Youngnam Statistical Letters
June, 1991, Vol. 2, 1-9

Sensitivity Analysis in Principal Component Regression : Numerical Investigation

Jae-Kyoung Shin*, Tomoyuki Tarumi**, and Yutaka Tanaka**

ABSTRACT

Shin, Tarumi and Tanaka(1989) discussed a method of sensitivity analysis in principal component regression(PCR) based on an influence function derived by Tanaka(1988). The present paper is its continuation. In this paper we first consider two new influence measures, then apply the proposed method to various data sets and discuss some properties of sensitivity analysis in PCR.

1. Introduction

We consider an ordinary regression model

$$\mathbf{y} = \mathbf{1}\beta_0 + X\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I), \tag{1}$$

where $\mathbf{y}$ is an $(n \times 1)$ vector of the dependent variable, $\mathbf{1}$ is an $(n \times 1)$ vector whose elements are all 1's, X is an $(n \times p)$ matrix of the independent variables and ϵ is an $(n \times 1)$ vector of error terms. Denote a mean vector and a covariance matrix by μ and Φ with subscripts indicating the related variables, *i.e.*, μ_x = the mean vector of $\mathbf{x}$, Φ_{xx} = the covariance matrix of $\mathbf{x}$, Φ_{xy} = the covariance matrix between $\mathbf{x}$ and y, etc.

It is well-known that, if we use the method of least squares to estimate (β_0, β^T), the estimate becomes poor when the matrix X^TX is nearly singular. This phenomenon is called *multicollinearity*. Principal component regression(PCR) is one of the methods developed to avoid this difficulty. Shin, Tarumi and Tanaka(1989) considered a method of sensitivity analysis in PCR based on an influence function

* Graduate School of Natural Science and Technology, Okayama University.
** Department of Statistics, Okayama University, Okayama, Japan.

derived by Tanaka(1988). The present paper is its continuation. In Section 2 two new influence measures are derived. In Section 3 we apply the proposed method to various data sets and discuss some properties of sensitivity analysis in PCR.

2. Sensitivity Analysis

We study the influence of a small change of data on the standardized regression coefficient vector β^* obtained by PCR. The influence function $I(\mathbf{x}, y; \beta^*)$, which we simply denote by $\beta^{*(1)}$, is given as follows.

$$\begin{aligned}\beta^{*(1)} =&(V_1\Lambda_1^{-1}V_1^T)^{(1)}(\Phi_{xx})_D^{-\frac{1}{2}}\Phi_{xy} + V_1\Lambda_1^{-1}V_1^T\{(\Phi_{xx})_D^{-\frac{1}{2}}\}^{(1)}\Phi_{xy} \\ &+ V_1\Lambda_1^{-1}V_1^T\{\Phi_{xx}\}_D^{-\frac{1}{2}}\Phi_{xy}^{(1)},\end{aligned} \tag{2}$$

where D implies "diagonal", Λ_1 and Λ_2 are the diagonal matrices of the eigenvalues of interest and the remaining eigenvalues, respectively, and V_1 and V_2 are the matrices of the associated eigenvectors. The quantity $(V_1\Lambda_1^{-1}V_1^T)^{(1)}$ in the right hand side can be calculated as follows(see Tanaka,1989).

$$\begin{aligned}(V_1\Lambda_1^{-1}V_1^T)^{(1)} =& -\sum_{s=1}^{q}\sum_{r=1}^{q}\lambda_s^{-1}\lambda_r^{-1}[\mathbf{v}_s^T\Gamma_{xx}^{(1)}\mathbf{v}_r]\mathbf{v}_s\mathbf{v}_r^T \\ &+\sum_{s=1}^{q}\sum_{r=q+1}^{p}\lambda_s^{-1}(\lambda_s-\lambda_r)^{-1}[\mathbf{v}_s^T\Gamma_{xx}^{(1)}\mathbf{v}_r](\mathbf{v}_s\mathbf{v}_r^T+\mathbf{v}_r\mathbf{v}_s^T),\end{aligned} \tag{3}$$

where Γ is a correlation matrix. Notice that the right hand side tends to be large, when there is an eigenvalue λ_s such that $\lambda_s \cong 0$ or there is a pair of eigenvalues (λ_s, λ_r) such that $\lambda_s - \lambda_r \cong 0$ where λ_s belongs to the set of the eigenvalues of interest and λ_r to the set of the remaining eigenvalues.

