한국통계학회 논문집 제1권 1호
1994년 12월 pp. 102-111

# Sensitivity Analysis in Latent Root Regression

Jae Kyoung Shin[1], Tomoyuki Tarumi[2] and Yutaka Tanaka[3]

## Abstract

We propose a method of sensitivity analysis in latent root regression analysis (LRRA). For this purpose we derive the quantities $\hat{\beta}_{LRR}^{(1)}$, which correspond to the theoretical influence function $I(x, y; \hat{\beta}_{LRR})$ for the regression coefficient $\hat{\beta}_{LRR}$ based on LRRA. We give a numerical example for illustration and also investigate numerically the relationship between the estimated values of $\hat{\beta}_{LRR}^{(1)}$ with the values of the other measures called sample influence curve(SIC) based on the recomputation for the data with a single observation deleted.We also discuss the comparision among the results of LRRA, ordinary least square regression analysis (OLSRA) and ridge regression analysis(RRA).

## 1. Introduction

We may consider that a statistical method is a system, a set of data is an input and the result of analysis is an output. We are interested in the sensitivity of this system, that is, how a small change of data(input) affects the result of analysis(output). The detection of influential observations in regression analysis is an example of sensitivity analysis.

Radhakrishnan and Kshirsagar(1981), Critchley(1985), Jolliffe(1986) and Pack,Jolliffe and Morgan(1987) disscussed sensitivity analysis in principal component analysis(PCA). The essential part of their approaches was to compute the influence functions for eigen- values and eigenvectors derived from the perturbation theory of eigenvalue problems. Tanaka(1987) also discussed sensitivity analysis in PCA. He gave explicitly some influence functions which represent the influence on the subspace spanned by a specified set of eigenvectors. Walker and Birch(1988) discussed sensitivity analysis in Ridge regression analysis(RRA) and the same topic in principal component regression analysis (PCRA) was discussed by Shin,Tarumi and Tanaka(1989). In this paper we propose a method of sensitivity analysis in LRRA.

1) Dept. of Statistics, Kyungpook National University, Taegu, 702-701, KOREA
2) Dept. of Statistics, Okayama University, Tsushima, Okayama 700, JAPAN
3) Dept. of Statistics, Okayama University, Tsushima, Okayama 700, JAPAN

## 2. Latent Root Regression

In latent root regression(LRR) we apply PCA to all of the dependent and independent variables. Here, as in principal component regression(PCR) we consider PCA based on the correlation matrix. Here the principal components corresponding to the smallest eigenvalues are examined, and those for which the coefficient of the dependent variable, $y$, is also small are called *non-predictive multicollinearities*, and are deemed to be of no use in predicting $y$.

Now we consider an ordinary regression model

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

where $\boldsymbol{y}$ is an *(nx1)* vector of dependent variable, $\boldsymbol{1}$ is an *(nx1)* vector whose elements are all *1's*, $X$ is an *(nxp)* matrix of standardized ( $\sum_i^n X_{ij} = 0, \sum_i^n X_{ij}^2 = 1, j = 1, \cdots, p$) independent variables, $\varepsilon$ is an *(nx1)* vector of error terms and let $Y_i^* = (Y_i - \overline{Y})/\eta$, where $\eta^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2$. Define the matrix $A = [\mathbf{Y}^* : \boldsymbol{X}]$, *i.e.* the *(nx(p+1))* matrix of standardized dependent and independent variables. Then $A^TA$ is the "correlation matrix" of dependent and independent variables. Denote the element of the *j-th* latent vector by $\gamma_j^T = (\gamma_{0j}\gamma_{1j}\cdots\gamma_{pj})$ and let $\gamma_j^{0T} = (\gamma_{1j}\gamma_{2j}\cdots\gamma_{pj})$, *i.e.* $\gamma_j^0$ contains all the elements of $\gamma_j$ except for the first one. Finally, let $\Gamma = (\gamma_0\gamma_1\cdots\gamma_p)$ and $\Lambda = diag(\lambda_0, \lambda_1, \cdots, \lambda_p)$.

