Computers & Operations Research 52 (2014) 349–354

Contents lists available at ScienceDirect

Computers & Operations Research

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

rs-Sparse principal component analysis: A mixed integer nonlinear programming approach with VNS[☆]

Emilio Carrizosa, Vanesa Guerrero[*]

Instituto de Matemáticas de la Universidad de Sevilla, Fac. de Matemáticas, Avda Reina Mercedes s/n, 41012 Sevilla, Spain

ARTICLE INFO

Available online 3 May 2013

Keywords:
Sparse principal component analysis
Variable neighborhood search
Nonlinear mixed integer programming

ABSTRACT

Principal component analysis is a popular data analysis dimensionality reduction technique, aiming to project with minimum error for a given dataset into a subspace of smaller number of dimensions.

In order to improve interpretability, different variants of the method have been proposed in the literature, in which, besides error minimization, sparsity is sought. In this paper we formulate as a mixed integer nonlinear program the problem of finding a subspace with a sparse basis minimizing the sum of squares of distances between the points and their projections. Contrary to other attempts in the literature, with our model the user can fix the level of sparseness of the resulting basis vectors. Variable neighborhood search is proposed to solve the problem obtained this way.

Our numerical experience on test sets shows that our procedure outperforms benchmark methods in the literature, both in terms of sparsity and errors.

1. Introduction and literature review

Principal component analysis (PCA) was first introduced by [20] as a method for projecting a set of points $u_1, \dots, u_p \in \mathbb{R}^n$ onto a lower dimensional space in such a way that the distances between the points and their projections are minimized, see e.g., [14,9].

Given k vectors $c_1, \dots, c_k \in \mathbb{R}^n$, let $\pi_{\{c_1,\dots,c_k\}}$ denote the projection onto the linear space $L(\{c_1,\dots,c_k\})$ spanned by the vectors $c_1, \dots, c_k$

$$\pi_{\{c_1,\dots,c_k\}}(u) = \arg\min\{\|u-z\| : z \in L(\{c_1,\dots,c_k\})\}.$$

The aim of PCA is to find a set of $k \le n$ orthonormal vectors $c_1, \dots, c_k$ such that the mean squared distance between the points in the dataset $\{u_1, \dots, u_p\}$ and their projections onto the vector space $L(\{c_1,\dots,c_k\})$ generated by $\{c_1,\dots,c_k\}$ is minimized. In other words, the following mixed integer nonlinear program (MINLP) is to be solved

$$\min_{c_1,\dots,c_k:\ \text{orthonormal}} \frac{1}{p} \sum_{h=1}^{p} \|u_h - \pi_{\{c_1,\dots,c_k\}}(u_h)\|^2$$

The optimal solutions, $c^* = (c_1^*, \dots, c_k^*)$, are called *principal components* (PC). The main drawback of this dimensionality-reduction technique is interpretability: interpreting the projections is usually quite difficult due to the fact that most of the original variables are present in each c_i^*, $i = 1,\dots,k$, i.e., the PC are not sparse. Interpretability is improved if some loadings (the coefficients of PCs) are zero, and this has been pursued in different papers.

A first attempt for achieving this is rounding, by considering all loadings smaller than some threshold absolute value as zero. However, this procedure has been shown to be misleading, see [4]. *Varimax rotation*, see [15], has also been proposed, but it hardly ever achieves the aim of an easier interpretation despite losing orthogonality of the loadings and uncorrelation of the components.

Some authors relate the notion of simplicity to the fact that loadings belong to the set of integers. Such idea was developed in [26] and later in [23], who called their method *simple component analysis* (SCA). SCA allows the user, under his or her criterion, to modify a simple system of components in order to improve interpretability. However, SCA does not yield either orthogonal or uncorrelated components. An exploratory approach to SCA was presented in [1] for achieving orthogonality. Also in [24], SCA is modified by using genetic algorithms.

