, 449–464.

Mansfield, E.R., Webster, J.T., and Gunst, R.F. (1977), "An Analytic Variable Selection Technique for Principal Component Regression," *Applied Statistics*, 36, 34–40.

Olaya, J. (1997), "Some Findings on an Existing Method of Using Principal Components Regression to Select Predictor Variables," unpublished manuscript, presented at the 1997 Joint Statistical Meetings.