Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.3 pp. 623~630

Sensitivity Analysis in Principal Component Regression with Quadratic Approximation[1)]

Jae-Kyoung Shin[2)] · Duk-Joon Chang[3)]

Abstract

Recently, Tanaka(1988) derived two influence functions related to an eigenvalue problem $(A-\lambda_s I)\,v_s = 0$ of real symmetric matrix A and used them for sensitivity analysis in principal component analysis. In this paper, we deal with the perturbation expansions up to quadratic terms of the same functions and discuss the application to sensitivity analysis in principal component regression analysis(PCRA). Numerical example is given to show how the approximation improves with the quadratic term.

Key words : influence function, quadratic term, perturbation theory of eigenvalue problems, principal component regression

1. Introduction

Many statisticians discussed sensitivity analysis in ordinary multiple regression, and some authors including Pregibons(1981), Williams(1987), Walker and Birch (1988), Shin et al.(1989, 1990, 1994) discussed the same topic in other types of regression such as logistic regression, generalized linear model, ridge type regression, principal component regression, latent root regression and logistic principal component regression. Radhakrishnan and Kshirsagar(1981), Tanaka (1984), Critchley(1985), Jolliffe(1986), Pack, Jolliffe and Morgan(1988), Bénasséni(1988) and others discussed sensitivity analysis in principal component analysis and related multivariate methods. The essential part of their approaches to compute the influence functions for eigenvalues and eigenvectors derived from

---

1) This research is financially supported by Changwon National University in 2001.

2) Associate Professor, Dept. of Statistics, Changwon National University, Changwon, 641-773, Korea.
E-mail : jkshin@sarim.changwon.ac.kr

3) Professor, Dept. of Statistics, Changwon National University, Changwon, 641-773, Korea.

the perturbation theory of eigenvalue problems. Tanaka(1988, 1989) has derived explicitly some influence functions, considering the influence on the subspace spanned by a specified set of eigenvectors. In this paper we derive quadratic perturbation expansions for two functions of eigenvalues and eigenvectors, and apply them to sensitivity analysis in PCRA.

In section 2 we briefly review the perturbation theory of eigenvalue problems and show the expansions up to the quadratic terms for the eigenvalues and eigenvectors. In section 3 for two functions of eigenvalues and eigenvectors, after reviewing the linear expansions obtained by Tanaka(1988), we derive the corresponding quadratic terms. In section 4 we treat their application to sensitivity analysis in PCRA. Numerical example is given for this method to show the usefulness of the quadratic term.

2. Influence functions related to eigenvalue problems

We consider an eigenvalue problem such as

$$(A-\lambda_s I)\,\boldsymbol{v}_s=\boldsymbol{0}, \tag{2.1}$$

where A is a $p\times p$ real symmetric matrix, λ_s is the s-th eigenvalue and $\boldsymbol{v}_s$ is the associated eigenvector. We consider a small perturbation in this eigenvalue problem, as follows :

$$A \mapsto A(\varepsilon)=A+\varepsilon A^{(1)}+(\varepsilon^2/2)A^{(2)}+O(\varepsilon^3). \tag{2.2}$$

Then the eigenvalues and eigenvectors can be expanded as

$$\lambda_s(\varepsilon)=\lambda_s+\varepsilon\lambda_s^{(1)}+(\varepsilon^2/2)\lambda_s^{(2)}+O(\varepsilon^3), \tag{2.3}$$

$$\boldsymbol{v}_s(\varepsilon)=\boldsymbol{v}_s+\varepsilon\,\boldsymbol{v}_s^{(1)}+(\varepsilon^2/2)\,\boldsymbol{v}_s^{(2)}+O(\varepsilon^3). \tag{2.4}$$

Here, we suppose, without any loss of generality, that we are interested in the first q eigenvalues ($q\le p$). Then we have the following formulas(see, Tanaka, 1984).