Another way of obtaining sparsity is by constraining the number of non-zero loadings in each PC. In this line, Ref. [5] proposes a convex relaxation method based on semidefinite programming, which does not preserve orthogonality or uncorrelation, as while [6] a branch-and-bound approach lets the user choose between keeping orthogonality or uncorrelation. A related approach is presented in [11], in which a bound on the sum of the absolute values of the loadings is added, combining this way the *lasso* paradigm, [25], with PCA.

In [28], PCA is formulated as a regression-type optimization problem. Sparse loadings are achieved by imposing *elasticnet* constraint, [27], a generalization of *lasso*, on the regression coefficients. The sparse PCs obtained from *sparse principal component analysis* (SPCA) are neither orthogonal nor uncorrelated. See another *lasso*-based model in [21].

[☆] This research is supported by Grants MTM2012-36136 (Ministerio de Economía y Competitividad, Spain) and FQM329 (Junta de Andalucía), both co-funded by EU ERD funds.

[*] Corresponding author. Tel.: +34 687153536.
E-mail addresses: ecarrizosa@us.es (E. Carrizosa), vguerrero@us.es (V. Guerrero).

0305-0548/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cor.2013.04.012

Some other authors address sparsity in PCA with tools of feature selection. This is the case of [17], who called *principal variables* those which are considered to be relevant. Rejection methods are introduced and tested in artificial and real datasets in [12,13]. Also the well-known *variable neighborhood search* (VNS), [18,8], is adapted in [3] for developing two heuristics for feature selection in PCA. More recently, an exact method (branch-and-bound) is presented in [19].

In this paper we present a new procedure for obtaining a set of orthonormal vectors which are as sparse as desired, and minimizing the sum of squared distances between the data points and their projections. This is achieved by heuristically solving a MINLP with VNS. Our procedure has been tested in five datasets and compared with some benchmark methods in the literature, namely, VARIMAX, SCA and SPCA. In all cases we have outperformed the results given by these other procedures, both in terms of sparsity and errors.

The remainder of the paper is structured as follows. In Section 2, our problem is stated. Section 3 analyses an important particular case, namely, the case in which sparsity is obtained by forcing each original variable to appear with nonzero value in at most one component. The model is extended in Section 4 to give more freedom to the user to control sparsity. Section 5 reports our numerical experience, in which our method is compared against state-of-the-art procedures for sparse PCA, showing the advantages of our procedure. The paper ends with Section 6, in which some concluding remarks and future lines of research are outlined.

2. Problem statement

The special feature of our sparse PCA method against the classical one is the fact that we force principal components to have the most of their coordinates equal to zero. Our proposal, as in the classical PCA, tries to minimize the sum of the squares of the distances between the dataset and its projection on the vector space generated by the principal components. However, those k principal components will be forced to satisfy some sparsity constraints, which are basically that each variable is nonzero in at most r components, and each component has at most s nonzero elements. Since we have two parameters to control the model and make principal components sparse, we call our proposal *rs-sparse principal component analysis*, *rs*-SPCA.

So, we formulate the sparse problem as a MINLP by defining the following set of variables. First, let $c_1,\ldots,c_k$ denote the k vectors in $\mathbb{R}^n$ which are sought as principal components. We call c_{il} the l-th coordinate of the i-th principal component c_i, where $i=1\ldots k$ and $l=1\ldots n$. Binary variables z_{il} are defined as

$$z_{il}=\begin{cases}1 & \text{if } c_{il}\neq 0\\ 0 & \text{else}\end{cases}\qquad i=1,\ldots,k,\quad l=1,\ldots,n.$$

This assignment will let us control the sparsity of the principal components $c_1,\ldots,c_k$.

With the previous notation, our aim is to solve the following MINLP:

$$\min \frac{1}{p}\sum_{h=1}^{p}\|u_h-\pi_{\{c_1,\ldots,c_k\}}(u_h)\|^2$$
$$\text{s.t.}\begin{cases} c_i^{\top}c_j=\delta_{ij} & \forall i,j=1,\ldots,k\\ |c_{il}|\le z_{il} & \forall i=1,\ldots,k,\quad l=1,\ldots,n\\ \sum_{i=1}^{k} z_{il}\le r & \forall l=1,\ldots,n\\ \sum_{l=1}^{n} z_{il}\le s & \forall i=1,\ldots,k\\ z_{il}\in\{0,1\} & \forall i=1,\ldots,k,\quad l=1,\ldots,n\end{cases}$$

where $\delta_{ij}=1$ if $i=j$ and 0 otherwise.

