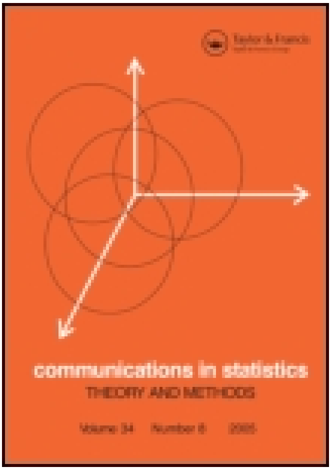

Communications in Statistics - Theory and Methods

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

More on local influence in pllincipal components analysis

Baibing Li [a] & Bart De Moor [a]

[a] ESAT-SISTA, Dept. of Electrical Engineering , Katholieke Universiteit Leuven , Kardinaal Mercierlaan, Leuven, 945 300, Belgium

Published online: 27 Jun 2007.

To cite this article: Baibing Li & Bart De Moor (1999) More on local influence in pllincipal components analysis, Communications in Statistics - Theory and Methods, 28:10, 2487-2495, DOI: 10.1080/03610929908832432

To link to this article: http://dx.doi.org/10.1080/03610929908832432

COMMUN. STATIST.—THEORY METH., 28(10), 2487–2495 (1999)

MORE ON LOCAL INFLUENCE IN PRINCIPAL COMPONENTS ANALYSIS

Baibing Li and Bart De Moor

ESAT-SISTA, Dept. of Electrical Engineering, Katholieke Universiteit Leuven
Kardinaal Mercierlaan 94, 3001 Leuven, Belgium

Key Words: local influence; perturbation scheme; principal components analysis.

ABSTRACT

Local influence on the eigenvalues of sample covariance matrices in principal components analysis is examined for a reasonable modification of Shi's (1997) perturbation scheme. The modification is suggested for samples from populations with both unknown mean vector and covariance matrix. While Shi's detection indexes (1997) consist of only quadratic terms, the modified perturbation scheme leads to detection indexes constituted by both linear and quadratic terms associated with centralized observations. These linear and quadratic terms reflect local influences on the first two sample moments. Examples are investigated based on the two detection indexes.

1. INTRODUCTION TO PERTURBATION SCHEMES

Consider an independent and identically distributed sample $x_1,\ldots,x_n\in R^p$ and its sample covariance matrix S. The purpose of this note is to suggest other detection indexes for local influence on the p distinct eigenvalues λ_j $(j=1,\ldots,p)$ of S. The indexes are produced by modifying the perturbation scheme of Shi (1997). Extensions of this approach to the eigenvectors of S are straightforward.

It is known that principal components analysis is quite sensitive to outliers and influential cases (Huber, 1981; Critchley, 1985; Shi, 1997). To avoid obtaining misleading results, identification of such cases is necessary. For this purpose, Critchley (1985) considered global influence analysis. Recently, another approach, local influence, was investigated by Shi (1997) as an extension from the likelihood approach for local influence analysis (Cook, 1986).

For a sample $x_1,\ldots,x_n\in R^p$ from a population with known mean μ and unknown covariance matrix, Shi (1997) considered following perturbation scheme

$$x_i(\omega)=\omega_i(x_i-\mu) \qquad \text{for} \quad i=1,\ldots,n \qquad (1)$$

with perturbation vector $\omega=[\omega_1,\ldots,\omega_n]^T$ and $\omega_i=1+\varepsilon h_i$ $(i=1,\ldots,n)$. $h=[h_1,\ldots,h_n]^T$ and $\|h\|^2=1$. When the population mean is unknown, Shi (1997) replaced μ by its sample version $\bar{x}=(x_1+\ldots+x_n)/n$ and considered $\bar{x}$ as perturbation-free.