As new influence measures we consider the influence $\hat{\sigma}^{2^{(1)}}$ on the error variance $\hat{\sigma}^2$ and the generalized variance $det(V(\hat{\beta}^*))$ of the estimated β^*.

The influence on $\hat{\sigma}^2$ is evaluated by

$$\hat{\sigma}^{2^{(1)}} = -(n-1)(s_{(i)}^2 - s^2), \tag{4}$$

where s^2 and $s_{(i)}^2$ are the estimated σ^2 based on the whole sample and the sample without the i-th observation (the quantities with the subscript (i) indicates the

estimate without the $i-th$ observation). The estimates s^2 and $s^2_{(i)}$ are calculated by

$$s^2 = (n-p-1)^{-1}(\mathbf{y} - X^*\hat{\beta}^*)^T(\mathbf{y} - X^*\hat{\beta}^*),$$

$$s^2_{(i)} = (n-p-2)^{-1}(\mathbf{y}_{(-i)} - X^*_{(i)}\hat{\beta}^*_{(i)})^T(\mathbf{y}_{(-i)} - X^*_{(i)}\hat{\beta}^*_{(i)}), \tag{5}$$

where the subscript $(-i)$ simply indicates the omission of the i-th observation. The elements of $X^*_{(i)}$ are calculated using the following relations between the two means and the two variances for the whole sample and the sample without the i-th observation.

$$\bar{x}_{(i)} = (n-1)^{-1}(n\bar{x} - x_i),$$

$$s^2_{x(i)} = (n-2)^{-1}((n-1)s^2_x - n(n-1)(x_i - \bar{x})^2)$$

To get $\hat{\sigma}^{2(1)}$ exactly we need to compute $\hat{\beta}^*_{(i)}$ for i=1 to n by solving the eigenvalue problems n times. It requires high computing cost. here, instead of computing exact $\hat{\beta}^*_{(i)}$ we use a linear approximation based on the perturbation expansion as

$$\widetilde{\beta}^*_{(i)} = \hat{\beta}^* - (n-1)^{-1}\hat{\beta}^{*(1)},$$

compute an approximate value $\widetilde{s}^2_{(i)}$ for $s^2_{(i)}$ by (5) with $\hat{\beta}^*_{(i)}$ replaced by $\widetilde{\beta}^*_{(i)}$, and finally obtain an approximate $\widetilde{\sigma^2}^{(1)}$ by (4) with $s^2_{(i)}$ replaced by $\widetilde{s}^2_{(i)}$.

The influence on $det(V(\hat{\beta}^*))$ is evaluated by the influence function

$$[det(V(\hat{\beta}^*))]^{(1)} = [det(V(\hat{\beta}^*))]tr([V(\hat{\beta}^*)]^{-1}[V(\hat{\beta}^*)]^{(1)}), \tag{6}$$

where

$$[V(\hat{\beta}^*)]^{(1)} = \frac{\sigma^{2(1)}}{n}(V_1\Lambda_1^{-1}V_1^T) + \frac{\sigma^2}{n}(V_1\Lambda_1^{-1}V_1^T)^{(1)}.$$

The first term of the right hand side of the last equation is approximated by using (4), and the second term can be obtained from (3).

3. Numerical Investigation

To investigate the properties of our procedure we applied our method of sensitivity analysis to the data sets shown in Table 1.

