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 $\mathbf{c}_1, \ldots, \mathbf{c}_k,\ k \le p$. 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.

Let us introduce the formulation of the biobjective problem of optimizing both sparseness and variance explained. To do that, let us denote by c_{il} the l-th coordinate of the ith PC, $i = 1, \ldots, k$, $l = 1, \ldots, p$. We define

$$z_{il} = \begin{cases} 1 & \text{if } c_{il} \neq 0 \\ 0 & \text{else.} \end{cases}$$

Notice that, thanks to the definition of z_{il}, one has that

$$\sum_{l=1}^{p} z_{il} = \#\text{nonzero loadings in the } i\text{th PC},$$

and thus,

$$\sum_{i=1}^{k} \sum_{l=1}^{p} z_{il} = \#\text{overall nonzero loadings}.$$

Thus, the problem can be written as follows:

$$\begin{array}{ll} \min\limits_{\mathbf{c}_1,\ldots,\mathbf{c}_k,\mathbf{z}_1,\ldots,\mathbf{z}_k} & \left(\dfrac{1}{n} \sum\limits_{h=1}^{n} \left\| u_h - \pi_{\{\mathbf{c}_1,\ldots,\mathbf{c}_k\}}(u_h) \right\|_2^2 , \sum\limits_{i=1}^{k} \sum\limits_{l=1}^{p} z_{il} \right) \\ \text{s.t.} & \begin{cases} \mathbf{c}_1, \ldots, \mathbf{c}_k & \text{orthonormal (or uncorrelated)} \\ |c_{il}| \leq z_{il} & \forall i = 1, \ldots, k \; l = 1, \ldots, p \\ z_{il} \in \{0, 1\} & \forall i = 1, \ldots, k \; l = 1, \ldots, p. \end{cases} \end{array} \tag{2}$$

The second constraint forces c_{il} to take zero value if the corresponding z_{il} is equal to zero. However, if z_{il} is equal to one, this constraint is redundant because every c_{il} is already less than or equal to one, since we are imposing $\|\mathbf{c}_i\|_2 = 1, \; i = 1, \ldots, k$.

Problem (2) can be reformulated as a parametric problem, parameterized by an integer α, in which the error between points and their projections is minimized among those vector spaces whose basis have overall at most α nonzero components. Thus, (2) can be written as follows:

$$\begin{array}{ll} \min\limits_{\mathbf{c}_1,\ldots,\mathbf{c}_k} & \dfrac{1}{n} \sum\limits_{h=1}^{n} \left\| u_h - \pi_{\{\mathbf{c}_1,\ldots,\mathbf{c}_k\}}(u_h) \right\|^2 \\ \text{s.t.} & \begin{cases} \mathbf{c}_1, \ldots, \mathbf{c}_k & \text{orthonormal (or uncorrelated)} \\ |c_{il}| \leq z_{il} & \forall i = 1, \ldots, k \; l = 1, \ldots, p \\ \sum\limits_{i=1}^{k} \sum\limits_{l=1}^{p} z_{il} \leq \alpha & \\ z_{il} \in \{0, 1\} & \forall i = 1, \ldots, k \; l = 1, \ldots, p. \end{cases} \end{array} \tag{3}$$

Simple algebraic manipulations give us the following:

Proposition 1. *Problem (3) is equivalent to finding* $\mathbf{c}_1, \ldots, \mathbf{c}_k$ *solution of*

$$\begin{array}{ll} & \max\limits_{\mathbf{c}_1,\ldots,\mathbf{c}_k} \operatorname{tr} \left\{ (C^\top C)^{-1} \cdot C^\top \cdot V \cdot C \right\} \\ \text{s.t.} & \begin{cases} \mathbf{c}_1, \ldots, \mathbf{c}_k & \text{orthonormal (or uncorrelated)} \\ |c_{il}| \leq z_{il} & \forall i = 1, \ldots, k \; l = 1, \ldots, p \\ \sum\limits_{i=1}^{k} \sum\limits_{l=1}^{p} z_{il} \leq \alpha & \\ z_{il} \in \{0, 1\} & \forall i = 1, \ldots, k \; l = 1, \ldots, p \end{cases} \end{array} \tag{4}$$

where $C = (\mathbf{c}_1 | \ldots | \mathbf{c}_k)$ *and* tr *represents the trace of a matrix.*

Normalizing the objective function of (4) we obtain the expression

$$\frac{\operatorname{tr} \left\{ (C^\top C)^{-1} \cdot C^\top \cdot V \cdot C \right\}}{\operatorname{tr}\{V\}} \cdot 100,$$

which will be called thereafter *percentage of explained variance*.

