Journal of the Korean
Data & Information Science Society
1999, Vol. 10, NO. 1, pp. 163 ~ 172

Numerical Investigations in Choosing the Number of Principal Components in Principal Component Regression - CASE II

Jae-Kyoung Shin[1] · Sung-Ho Moon[2]

Abstract

We propose a cross-validatory method for the choice of the number of principal components in principal component regression based on the magnitudes of correlations with y. There are two different manners in choosing principal components, one is the order of eigenvalues(Shin and Moon, 1997) and the other is that of correlations with y. We apply our method to various data sets and compare results of those two methods.

Key Words and Phrases: Principal component regression, Influence function, Predicted error sum of squares(PRESS), Perturbation expansion, Cross-validatory method

1. Introduction

It is important to choose an adequate number of principal components(PCs) in principal component regression(PCR).

Tanaka(1988) derived influence functions related to an ordinary eigenvalue problem $(A - \lambda_s I)\mathrm{v}_s = 0$, where A is a $p \times p$ real symmetric matrix , λ_s is the *s-th*

[1] Associate Professor, Dept. of Statistics, Changwon National University, Changwon, Kyongnam, 641-773, Korea

[2] Assistant Professor, Dept. of Statistics, Pusan University of Foreign Studies, Pusan, 608-738, Korea

eigenvalue and v_s is the associated eigenvector, and used them for sensitivity analysis in PCR. Using the expansion of $\sum_s \lambda_s^{-1}\mathrm{v}_s\mathrm{v}_s^T$, Shin et al.(1989) tackled sensitivity analysis in PCR and they selected PCs associated with the preassigned number of largest eigenvalues. Recently, Shin and Tanaka(1996) proposed a cross-validatory method to choose the number of PCs in PCR based on the predicted error sum of squares(PRESS). Shin and Moon(1997) applied the above cross-validatory method to various data sets and discuss some properties of the choice for the number of PCs in PCR based on eigenvalues. The present paper is its continuation and we want to compare results of those two methods based on eigenvalues and correlations with y.

2. Cross-Validatory Choice of Principal Components in PCR

The procedure of PCR was proposed by Shin and Tanaka(1996) and in PCR *we have to determine how many and which PCs are to be selected.* There are two different manners in choosing PCs, one is the order of the magnitudes of eigenvalues and the other is that of the magnitudes of correlations with y. In either of these two orders we compute PRESS values for PCR by assigning the number of PCs from 1 to p, and search for PCR with the smallest PRESS value. Then the model with the minimum value of PRESS is regarded as the "best" model.

To evaluate PRESS exactly, we have to compute $\widehat{\beta}_{[i]}$ n times by omitting every one observation in turn. In general, it requires much computing time. Here, we evalute PRESS approximately using a linear approximation based on the perturbation expansion to reduce the computing time in stead of exact computing.

For computing PRESS values , it is necessary to compute

$$\widehat{\beta}_{[i]} = (\Phi_{xx[i]})_D^{-\frac{1}{2}}(V_1\Lambda_1^{-1}V_1^T)_{[i]}(\Phi_{xx[i]})_D^{-\frac{1}{2}}\Phi_{xy[i]},$$

where subscript $[i]$ indicates the omission of the $i\text{-}th$ observation, the subscript D implies "diagonal", Λ_1=diag$(\lambda_1, \cdots, \lambda_q)$ consist of eigenvalues corresponding to the adopted PCs, and V_1=$(\mathbf{v}_1, \cdots, \mathbf{v}_q)$ is the matrix of its associated eigenvectors, Φ_{xx} = the covariance matrix of $\mathbf{x}$ and Φ_{xy} = the covariance matrix of $\mathbf{x}$ and y. Instead of computing above equation exactly, we compute $\widehat{\beta}_{[i]}$ approximately by using the approximation formulas defined in Shin and Moon(1997).