The generalized local influence functions of λ_j are given by Shi (1997) based on the perturbation scheme (1):

$$GIF_S(\lambda_j;\, h)=(2/n)\sum_{i=1}^{n} y_{ij}^2 h_i\ , \quad j=1,\ldots,p$$

where $y_{ij}=(x_i-\bar{x})^T\alpha_j$. α_j is an eigenvector associated with the eigenvalue λ_j $(j=1,\ldots,p)$.

Consequently, $h_{max}(\lambda_j)$ which maximizes $[GIF_S(\lambda_j;h)]^2$ satisfies $h_{max}(\lambda_j) \propto [y_{1j}^2,\dots, y_{nj}^2]^T$. This gives an index $I_S(x_i; \lambda_j)$ used to identify influential cases by plots of $I_S(x_i; \lambda_j)$ against case number:

$$I_S(x_i; \lambda_j)=y_{ij}^2, \qquad i=1,\dots,n \quad \text{and} \quad j=1,\dots,p.$$

In this note, for a sample $x_1,\dots,x_n \in R^p$ from a population with both unknown population mean μ and unknown covariance matrix, we suggest following perturbation scheme

$$x_i(\omega)=\omega_i x_i \qquad \text{for} \quad i=1,\dots,n. \tag{2}$$

The main difference between the two perturbation schemes (1) and (2) is that perturbation characterized by (2) may influence any of the sample moments, while the scheme (1) with the replacement of μ by $\bar{x}$ is a perturbation-free scheme for sample means (i.e. it is assumed that perturbation of each observation case does not affect sample means). We believe that when no prior information is available, the perturbation scheme (2) is more reasonable since in general, minor changes of cases may influence sample means as well as other sample moments.

Similar to Shi's approach, letting $\omega_i=1+\varepsilon h_i$ in (2), we consider the sample covariance matrix from the perturbed data (2) and its eigenvalues $\lambda_j(\omega)$. Local influence functions of λ_j are then given by $[\partial\lambda_j(\omega)/\partial\omega]^T h$ at $\omega=[1,\dots,1]^T$, that is

$$GIF_M(\lambda_j; h)=GIF_S(\lambda_j; h)+(2/n)\sum_{i=1}^{n} v_j y_{ij} h_i , \quad j=1,\dots,p$$

with $v_j=\bar{x}^T\alpha_j$. Consequently, $h_{max}(\lambda_j)$ obtained by maximizing $[GIF_M(\lambda_j; h)]^2$ can be used for detecting locally influential cases. Since

$$h_{max}(\lambda_j) \propto [y_{1j}^2+v_j y_{1j},\dots, y_{nj}^2+v_j y_{nj}]^T$$

it leads to a detection index for eigenvalue λ_j as follows:

$$I_M(x_i; \lambda_j)=y_{ij}^2+v_j y_{ij} , \qquad i=1,\dots,n \quad \text{and} \quad j=1,\dots,p.$$

It is interesting to compare $I_M(x_i;\ \lambda_j)$ and $I_S(x_i;\ \lambda_j)$. $I_M(x_i;\ \lambda_j)=y_{ij}^2+v_jy_{ij}$ consists of both linear and quadratic terms of y_{ij} associated with centralized observation $(x_i-\bar{x})$, while $I_S(x_i;\ \lambda_j)=y_{ij}^2$ has only a quadratic term. $I_M(x_i;\ \lambda_j)$ reflects two types of effects by minor changes of cases: one for sample mean and another for sample second moment about the origin. The added effect depends on their relative magnitudes and signs. In general, for a small linear term v_jy_{ij}, $I_M(x_i;\ \lambda_j)$ and $I_S(x_i;\ \lambda_j)$ give similar detection results, while for a large linear term, their results may be quite different. It is clear that the difference between $I_M(x_i;\ \lambda_j)$ and $I_S(x_i;\ \lambda_j)$ originates from the perturbation schemes (i.e. whether sample means are perturbation-free).

2. EXAMPLES

In this section, we give two examples to illustrate the detection index $I_M(x_i;\ \lambda_j)$ developed above. We firstly discuss a two-dimensional problem to have a graphical view for scatters of observation cases.