Problem (4) is a nonconvex mixed integer program, and thus unlikely to be solved by exact methods, even for problems of moderate size, see [2]. For this reason, instead of attempting to find an optimal solution to (4) we suggest a heuristic procedure, which is much less time demanding than exact methods. The heuristic is described in Section 3.

3. The optimization process

Problem (4) is a hard MINLP, unlikely to be solvable to optimality by standard numerical routines, see e.g. [2] for a survey on Mixed Integer Nonlinear Programming. Instead we propose to solve it heuristically, firstly fixing an initial solution, as described in Section 3.1, and then improving it via a Pairwise Exchange Algorithm, as presented in Section 3.2. We stress that our approximation is heuristic, and thus a suboptimal solution to Problem (4) may be obtained. Nevertheless, as shown in our numerical testing, such heuristic yields competitive results against state-of-the-art methods.

3.1. Fixing an initial solution

In order to build a starting solution to Problem (4), we have to build a pair $(\mathbf{z}^0 = (\mathbf{z}_1^0, \ldots, \mathbf{z}_k^0), \mathbf{c}^0 = (\mathbf{c}_1^0, \ldots, \mathbf{c}_k^0))$. This is done in a two-stage process, by first finding a sensible guess for $\mathbf{z}^0$, and later finding the best-possible $\mathbf{c}^0$ for such $\mathbf{z}^0$.

Firstly, binary variables $z_{il},\ i = 1, \ldots, k,\ l = 1, \ldots, p$, are fixed by solving a linear program. In order to formulate the problem, classical principal components $\mathbf{c}^* = (\mathbf{c}_1^*, \ldots, \mathbf{c}_k^*)$ need to be computed by solving Problem (1). Then, we find $z_{il},\ i = 1, \ldots, k,\ l = 1, \ldots, p$, such that the sum of the absolute values of loadings of principal components are maximized by imposing constraints which are required in Problem (4). A constraint forcing at least one variable to appear in each principal component is included in this formulation. So, $z_{il},\ i = 1, \ldots, k,\ l = 1, \ldots, p$, are sought so that they satisfy

$$\begin{aligned} \max_{\mathbf{z}_1,\ldots,\mathbf{z}_k} \quad & \sum_{i=1}^{k}\sum_{l=1}^{p} |c_{il}^*| z_{il} \\ \text{s.t.} \quad & \begin{cases} \sum_{l=1}^{p} z_{il} \geq 1 & \forall i = 1, \ldots, k \\ \sum_{i=1}^{k}\sum_{l=1}^{p} z_{il} \leq \alpha & \\ 0 \leq z_{il} \leq 1 & \forall i = 1, \ldots, k\, l = 1, \ldots, p \end{cases} \end{aligned} \tag{5}$$

where c_{il}^* denotes the l-th coordinate of the ith classical PC, i.e., solution of (1). Problem (5) has a flow problem constraint structure, see [1], which means that constraint matrix is totally unimodular, and it attains its optimal value at some point $\mathbf{z}^0 = (\mathbf{z}_1^0, \ldots, \mathbf{z}_k^0)$ with all coordinates $z_{il}^0 \in \{0, 1\}$.

Once the values z_{il}^0 are fixed, we come back to problem (4) by adding the constraints controlling which coordinates must be equal to zero. This amounts to solving the nonlinear optimization problem

$$\begin{aligned} \max_{\mathbf{c}_1,\ldots,\mathbf{c}_k} \quad & \operatorname{tr}\left\{C^\top \cdot V \cdot C\right\} \\ \text{s.t.} \quad & \begin{cases} \mathbf{c}_i^\top \cdot \mathbf{c}_j = \delta_{ij} & \forall i, j = 1, \ldots, k \\ c_{il} = 0 & \forall i = 1, \ldots, k,\ l = 1, \ldots, p : z_{il}^0 = 0, \end{cases} \end{aligned} \tag{6}$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases}$$

