Journal of Multivariate Analysis 132 (2014) 151–159

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Biobjective sparse principal component analysis☆

Emilio Carrizosa, Vanesa Guerrero *

Instituto de Matemáticas de la Universidad de Sevilla (IMUS), Edificio Celestino Mutis, Avda Reina Mercedes s/n, 41012 Sevilla, Spain

ARTICLE INFO

Article history:
Received 14 May 2013
Available online 21 August 2014

AMS subject classifications:
62H25
90C26

Keywords:
Principal component analysis
Mixed Integer Nonlinear Programming
Biobjective optimization
Sparseness

ABSTRACT

Principal Components are usually hard to interpret. Sparseness is considered as one way to improve interpretability, and thus a trade-off between variance explained by the components and sparseness is frequently sought. In this note we address the problem of simultaneous maximization of variance explained and sparseness, and a heuristic method is proposed. It is shown that recent proposals in the literature may yield dominated solutions, in the sense that other components, found with our procedure, may exist which explain more variance and at the same time are sparser.

1. Introduction

Principal Components Analysis (PCA) is a classical dimension reduction technique in multivariate data analysis, introduced by Pearson [13]. The goal of PCA is to find a set of orthonormal vectors minimizing the sum of squares of distances between a set of points and their projections on the vector space spanned by such orthonormal vectors. In other words, in PCA k orthonormal vectors $\mathbf{c}_1, \ldots, \mathbf{c}_k$ are sought by solving the following optimization problem:

$$\min_{\mathbf{c}_1,\ldots,\mathbf{c}_k:\ \text{orthonormal}} \frac{1}{n}\sum_{h=1}^{n} \left\| u_h - \pi_{\{\mathbf{c}_1,\ldots,\mathbf{c}_k\}}(u_h) \right\|_2^2, \tag{1}$$

where $\{u_1, \ldots, u_n\} \in \mathbb{R}^p$ is the set of points, and $\pi_{\{\mathbf{c}_1,\ldots,\mathbf{c}_k\}}$ denotes the projection onto the linear space $L(\{\mathbf{c}_1, \ldots, \mathbf{c}_k\})$ spanned by the vectors $\mathbf{c}_1, \ldots, \mathbf{c}_k,\ k \leq p$. The optimal solutions, $\mathbf{c}_1, \ldots, \mathbf{c}_k$, are called the Principal Components (PCs). In this work we suppose that k has been fixed in advance using any method, such as e.g. the Scree Plot, PC rank trace, and Kaiser's rule, Izenman [6].

The so-obtained PCs enjoy important properties, such as the fact that the projection of the points $u_1, \ldots, u_n$ have uncorrelated components, see [8]. Moreover, the optimal solution of Problem (1) admits an interpretation in terms of the variance explained by the projections. However, the most important problem of PCA is the lack of interpretability of the results, e.g. [15,16], in part due to the fact that PCs have most components at nonzero value, i.e., most original variables are related with each PC. For this reason, several authors have advocated the use of simpler components (i.e., orthonormal vectors with a few nonzero entries), at the expense of losing variance explained or other properties, such as uncorrelation of the projections. Sparseness is usually considered to be a tool to make interpretation of the components easier.

☆ This research is funded in part by projects MTM2012-36136 (Ministerio de Economía y Competitividad, Spain), P11-FQM-7603 and FQM329 (Junta de Andalucía).

* Corresponding author.
E-mail addresses: ecarrizosa@us.es (E. Carrizosa), vguerrero@us.es (V. Guerrero).

http://dx.doi.org/10.1016/j.jmva.2014.07.010
0047-259X/

A first proposal, following the idea of making PCs sparser, and thus potentially more interpretable, is to build components from PCs, but setting to zero all coefficients which, in absolute value, are below a threshold. However, this intuitively appealing idea may be misleading, see [3]. Other related criteria are presented in [9], called SCoTLASS, Zou et al. [18] or d'Aspremont et al. [5]. Since interpretability is a subjective criterion, these papers, as we also do, assume that sparseness can be seen as potential interpretability, and thus maximizing sparseness will make it likely to understand and interpret the output of the procedure.