The ordinary least square regression(OLSR) estimator of $\beta$ can be written in terms of the latent roots and latent vectors of $A^TA$ as

$$\hat{\beta} = -\eta \sum_{j=0}^{p} (a_j \gamma_j^0), \tag{2}$$

where

$$a_j = \gamma_{oj}\lambda_j^{-1} / (\sum_{r=0}^{p} \gamma_{or}^2/\lambda_r), \qquad j = 0, \cdots, p, \tag{3}$$

and the residual sum of squares is given by

$$SSE=(Y-\hat{Y})^T(Y-\hat{Y})=\eta^2(\sum_{j=0}^{p}\gamma_{oj}^2/\lambda_j)^{-1}.$$

We suppose that $q$ latent vectors $\gamma_0,\gamma_1,\cdots,\gamma_{q-1}$ correspond to non-predictive multicollinearities. The above OLSR estimator can be modified by setting $a_0=a_1=\cdots=a_{q-1}=0$. This modified estimator was discussed in Webster, *et al.*(1974) and Hawkins(1973). It is expressed as

$$\hat{\beta}_{LRR}=-\eta\sum_{j=q}^{p}a_j\gamma_j^0, \tag{4}$$

where

$$a_j=\gamma_{0j}\lambda_j^{-1}/(\sum_{r=q}^{p}\gamma_{or}^2/\lambda_r), \qquad j=q,q+1,\cdots,p. \tag{5}$$

In this case the residual sum of squares is given by

$$SSE_{LRR}=(Y-\hat{Y})^T(Y-\hat{Y})=\eta^2(\sum_{j=q}^{p}\gamma_{oj}^2/\lambda_j)^{-1}.$$

## 3. Sensitivity Analysis

It is obvious that quantities appeared in equations (2) and (3) are the functionals of the joint distribution function $F$ of $y$ and $\boldsymbol{x}$.

In LRR the influence function $I(\mathrm{x},y;\hat{\beta}_{LRR})$, $\hat{\beta}_{LRR}^{(1)}$, can be derived as follows.

$$\hat{\beta}_{LRR}^{(1)}=(\sum_{j=q}^{p}f_j\gamma_j^0)^{(1)}=\sum_{j=q}^{p}(f_j^{(1)}\gamma_j^0+f_j\gamma_j^{0(1)}), \tag{6}$$

where $f_j=-\eta a_j$ and

$$\begin{aligned} f_j^{(1)}=-\ & (\gamma_{oj}^{(1)}\eta\lambda_j^{-1}+\gamma_{0j}\eta^{(1)}\lambda_j^{-1}+\gamma_{0j}\eta\lambda_j^{-1^{(1)}})(\sum_{j=q}^{p}\gamma_{0j}^2\lambda_j^{-1})^{-1} \\ &+\gamma_{0j}\eta\lambda_j^{-1}\sum_{j=q}^{p}(\gamma_{0j}^{2^{(1)}}\lambda_j^{-1}+\gamma_{0j}^2\lambda_j^{-1^{(1)}})(\sum_{j=q}^{p}\gamma_{0j}^2\lambda_j^{-1})^{-2}. \end{aligned} \tag{7}$$

The quantities $\gamma_j^{0^{(1)}}$, $\gamma_{oj}^{(1)}$, $\gamma_{oj}^{2^{(1)}}$, $\lambda_j^{(1)}$, and $\lambda_j^{-1^{(1)}}$ are obtained by using the influence func- tions

for the eigenvalues and eigenvectors(see Tanaka(1988)).

By using the influence function $\hat{\beta}_{LRR}^{(1)}$ we can evaluate the influence of a small perturbation of data. However, as in PCR, we consider two, scalar-valued summary statistics, one is the Euclidean norm of a vector $\hat{\beta}_{LRR}^{(1)}$, *i.e.* $\| \hat{\beta}_{LRR}^{(1)} \|$, the other is a statistic similar to Cook's D in regression diagnostic, *i.e.*

$$D^* = [\hat{\beta}_{LRR}^{(1)}]^T \widehat{Var}(\hat{\beta}_{LRR})^{-1} [\hat{\beta}_{LRR}^{(1)}] / \hat{\sigma}^2, \tag{8}$$

where $\widehat{Var}(\hat{\beta}_{LRR})$ is an estimate of the variance-covariance matrix of an estimator $\hat{\beta}_{LRR}$.

## 4. Numerical Example

We analyze the EEO data, which was collected from a cross section of school districts throughout the country(Chatterjee and Price; 1977). The data consists of 70 observations (schools) of three independents and one dependent variables.