if orthogonal sparse components are sought, and if, instead, uncorrelated projections are sought, one solves

$$\begin{aligned} \max_{\mathbf{c}_1,\ldots,\mathbf{c}_k} \quad & \operatorname{tr}\left\{(C^\top C)^{-1} \cdot C^\top \cdot V \cdot C\right\} \\ \text{s.t.} \quad & \begin{cases} \mathbf{c}_i^\top \cdot V \cdot \mathbf{c}_j = 0 & \forall i, j = 1, \ldots, k,\ i \neq j \\ \|\mathbf{c}_i\|_2 = 1 & \forall i = 1, \ldots, k \\ c_{il} = 0 & \forall i = 1, \ldots, k,\ l = 1, \ldots, p : z_{il}^0 = 0. \end{cases} \end{aligned} \tag{7}$$

This way, an approximation with at most α coordinates at a nonzero value is found. The solution is likely to be similar to the one provided by PCA, but without forcing that every coefficient lower than some threshold value sets to zero. In fact, this approach is proven to be deceived in [3].

3.2. Improving the solution via Pairwise Exchange Algorithm

The solution found by the procedure described in Section 3.1 may not be optimal for Problem (4), but it can be a good starting point for a search procedure such as Pairwise Exchange Algorithm, Nahar et al. [12]. Other more sophisticated heuristic approaches, such as Simulated Annealing, see [10], or Variable Neighborhood Search, see [11], might also be used. However, our numerical experience shows that the simplest approach, namely, the Pairwise Exchange, yields in general excellent results and is thus the one we suggest and describe below.

- **Initialization:** Let x_0 be an initial solution.
- **Repeat:** While the stopping condition is not satisfied:
  - (a) Construct a solution $\tilde{x}$ from x_0 by making a pairwise exchange.
  - (b) If $\tilde{x}$ improves the objective function, update it ($x_0 \leftarrow \tilde{x}$).

Fig. 1. Pairwise Exchange Algorithm.

(i) Initialization:

- find $\mathbf{c}^* = (\mathbf{c}_1^*, \ldots, \mathbf{c}_k^*)$, principal components.
- find $\mathbf{z}^0 = (\mathbf{z}_1^0, \ldots, \mathbf{z}_k^0)$, optimal solution to (5).
- find $\mathbf{c}^0 = (\mathbf{c}_1^0, \ldots, \mathbf{c}_k^0)$, optimal solution to (6) (or (7)).

(ii) Repeat the following steps until a maximum number of iterations $Nmax$ is achieved.

(a) $N \leftarrow 1$

(b) Construct a solution $znew$ from $\mathbf{z}^0$ by making a pairwise exchange and find $\mathbf{c}(znew) = (\mathbf{c}_1(znew), \ldots, \mathbf{c}_k(znew))$, solution of (6) (or (7)).

(c) If $tr\{\mathbf{c}(znew)^\top \cdot V \cdot \mathbf{c}(new)\} > tr\{(\mathbf{c}^0)^\top \cdot V \cdot \mathbf{c}^0\}$
$\Big($or $tr\{(\mathbf{c}(znew)^\top \mathbf{c}(znew))^{-1} \mathbf{c}(znew)^\top \cdot V \cdot \mathbf{c}(new)\} > tr\{(\mathbf{c}^{0\top} \mathbf{c}^0)^{-1} \mathbf{c}^{0\top} \cdot V \cdot \mathbf{c}^0\}\Big)$,
then move: $\mathbf{z}^0 \leftarrow znew$, $\mathbf{c}^0 \leftarrow \mathbf{c}(znew)$.
$N \leftarrow N + 1$.
Go to (b).

Fig. 2. Pairwise Exchange Algorithm for solving problem (6) (or (7)).

Firstly, let us describe Pairwise Exchange Algorithm for a general problem, Fig. 1, and then it will be customized to our problem, Fig. 2.

Having a solution of the optimization problem under consideration, Pairwise Exchange Algorithm consists of interchanging the positions of two randomly chosen elements. Such perturbation is accepted only if it improves the objective function of the problem. A general scheme of the algorithm is presented in Fig. 1.

In order to solve Problems (6) and (7), we customize the Pairwise Exchange Algorithm described above to our particular problem. In our method, we use a few trial values for the binary variables z_{il}, and once they are fixed, the vectors $\mathbf{c}_1, \ldots, \mathbf{c}_k$ are obtained by solving (6) or (7).