EXAMPLE 1. Artificial data with 15 cases and 2 variables given by TABLE I.

FIG. 1 gives a scatter plot for the data. For the largest eigenvalue λ_1 of the data, the only outlier, case 10 at the position (1.5,2.5), is globally influential by Critchley's global influence function. FIG. 2 gives the plot of $I_M(x_i;\ \lambda_j)$ and $I_S(x_i;\lambda_j)$ against case number i for local influence analysis. Obviously, from $I_S(x_i;\lambda_j)$, case 10 is a locally influential case, while it is not by $I_M(x_i;\ \lambda_j)$.

In order to gain insight for this problem, we start out from fundamental ideas of local influence, and for minor changes of a case, consider the impacts on the largest eigenvalue λ_1. Specifically, for an increment ε of case i such that $x_i(\varepsilon)=(1+\varepsilon)x_i$, we directly compute the largest eigenvalue $\lambda_1^{(i)}(\varepsilon)$ and its relative

TABLE I. A set of artificial data

Case No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_1	2.1	2.4	2.9	2.7	3.6	3.0	2.9	2.2	2.4	1.5	3.2	3.3	3.3	3.5	2.8
x_2	1.0	1.3	1.1	0.4	0.6	0.5	0.9	0.6	0.9	2.5	0.8	1.2	0.4	0.9	0.8

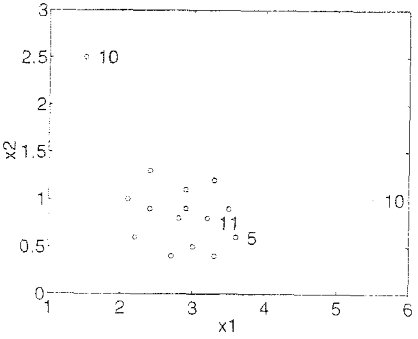

FIG.1. Artificial data: scatter plot

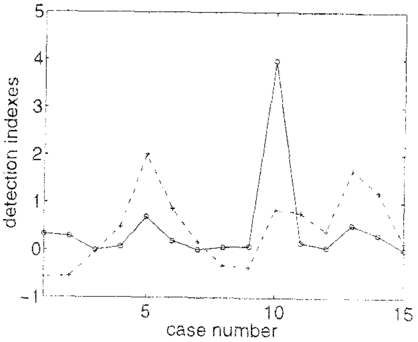

FIG. 2. $I_S(x_i; \lambda_j)$(—) and $I_M(x_i; \lambda_j)$ (–·–·)

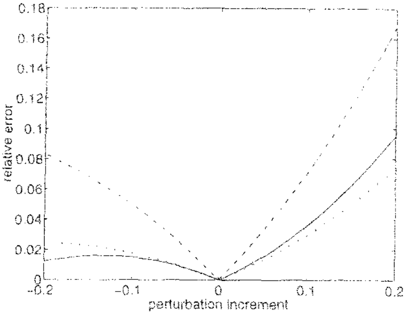

FIG.3. Relative error versus perturbation increment ε: case 5(—·—), 10(—) and 11(.....)

error $J_i(\varepsilon)=|\lambda_1-\lambda_1^{(i)}(\varepsilon)|/\lambda_1$. Similar treatment was adopted by Cook (1986). FIG. 3 gives the plot of relative errors against the perturbation increment ε for cases 5, 10 and 11. It can be seen that case 5 has much stronger impact than cases 10 and 11. Moreover, cases 10 and 11 have almost equal effects. These observations agree with what $I_M(x_i; \lambda_j)$ indicates.

From the above analysis we conclude that minor changes of case 10 do not have strong impacts on λ_1 or equivalently, λ_1 is not sensitive to minor changes of case 10.