Table 1 Correlation matrix of EEO data

| | $Y$ | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|---|
| $Y$ | 1.0000 | | | |
| $X_1$ | 0.4195 | 1.0000 | | |
| $X_2$ | 0.4398 | 0.9601 | 1.0000 | |
| $X_3$ | 0.4181 | 0.9857 | 0.9822 | 1.0000 |

First, we applied PCA based on the dependent and independent variables. The correlation matrix and the latent vectors of $A^TA$ are shown in Table 1 and 2, respectively. From Table 2 we can see that latent vectors $\gamma_0$ and $\gamma_1$ correspond to non-predictive multicollinearities, because the values of $\lambda_0$(=0.0078) and $\lambda_1$(=0.0397) are small and the magnitudes of $|\gamma_{00}|$ (=0.0141) and $|\gamma_{01}|$ (=0.0200) are also small(cut-off value : $\lambda_j < 0.05$ and $|\gamma_{0j}| < 0.1$). Then we calculated the LRR replacing "q" by "2" in (4). The coefficient vector ( $\hat{\beta}_{LRR}$ ) and the standardized coefficient vector ( $\hat{\beta}^*_{LRR}$ ) are given in Table 3.

Table 2 The result of PCA based on $\Gamma_{AA}$

| | $Z_0$ | $Z_1$ | $Z_2$ | $Z_3$ |
|---|---|---|---|---|
| $Y$ | 0.0141 | 0.0200 | 0.9478 | 0.3179 |
| $X_1$ | -0.4518 | 0.6795 | -0.1906 | 0.5457 |
| $X_2$ | -0.3718 | -0.7321 | -0.1626 | 0.5471 |
| $X_3$ | 0.8109 | 0.0426 | -0.1972 | 0.5493 |
| Eigen val. | 0.0078 | 0.0397 | 0.7532 | 3.1993 |
| Prop. | 0.0020 | 0.0099 | 0.1883 | 0.7998 |
| C.Prop. | 0.0020 | 0.0119 | 0.2002 | 1.0000 |

Table 3 The result of LRR

| Variable | $\hat{\beta}_{LRR}$ | $\hat{\beta}^*_{LRR}$ |
|---|---|---|
| $X_1$ | 2.8630 | 0.1517 |
| $X_2$ | 2.3162 | 0.1227 |
| $X_3$ | 2.9855 | 0.1582 |

Next, to investigate the influence of each individual on the result we calculated the empirical influence curves $(EIC_i)$ $\beta^{(1)}_{LRR_i}$. The index plots of $\beta^{*(1)}_{LRR_i}$'s and $\| \beta^{*(1)}_{LRR_i} \|$ are shown in Figure 1.