$$\begin{cases} \lambda_s^{(1)}=a_{ss}^{(1)}, \\ \boldsymbol{v}_s^{(1)}=\sum_{r\neq s}(\lambda_s-\lambda_r)^{-1}a_{rs}^{(1)}\boldsymbol{v}_r, \\ \lambda_s^{(2)}=a_{ss}^{(2)}+2\sum_{r\neq s}(\lambda_s-\lambda_r)^{-1}(a_{rs}^{(1)})^2, \\ \boldsymbol{v}_s^{(2)}=\sum_{r\neq s}(\lambda_s-\lambda_r)^{-1}\{a_{rs}^{(2)} \\ \qquad +2\sum_{t\neq s}(\lambda_s-\lambda_t)^{-1}a_{st}^{(1)}(a_{rt}^{(1)}-a_{ss}^{(1)}\delta_{rt})\}\boldsymbol{v}_r-\|\boldsymbol{v}_s^{(1)}\|^2\boldsymbol{v}_s, \end{cases} \tag{2.5}$$

where

$$a_{rs}^{(k)}=\boldsymbol{v}_r^T A^{(k)}\boldsymbol{v}_s, \qquad k=1,2, \tag{2.6}$$

and δ_{rt} the Kronecker delta. Note that the right-hand sides from the second to

the fourth equations of (2.5) contain the terms $(\lambda_s-\lambda_r)^{-1}$, $r=1,\cdots,p$, $r\neq s$, and then these formulas become numerically inappropriate when there exist some $\lambda_r's$ which are nearly equal to λ_s.

3. Quadratic expansion of the functions of eigenvalues and eigenvectors

Shin et al.(1989) studied sensitivity analysis in PCRA. There the following matrix, which is function of eigenvalues and eigenvectors, plays an important role.

$$R=\sum_{s=1}^{q}\lambda_s^{-1}\boldsymbol{v}_s\boldsymbol{v}_s^T. \tag{3.1}$$

We shall consider the perturbation expansion of this matrix, and try to derive the quadratic expansion

$$R(\varepsilon)=R+\varepsilon R^{(1)}+(\varepsilon^2/2)R^{(2)}+O(\varepsilon^3), \tag{3.2}$$

corresponding to the perturbation given by (2.2)

The coefficient $R^{(1)}$, which is equivalent to the influence function, was already obtained by Shin et al.(1989) as

$$\begin{aligned}R^{(1)}=&-\sum_{s=1}^{q}\sum_{r=1}^{q}\lambda_s^{-1}\lambda_r^{-1}(\boldsymbol{v}_s^TA^{(1)}\boldsymbol{v}_r)\boldsymbol{v}_s\boldsymbol{v}_r^T\\&+\sum_{s=1}^{q}\sum_{r=q+1}^{p}\lambda_s^{-1}(\lambda_s-\lambda_r)^{-1}(\boldsymbol{v}_s^TA^{(1)}\boldsymbol{v}_r)(\boldsymbol{v}_s\boldsymbol{v}_r^T+\boldsymbol{v}_r\boldsymbol{v}_s^T).\end{aligned} \tag{3.3}$$

Now we shall derive the quadratic terms. By means of (2.5) the second derivative can be formulated as

$$\begin{aligned}R^{(2)}&=\Big(\sum_{s=1}^{q}\lambda_s^{-1}\boldsymbol{v}_s\boldsymbol{v}_s^T\Big)^{(2)}\\&=\sum_{s=1}^{q}\Big\{(\lambda_s^{-1})^{(2)}\boldsymbol{v}_s\boldsymbol{v}_s^T+2(\lambda_s^{-1})^{(1)}(\boldsymbol{v}_s^{(1)}\boldsymbol{v}_s^T+\boldsymbol{v}_s\boldsymbol{v}_s^{(1)^T})\\&\quad+2\lambda_s^{-1}\boldsymbol{v}_s^{(1)}\boldsymbol{v}_s^{(1)^T}+\lambda_s^{-1}(\boldsymbol{v}_s^{(2)}\boldsymbol{v}_s^T+\boldsymbol{v}_s\boldsymbol{v}_s^{(2)^T})\Big\}\\&=\sum_{s=1}^{q}\Big\{(2\lambda_s^{-3}(\lambda_s^{(1)})^2-\lambda_s^{-2}\lambda_s^{(2)})\boldsymbol{v}_s\boldsymbol{v}_s^T-2\lambda_s^{-2}\lambda_s^{(1)}(\boldsymbol{v}_s^{(1)}\boldsymbol{v}_s^T+\boldsymbol{v}_s\boldsymbol{v}_s^{(1)^T})\\&\quad+2\lambda_s^{-1}\boldsymbol{v}_s^{(1)}\boldsymbol{v}_s^{(1)^T}+\lambda_s^{-1}(\boldsymbol{v}_s^{(2)}\boldsymbol{v}_s^T+\boldsymbol{v}_s\boldsymbol{v}_s^{(2)^T})\Big\},\end{aligned} \tag{3.4}$$

where the superscript (2) indicates the second derivative.