Every constraint plays an important role in the achievement of the sparsity. The first one is given by the classical PCA, which forces orthonormality. The second constraint is obtained by imposing that, if $z_{il}=0$, then $c_{il}=0$, $i=1,\ldots,k$, $l=1,\ldots,n$. Finally, the constraints which really control the sparsity are the last two. In the third one, each variable is forced to appear in at most r components, with the fourth one every component guarantees to have no more than s nonzero loadings.

Denote the variance–covariance matrix (or correlation matrix) by V. The previous problem is equivalent to

$$\max \sum_{i=1}^{k} c_i^{\top}\cdot V\cdot c_i$$
$$\text{s.t.}\begin{cases} c_i^{\top}c_j=\delta_{ij} & \forall i,j=1,\ldots.k\\ |c_{il}|\le z_{il} & \forall i=1,\ldots,k,\quad l=1,\ldots,n\\ \sum_{i=1}^{k} z_{il}\le r & \forall l=1,\ldots,n\\ \sum_{l=1}^{n} z_{il}\le s & \forall i=1,\ldots,k\\ z_{il}\in\{0,1\} & \forall i=1,\ldots,k,\quad l=1,\ldots,n\end{cases}\tag{1}$$

We must point out that principal components given by (1) will be orthonormal but no longer uncorrelated.

In order to compare the results obtained with (1) against other procedures, a goodness criterion is introduced. We suggest proceeding in the same way as in the classical problem, which means to calculate the percentage of total variance explained by, in this case, *rs*-sparse principal components, as follows:

$$f=\frac{1}{\mathrm{tr(V)}}\sum_{i=1}^{k} c_i^{\top}\cdot V\cdot c_i\cdot 100,\tag{2}$$

where $\mathrm{tr}(V)$ is the trace of the variance–covariance or correlation matrix V. Other related measures can be found in [7].

In the following sections two special cases of (1), namely, $r=1$, and the general case, $r\geq 1$, will be analyzed.

3. The case $r=1$

Imposing in (1) the condition $r=1$ we are forcing that each variable appears just once. This induces a partition in the set of variables, i.e., those assigned to each component representing one cluster. Hence, interpretability will be high, though perhaps at expense of a decrease in variance explained.

In this case we have directly the orthonogonality of the components, so the constraint $c_i^{\top}\cdot c_j=\delta_{ij}$ becomes in $c_i^{\top}\cdot c_i=1$, $\forall i=1\ldots k$ and (1) is reduced to

$$\max \sum_{i=1}^{k} c_i^{\top}\cdot V\cdot c_i$$
$$\text{s.t.}\begin{cases} c_i^{\top}c_i=1 & \forall i=1,\ldots.k\\ |c_{il}|\le z_{il} & \forall i=1,\ldots,k,\quad l=1,\ldots,n\\ \sum_{i=1}^{k} z_{il}\le 1 & \forall l=1,\ldots,n\\ \sum_{l=1}^{n} z_{il}\le s & \forall i=1,\ldots,k\\ z_{il}\in\{0,1\} & \forall i=1,\ldots,k,\quad l=1,\ldots,n\end{cases}\tag{3}$$

However, if z_{il} could be fixed by any procedure, the resulting problem, *P(z)*, turns into a very familiar one, namely

$$\max \sum_{i=1}^{k} c_i^{\top}\cdot V\cdot c_i$$
$$\text{s.t.}\begin{cases} c_i^{\top}c_i=1 & \forall i=1,\ldots,k\\ c_{il}=0 & \forall i=1,\ldots,k,\quad l=1,\ldots,n: z_{il}=0\end{cases}\tag*{P(z)}$$