Table 1. The data sets of example

Data set	sample size n	variables p	condition number	source of data
Longley	16	7	12114.158	Longley(1967)
Hill	15	7	119.685	Hill(1977)
Equal Educational Opportunity(EEO)	70	4	370.853	Chatterjee and Price(1977)
Rat	19	4	242.035	Weisberg(1980)
Coleman	20	6	41.149	Rousseeuw and Leroy(1987)
Heart Catheterization	12	3	50.401	Rousseeuw and Leroy(1987)
Aircraft	23	5	37.799	Rousseeuw and Leroy(1987)
Wood Specific Gravity	20	6	30.633	Rousseeuw and Leroy(1987)

In PCR, we first apply principal component analysis(PCA) based on the correlation matrix to the independent variables, then we select some principal components(PCs) and take regression on the PCs. In sensitivity analysis we compute the empirical influence curve $\hat{\beta}^{*(1)}$ based on the proposed procedure and summarize it into scalar valued measure $\|\hat{\beta}^{*(1)}\|$, D and newly proposed $\hat{\sigma}^{2(1)}$ and $[det(V(\hat{\beta}^{*}))]^{(1)}$.

The aims of this numerical study are :

1) to investigate the usefulness of the empirical influence curve(EIC) $\hat{\beta}^{*(1)}$ based on the proposed procedure, under different conditions of selecting PCs, by checking its relationship with the sample influence curves(SIC),
2) to investigate how the result of sensitivity analysis changes when the PCs selected change,

and

3) to investigate the properties of the observations which are found to be influential with by $\|\hat{\beta}^{*(1)}\|$, D^{*}, $\hat{\sigma}^{2(1)}$ and $[det(V(\hat{\beta}^{*}))]^{(1)}$.

The scatter diagrams in Fig. 1 are some examples of the scatter diagrams of EIC versus SIC, under some different conditions of selecting PCs. It is clear that most of the points are located near the straight line $SIC = EIC$. We can observe

the similar tendencies in the other cases, *i.e.* other variables and/or other data sets. From this we may conclude that the quantity EIC can be used practically instead of SIC under various conditions of selecting PCs.

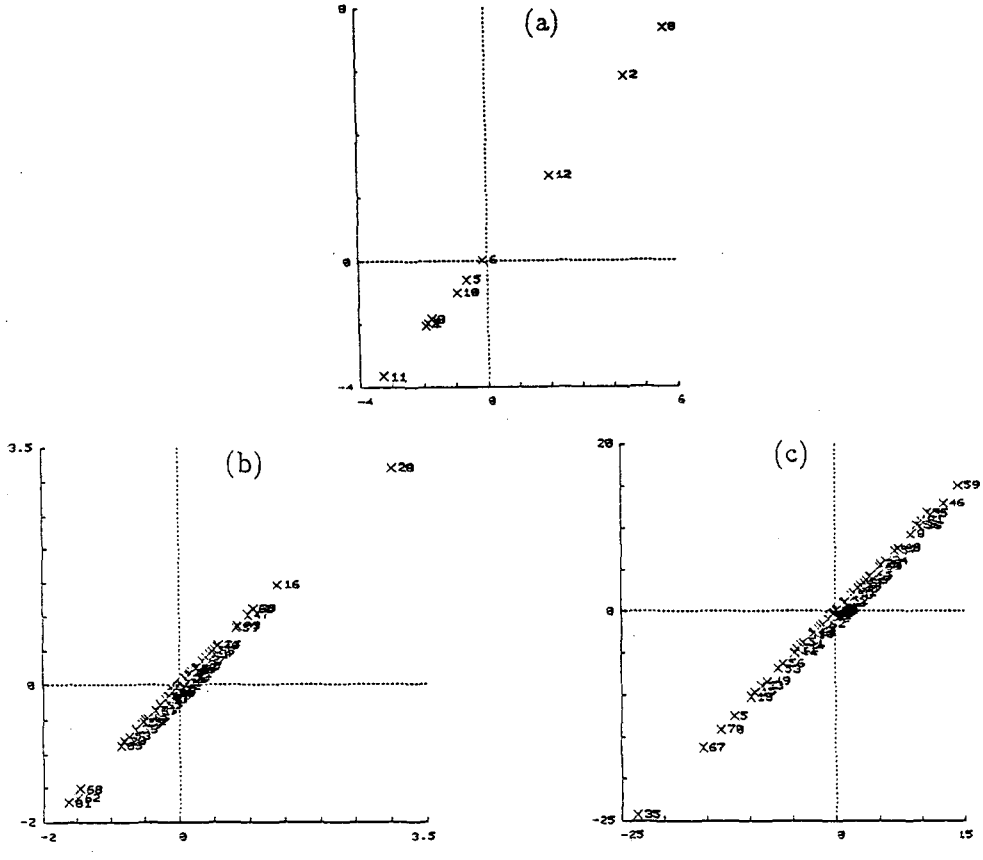

Fig.1 Scatter Diagrams of EIC(horizontal) vs. SIC(vertical) for $\beta_1^{*(1)}$ (a)heart catheterization data(pc=1), (b)EEO(pc=1), (c)EEO(pc=2)