4. Application to Sensitivity Analysis in Principal Component Regression Analysis

4.1 Principal Component Regression

We consider ordinary regression model, which is expressed as

$$\boldsymbol{y} = \mathbf{1}\beta_0 + X\beta + \varepsilon, \qquad \varepsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}), \tag{4.1}$$

where $\boldsymbol{y}$ is an $(n\times 1)$ vector of the dependent variable, $\mathbf{1}$ is an $(n\times 1)$ vector whose element 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.* $\mu_x =$ the mean vector of $\boldsymbol{x}$, $\Phi_{xx} =$ the covariance matrix of $\boldsymbol{x}$, $\Phi_{xy} =$ the covariance matrix between $\boldsymbol{x}$ and y, etc.

The correlation matrix is decomposed as

$$\Gamma_{xx} = (\Phi_{xx})_D^{-\frac{1}{2}} \Phi_{xx} (\Phi_{xx})_D^{-\frac{1}{2}} = V_1 \Lambda_1 V_1^T + V_2 \Lambda_2 V_2^T, \tag{4.2}$$

by using the spectral decomposition, where subscript 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 q principal components which we are interested in can be expressed by

$$\boldsymbol{z} = V_1^T (\Phi_{xx})_D^{-\frac{1}{2}} (\boldsymbol{x} - \mu_x).$$

Consider the regression of $\boldsymbol{y}$ on $(\mathbf{1}, \boldsymbol{z}^T)$. Using the ordinary least square method, the coefficients become

$$\begin{pmatrix} \alpha_0 \\ \alpha \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \Lambda_1 \end{pmatrix}^{-1} \begin{pmatrix} \mu_y \\ V_1^T (\Phi_{xx})_D^{-\frac{1}{2}} \Phi_{xy} \end{pmatrix} = \begin{pmatrix} \mu_y \\ \Lambda_1^{-1} V_1^T (\Phi_{xx})_D^{-\frac{1}{2}} \Phi_{xy} \end{pmatrix}.$$

The coefficient vector for the standardized original variables $\boldsymbol{x}^* = (\Phi_{xx})_D^{-\frac{1}{2}} (\boldsymbol{x} - \mu_x)$ is obtained as

$$\beta^* = V_1 \Lambda_1^{-1} V_1^T (\Phi_{xx})_D^{-\frac{1}{2}} \Phi_{xy}. \tag{4.3}$$

4.2 Sensitivity Analysis

In the preceding section, quadratic perturbation expansions was derived for the function of eigenvalues and eigenvectors of a real symmetric matrix.

Using the influence function $R^{(1)}$based on the linear perturbation expansion Shin et al.(1989) proposed a method of sensitivity analysis in PCRA. Here we apply the quadratic expansion derived in the preceding section to PCRA and investigate how the approximation improves.

Consider PCRA based on a $p\times p$ correlation matrix Γ. Let $\lambda_1,\cdots,\lambda_p$ $(\lambda_1\geq\cdots\geq\lambda_p)$ be the eigenvalues and $\boldsymbol{v}_1,\cdots,\boldsymbol{v}_p$ be the associated eigenvectors of Γ, and suppose we are interested in the largest q eigenvalues. Then the matrix R defined by (3.1) indicates the part corresponding to the q eigenvalues of interest in the spectral decomposition of Γ.

First we study the influence of a small change of data on the standard regression coefficient vector β^*. Taking the first derivatives of the both sides of (4.3) with respect to ε, we obtain

$$\beta^{*(1)} = (V_1\Lambda_1^{-1}V_1^T)^{(1)}(\Phi_{xx})_D^{-\frac{1}{2}}\Phi_{xy} + V_1\Lambda_1^{-1}V_1^T\left\{(\Phi_{xx})_D^{-\frac{1}{2}}\right\}^{(1)}\Phi_{xy} + V_1\Lambda_1^{-1}V_1^T(\Phi_{xx})_D^{-\frac{1}{2}}\Phi_{xy}^{(1)}, \tag{4.4}$$

where the superscript (1) indicates the first derivative or influence function. The quantity $(V_1\Lambda_1^{-1}V_1^T)^{(1)}$ in the right hand side can be calculated as follows.