Using the proposed approximation formulas, the cross-validated predicted values can be calculated by substituting the approximate values to the right hand side of the following equation :

$$\widehat{y}_{i[i]} = \bar{y}_{[i]} + (\mathbf{x}_i - \bar{\mathbf{x}}_{[i]})^T (\Phi_{xx[i]})_D^{-\frac{1}{2}} (V_1 \Lambda_1^{-1} V_1^T)_{[i]} (\Phi_{xx[i]})_D^{-\frac{1}{2}} \Phi_{xy[i]},$$

where $\bar{\mathbf{x}}_{[i]} = (n\bar{\mathbf{x}} - \mathbf{x}_i)/(n-1) = \bar{\mathbf{x}} - (n-1)^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}), \quad \bar{y}_{[i]} = \bar{y} - (n-1)^{-1}(y_i - \bar{y}).$

The aims of this study are to compare and discuss some properties of the two manners of ordering and to investigate how well the approximation method works. In the numerical examples we compute PRESS values in the exact and approximate methods, respectively. Then, we show the validity of approximation and discuss some properties of the choice for the number of PCs in PCR.

3. Numerical Examples

To illustrate some properties of our proposed procedure we applied our method of cross-validatory choice of PCs in PCR to some data sets listed in Table 3.1.

Table 3.1 The list of data sets

Data set	sample size n	variables $p(R^2)$	condition number	source of data
Longley	16	7 (**0.996**)	12114.158	Longley(1967)
Equal Educational Opportunity(EEO)	70	4 (0.206)	370.853	Chatterjee and Price(1977)
Rat	19	4 (0.364)	204.035	Cook and Weisberg(1980)
Stack and Loss	21	4 (**0.914**)	10.299	Brownlee(1965)
Import	18	4 (**0.973**)	1012.897	Chatterjee and Price(1977)

First, we applied PCA based on the correlation matrix to the independent variables, then we compute all of the PRESS in the order of eigenvalues and that of correlations with y.

As explained in the previous section, PCs are entered into regression one by one in the order of the magnitudes of eigenvalues or that of the magnitudes of correlations with y. The ordering of the latter case is equivalent to that of F-values for testing the coefficients of PCs, because PCs are uncorrelated with each other. The exact and approximate PRESS values are plotted in Figures 3.1 and 3.2 and their plotted values are given in Tables 3.2 and 3.3. These results show that the PCR with the approprite PC(s) selected by the eigenvalues and F-values gives the best model.

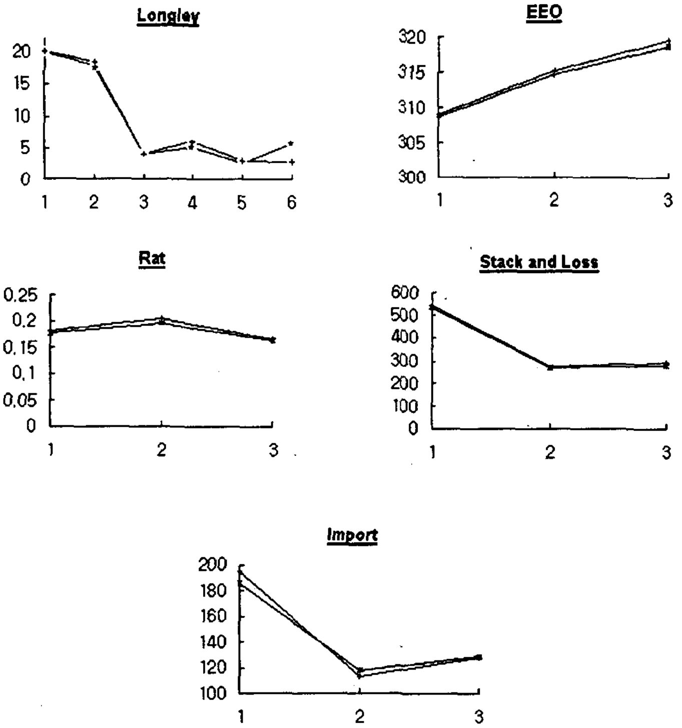

Figure 3.1 Index plots of PRESS based on eigenvalues : (+) exact , (*) approximate (horizontal : # of PCs, vertical : PRESS)