In order to build the trial values of $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_k)$, we first construct a starting value $\mathbf{z}^0 = (\mathbf{z}_1^0, \ldots, \mathbf{z}_k^0)$, obtained by solving the linear program (5). From such starting value $\mathbf{z}^0$, a Pairwise Exchange is implemented. The steps of our procedure are summarized in Fig. 2.

Observe that the procedure stops when the number of iteration attains its limit *Nmax*. Such *Nmax* is fixed in advance by the user so that (s)he avoids the running times of this combinatorial problem exploit. One should keep in mind that, generally speaking, the lower the running time allowed, the lower the expected quality of the solution provided by the procedure.

4. Computational experiments

We compare our approach with the one by Qi et al. [14] outlined in Section 1 on two different settings, identified by their covariance matrix V. The first model tested corresponds to the *Pitprops* data set, see [7]. The *Pitprops* data has become a benchmark data set for testing the performance of sparse PCA methods due to the difficulty of achieving simplicity in PCs without losing information. 13 variables arose from a study on the strength of pitprops cut from home-grown timber, measured in 180 observations. Six principal components are calculated. The correlation matrix can be found in R package `elasticnet`.

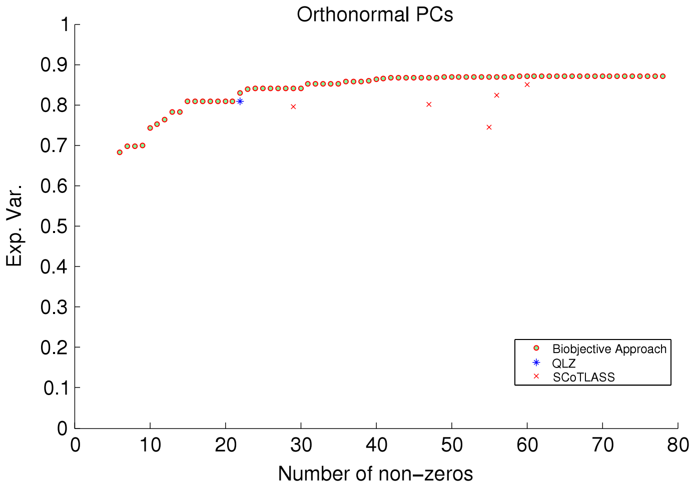

Fig. 3. Orthonormal PCs shown in [14] against biobjective approach.

The second data set is an artificial set used to test the behavior of both methods on larger dimensions. For this reason we have generated a $p = 200$-dimensional covariance matrix V from a data model

$$\mathbf{X} = s\mathbf{e} + t\mathbf{f} + \varepsilon \in \mathbb{R}^p,$$

with $e = (e_1, \ldots, e_p), f = (f_1, \ldots, f_p), \varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)$, s and t given as follows:

- $e_i = \begin{cases} 1 & \text{if } i < \lfloor \frac{p}{2} \rfloor \\ 0 & \text{else} \end{cases} \quad i = 1, \ldots, p.$
- $f_i = 1, \; i = 1, \ldots, p.$
- ε_i is a random variable with $\text{var}(\varepsilon_i) = \sigma^2 = 0.1, \; i = 1, \ldots, p.$
- s and t are independent random variables with $\text{var}(s) = \text{var}(t) = 1$.

Roughly speaking the random vector $\mathbf{X}$ is around the 2-dimensional space generated by the vectors e and f.
Hence, the covariance matrix V obtained has the form

$$V = \begin{pmatrix} e_1^2 + f_1^2 + \sigma^2 & e_1e_2 + f_1f_2 & \cdots & e_1e_p + f_1f_p \\ e_2e_1 + f_2f_1 & e_2^2 + f_2^2 + \sigma^2 & \cdots & e_2e_p + f_2f_p \\ \vdots & \vdots & \ddots & \vdots \\ e_pe_1 + f_pf_1 & e_pe_2 + f_pf_2 & \cdots & e_p^2 + f_p^2 + \sigma^2 \end{pmatrix}.$$

Two principal components are extracted. Observe that, by construction, the (nonsparse) components are close to the generating vectors e and f above, and thus the total number of zeros is around $p + \frac{p}{2}$.

