Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Multivariate Analysis 95 (2005) 206–226
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

High breakdown estimators for principal components: the projection-pursuit approach revisited

Christophe Croux[a], Anne Ruiz-Gazen[b, c, *]

[a] *Department of Applied Economics, K.U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium*
[b] *GREMAQ (U.M.R. CNRS 5604), University Toulouse I, Manufacture des Tabacs, 21, al. de Brienne 31042 Toulouse Cedex, France*
[c] *L.S.P. (U.M.R. CNRS 5583), University Toulouse III, 118 route de Narbonne, 31062 Toulouse Cedex, France*

Received 7 November 2003
Available online 18 September 2004

Abstract

Li and Chen (J. Amer. Statist. Assoc. 80 (1985) 759) proposed a method for principal components using projection-pursuit techniques. In classical principal components one searches for directions with maximal variance, and their approach consists of replacing this variance by a robust scale measure. Li and Chen showed that this estimator is consistent, qualitative robust and inherits the breakdown point of the robust scale estimator. We complete their study by deriving the influence function of the estimators for the eigenvectors, eigenvalues and the associated dispersion matrix. Corresponding Gaussian efficiencies are presented as well. Asymptotic normality of the estimators has been treated in a paper of Cui et al. (Biometrika 90 (2003) 953), complementing the results of this paper. Furthermore, a simple explicit version of the projection-pursuit based estimator is proposed and shown to be fast to compute, orthogonally equivariant, and having the maximal finite-sample breakdown point property. We will illustrate the method with a real data example.

AMS 1991 subject classification: 62F35; 62G35

Keywords: Breakdown point; Dispersion matrix; Influence function; Principal components analysis; Projection-pursuit; Robustness

* Corresponding author. University Toulouse III, 118 route de Narbonne, 31062 Toulouse Cedex, France. Fax: +33-05-61-22-55-63.
E-mail addresses: christophe.croux@econ.kuleuven.ac.be (C. Croux), ruiz@cict.fr (A. Ruiz-Gazen).

0047-259X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2004.08.002

1. Introduction

Classical principal components analysis (PCA) is very sensitive to outlying observations, since it is computed from eigenvectors and eigenvalues of the non-robust sample covariance or correlation matrix. Practitioners interpreting multivariate data solely on a classical PCA may therefore end up with wrong conclusions. This fact has been pointed out by many authors and has led to several robustifications of PCA (cf. [22, Chapter 10] for an overview). One may distinguish between two major types of approaches.

The first one calculates eigenvalues and eigenvectors based on a robust estimate of the covariance matrix. Originally, M-estimators for the covariance matrix were used for this (e.g. [13]). Their computation is not time consuming but they have a very low breakdown point in high dimensions. The breakdown point of an estimator measures the maximal percentage of the data points that may be contaminated before the estimate becomes completely corrupted and is very often used as a measure of robustness. Hence, high breakdown estimators for the covariance matrix are to be preferred. As such, the *minimum volume ellipsoid* estimator [29] was used by Naga and Antille [27]. The question of which robust covariance matrix estimator to use has recently been addressed by Croux and Haesbroeck [9]. They also computed influence functions and efficiencies for PCA based on robust estimators of the covariance or correlation matrix.

The second approach consists in calculating directly robust estimates of the eigenvalues and eigenvectors, without passing by a robust estimate of the covariance matrix. A projection-pursuit (PP) based method has been developed by Li and Chen [23] and was already mentioned by Huber [21]. Like classical PCA, they search for directions with maximal dispersion of the data projected on it. But instead of using the variance as a measure of dispersion, they use a robust scale estimator S_n as *projection-pursuit index*. For a sequence of observations $x_1, \ldots, x_n \in \mathbb{R}^p$, the first "eigenvector" is defined as

$$v_{S_n,1} = \underset{\|a\|=1}{\operatorname{argmax}} S_n(a^t x_1, \ldots, a^t x_n). \tag{1.1}$$

The associated "eigenvalue" is then by definition $\lambda_{S_n,1} = S_n^2((v_{S_n,1})^t x_1, \ldots, (v_{S_n,1})^t x_n)$. Suppose now that the first $k-1$ eigenvectors have already been found ($k > 1$). Then the kth eigenvector is defined as

$$v_{S_n,k} = \underset{\|a\|=1, a \perp v_{S_n,1}, \ldots, a \perp v_{S_n,(k-1)}}{\operatorname{argmax}} S_n(a^t x_1, \ldots, a^t x_n), \tag{1.2}$$

while the kth eigenvalue is defined as

$$\lambda_{S_n,k} = S_n^2((v_{S_n,k})^t x_1, \ldots, (v_{S_n,k})^t x_n). \tag{1.3}$$

Principal components scores are then given by the projections of the observations on the eigenvectors. Li and Chen [23] showed that the estimates inherit the breakdown point of the scale estimator S_n and are qualitative robust. As a by-product, a robust covariance estimate