$P(z)$ can be expressed in a much more manageable form. Indeed, the problem is separable and can thus be split into k independent subproblems. In other words, $P(z)$ can be written as

$$\begin{aligned} &\sum_{i=1}^{k} \quad \{\max\ c_i^\top \cdot V \cdot c_i\} \\ &\text{s.t.} \begin{cases} c_i^\top \cdot c_i = 1 & \forall i = 1, \ldots, k \\ c_{il} = 0 & \forall i, l : z_{il} = 0 \end{cases} \end{aligned} \tag{4}$$

For $z = (z_1, \ldots, z_k)$ given, denote by $V_i^{z_i}$, $i=1,\ldots,k$, the symmetric matrix obtained from V by deleting all rows and columns l such that $z_{il} = 0$, $l=1,\ldots,n$. Let $d_i^{z_i}$ denote an optimal solution

$$\begin{aligned} &\max \quad d_i^\top \cdot V_i^{z_i} \cdot d_i \\ &\text{s.t.} \quad \{d_i^\top \cdot d_i = 1 \end{aligned} \tag{5}$$

Moreover, define the vector $c_i^{z_i}$ as

$$c_{il}^{z_i} = \begin{cases} d_{il}^{z_i} & \text{if } z_{il} \neq 0 \\ 0 & \text{else} \end{cases}$$

It immediately follows that the so-constructed $c_i^{z_i}$, $i=1,\ldots,k$ solve (4), and thus $P(z)$. In other words, $P(z)$ is solved by solving k problems of type (5), which, since $V_i^{z_i}$ is symmetric, its optimal solution corresponds to any unit eigenvector $d_i^{z_i}$ associated with the highest eigenvalue $\lambda_i^{z_i}$ of $V_i^{z_i}$.

Nevertheless, we must find a way for fixing those binary variables which make our initial problem turn into $P(z)$.

3.1. Problem resolution: fixing an initial feasible solution

Given V, the correlation or variance–covariance matrix, classical principal components $c^* = (c_1^*, \ldots, c_k^*)$ can be easily computed. We will find z_{il}, $i=1,\ldots,k$, $l=1,\ldots,n$, such that the sum of the absolute values of loadings of principal components are maximized by imposing constraints which are required in (3). A constraint forcing at least one variable appears in each principal component will be also included in this formulation. So, we want z_{il} to satisfy

$$\begin{aligned} &\max \sum_{i=1}^{k} \sum_{l=1}^{n} |c_{il}^*| z_{il} \\ &\text{s.t.} \begin{cases} \sum_{i=1}^{k} z_{il} \leq 1 & \forall l = 1, \ldots, n \\ \sum_{l=1}^{n} z_{il} \leq s & \forall i = 1, \ldots, k \\ \sum_{l=1}^{n} z_{il} \geq 1 & \forall i = 1, \ldots, k \\ z_{il} \geq 0 & \forall i = 1, \ldots, k, \quad l = 1, \ldots, n \end{cases} \end{aligned} \tag{6}$$

Problem (6) has a flow problem constraint structure, and it attains its optimal value at some point z^* with all coordinates $z_{il}^* \in \{0, 1\}$. Such z^* will be used as values for z in $P(z)$.

Combining problems (6) and $P(z)$, a solution (c,z), feasible in terms of sparsity and orthonormality, is achieved. However, it may not be optimal for (3), but it could be a good starting point for a search procedure such as VNS.

3.2. Problem resolution: improving the solution via a VNS algorithm

VNS is a metaheuristic for avoiding been trapped in a local optimum. It does not assure reaching the global one, but it sometimes achieves a good improvement on the solution initially given. Now we describe how VNS is customized for our problem (3).

As we have seen above, problem (3) is a MINLP in the variables (c,z). For $z = (z_1, \ldots, z_k)$ given, the optimal solution to $P(z)$, $c(z) = (c_1(z_1), \ldots, c_k(z_k))$, is obtained by calculating eigenvectors of the matrices $V_i^{z_i}$ defined above.

Our algorithm will perform a neighborhood search in the z-space, yielding $(c(z), z)$ as solution to problem (3).

Since z consists of the binary variables z_{il}, a natural neighborhood structure in the z-space is defined as follows: given z, satisfying the constraints in (6), for each radius $\rho = 1, \ldots, n$, the neighborhood $\mathcal{N}_\rho(z)$ of z is defined as the set of all z' fulfilling the constraints in (6) and obtained form z by exchanging ρ rows.

