Journal of Statistical
Theory & Methods
1997, Vol. 8, No. 2, pp. 127 ~ 134

Numerical Investigations in Choosing the Number of Principal Components in Principal Component Regression - CASE I [1]

Jae-Kyoung Shin[2] and Sung-Ho Moon[3]

Abstract

A method is proposed for the choice of the number of principal components in principal component regression based on the predicted error sum of squares. To do this, we approximately evaluate that statistic using a linear approximation based on the perturbation expansion. In this paper, we apply the proposed method to various data sets and discuss some properties in choosing the number of principal components in principal component regression.

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

1. Introduction

It is important that the problem of the choice of the number of principal components(PCs) in principal component regression(PCR). There are two partially conflicting objectives, in choosing the PCs. In order to eliminate large variances of the estimates due to multicollinearities, it is essential to delete components whose variances are very small but, at the same time, it is undesirable to delete components which are effective for the prediction of the dependent variable.

Shin, Tarumi and Tanaka(1989) discussed a method of sensitivity analysis in PCR based on an influence function derived by Tanaka(1988). In the above paper we selected PCs associated with the preassigned number of largest eigenvalues. Shin and Tanaka(1997) proposed a cross-validatory method to choose the number of PCs in PCR based on the predicted error sum of squares(PRESS). In the present paper, we apply the proposed method to various data sets and discuss some properties of the choice for the number of PCs in PCR.

[1]This Research was supported by the Pusan University of Foreign Studies research grants in 1997
[2]Dept. of Statistics, Changwon National University, Changwon, Korea.
[3]Dept. of Statistics, Pusan University of Foreign Studies, Pusan, Korea.

2. Principal Component Regression

In PCR, we first apply principal component analysis(PCA) to the independent variables. There are two variations of PCA. One is based on the covariance matrix and the other is based on the correlation matrix. In this paper we mainly use the latter type, because the standardization of all the independent variables is usually the first choice to avoid multicollinearity.

The procedure of PCR can be summarized in the following steps.

Step 1. Compute the covariance matrices Φ_{xx} and Φ_{xy} from the data matrix.

Step 2. Compute the correlation matrix

$$\Gamma_{xx} = (\Phi_{xx})_D^{-\frac{1}{2}} \Phi_{xx} (\Phi_{xx})_D^{-\frac{1}{2}},$$

where the subscript D implies "diagonal".

Step 3. Apply the spectral decomposition to the correlation matrix

$$\Gamma_{xx} = V_1 \Lambda_1 V_1^T + V_2 \Lambda_2 V_2^T$$

where Λ_1=diag$(\lambda_1, \cdots, \lambda_q)$ and Λ_2=diag$(\lambda_{q+1}, \cdots, \lambda_p)$ consist of eigenvalues corresponding to the adopted PCs and the remaining eigenvalues, respectively, and V_1=$(\mathbf{v}_1, \cdots, \mathbf{v}_q)$ and V_2=$(\mathbf{v}_{q+1}, \cdots, \mathbf{v}_p)$ are the matrices of their associated eigenvectors.

Step 4. Compute the regression coefficients $\hat{\alpha}$ on the q PCs and transform them to the coefficients $\hat{\beta}$ on the original variables

$$\widehat{\alpha} = \Lambda_1^{-1} V_1^T (\Phi_{xx})_D^{-\frac{1}{2}} \Phi_{xy},$$

$$\widehat{\beta} = (\Phi_{xx})_D^{-\frac{1}{2}} V_1 \Lambda_1^{-1} V_1^T (\Phi_{xx})_D^{-\frac{1}{2}} \Phi_{xy}.$$

Step 5. Obtain the regression equation of y on $\mathbf{x}$

$$\widehat{\mu_y} = \bar{y} + (\mathbf{x} - \bar{\mathbf{x}})^T \widehat{\beta}.$$

3. Cross-Validatory Choice of Principal Components

In actual data analysis, we have to determine how many and what PCs are to be adopted. In doing this, we use stepwise procedure. PCs are entered into regression one by one in the order of the magnitudes of eigenvalues. 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. In computation of PRESS values, Shin and Tanaka(1997) proposed to use an approximation formula to save computing cost and considered a method of the choice of the number of PCs in PCR based on the PRESS. The present study is its continuation.