The Mixed Integer Programming Problem (4) is solved as described in Section 3. The algorithm has been implemented in MATLAB, using the routines `linprog` to numerically solve the flow problem (5) and routine `fmincon` to solve nonlinear problems (6) and (7). The process is repeated until $N_{\max}$ iterations are performed, where $N_{\max} = 300$ for *Pitprops* and $N_{\max} = 50$ in the high dimensional example.

We compare now our results with those obtained by Qi et al. [14] with their method for the *Pitprops* data set. In Fig. 3 the explained variance, obtained heuristically by means of our procedure, is shown for different values α of nonzero coordinates of Problem (6), i.e. Problem (4) under the constraint that the PCs must be orthonormal. In the same way, Fig. 4 shows the percentages of explained variance in Problem (7) for varying α when uncorrelation is imposed on the PCs, i.e. Problem (4) under the constraint that the PCs must be uncorrelated.

We also plot the solution(s) proposed by Qi et al. [14], which are marked as stars ('$*$'), and denoted QLZ, and the ones given by SCoTLASS approach, Jolliffe et al. [9], and denoted by crosses ('$\times$'), in the case of orthonormal components. All these solutions are included in the cited papers. Notice that we have found better solutions, which have higher sparseness than the ones presented in [14] and, at the same time, also better percentage of explained variance.

In order to compare with Qi et al. [14], we include the PCs obtained by their method, which appear in the cited paper with their corresponding parameters, and the ones we obtain for the very same value α of nonzero coordinates, see Tables 1–3.

Since the performance of QLZ algorithm is strongly affected by tuning parameters and results shown in [14] are only for a few of them, we have computed the principal components for different values of such tuning parameters. We have chosen vector of parameters $\lambda = (\lambda_1, \ldots, \lambda_6)$ such that $\lambda_i \in \{0.2, 0.4, 0.6, 0.8, 1\}, \; i = 1, \ldots, 6$. Thus, QLZ has been executed with 15 625 different values of the tuning parameters under the constraint that components must be orthonormal. Best solutions found, in terms of explained variance, have been stored. A comparison between the results of the biobjective

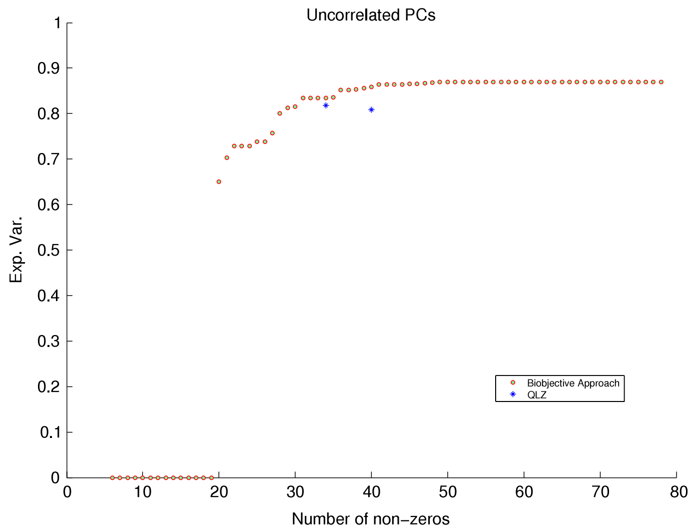

Fig. 4. Uncorrelated PCs shown in [14] against biobjective approach.

Table 1
Orthonormal PCs, $\alpha = 22$.