For instance, consider the case $n=4$, $k=2$, $r=1$, $s=2$, and z given by.

$$z = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \tag{7}$$

The elements in the neighborhood $\mathcal{N}_1(z)$ of radius 1 of z are the five matrices z' given in Table 1.

Observe that each element z induces a partition of the features space $1, 2, \ldots, n$ into $k+1$ clusters: For $l=1,\ldots,n$, variable l is associated with cluster i ($i=1,\ldots,k$) if $z_{il} = 1$; if $z_{il} = 0$ for all $i=1, \ldots, k$, then variable l is associated with the extra cluster $k+1$.

Hence, the elements z' in a neighborhood of radius ρ of z correspond to those partitions of the features obtained by exchanging ρ features form the partition in z.

For instance, z given in (7) induces the clustering {1}, {3,4}, {2} for the features {1,2,3,4}; the partitions in the neighborhood of radius 1 are shown in Table 2.

As a stopping criterion, let the algorithm looks for no more than Q_{max} solutions in each neighborhood and impose that the total running time does not exceed a given T_{max}.

Assuming these conditions, the implementation of the algorithm is done as is shown in Fig. 1.

4. General *r*

Let us now consider problem (1). We propose to proceed as we did in the case $r=1$. Firstly, we obtain the principal components $c^* = (c_1^*, \ldots, c_k^*)$ and then we obtain an approximation to c^* which is feasible (in terms of the number of nonzero elements) and as close as possible to c^*. To do that, we find an initial z by solving the following flow problem (which is identical to problem (6), except that r is an arbitrary integer number, $r \leq k$)

$$\max \sum_{i=1}^{k} \sum_{l=1}^{n} |c_{il}^*| z_{il}$$

Table 1
Example $\mathcal{N}_1(z)$.

z'	Perturbed row
–	1
$z'_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$	2
$z'_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}, z'_3 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$	3
$z'_4 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, z'_5 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$	4

$$\text{s.t.}\begin{cases}\sum_{i=1}^{k} z_{il} \le r & \forall l = 1\ldots n\\ \sum_{l=1}^{n} z_{il} \le s & \forall i = 1\ldots k\\ \sum_{l=1}^{n} z_{il} \ge 1 & \forall i = 1\ldots k\\ z_{il} \ge 0 & \forall i, l\end{cases} \tag{8}$$

Let z^* be the optimal solution of the flow linear program (8), and then we find $c(z^*)$ by solving the nonlinear continuous program

$$\max \sum_{i=1}^{k} c_i^\top \cdot V \cdot c_i$$
$$\text{s.t.}\begin{cases} c_i^\top c_j = \delta_{ij} & \forall i, j\\ c_{il} = 0 & \forall i, l : z^*_{il} = 0\end{cases} \tag{9}$$

Orthogonality constraints cannot be simplified anymore, contrary to what we did in the previous section. Hence, a general-purpose optimization routine is to be used to solve this problem.

5. Computational experiments

In this section we analyze the empirical behavior of our *rs*-SPCA, which is compared against benchmark procedures in the literature, namely rotation methods (varimax criterion) [15], SCA, [23] and SPCA, [28]. Those different procedures for achieving simplicity will be carried out on datasets proposed in [22]. The numerical experience will show that *rs*-SPCA provides principal components which outperform, in terms of sparsity and error minimization, the ones given by the procedures cited above.

Table 2
Partition.

z'	Partition
z'_1	{1,2}, {3,4}, {∅}
z'_2	{1}, {4}, {2,3}
z'_3	({1,3}, {4}, {2}
z'_4	{1}, {3}, {2,4}
z'_5	{1,4}, {3}, {2}

5.1. Data description

`Hearing loss` data, come from [9], who considered the first four principal components in their study. It consists of eight measurements of hearing loss taken on 100 males who had no indication of hearing difficulties. The correlation matrix is available in `R` package `sca`.