3.1 Criterion for the Choice of PCs

In choosing the number of PCs we use the PRESS statistic proposed by Allen(1971), as a criterion. A model with the minimum value of PRESS is regarded as the "best" model. The PRESS statistic is defined as follows :

$$PRESS = \sum_{i=1}^{n} \hat{e}_{i[i]}^2,$$

where $\hat{e}_{i[i]} = y_i - \hat{y}_{i[i]}$, y_i is the observed value of the dependent variable for $\mathbf{x} = \mathbf{x}_i$, $\hat{y}_{i[i]}$ is the predicted value for $\mathbf{x} = \mathbf{x}_i$ using the estimated regression coefficient coefficient $\hat{\beta}_{[i]}$, and $\hat{\beta}_{[i]}$ is the estimate based on the sample without the i-th observation.

To evaluate PRESS exactly, we have to compute $\hat{\beta}_{[i]}$ n times by omitting every one observation in turn. In general, it requires high computing cost. In order to reduce computing cost, we use a linear approximation based on the perturbation expansion in evaluating PRESS approximately in stead of exact computing.

3.2 Approximate Cross-Validatory Method

To evaluate PRESS it is necessary to compute

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

where subscript $[i]$ indicates the omission of the i-th observation. Instead of computing $\hat{\beta}_{[i]}$ exactly, we compute it approximately using the following approximation formulas :

$$\Phi_{xx[i]} \cong \Phi_{xx} - (n-1)^{-1} \Phi_{xx}^{(1)}, \quad \Phi_{xy[i]} \cong \Phi_{xy} - (n-1)^{-1} \Phi_{xy}^{(1)},$$

$$(V_1 \Lambda_1^{-1} V_1^T)_{[i]} \simeq V_1 \Lambda_1^{-1} V_1^T - (n-1)^{-1} (V_1 \Lambda_1^{-1} V_1^T)^{(1)},$$

where superscript (1) indicates the empirical influence function, and $\Phi_{xx}^{(1)}$, $\Phi_{xy}^{(1)}$ and $(V_1\Lambda_1^{-1}V_1^T)^{(1)}$ are given as follows :

$$\Phi_{xx}^{(1)} = (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T - \Phi_{xx}, \quad \Phi_{xy}^{(1)} = (\mathbf{x}_i - \bar{\mathbf{x}})(y_i - \bar{y}) - \Phi_{xy},$$

$$\begin{aligned}(V_1\Lambda_1^{-1}V_1^T)^{(1)} = & - \sum_{s=1}^{q}\sum_{r=1}^{q} \lambda_s^{-1}\lambda_r^{-1}[\mathbf{v}_s^T\Gamma_{xx}^{(1)}\mathbf{v}_r]\mathbf{v}_s\mathbf{v}_r^T \\ & + \sum_{s=1}^{q}\sum_{r=q+1}^{p} \lambda_s^{-1}(\lambda_s - \lambda_r)^{-1}[\mathbf{v}_s^T\Gamma_{xx}^{(1)}\mathbf{v}_r](\mathbf{v}_s\mathbf{v}_r^T + \mathbf{v}_r\mathbf{v}_s^T).\end{aligned}$$

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

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

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

The "best" model is obtained as a model with the minimum value of PRESS. In the numerical investigation, we compute PRESS exactly and approximately and discuss the reliability of approximation.

4. Numerical Investigation

To investigate 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 4.1.

Table 4.1 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.