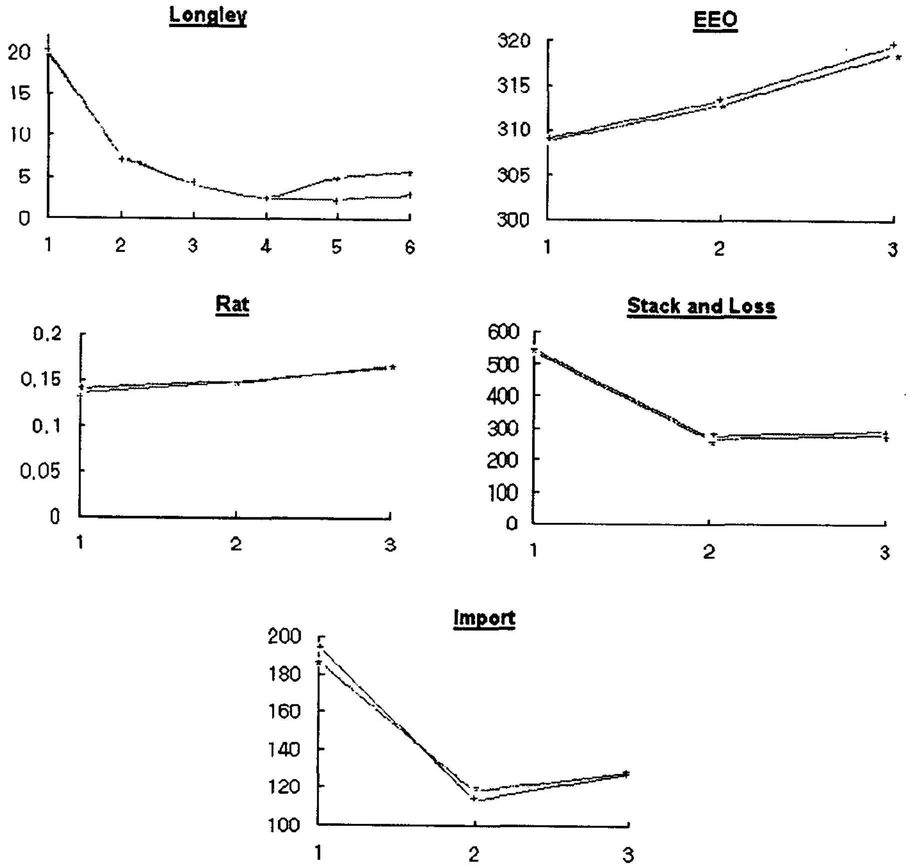

Figure 3.2 Index plots of PRESS based on correlations with y : (+) exact , (∗) approximate (horizontal : # of PCs, vertical : PRESS)

Table 3.2 The values of PRESS of all data sets by using the eigenvalues

	Number of PCs					
Data set	1	2	3	4	5	6
Longley	19.9673	17.8791	**4.0987**	5.2087	**2.7054**	5.6668
EEO	308.7043	314.7049	318.6351	***	***	***
Rat	0.1768	0.1966	**0.1647**	***	***	***
Stack and Loss	534.5933	266.9562	276.1991	***	***	***
Import	186.3671	117.9507	128.4166	***	***	***

Table 3.3 The values of PRESS of all data sets by using the correlation with y

	Number of PCs					
Data set	1	2	3	4	5	6
Longley	19.9673	7.3854	4.0987	**2.2509**	5.1153	5.6668
EEO	308.7043	312.7577	318.6351	***	***	***
Rat	**0.1411**	0.1485	0.1647	***	***	***
Stack and Loss	534.5933	266.9562	276.1991	***	***	***
Import	186.3671	117.9507	128.4166	***	***	***

4. Summary and Discussion

In the preceding section, we compared the PRESS values for the models obtained by the PC selection procedures based on two criteria.

In Figures 3.1 and 3.2 of the above numerical example, we can observe the minimum of the PRESS values for all the data sets. For the cases of Longley, Import and Stack & Loss data sets in both procedures, we can see the concave typed minimum points. But, we can find the minimum points on condition that we choose only one PC(EEO and Rat data sets) or all PCs(Rat data set). So it may be concluded that for the data sets with relatively high value of R^2(around 0.9 ; see, Table 3.1), we can find the concave typed minimum points.

Particularly, for the Longley data set in the case of procedure based on eigenvalues we can observe that there are two local minima of the PRESS values at three PCs and five PCs and that the latter gives the global minimum. But, in the case of procedure based on correlation there is only one local(and global) minimum at four PCs. It might be caused by the F-value (Shin and Tanaka, 1996). This four PCs model gives the smallest PRESS value among all the models which have been investigated with the two PC selection procedures, and it is regarded as the best model.