Next, we investigate how the results of sensitivity analysis change when the selected PCs change. The index plots of $\|\hat{\beta}^{*(1)}\|$ are shown in Fig. 2 when the selected PCs change in Hill's data. Table 2 shows influential observations found with $\|\hat{\beta}^{*(1)}\|$ when the PCs corresponding to the largest q eigenvalues are selected. We can see that different observations are found to be influential when the selected PCs are different.

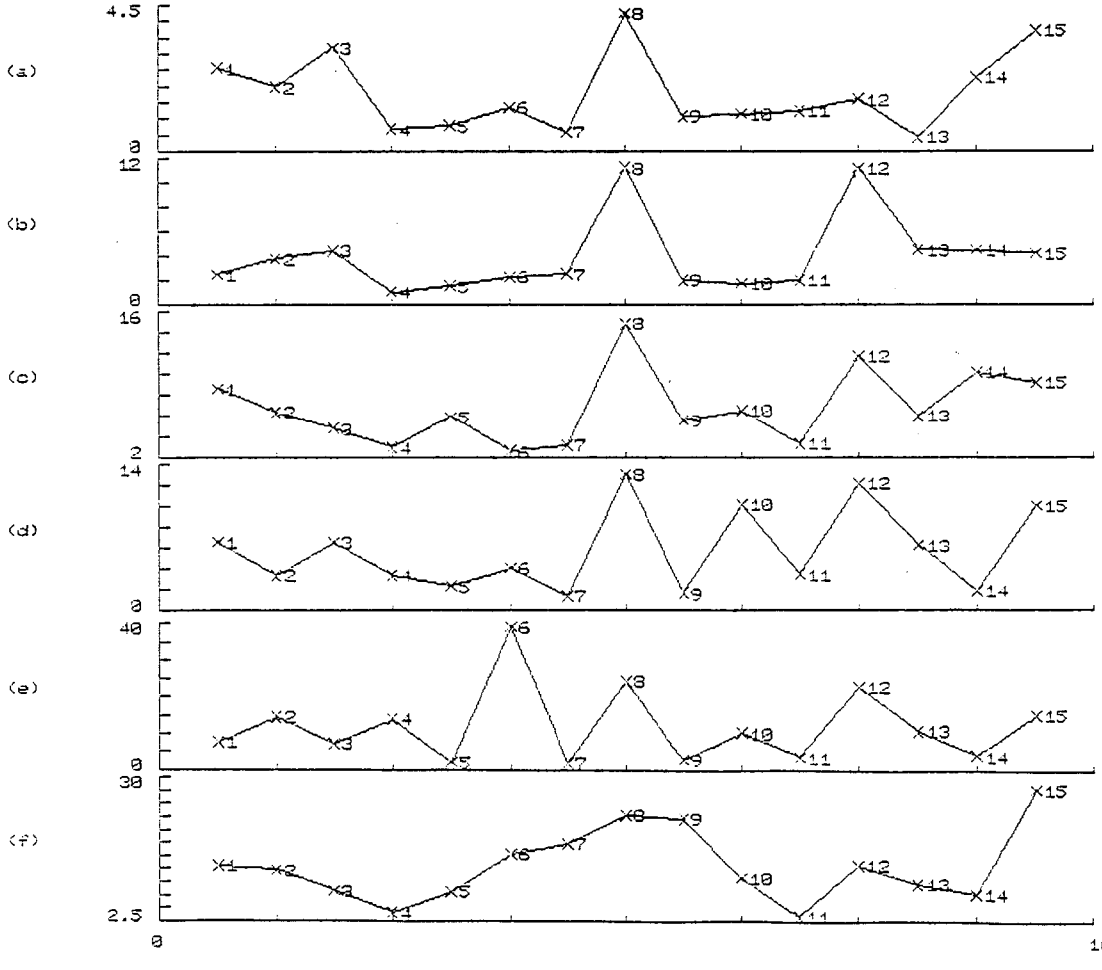

Fig. 2 Index Plots of $\|\hat{\beta}^{*(1)}\|$ (Hill' data):
(a) pc=1, (b) pc=2, (c) pc=3, (d) pc=4, (e) pc=5, (f) pc=6.

Table 2 Cumulative proportion, coefficient of determination and influential observations found with $\|\hat{\beta}^{*(1)}\|$ when the PCs corresponding to the largest q eigenvalues are selected.

Data set	PC (q)	eigenvalue	cumulative proportion	coefficient of determination	influential observations measures with $\|\hat{\beta}^{*(1)}\|$
Longley	1	4.60338	0.76723	0.91425	3
	2	1.17534	0.96312	0.92888	3
	3	0.20343	0.99702	0.98597	1
	4	0.01493	0.99951	0.98612	16
	5	0.00255	0.99994	0.99397	10
	6	0.00038	1.00000	0.99547	5(OLS)
Hill	1	3.79999	0.63333	0.66081	8, 15
	2	1.05511	0.80918	0.69471	8, 12
	3	0.62357	0.91311	0.73001	8