As explained in the previous section, PCs are entered into regression one by one in the order of the magnitudes of eigenvalues. The exact and approximate PRESS values are plotted in Figure 4.1. and their plotted values are given in Table 4.2. These results show that the PCR with the approprite PC(s) selected by the eigenvalues gives the best model.

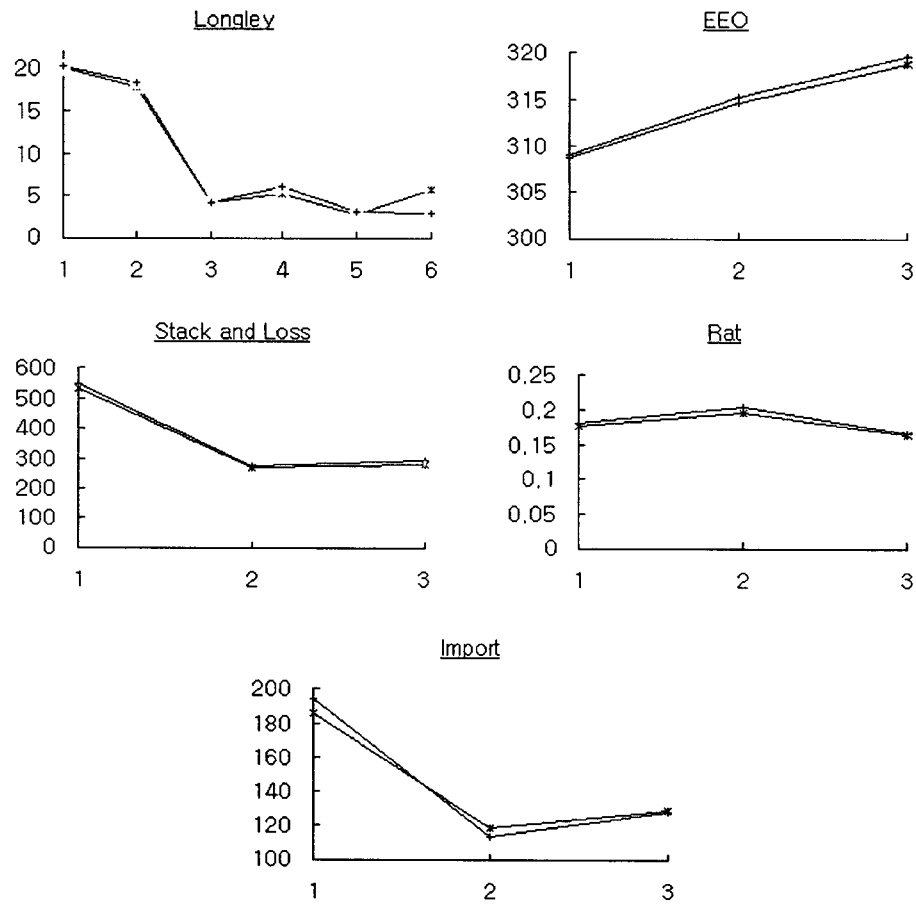

Figure 4.1 Index plots of PRESS : (+) exact , (∗) approximate

Table 4.2 The values of PRESS of all data sets

	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	308.7042	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	***	***	***

In Figure 4.1 of the above numerical example, we can observe the minimum of the PRESS values for all the data sets. Especially for the cases of Longley, Import and Stack & Loss data sets, we can see the concave typed minimum points as we expected. But for EEO and Rat data sets, we can find the minimum points on condition that we choose only one or all PCs. So it may be concluded that for the data sets with relatively high value of R^2(around 0.9 ; see, Table 4.1), we can find the concave typed minimum points.

Particularly for the Longley data set, 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. It might be caused by the F-value (Shin and Tanaka, 1996).

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 4.1).

Figure 4.2 shows 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" 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.

In present study, we apply the method of Shin and Tanaka(1996) to various data sets based on the order of the magnitudes of eigenvalues and investigate some properties from the results. We would like to study in the future more precisely about this matter by applying their method based on the order of the magnitudes of correlations with y.