Variables	Biobjective approach						QLZ					
	PC1	PC2	PC3	PC4	PC5	PC6	PC1	PC2	PC3	PC4	PC5	PC6
Topdiam	−0.40	−0.33	0.00	0.00	0.00	0.00	0.47	0.00	−0.16	0.00	0.00	0.00
Length	−0.43	−0.37	0.00	0.00	0.14	0.00	0.48	0.00	−0.18	0.00	0.00	0.00
Moist	0.00	0.00	−0.71	0.00	0.00	0.00	0.00	−0.71	−0.01	0.00	0.00	0.00
Testsg	0.00	0.00	−0.71	0.00	0.00	0.00	0.00	−0.71	0.00	0.00	0.00	0.00
Ovensg	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.54	0.00	0.00	−0.74
Ringtop	0.00	−0.02	0.00	−0.70	0.03	0.00	0.13	−0.02	0.47	0.00	0.00	0.00
Ringbut	0.00	−0.20	0.00	−0.49	−0.23	0.00	0.38	0.00	0.25	0.00	0.00	0.00
Bowmax	−0.33	−0.16	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00
Bowdist	−0.49	0.00	0.00	0.00	−0.12	0.00	0.38	0.00	0.00	0.00	0.00	0.00
Whorls	0.00	−0.45	0.00	0.00	−0.37	0.00	0.41	0.01	0.00	0.00	0.00	0.00
Clear	−0.56	0.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Knots	0.00	0.00	0.00	−0.37	0.73	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Diaknot	0.00	−0.33	0.00	0.36	0.49	0.00	0.00	0.00	−0.60	0.00	0.00	−0.67
Variance (%)	18.2	16.0	14.6	13.5	13.1	7.7	30.1	14.6	14.6	7.7	7.7	6.1
Cum. Var. (%)	18.2	34.2	48.8	62.3	75.4	83.1	30.1	44.7	59.3	67.1	74.8	80.9

Table 2
Uncorrelated PCs, $\alpha = 34$.

Variables	Biobjective approach						QLZ					
	PC1	PC2	PC3	PC4	PC5	PC6	PC1	PC2	PC3	PC4	PC5	PC6
Topdiam	0.00	−0.40	0.00	0.19	0.00	0.07	0.47	0.00	0.25	0.00	0.00	0.00
Length	0.00	−0.41	0.00	0.19	0.00	0.00	0.48	0.00	0.27	0.00	0.00	−0.27
Moist	0.00	0.00	−0.67	−0.28	−0.13	0.16	0.00	−0.71	0.00	0.09	0.00	0.33
Testsg	−0.20	0.15	−0.53	0.00	0.00	0.00	0.00	−0.62	−0.11	0.09	0.00	0.00
Ovensg	0.00	−0.12	0.00	0.06	0.90	0.00	0.00	0.00	−0.58	0.00	−0.56	−0.40
Ringtop	−0.52	0.20	0.00	0.11	0.00	0.00	0.13	0.00	−0.45	0.00	0.00	0.00
Ringbut	−0.45	0.01	0.00	0.00	0.00	0.00	0.38	0.00	−0.26	0.00	0.00	0.00
Bowmax	0.00	−0.28	0.00	−0.26	0.00	0.00	0.25	0.00	0.00	−0.06	0.00	0.00
Bowdist	0.00	−0.35	0.00	0.00	0.00	0.00	0.38	0.00	0.09	0.00	0.00	0.00
Whorls	0.00	−0.40	0.33	0.00	0.00	0.00	0.41	0.32	0.00	0.00	0.02	0.00
Clear	0.00	0.00	0.00	0.52	0.00	−0.84	0.00	0.00	0.00	−0.98	−0.13	−0.09
Knots	0.00	0.00	0.35	0.71	0.00	0.52	0.00	0.00	0.00	0.00	0.48	−0.78
Diaknot	0.69	−0.49	−0.19	0.00	0.42	0.00	0.00	0.00	0.49	0.16	−0.67	−0.21
Variance (%)	25.5	17.2	11.6	11.5	9.4	8.1	30.1	14.0	14.5	7.6	6.1	5.8
Cum. Var. (%)	25.5	42.7	54.3	65.8	75.2	83.3	30.1	44.1	58.7	66.2	72.3	78.2

Table 3
Uncorrelated PCs, $\alpha = 40$.

