Computational Statistics and Data Analysis 55 (2011) 752–764

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A new and practical influence measure for subsets of covariance matrix sample principal components with applications to high dimensional datasets

Luke A. Prendergast *, Connie Li Wai Suen

Department of Mathematics and Statistics, La Trobe University, Postcode 3086, Victoria, Australia

ARTICLE INFO

Article history:
Received 19 June 2009
Received in revised form 23 June 2010
Accepted 24 June 2010
Available online 3 July 2010

Keywords:
Canonical correlations
Influence function
Influential observations
Microarray data
Principal component analysis

ABSTRACT

Principal Component Analysis (PCA) is an important tool in multivariate analysis, in particular when faced with high dimensional data. There has been much done with regard to sensitivity analysis and the development of influence diagnostics for the eigenvector estimators that define the sample principal components. However, little, if any, has been done in this setting with regard to the sample principal components themselves. In this paper we develop a sensitivity measure for principal components associated with the covariance matrix that is very much related to the influence function (Hampel, 1974). This influence measure is based on the average squared canonical correlation and differs from the existing measures in that it assesses the influence of certain observational types on the sample principal components. We use this measure to derive an influence diagnostic that satisfies two key criteria being (i) it detects influential observations with respect to subsets of sample principal components and (ii) is efficient to calculate even in high dimensions. We use several microarray datasets to show that our measure satisfies both criteria.

1. Introduction

With many areas of research requiring analysis of data consisting of a large number of measurement variables, principal component analysis (PCA) has never been more popular. Seeking to explain variability of high dimensional datasets in just a few summary measures, PCA provides a useful means to explore such data for which standard analysis would otherwise be daunting, if not impossible. As an example for which there has been an abundance of recent research, consider DNA microarray data which typically consists of thousands of measurement variables associated with gene expression. For almost all microarray datasets, the number of measurement variables (p) far exceeds the number of samples (n). Whilst we are limited with respect to our own inability to visualize data in high dimensions, it is also true that many popular methods of analysis are not capable of treating data of this type. As an example consider Fisher's Linear Discriminant (Fisher, 1936) for discriminant analysis which is not applicable here due to the singularity of the sample covariance matrices. However, if just a few summary measures are instead available and used in place of the p measurement variables, then the method could be utilized. As such PCA is an appealing dimension reduction technique for many areas of research. A useful reference that discusses PCA in the framework of microarray data is Wall et al. (2003).

Let $\{\mathbf{x}_i\}_{i=1}^n$ denote a sample of n p-dimensional vectors of observed values where $\overline{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i/n$ and $\mathbf{S} = \sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top/(n-1)$ denote the sample mean and sample covariance matrix respectively. Let $\{\widehat{\lambda}_j, \widehat{\boldsymbol{\eta}}_j\}_{j=1}^p$ denote the set of eigenvalue–eigenvector pairs associated with the spectral decomposition of $\mathbf{S}$ such that $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_p$ and

* Corresponding author. Tel.: +61 3 9479 2610.
E-mail address: luke.prendergast@latrobe.edu.au (L.A. Prendergast).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.06.022

where $\{\widehat{\boldsymbol{\eta}}_j\}_{j=1}^p$ forms an orthonormal basis for $\mathbb{R}^p$. Then the jth sample principal component (PC) for the ith observation is defined as $\widehat{\alpha}_{ji} = \widehat{\boldsymbol{\eta}}_j^\top \mathbf{x}_i$ (or $\widehat{\omega}_{ji} = \widehat{\boldsymbol{\eta}}_j^\top(\mathbf{x}_i - \overline{\mathbf{x}})$ if centering is preferred) where we will refer to $\widehat{\boldsymbol{\alpha}}_j = [\widehat{\alpha}_{ji}]_{i=1}^n$ or $\widehat{\boldsymbol{\omega}}_j = [\widehat{\omega}_{ji}]_{i=1}^n$ collectively as the jth sample principal component. Given that $\widehat{\lambda}_j$ equals the variance of the jth sample PC, typically the first L sample PCs are retained as useful summary measurements that explain a large proportion of the total sample variability. It is also true, however, that L is often simply chosen to be 1, 2 or 3 so as to allow for simple visualization of the data. For more on PCA see, for example, Jolliffe (2002).

There has been much work with regard to sensitivity of PCA via influence analysis and subsequent creation of influence diagnostics that may be used to detect influential observations (see, for e.g. Critchley, 1985; Tanaka, 1988; Tanaka and Castaño-Tostado, 1990; Bénasséni, 1990; Prendergast, 2008). However, these studies are primarily focused on influence associated with the eigenvector estimates that are used to define the sample PCs, rather than the sample PCs themselves. It is therefore our aim to construct an influence diagnostic that satisfies the following two key criteria:

Criterion 1: Can be used to detect influential observations with respect to the sample principal components rather than the directions used to define them.