`Reflexes` data were described in [14], gives measurements of strength of reflexes at 10 sites of the body, taken form 143 individuals. As the previous data, reflexes correlation matrix can also be found in `R` package `sca`.

`Pitprop`, data [10], is an example which has been proven the difficult achievement of simplicity in the past. About 13 variables arised from a study on the strength of pitprops cut from home-grown timber, measured in 180 observations. Six principal components were calculated. Correlation matrix can be found in `R` package `elasticnet`.

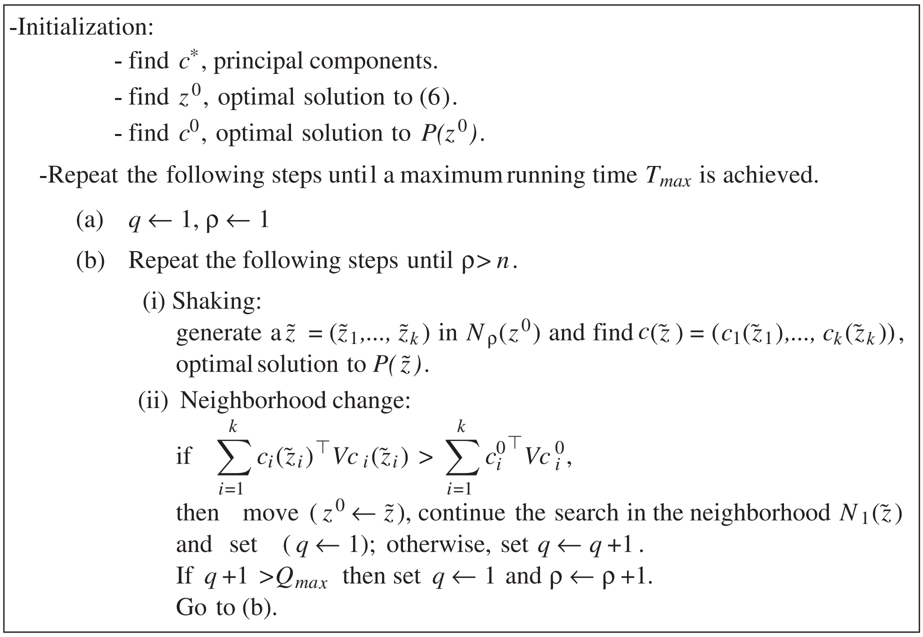
`Associated Movements` data were introduced in [16], where 10 measures of the amount of associated movements observed during some finemotor and grossmotor tasks in 484 children and adolescents were taken. Five principal components are proposed to be calculated.

`Muscle Strength` data consist of 51 maximal isometric muscle strength measurements made on all areas of the body from 569 healthy subjects. The first six principal components will be calculated. See [22].

5.2. Experiments description and results

In order to show the power of *rs*-SPCA, several computational experiments have been carried out. The measure of goodness used is the percentage of the variance explained by the basis selected, as given by (2).

We compare our *rs*-SPCA ($r=1$) against benchmark methods. We have implemented our procedure in Matlab, using the routines `linprog` to numerically solve the flow problems (6) and (8), while the routine `fmincon` is used to solve (9). VNS was implemented as described in Section 3.2, with $Q_{max} = 5$ trials within each neighborhood, and a time limit



```
-Initialization:
        - find $c^*$, principal components.
        - find $z^0$, optimal solution to (6).
        - find $c^0$, optimal solution to $P(z^0)$.
  -Repeat the following steps until a maximum running time $T_{max}$ is achieved.
    (a)  $q \leftarrow 1, \rho \leftarrow 1$
    (b)  Repeat the following steps until $\rho > n$.
         (i) Shaking:
             generate a $\tilde{z} = (\tilde{z}_1,\ldots,\tilde{z}_k)$ in $N_\rho(z^0)$ and find $c(\tilde{z}) = (c_1(\tilde{z}_1),\ldots,c_k(\tilde{z}_k))$,
             optimal solution to $P(\tilde{z})$.
         (ii) Neighborhood change:
             if $\sum_{i=1}^{k} c_i(\tilde{z}_i)^\top V c_i(\tilde{z}_i) > \sum_{i=1}^{k} c_i^{0\top} V c_i^0$,
             then move $(z^0 \leftarrow \tilde{z})$, continue the search in the neighborhood $N_1(\tilde{z})$
             and set $(q \leftarrow 1)$; otherwise, set $q \leftarrow q+1$.
             If $q+1 > Q_{max}$ then set $q \leftarrow 1$ and $\rho \leftarrow \rho+1$.
             Go to (b).
```

Fig. 1. VNS algorithm.