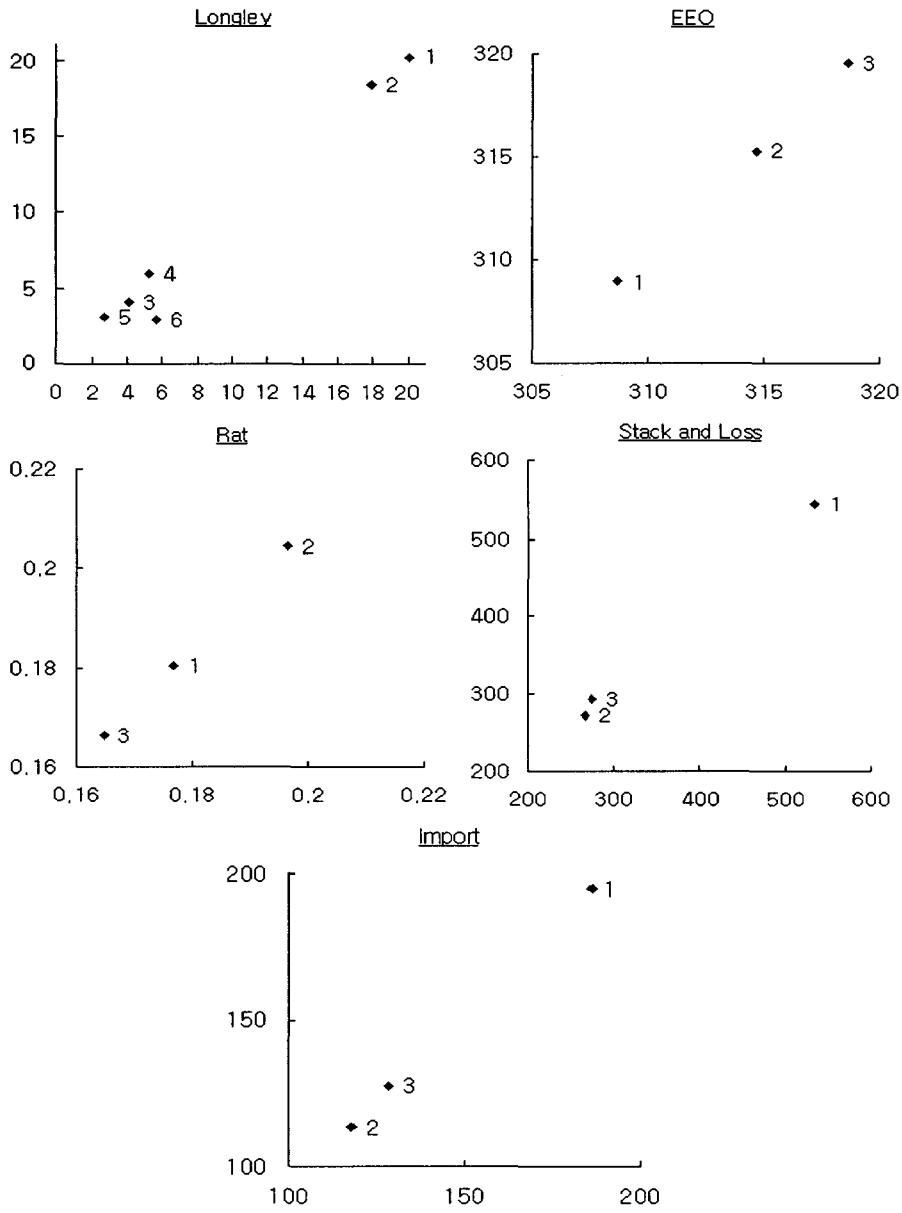

Figure 4.2 Scatter diagrams of proposed method(horizontal) versus exact method(vertical)

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 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.

6. 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.

7. 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.)

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