Qi et al. [14] proposed another methodology in which the trade off between sparseness and explanation of the variance is obtained by introducing a new norm, $\|\cdot\|_{(\lambda)}$, which depends on a parameter $\lambda \in [0, 1]$. The extreme values of $\|\cdot\|_{(\lambda)}$ correspond to the ℓ_2 distance (associated with error minimization) and the ℓ_1 distance (associated with sparseness, as in the lasso, Tibshirani [17]). Sparse PCs are calculated sequentially by imposing either orthonormality or uncorrelation of the components. Their approach can be summarized as follows.

Let us introduce the following norm in $\mathbb{R}^p$, $\|\cdot\|_{(\lambda)}$, $\lambda \in [0, 1]$ defined as

$$\|\mathbf{c}\|_{(\lambda)} = \left[(1-\lambda)\|\mathbf{c}\|_2^2 + \lambda\|\mathbf{c}\|_1^2\right]^{1/2}, \quad \mathbf{c} \in \mathbb{R}^p$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm and $\|\cdot\|_1$ denotes the ℓ_1-norm.

The first sparse PC is obtained by solving the optimization problem

$$\begin{array}{ll} \max & \dfrac{\mathbf{c}_1^\top \cdot V \cdot \mathbf{c}_1}{\|c_1\|_{(\lambda_1)}^2} \\ \text{s.t.} & \{\|\mathbf{c}_1\|_2 = 1, \end{array}$$

where V is the $p \times p$ covariance (or correlation) matrix and $\lambda_1 \in [0, 1]$ is a parameter which must be fixed and trades off somehow variance explained and sparseness.

Higher order sparse PCs are obtained as the solution of two alternative problems, depending whether orthogonality or uncorrelation on the PCs is imposed.

Firstly, if orthogonality is sought and $\mathbf{c}_j,\ j = 1, \ldots, k-1$, are known, the k-th PC is obtained by solving

$$\begin{array}{ll} \max & \dfrac{\mathbf{c}_k^\top \cdot V \cdot \mathbf{c}_k}{\|\mathbf{c}_k\|_{(\lambda_k)}^2} \\ \text{s.t.} & \begin{cases} \|\mathbf{c}_k\|_2 = 1 \\ \mathbf{c}_j^\top \cdot \mathbf{c}_k = 0 \quad \forall j = 1, \ldots, k-1. \end{cases} \end{array}$$

On the other hand, if uncorrelation of the components is required, the problem considered is

$$\begin{array}{ll} \max & \dfrac{\mathbf{c}_k^\top \cdot V \cdot \mathbf{c}_k}{\|\mathbf{c}_k\|_{(\lambda_k)}^2} \\ \text{s.t.} & \begin{cases} \|\mathbf{c}_k\|_2 = 1 \\ \mathbf{c}_j^\top \cdot V \cdot \mathbf{c}_k = 0 \quad \forall j = 1, \ldots, k-1. \end{cases} \end{array}$$

An algorithm with good theoretical properties is studied, as well as numerical illustrations using the classical *Pitprops* data set, see [7], and a large artificial data set are shown. However, as we show below, solutions obtained with this approach may be dominated, in the sense that other components may exist being sparser and, at the same time, explaining a higher percentage of variance.

The paper is organized in four more sections. In Section 2 we formulate a sparse version of problem (1) as a biobjective Mixed Integer Nonlinear Problem (MINLP). The reader is referred to Burer and Letchford [2] for an updated review on Mixed Integer Nonlinear Programming. Problem resolution is described in Section 3. Numerical results are included in Section 4. Finally, Section 5 includes some conclusions and extensions.

2. Problem statement

We address the problem under consideration by formulating an optimization problem, for which a heuristic method is proposed. This approach is similar to the methodology proposed in [4]. In such work, a new procedure for achieving sparseness in PCs is proposed by writing an optimization problem in which, on top of deciding the loadings (which are continuous variables), one has to decide which ones are allowed to take nonzero values. This is done by introducing some binary variables which allow the user to control how sparse PCs are. The so-obtained nonlinear optimization problem with continuous and binary variables is heuristically solved via a Variable Neighborhood Search, Mladenović and Hansen [11].

In this paper, we extend that idea to a more challenging problem, in which both the sparseness of the PCs and variance explained are simultaneously optimized. This leads to a biobjective problem, which is solved heuristically with a Pairwise Exchange Method, Nahar et al. [12]. Problem resolution will be dealt in Section 3.