Electronic Journal of Statistics
Vol. 8 (2014) 927–930
ISSN: 1935-7524
DOI: 10.1214/14-EJS915

Sensitivity of principal component subspaces: A comment on Prendergast's paper*

Jacques Bénasséni

Université Rennes 2
IRMAR, UMR CNRS 6625
Place du Recteur Henri Le Moal
CS 24307, 35043 Rennes Cedex, France
e-mail: jacques.benasseni@univ-rennes2.fr

Abstract: In a recent paper on sensitivity of subspaces spanned by principal components, Prendergast [5] introduces an influence measure based on second order expansion of the RV and GCD coefficients which are commonly used as measures of similarity between two matrices. The goal of this short note is to point out that the paper of Castaño-Tostado and Tanaka [2] is based on a similar approach. However this work seems unknown to Prendergast since it is missing in his references. A comparison of the two papers is provided together with a brief review of some related works.

MSC 2010 subject classifications: Primary 62F35; secondary 62H12.
Keywords and phrases: Principal components, RV-coefficient, sensitivity measures, subspaces.

Received May 2014.

Contents

1. Introduction

Over the past three decades, a large amount of work has been devoted to sensitivity studies for a wide range of statistical methods. When considering principal component subpsaces Tanaka [7], Tanaka and Castaño-Tostado [8],

*Comment on Prendergast, L. A. (2008). A note on sensitivity of principal component subspaces and the efficient detection of influential observations in high dimensions. *Electron. J. Statist.* **2** 454–467. doi:10.1214/08-EJS201.

Castaño-Tostado and Tanaka [2], Bénasséni [1], Prendergast [5], Prendergast and Li Wai Suen [6] use influence functions introduced by Hampel [4] in order to detect influential observations. The goal of this short note is to emphasize that the approaches of Castaño-Tostado and Tanaka [2] and Prendergast [5] are similar.

Throughout this note, we consider a c.d.f. F defined on $\mathbb{R}^p$. We assume that the mean $\boldsymbol{\mu} = \int \boldsymbol{z} dF(\boldsymbol{z})$ and the $p \times p$ covariance matrix $\boldsymbol{\Sigma} = \int (\boldsymbol{z} - \boldsymbol{\mu})(\boldsymbol{z} - \boldsymbol{\mu})^t dF(\boldsymbol{z})$ exist and that $\boldsymbol{\Sigma}$ has distinct eigenvalues $\lambda_1 > \cdots > \lambda_p$ associated to normalized eigenvectors $\mathbf{v}_k$ for $k = 1, \ldots, p$. Letting S denote an arbitrary subset of $\{1, \ldots, p\}$ with K elements ($K < p$) and $\mathbf{V}$ the $p \times K$ matrix whose columns are the vectors $\mathbf{v}_k$ for $k \in S$, we focus on the modification of the column space of $\mathbf{V}$ as F is shifted to F_ϵ defined as $F_\epsilon = (1 - \epsilon)F + \epsilon \delta_{\boldsymbol{x}}$ where $\delta_{\boldsymbol{x}}$ is the Dirac distribution giving mass one at some $\mathbf{x} \in \mathbb{R}^p$.

2. Tanaka's approach

Tanaka [7] considers the projection operator $\mathbf{P} = \mathbf{V}\mathbf{V}^t$ onto the column space of $\mathbf{V}$. When F is modified to F_ϵ, $\mathbf{P}$ is modified to $\mathbf{P}_\epsilon = \mathbf{V}_\epsilon {\mathbf{V}_\epsilon}^t$ which can be expressed in a convergent power series:

$$\mathbf{P}_\epsilon = \mathbf{P} + \epsilon \mathbf{P^{(1)}} + \frac{\epsilon^2}{2} \mathbf{P^{(2)}} + O(\epsilon^3)$$

for sufficiently small ϵ. Letting S' denote the complement of S and $y_k = \mathbf{v}_k^t (x - \mu)$, Tanaka derives the influence function $\mathbf{P^{(1)}}$ of $\mathbf{P}$ as:

$$\mathbf{P^{(1)}} = \sum_{k \in S} \sum_{j \in S'} \frac{y_k y_j}{(\lambda_k - \lambda_j)} \left(\mathbf{v}_k \mathbf{v}_j^t + \mathbf{v}_j \mathbf{v}_k^t \right) \tag{1}$$

so that $\|\mathbf{P^{(1)}}\|$ can be used as sensitivity measure. However, in practice, F is generally unknown and must be estimated by the empirical c.d.f. $\hat{F}$ based on a sample. In his numerical study, following Critchley [3], Tanaka constructs three sample versions of this influence function. The reader is referred to these two papers for further details.