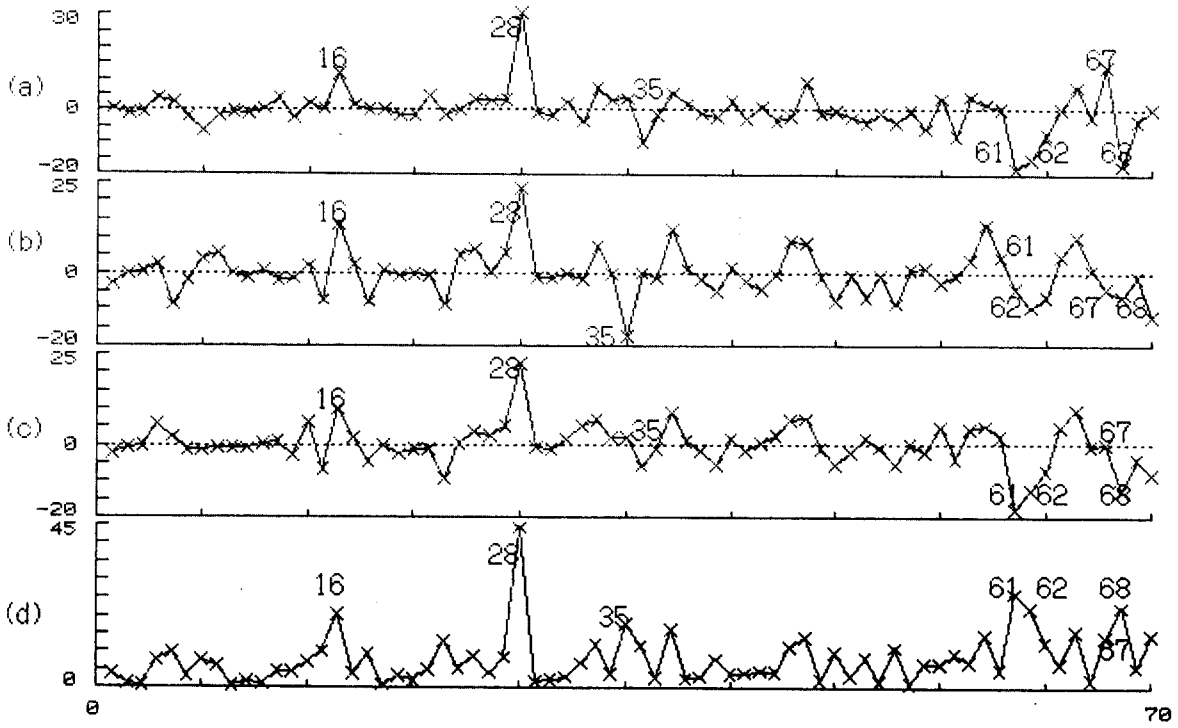


Figure 1 Index plots of $\beta^{*(1)}_{LRR_i}$'s and $\| \beta^{*(1)}_{LRR} \|$ (EEO data)

((a)variable 1, (b)variable 2, (c)variable 3, (d)norm)

The inspection of figure suggest that the individual number 28 is the most influential among the all individuals. Table 4 shows the result of LRR for the data with the 28th individual omitted.

Table 4 The result of LRR for the full and reduced data

| Variable | all data | | No.28 omitted | |
|---|---|---|---|---|
| | $\hat{\beta}_{LRR}$ | $\hat{\beta}^{*}_{LRR}$ | $\hat{\beta}_{LRR(28)}$ | $\hat{\beta}^{*}_{LRR(28)}$ |
| $X_1$ | 2.8630 | 0.1517 | 2.3929 | 0.1355 |
| $X_2$ | 2.3162 | 0.1227 | 1.9438 | 0.1101 |
| $X_3$ | 2.9855 | 0.1582 | 2.6272 | 0.1488 |

## 5. Discussion

First, we study the validity of the proposed quantity $\beta^{(1)}_{LRR}$. For this purpose, we investigate the relationship between $\hat{\beta}^{(1)}_{LRR_i}(EIC_i)$ and the so-called $SIC_i$(sample influence curve) defined as follows.

$$SIC_i(\beta_{LRR}) = -(n-1)(\hat{\beta}_{LRR_i} - \hat{\beta}_{LRR}), \qquad i=1,\cdots,n,$$

where $\hat{\beta}_{LRR_i}$ and $\hat{\beta}_{LRR}$ indicate the estimates of the standardized coefficient $\beta^*$ based on the whole sample and the sample without the $i-th$ observation, respectively. It is obvious that $SIC$ can be interpreted very clearly but it requires high computing cost. Figure 2 shows the scatter diagram of $\hat{\beta}_{LRR_i}$ and $\| \beta^{(1)}_{LRR} \|$ based on $EIC_i(\beta_{LRR})$ and $SIC_i$ $(\beta_{LRR})$. In these scatter diagram, most of the points are located near the straightline $SIC = EIC$. We can observe the similar tendency for the other variables. These scatter diagrams indicate that the quantities $EIC_i$'s can be used practically instead of $SIC_i$'s.