Variables	Biobjective approach						QLZ					
	PC1	PC2	PC3	PC4	PC5	PC6	PC1	PC2	PC3	PC4	PC5	PC6
Topdiam	−0.42	0.00	0.00	0.13	0.00	−0.26	−0.47	0.00	0.20	0.00	0.00	0.00
Length	−0.43	0.00	0.00	0.16	0.00	−0.24	−0.48	0.00	0.22	0.00	−0.05	0.00
Moist	0.00	−0.60	0.00	−0.15	0.00	0.00	0.00	−0.68	0.00	0.06	0.26	0.00
Testsg	0.00	−0.52	0.00	−0.32	0.00	0.00	0.00	−0.66	−0.07	0.06	0.19	−0.12
Ovensg	0.00	−0.03	0.00	0.26	0.96	0.00	0.00	0.00	−0.75	0.00	0.00	−0.46
Ringtop	−0.33	0.31	0.35	−0.40	0.00	0.00	−0.13	0.00	−0.40	0.00	−0.14	0.35
Ringbut	−0.35	0.22	0.00	−0.28	0.00	0.00	−0.38	0.00	−0.11	0.00	−0.14	0.30
Bowmax	−0.34	0.00	0.00	0.27	0.00	0.27	−0.25	0.14	0.00	−0.09	0.00	−0.68
Bowdist	−0.38	0.00	0.00	0.19	0.00	0.00	−0.38	0.00	0.00	0.00	−0.08	0.00
Whorls	−0.24	0.00	−0.35	−0.06	0.00	0.00	−0.41	0.16	0.00	0.03	0.00	0.00
Clear	−0.30	0.00	0.76	0.42	0.00	0.40	0.00	0.00	0.00	−0.98	−0.04	−0.09
Knots	0.00	0.47	0.40	−0.15	0.00	−0.58	0.00	−0.23	0.00	0.00	−0.92	−0.32
Diaknot	0.00	0.00	−0.08	0.47	0.28	−0.55	0.00	0.00	0.42	0.16	0.00	0.00
Variance (%)	27.9	14.3	12.2	10.8	10.6	10.2	30.1	15.6	13.2	7.8	6.5	4.6
Cum. Var. (%)	27.9	42.2	54.4	65.2	75.8	86.0	30.1	45.7	58.9	66.6	73.1	77.8

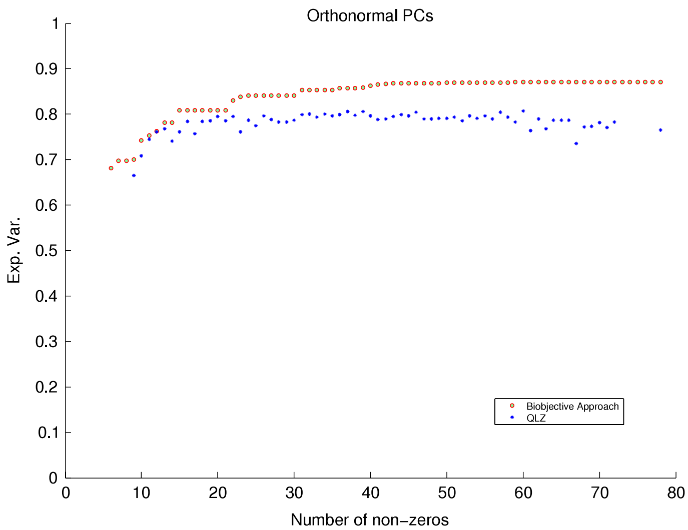

Fig. 5. Comparison between biobjective approach and QLZ in Pitprops.

approach and QLZ is shown in Fig. 5. This figure shows how our biobjective approach is never worse than QLZ algorithm and, in several cases, a higher percentage of explained variance with the same sparseness is achieved by our procedure. Results are equivalent in case that uncorrelation is imposed.

We describe now the results obtained for the second data set. To obtain explained variances with two orthogonal components for different values α of nonzero coordinates in the biobjective approach, Problem (6) has been solved and a pairwise exchange has been performed. On the other hand, QLZ has been executed for tuning parameters generated randomly as many times as its execution time has been the same than the time employed by our approach to obtain solutions for $\alpha = 2, \ldots, 50$. In other words, both procedures are running exactly the same time.

Fig. 6 shows the results for this high dimensional problem and both methods. Observe that almost in all cases our method obtains solutions which have higher or equal explained variance with the same sparseness than QLZ. Results are equivalent in case that uncorrelation is imposed.

5. Conclusions and extensions

A biobjective approach for finding sparse PCs has been proposed. The problem is expressed as solving a series of MINLPs, which are solved by means of a simple exchange heuristic. In order to evaluate the performance of our procedure, a comparison with the methodology described in [14] has been done. Better results, in terms of sparseness and error minimization, have been achieved with our procedure, which also gives higher freedom in trading off sparseness (by varying α) and variance explained.