Further analysis on $I_M(x_i; \lambda_j)$ shows that for case 10, its first and second order effects $v_j y_{ij}$ and y_{ij}^2 have almost the same (relatively large) absolute values but opposite signs, which leads their impacts to cancel out.

Somewhat interesting is that by similar analysis, we can see that case 10 is a globally as well as a locally influential case if it is moved to (5.5,1).

Next, we consider a more complicated practical example.

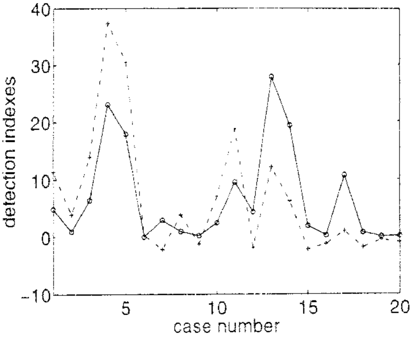

FIG. 4. Detection indexes $I_S(x_i; \lambda_j)$ (—) and $I_M(x_i; \lambda_j)$ (-·-·).

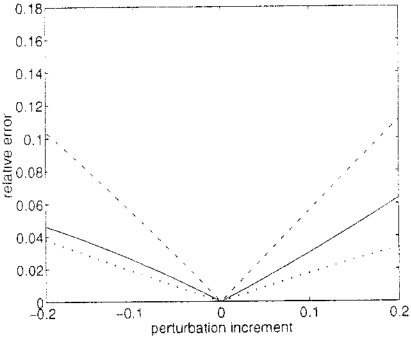

FIG. 5. Relative errors versus perturbation ε: case 4 (-·-·), 11 (—) and 13 (......).

EXAMPLE 2. Kendell's soil composition data (1975, Table 2.1).

Kendell (1975) investigated a set of soil composition data. There are 20 observations and 4 variables including silt content, clay content, organic matter, and acidity on the pH scale. This set of data was also investigated by Critchley (1985) and Shi (1997).

For brevity we concentrate on the second largest eigenvalue λ_2. FIG. 4 gives the plot of the two detection indexes for local influence. $I_S(x_i; \lambda_j)$ indicates that locally influential cases are cases 13 and 4 (Shi, 1997), while by $I_M(x_i; \lambda_j)$, case 13 is not locally influential. FIG. 4 shows that even case 11 has stronger local influence than case 13. Again, we consider the plot of relative errors displayed in FIG 5. It is clear that it gives consistent results with $I_M(x_i; \lambda_j)$. Thus, although case 13 is a globally influential case (Critchley, 1985), the second largest eigenvalue λ_2 is not sensitive to its minor changes. In contrast to this, case 4 is a globally as well as a locally influential case.

These examples clearly show that the detection index $I_M(x_i; \lambda_j)$ is reasonable for identifying the cases whose minor changes have strong impacts on eigenvalues of sample covariance matrices. Similar discussion can be given to eigenvectors.

ACKNOWLEDGEMENTS

The authors are grateful for the reviewers' valuable suggestions for the earlier version of this paper. The second author is a senior Research Associate with the FWO-Flanders. This work was supported by Concerted Research Action GOA-MIPS, the FWO project G.0256.97: Numerical Algorithms for Subspace System Identification, the FWO Research Communities: ICCOS and Advanced Numerical Methods for Mathematical Modeling, and the Belgian Government: Interuniversity Attraction Poles (IUAP P4-02 and P4-24).

BIBLIOGRAPHY

Cook, R. D. (1986). "Assessment of local influence," *J. R. Statist. Soc.* **B48**, 133-169.

Critchley, F. (1985). "Influence on principal components analysis," *Biometrika* **72**, 627-636.

Huber, P. (1981). *Robust Statistics*, New York: Wiley.

Kendell, M. G. (1975). *Multivariate Analysis*, London: Griffin.

Shi, L. (1997). "Local influence in principal components analysis," *Biometrika*, **84**, 175-186.

Received January, 1998; Revised March, 1999.