	4	0.42661	0.98421	0.77777	8
	5	0.06298	0.99471	0.78391	6, 8
	6	0.03175	1.00000	0.86457	15(OLS)
Equal Educational opportunity(EEO)	1	2.95199	0.98400	0.18423	28
	2	0.04005	0.99735	0.19025	35
	3	0.00796	1.00000	0.20626	35(OLS)
Rat	1	2.35258	0.78420	0.04662	3
	2	0.63770	0.99676	0.05170	5
	3	0.00972	1.00000	0.36390	5(OLS)
Coleman	1	2.83681	0.56736	0.72559	10, 11
	2	1.39508	0.84638	0.72587	10
	3	0.49664	0.94571	0.78730	10
	4	0.20253	0.98621	0.90212	15, 18
	5	0.06894	1.00000	0.90631	18(OLS)
Heart Catheterization	1	1.96109	0.98055	0.82359	8
	2	0.03891	1.00000	0.82536	6, 8(OLS)
Aircraft	1	2.67464	0.66866	0.64415	22
	2	0.87665	0.88782	0.65524	22
	3	0.37795	0.98231	0.68149	22
	4	0.07076	1.00000	0.88364	22(OLS)
Wood Specific Gravity	1	2.73984	0.54797	0.51896	19
	2	1.03632	0.75523	0.58532	11
	3	0.70911	0.89705	0.58760	7, 9, 11
	4	0.42529	0.98211	0.74226	11
	5	0.08944	1.00000	0.80840	11(OLS)

Finally, we investigate the properties of the observations which are found to be influential with by $\|\hat{\beta}^{*(1)}\|$, D^*, $\hat{\sigma}^{2^{(1)}}$ and $[det(V(\hat{\beta}^*))]^{(1)}$. Fig. 3 shows the index plots of $\|\hat{\beta}^{*(1)}\|$, D^*, $\hat{\sigma}^{2^{(1)}}$ and $[V(\hat{\beta}^*_{1,1})]^{(1)}$ in Hill's data.