$$(V_1\Lambda_1^{-1}V_1^T)^{(1)} = -\sum_{s=1}^{q}\sum_{r=1}^{q}\lambda_s^{-1}\lambda_r^{-1}[\boldsymbol{v}_s^T\Gamma_{xx}^{(1)}\boldsymbol{v}_r]\boldsymbol{v}_s\boldsymbol{v}_r^T + \sum_{s=1}^{q}\sum_{r=q+1}^{p}\lambda_s^{-1}(\lambda_s-\lambda_r)^{-1}[\boldsymbol{v}_s^T\Gamma_{xx}^{(1)}\boldsymbol{v}_r](\boldsymbol{v}_s\boldsymbol{v}_r^T+\boldsymbol{v}_r\boldsymbol{v}_s^T). \tag{4.5}$$

As in Tanaka(1988), he introduced a perturbation

$$F\mapsto\tilde{F}=(1-\varepsilon)F+\varepsilon\delta_x, \tag{4.6}$$

where F is the unperturbed cdf and δ_x is the cdf of a unit point mass at $\boldsymbol{x}$, and try to evalute the corresponding changes from R to $R(\varepsilon)$.

From the preceding section, $R(\varepsilon)$ can be expanded as the power series (3.2) with the coefficient given by (3.3) through (3.4), where $a_{rs}^{(k)}$ is defined as (2.6) with $H^{(1)}$ and $H^{(2)}$ replaced by the coefficients $\Sigma^{(1)}$ and $\Sigma^{(2)}$ of the linear and quadratic terms of the expansion of the perturbed covariance matrix $\Sigma(\varepsilon)$. As shown in Critchley(1985), $\Sigma(\varepsilon)$ is expanded as follows.

$$\Sigma(\varepsilon)=\Sigma+\varepsilon\Sigma^{(1)}+(\varepsilon^2/2)\Sigma^{(2)}, \tag{4.7}$$

where

$$\Sigma^{(1)} = (x-\mu)(x-\mu)^T - \Sigma, \tag{4.8}$$

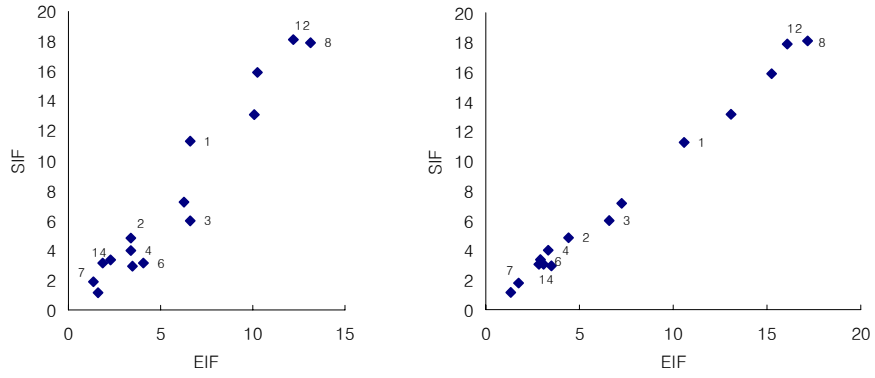
$$\Sigma^{(2)} = -2(x-\mu)(x-\mu)^T. \tag{4.9}$$

The formulation above is based on the theoretical cdf F and contains population parameters μ and Σ. If our concern is, for instance, the change produced by the omission of a certain x_i, the theoretical cdf F and the population parameters μ and Σ should be replaced by the empirical cdf $\tilde{F}$ and the sample counterparts $\bar{x}$ and S, respectively, and the perturbation parameter ε should be set $\varepsilon = -1/(n-1)$.

5. Numerical Performances and Discussion

To illustrate our procedure we apply our method of sensitivity analysis to the Hill's data(1977), which was analyzed by Walker and Birch(1988) with Ridge regression. The data set is related to the performance of a computerized system for processing military personnel action forms. There are 15 observations of six independent and one dependent variables.