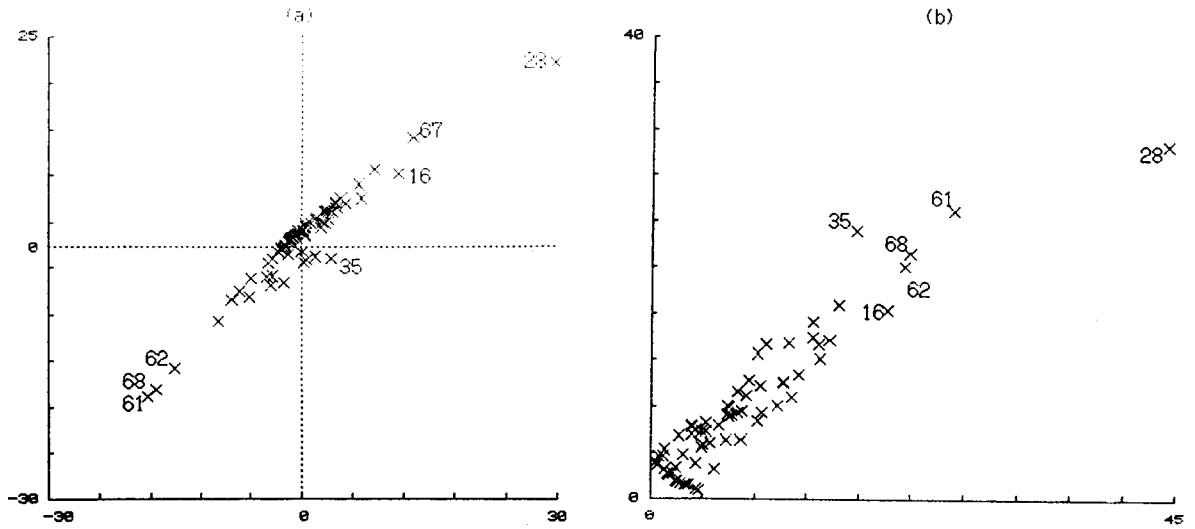


Figure 2 Scatter diagrams of $EIC$(horizontal) versus $SIC$(vertical) for $\hat{\beta}^{(1)}_{LRR_i}$ and $\| \hat{\beta}^{(1)}_{LRR} \|$ ((a) $\hat{\beta}^{(1)}_{LRR_i}$,(b) $\| \hat{\beta}^{(1)}_{LRR} \|$ )

Belsley, Kuh and Welsch(1980), Cook and Weisberg(1982), Chatterjee and Hadi(1988) among others discussed various methods of sensitivity analysis in ordinary multiple regression. Recently, Walker and Birch(1988), Shin, Tarumi and Tanaka(1989) proposed some influence measures with respect to RRA and PCRA, respectively. We are interested in whether these three methods, *i.e.* OLSR, PCR and LRR, gave similar results or not in sensitivity analysis. For this purpose we also applied OLSR with sensitivity analysis procedures to the data shown in Shin,Tarumi and Tanaka(1989).

Figure 3 shows the index plots of Euclidean norm for these three methods. For OLSR we show two cases. One is the result for the full model and the other is that for the reduced model obtained from the backward elimination technique with $F_{out}=2.0$. The estimated regression equations based on the above methods are given in Table 5. If we compare the values of Euclidean norm, the magnitude is OLSR(full), OLSR(reduced), LRR and PCR in descending order. The patterns of the influence are very similar among OLSR(full) and OLSR(reduced), but their patterns are quite different from that of PCR and LRR.