Table 4.1 The F-values(correlation with y) of all data sets

	Number of PCs					
Data set	1	2	3	4	5	6
Longley	1819.6205	29.1183	113.6215	0.3078	15.6764	2.9455
EEO	15.3189	0.5007	1.3309	***	***	***
Rat	1.0993	0.1199	7.3621	***	***	***
Stack and Loss	156.7422	20.9445	2.0197	***	***	***
Import	497.1567	8.0752	0.1150	***	***	***

Table 4.2 The Eigenvalues of all data sets

	Number of PCs					
Data set	1	2	3	4	5	6
Longley	4.6039	1.1753	0.2034	0.0149	0.0026	0.0004
EEO	2.9520	0.0401	0.0080	***	***	***
Rat	2.3526	0.6377	0.0097	***	***	***
Stack and Loss	2.1332	0.6597	0.2071	***	***	***
Import	2.0905	0.8951	0.0145	***	***	***

In the Rat data set, we can find one local minimum with the PC selection procedures based on two criteria. In this data set, model with all PCs gives the minimum PRESS value in the procedure based on eigenvalues and the other hand the one PC which the order of third PC in magnitude of eigenvalues gives the minimum PRESS values in the procedure based on correlation(see, Table 4.1 and Table 4.2). The latter case model gives the smallest PRESS value for PC selection manners based on two criteria, and also it is regarded as the best model.

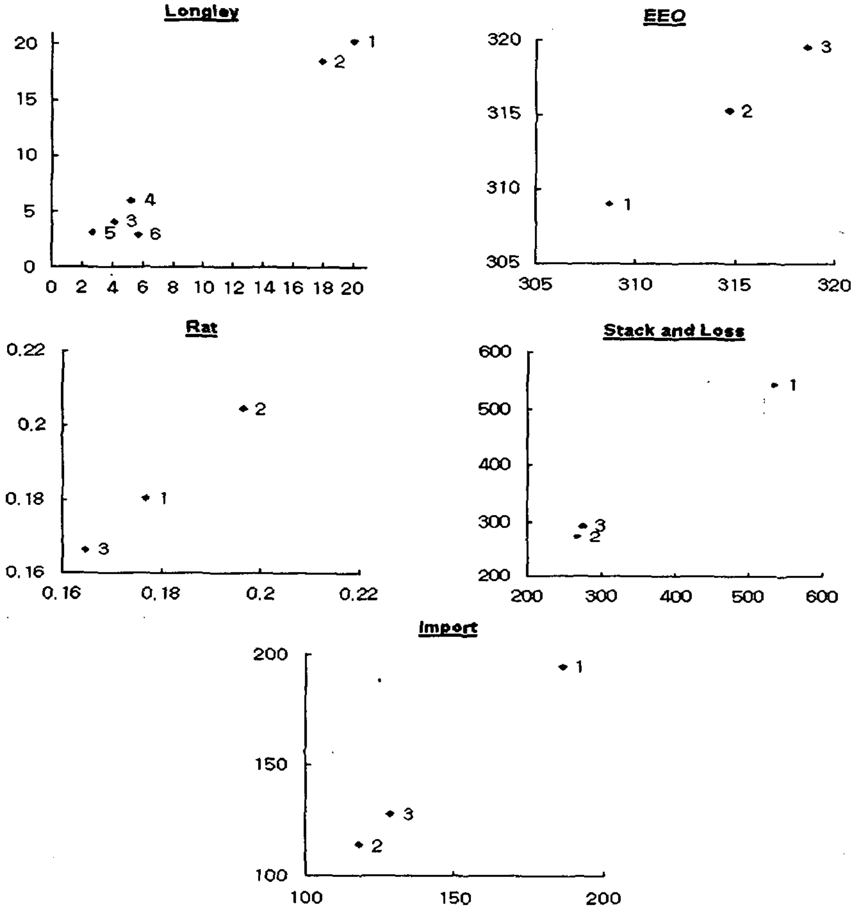

Figure 4.1 Scatter diagrams of proposed method(horizontal) versus exact method(vertical) : *based on eigenvalues*

Finally, it may be concluded that we can determine the number of PCs by the proposed method regardless of the value of R^2, sample size n or number of variables and the conditional number (see, Table 3.1).