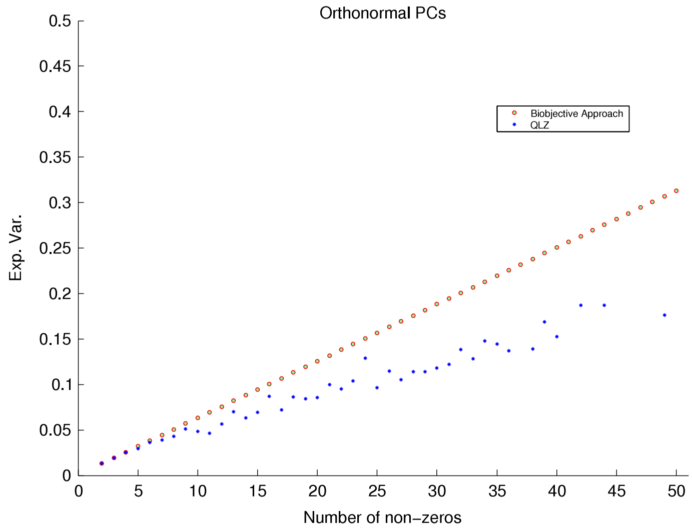

Fig. 6. Comparison between biobjective approach and QLZ in the artificial data set.

Sparseness is usually considered as the main property to make the interpretation of the components easier. The usefulness of the proposed method consists of the possibility of finding sparse components at the same time that explained variance is being maximized. However, interpretability is a subjective property which is subject to the user criterion, which is usually related to sparseness due to the belief that if PCs are sparser then they may be easier understandable.

The results obtained in this paper are only of algorithmic nature. Statistical properties of the so-obtained sparse components remain unexplored.

Acknowledgments

We would like to thank the referees and the editor for their helpful comments, which have made the paper significantly improve.

References

[1] M.S. Bazaraa, J.J. Jarvis, H.D. Sherali, Linear Programming and Network Flows, Wiley, 1977.
[2] S. Burer, A.N. Letchford, Non-convex mixed-integer nonlinear programming: a survey, Surveys in Operations Research and Management Science 17 (2012) 97–106.
[3] J. Cadima, T. Jolliffe, Loading and correlation in the interpretation of principal components, J. Appl. Stat. 22 (1995) 203–214.
[4] E. Carrizosa, V. Guerrero, rs-Sparse principal component analysis: a mixed integer nonlinear programming approach with VNS, Comput. Oper. Res. (2013) http://dx.doi.org/10.1016/j.cor.2013.04.012.
[5] A. d'Aspremont, L. Ghaoui, M. Jordan, G. Lanckriet, Direct formulation of sparse PCA using semidefinite programming, SIAM Rev. 49 (2007) 434–448.
[6] A.J. Izenman, Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, Springer, New York, 2008.
[7] J. Jeffers, Two case studies in the application of principal component analysis, Appl. Stat. 16 (1967) 225–236.
[8] I.T. Jolliffe, Principal Component Analysis, second ed., Springer, New York, 2002.
[9] I.T. Jolliffe, N.T. Trendafilov, M. Uddin, A modified principal component technique based on the Lasso, J. Comput. Graph. Statist. 12 (2003) 531–547.
[10] S. Kirkpatrick, C.D. Gelatt Jr., M.P. Vecchi, Optimization by simulated annealing, Science 220 (1983) 671–680.
[11] N. Mladenović, P. Hansen, Variable neighborhood search, Comput. Oper. Res. 24 (1997) 1097–1100.
[12] S. Nahar, S. Sahni, E. Shragowitz, Simulated annealing and combinatorial optimization, in: Proceedings of the 23rd ACM/IEEE Design Automation Conference, IEEE Press, 1986, pp. 293–299.
[13] K. Pearson, On lines and planes of closest fit to systems of points in space, Phil. Mag. 2 (1901) 559–572.
[14] X. Qi, R. Luo, H. Zhao, Sparse principal component analysis by choice of norm, J. Multivariate Anal. 114 (2013) 127–160.
[15] A.C. Rencher, Interpretation of canonical discriminant functions, canonical variates and principal components, Amer. Statist. 46 (1992) 217–225.
[16] H. Shen, J.Z. Huang, Sparse principal component analysis via regularized low rank matrix approximation, J. Multivariate Anal. 99 (2008) 1015–1034.
[17] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.
[18] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Statist. 15 (2006) 265–286.