Criterion 2: Is efficient and hence practically applicable to high dimensional datasets.

In creating a diagnostic that satisfies Criteria 1 and 2, we formally identify the model based measure that is very much linked to the robustness tool known as the influence function (Hampel, 1974) and highlight how this provides insight into the sensitivity of PCA in general that differs from the aforementioned sensitivity studies.

In the next section we will briefly review currently available influence diagnostics associated with PCA before developing our own measure in Section 3. A sample based version of this measure will be considered in Section 4 where we apply the diagnostic to several microarray examples. Of critical importance in Section 4 is, firstly, that our measure is capable of detecting influential observations (to satisfy Criterion 1) and secondly that it is efficient to compute for large datasets (to satisfy Criterion 2). We will conclude this paper with a discussion in Section 5.

2. Influence measures for eigenvector estimators

2.1. Model based influence diagnostics

Consider the contamination distribution given as

$$F_\epsilon = (1-\epsilon)F + \epsilon \Delta_{\mathbf{x}_0} \tag{1}$$

where $\Delta_{\mathbf{x}_0}$ puts all of its probability mass at the contaminant $\mathbf{x}_0$ and ϵ is the proportion of contamination contributing to F_ϵ. Throughout we will suppose that, for a random $\mathbf{X} \sim F$, we have $E(\mathbf{X}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\mathbf{X}) = \boldsymbol{\Sigma}$. We will also suppose that $\{\lambda_j, \boldsymbol{\eta}_j\}_{j=1}^p$ is the set of eigenvalue–eigenvector pairs of $\boldsymbol{\Sigma}$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

Let t denote an arbitrary statistical estimator defined at F and F_ϵ. Then Hampel (1974) introduced the influence function defined to be

$$\mathrm{IF}(t, F; \mathbf{x}_0) = \lim_{\epsilon \downarrow 0} \frac{t(F_\epsilon) - t(F)}{\epsilon} = \frac{\partial}{\partial \epsilon} t(F_\epsilon)\bigg|_{\epsilon=0} \tag{2}$$

which provides the relative influence of an infinitesimally small proportion of contamination. The usefulness of this measure will become more apparent when we discuss links with sample versions later.

Let η_j denote the functional for the jth eigenvector estimator associated with the usual covariance matrix estimator. Critchley (1985) provides the perturbation

$$\eta_j(F_\epsilon) = \boldsymbol{\eta}_j + \epsilon \boldsymbol{\beta}_j + \frac{1}{2}\epsilon^2 \boldsymbol{\gamma}_j + O(\epsilon^3) \tag{3}$$

where, for $\omega_k = \boldsymbol{\eta}_k^\top(\mathbf{x}_0 - \boldsymbol{\mu})$,

$$\boldsymbol{\beta}_j = \mathrm{IF}(\eta_j, F; \mathbf{x}_0) = \omega_j \sum_{\substack{k=1 \\ k \neq j}}^p \frac{\omega_k}{\lambda_j - \lambda_k} \boldsymbol{\eta}_k$$

and

$$\boldsymbol{\gamma}_j = -\sum_{\substack{r=1 \\ r \neq j}}^p \left\{ \frac{\omega_j^2 \omega_r^2}{(\lambda_j - \lambda_r)^2} \boldsymbol{\eta}_r - \frac{2\omega_r^2}{\lambda_j - \lambda_r} \boldsymbol{\beta}_j + \frac{2\omega_j^3 \omega_r}{(\lambda_j - \lambda_r)^2} \boldsymbol{\eta}_r \right\}.$$

Hence, for a small ϵ, $\mathrm{IF}(\eta_j, F; \mathbf{x}_0)$ provides a useful measure of influence for the jth eigenvector estimator since a large $\mathrm{IF}(\eta_j, F; \mathbf{x}_0)$ results in a large difference in $\eta_j(F_\epsilon)$ from $\boldsymbol{\eta}_j$. An obvious measure of influence for the contaminant $\mathbf{x}_0$ is $\|\mathrm{IF}(\eta_j, F; \mathbf{x}_0)\|$, the length of the influence function vector for $\mathbf{x}_0$, and this can be used to assess the individual influence of the contaminant on each of the required eigenvector estimators.

We will now look at measures that can be used to detect influence on the span of the eigenvector estimators of interest. Let $S \subset \{1, 2, \ldots, p\}$ be the set of indices indicating the eigenvectors to be retained and let $\boldsymbol{\Gamma}_S = [\boldsymbol{\eta}_j]_{j \in S}$ denote the matrix whose jth column is the jth eigenvector indexed in S. Usually one would choose $S = \{1, \ldots, L\}$ so that $\boldsymbol{\Gamma}_S$ is simply $[\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_L]$.