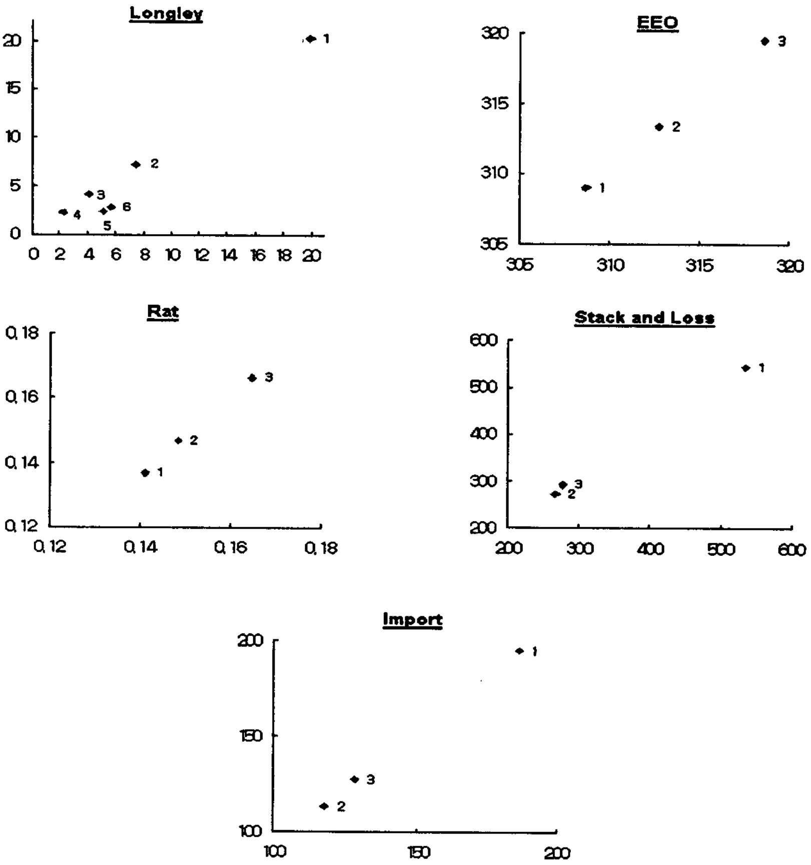

Figure 4.2 Scatter diagrams of proposed method(horizontal) versus exact method(vertical) : based on correlations with y

Figure 4.1 and 4.2 show the scatter diagrams of the PRESS values obtained by the proposed approximate method and the exact method. In these scatter diagrams, most of the points, except for "6"(eigenvalue) and "5 and 6"(correlation) of Longley data, are located near the straight line so that we may conclude that the proposed method can be used practically instead of the exact method for selecting PCs in PCR.

References

1. Allen, D.M. (1971). Mean square error of predictings as a criterion for selecting variables. *Technometrics, 13, 469-475.*

2. Jolliffe, I.T. (1986). *Principal Component Analysis.* Springer-Verlag.

3. Longley, J.W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of American Statistical Association, 62, 819-841.*

4. Massy, W.F. (1965). Principal components regression in exploratory statistical research. *Journal of American Statistical Association, 60, 234-256.*

5. Shin, J.K. and Moon, S.H. (1997). Numerical investigations in choosing the number of principal components in principal component regression-Case I. *Journal of Statistical Theory & Methods, 8, No.2, 127-134.*

6. Shin, J.K. and Tanaka, Y. (1996). Cross-validatory choice for the number of principal components in principal component regression. *Journal of the Japanese Society of Computational Statistics, 9, 53-59.*

7. Shin, J.K., Tarumi, T. and Tanaka, Y. (1989). Sensitivity analysis in principal component regression. *Bulletin of the Biometric Society of Japan, 10, 57-68.*

8. Tanaka, Y. (1988). Sensitivity analysis in principal component analysis : Influence on the subspace spanned by principal components. *Communication in Statiststics : Theory and Methods, 17, 3157-3175. (Corrections, A 18(1989), 4305.)*

9. Tanaka, Y. (1989). Influence functions related to eigenvalue problems which appear in multivariate methods. *Communication in Statistics : Theory and Methods, 18, 3991-4010.*