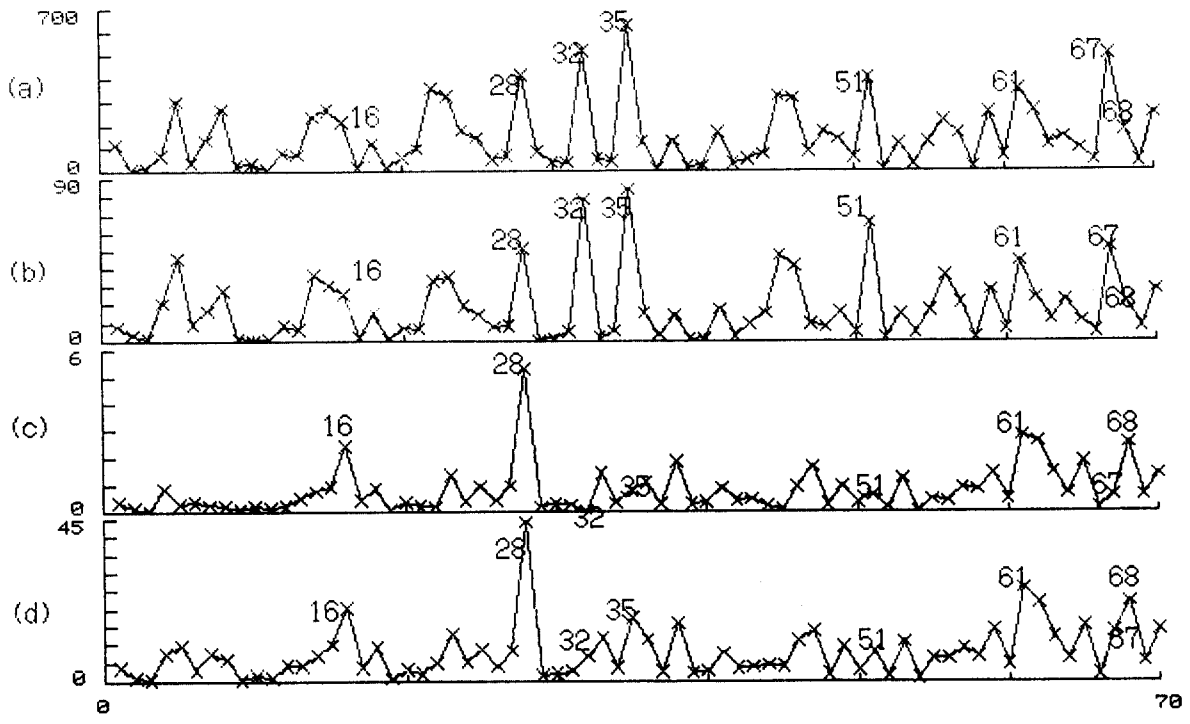


Figure 3 Index plots of Euclidean norm for OLSR(full and reduced model), PCR and LRR
((a)OLSR(full), (b)OLSR(reduced), (c)PCR, (d)LRR)

The EEO data, which was illustrated as a numerical example in the preceding section, can be also analyzed by PCR. If we do so, the eigenvalues are ordered so that $\lambda_1 = 2.9520 > 0.0401 > 0.0080 = \lambda_3$. Thus it may be reasonable to select the first one PC and calculate the PCR. The result of PCR is given in Table 5. On the contrary, we may apply LRR to the Hill's data which was analyzed in Shin, Tarumi and Tanaka(1989). We can not find latent vectors corresponding to non-predictive multicollinearities, because for the only one eigenvalue ($\lambda_0 = 0.0182$) which is smaller than cut-off value ($\lambda_j = 0.05$), coefficient $\gamma_{00}$ for the dependent variable of the latent vector is greater than cut-off value ($\gamma_{00} = 0.2543 > 0.1$).

Table 5 The estimated regression equations based on LRR, OLSR(full),OLSR(reduced) and PCR

| Variable | LRR (q=2) | OLSR (full) | OLSR (reduced) | PCR (pc=1) |
|---|---|---|---|---|
| 1 | 0.1517 | 0.5247 | *** | -0.1195 |
| 2 | 0.1227 | 0.9449 | 0.4398 | 0.4276 |
| 3 | 0.1582 | -1.0273 | *** | 0.1251 |

## References

[1] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics : Identifying influential data and sources of collinearity.* John Wiley & Sons, New York.

[2] Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity Analysis in Linear Regression.* John Wiley & Sons, New York.

[3] Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example.* John Wiley & Sons, New York.

[4] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall.

[5] Critchley, F.(1985). Influence in principal component analysis. *Biometrika*,72,627-636.

[6] Hill, R.W. (1977). *Robust Regression When There are Outliers in the Carriers.*Unpublished Ph.D. dissertation, Harvard University, Dept. of Statistics.

[7] Hoerl, A.E., Kennard, R.W. and Baldwin, K.F.(1975). Ridge regression:Some simulations. *Communications in Statistics*, A 4, 105-123.

[8] Jolliffe, I. T. (1986). *Principal Component Analysis.* Springer-Verlag.

[9] Pack, P., Jolliffe, I.T. and Morgan, B.J.T.(1987). Influential observations in principal component analysis. : A case study. *A draft for publication in it Appl. Statist.*

[9] Radhakrishnan, R. and Kshirsagar, A.M. (1981). Influence functions for certain parameters in multivariate analysis. *Communications in Statistics*, A 10, 515-529.

[10] 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.

[11] Tanaka, Y. (1988). Sensitivity analysis in principal component analysis : Influence on the subspace spanned by principal components. *Communications in Statistics*, A 17, 3157-3175.

[12] Tanaka, Y. (1989). Influence functions related to eigenvalue problems which appear in multivariate methods. *Communications in Statistics*, A 18, 3991-4010.

[13] Walker, E. and Birch, J.B. (1988). Influence measure in ridge regression.*Technometrics*, 30, 221-227.