3. Bénasséni's influence measure

The idea of Bénasséni [1] is to consider matrix measures such as the RV-measure between $\mathbf{V}$ and $\mathbf{V}_\epsilon$. However he notes that $RV(\mathbf{V}, \mathbf{V}_\epsilon) = 1 + O(\epsilon^2)$ so that the first order coefficient of ϵ vanishes. That is why he introduces the following measure of closeness between $\mathbf{V}$ and $\mathbf{V}_\epsilon$:

$$\rho_1 = 1 - \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{v_k} - \mathbf{P}_\epsilon \mathbf{v_k}\|. \tag{2}$$

and suggests using the coefficient of ϵ in the expansion of ρ_1 as a sensitivity indicator which can be expressed as:

$$\frac{1}{K}\sum_{k\in S}\|\mathbf{P}^{(\mathbf{1})}\mathbf{v_k}\| = \frac{1}{K}\sum_{k\in S}\left[\sum_{j\in S'}\frac{y_k^2 y_j^2}{(\lambda_k-\lambda_j)^2}\right]^{1/2} \tag{3}$$

4. Comments on Bénasséni's approach

4.1. Castaño-Tostado and Tanaka comment

Castaño-Tostado and Tanaka [2] consider the expansion of the RV measure up to the second order:

$$RV(\mathbf{V},\mathbf{V}_\epsilon) = 1 - \left(\epsilon^2/2K\right)\mathrm{Tr}(\mathbf{P}^{(\mathbf{1})})^2 + O(\epsilon^3) \tag{4}$$

and suggest using $[1-RV(\mathbf{V},\mathbf{V}_\epsilon)]^{1/2}$ as a sensitivity measure.

4.2. Prendergast's comment

Prendergast [5] introduces the influence measure defined as:

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon^2}\left[1-\frac{1}{K}\mathrm{Tr}(\mathbf{P}\mathbf{P}_\epsilon)\right] \tag{5}$$

However since $RV(\mathbf{V},\mathbf{V}_\epsilon) = \frac{1}{K}\mathrm{Tr}(\mathbf{P}\mathbf{P}_\epsilon)$, this influence measure is simply equal to the second order coefficient of ϵ in the expansion of $1-RV(\mathbf{V},\mathbf{V}_\epsilon)$, emphasizing that Castaño-Tostado and Tanaka measure is simply the square root of Prendergast's one which can be expressed as:

$$\frac{1}{K}\sum_{k\in S}\sum_{j\in S'}\frac{y_k^2 y_j^2}{(\lambda_k-\lambda_j)^2} \tag{6}$$

by developing (5). Prendergast also considers the correlation case, discuss other applications and shows the interest of an approximate sample version of his measure when considering a high-dimensional data set.

More generally it should be noted as a concluding remark that all the measures discussed in this paper contain similar sensitivity information as illustrated by the comparison of (3) and (6).

References

[1] Bénasséni, J. (1990). Sensitivity coefficients for the subspaces spanned by principal components. *Commun. Statist.-Theory Methods* **19** 2021–2034. MR1086218

[2] Castaño-Tostado, E. and Tanaka, Y. (1990). Some comments on Escoufier's RV-coefficient as a sensitivity measure in principal component analysis. *Commun. Statist.-Theory Methods* **19** 4619–4626. MR1114862
[3] Critchley, F. (1985). Influence in principal component analysis. *Biometrika* **72** 627–636. MR0817577
[4] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393. MR0362657
[5] Prendergast, L. A. (2008). A note on sensitivity of principal component subspaces and the efficient detection of influential observations in high dimensions. *Electron. J. Statist.* **2** 454–467. MR2417389
[6] Prendergast, L. A. and Li Wai Suen, C. (2011). A new and practical influence measure for subsets of covariance matrix sample principal components with applications to high dimensional datasets. *Comput. Statist. Data Anal.* **55** 752–764. MR2736594
[7] Tanaka, Y. (1988). Sensitivity analysis in principal component analysis: influence on the subspace spanned by principal components. *Commun. Statist.-Theory Methods* **17** 3157–3175. MR0963758
[8] Tanaka, Y. and Castaño-Tostado, E. (1990). Quadratic perturbation expansions of certain functions of eigenvalues and eigenvectors and their application to sensitivity analysis in multivariate methods. *Commun. Statist.-Theory Methods* **19** 2943–2965. MR1088060