First, we applied PCA based on the correlation matrix to the independent variables. Then we select the first four PC's, because the remaining eigenvalues were very small($\lambda_1 = 3.80$, $\lambda_2 = 1.06$, $\lambda_3 = 0.62$, $\lambda_4 = 0.43 \gg \lambda_5 = 0.06$, $\lambda_6 = 0.03$).



(i) linear approximation (ii) quadratic approximation

Fig. 1. Exact versus approximate values of $\| (R(\varepsilon) - R)/\varepsilon \|$ (exact : vertical, approximate : horizontal)

Next, we study the usefulness of the proposed quadratic terms. For this purpose, we investigate the relationship between EIF(Empirical Influence Function :

$(V_1 \Lambda_1^{-1} V_1^T)^{(1)}$; first derivative, and $(V_1 \Lambda_1^{-1} V_1^T)^{(2)}$; second derivative) and the so called SIF(Sample Influence Function). By using the influence function, we can evaluate the influence of a small perturbation of data. However, since $\beta^{*(1)}$ is vector-valued, it is useful to construct scalar-valued summary statistics based on it. the simplest way of summarization may be to take the Euclidean norm, *i.e.* $\| \beta^{*(1)} \|$.

Now, let us consider the relative changes $(R(\varepsilon) - R)/\varepsilon$, when each observation is omitted in turn. For this purpose from (3.2), the approximate relative changes based on the linear approximation $R^{(1)}$, and this based on the quadratic approximation $R^{(1)} + (\varepsilon/2) R^{(2)}$, along with the exact values, were computed for each observation. Figure 1 shows the scatter diagram of the Euclidean norms of the exact versus approximate changes. From this figure we can say that, though the linear approximation maybe enough for the purpose to detect influential observations, the approximation improves considerably by taking the quadratic term into account.

References

1. Allen, D.M. (1971). Mean square error of predictings as a criterion for selecting variables. *Technometrics*, Vol. 13, 469-475.
2. Bénnasséni, J. (1988). Sensitivity of principal component analysis to data perturbation. *Data Analysis and Informatics, Edited by Diday, E.,* 303-310.
3. Critchley, F. (1985). Influence in principal component analysis. *Biometrika,* 72, 627-636.
4. Hill, R.W. (1977). Robust regression when there are outliers in the carriers. *Unpublished Ph.D. dissertation, Harvard University, Dept. of Statistics.*
5. Jolliffe, I.T. (1986). *Principal Component Analysis*. Springer-Verlag.
6. Massy, W.F. (1965). Principal components regression in exploratory statistical research. *Journal of American Statistical Association*, 60, 234-256.
7. Pack, P., Jolliffe, I.T. and Morgan, B.J.T. (1988). Influential observations in principal component analysis : A case study. *J. Appl. Statist.,* 15, 37-50.
8. Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics,* 9, 705-724.
9. Radhakrishnan, R. and Kshirsagar, A.M. (1981). Influence functions for certain parameters in multivariate analysis. *Communication in Statistics :*

Theory and Methods, 10, 515-529.
10. 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 and Methods*, 8, No.2, 127-134.
11. 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.
12. 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.
13. Shin, J.K., Tarumi, T. and Tanaka, Y. (1990). Sensitivity analysis in logistic principal component regression. *Statistical Methods and Data Analysis, Edited by N. Niki,* 89-97.
14. Shin, J.K., Tarumi, T. and Tanaka, Y. (1994). Sensitivity analysis in Latent Root regression. *The Korean Communications in Statistics,* Vol. 1, No. 1, 102-111.
15. Tanaka, Y. (1984). Sensitivity analysis in Hayashi's third method of quantification. *Behaviormetrika*, 16, 31-44.
16. 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.)
17. Tanaka, Y. (1989). Influence functions related to eigenvalue problems which appear in multivariate methods. *Communication in Statistics : Theory and Methods*, 18, 3991-4010.
18. Tanaka, Y. (1990). Quadratic Perturbation Expansions of Certain Functions of Eigenvalues and Eigenvectors and Their Application to Sensitivity Analysis in Multivariate Methods. *Communication in Statistics : Theory and Methods*, 19, 2943-2965.
19. Walker, E. and Birch, J.B. (1988). Influence measure in Ridge regression. *Technometrics,* 30, 221-227.
20. Williams, D.A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Appl. Statist.,* 36, 181-191.

[received date : May. 2003, accepted date : Jul. 2003]