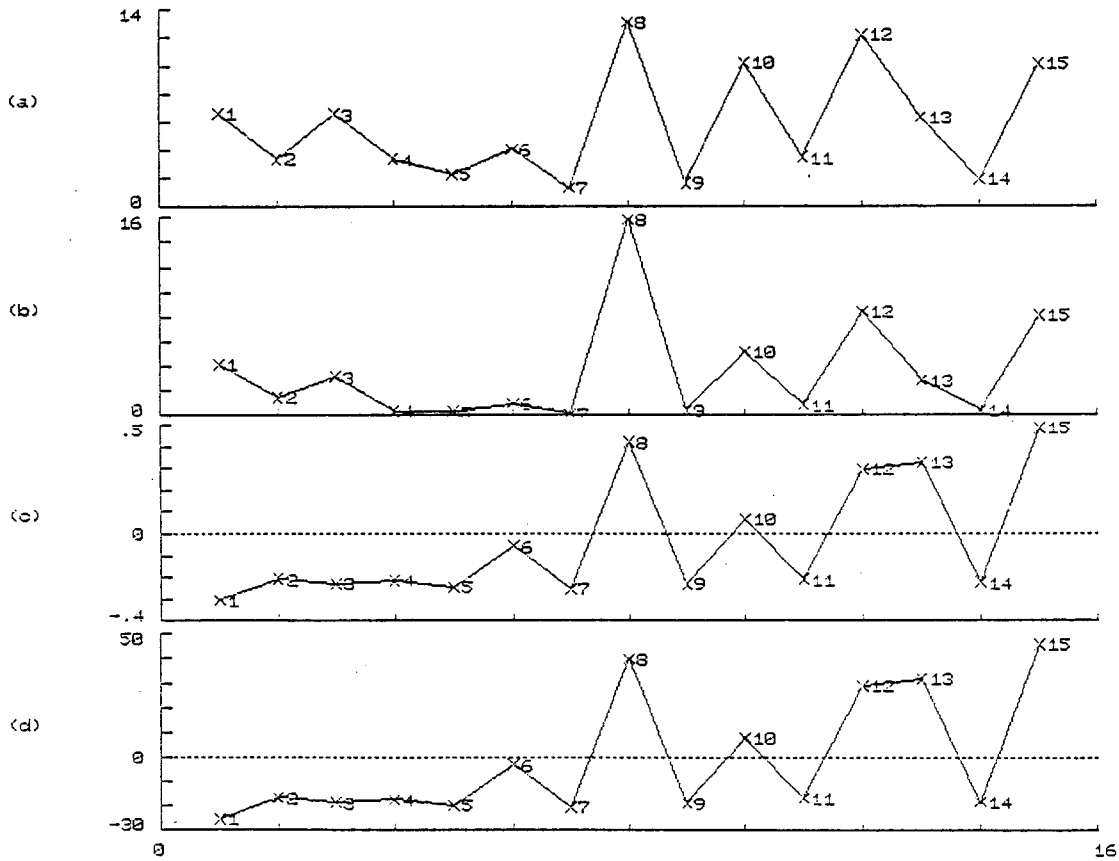

Fig. 3 Index plots of $\|\hat{\beta}^{*(1)}\|$, D^*, $[V(\hat{\beta}^*_{1,1})]^{(1)}$ and $\hat{\sigma}^{2^{(1)}}$ (Hill's data) :
(a) $\|\hat{\beta}^{*(1)}\|$, (b) D^*, (c) $[V(\hat{\beta}^*_{1,1})]^{(1)}$, (d) $\hat{\sigma}^{2^{(1)}}$

From this Fig. 3, we can see that the patterns of influence are very similar among four measures.

References

1. Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example.* John Wiley & Sons, New York.
2. Hill, R.W. (1977). Robust Regression when there are Outliers in the Carriers. Unpublished Ph.D. dissertation, Harvard University, Dept. of Statistics.
3. Longley, J.W. (1967). An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User. *J. Amer. Statist. Assoc.*, **62**, 819-841.
4. Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection.* John Wiley & Sons, New York.
5. Shin, J.K., Tarumi, T. and Tanaka, Y. (1989). Sensitivity Analysis in Principal Component Regression. *Bulletin of the Biometric Soc. of Japan.*, **10**, 57-68.
6. Tanaka, Y. (1988). Sensitivity Analysis in Principal Component Analysis: Influence on the Subspace spanned by Principal Components. *Comm. Statist.* A **17**, 3157-3175.
7. Tanaka, Y. (1989). Influence Functions related to Eigenvalue Problems which appear in Multivariate Methods. *Comm. Stastist.* A **18**, 3991-4010.
8. Weisberg, S. (1980). *Applied Linear Regression.* John Wiley & Sons, New York.