T_{max} of 15, 30, 45, 60 and 100 s respectively for the different datasets.

Varimax, SCA and SPCA algorithms are included in `stats`, `sca` and `elasticnet` packages of `R` software, respectively. For SCA, two parameters, namely, the number of block and difference components, must be chosen a priori. We have followed [22] to chose such parameters.

SPCA also depends on two parameters. For the quadratic penalty, we follow [28], and fix it to 0, since the number of variables is lower than the number of observations in all our datasets. The second parameter has been chosen so that the so-obtained principal components have the same non-zero coordinates than the ones given by our method. In other words, we are giving valuable information to the user, since we are already giving a clue on the distribution of non-zeros.

The results are given in Table 3. The first column gives the name of the dataset. Columns 2 and 3 give respectively n (the number of variables) and k (the number of components to be obtained). Then, for the different methods tested (PCA, VARIMAX, SCA, SPCA and our *rs*-SPCA), we represent the sparsity (as the percentage of zeros in the components) and the percentage of variance explained by the components, f, as given by (2).

We conclude from Table 3 that, in terms of sparsity, our procedure clearly outperforms PCA as well as the methods seeking sparseness. This happens for all methods except SPCA, for which, thanks to the extra information we provided, gives identical sparsity results to ours, but at the expense of a lower variance explained. Unlike in our proposal, orthogonality is not ensured in SPCA or the remaining sparse procedures. In other words we have proposed a method with a high sparsity and a low loss in terms of the variance explained, as compared with the classical PCA.

Table 3
Percentages of zeros of the components and percentages of explained variance.

Datasets	n	k	Sparsity	PCA	VARIMAX	SCA	SPCA	*rs*-SPCA
`Hearing Loss`	8	4	%zeros	0	0	12.50	75.00	75.00
			f	87.37	68.40	85.40	84.08	85.37
`Reflexes`	10	5	%zeros	0	16.00	52.00	80.00	80.00
			f	97.05	72.20	91.50	96.17	96.70
`Pitprop`	13	6	%zeros	1.28	7.69	80.77	83.33	83.33
			f	87.00	78.90	74.80	71.99	76.85
`Movements`	22	5	%zeros	0	2.27	20.45	75.00	75.00
			f	55.00	43.20	53.80	49.84	53.60
`Musclestrength`	51	6	%zeros	1.63	8.50	34.64	80.07	80.07
			f	70.40	70.39	68.10	60.00	61.39

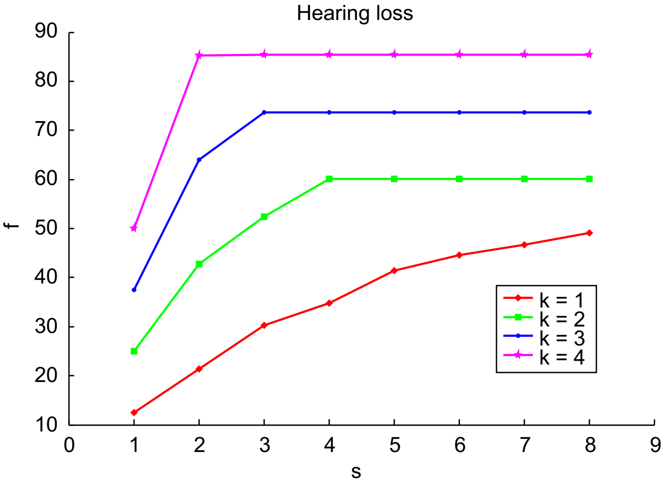


Fig. 2. f varying s and k for Hearingloss dataset.

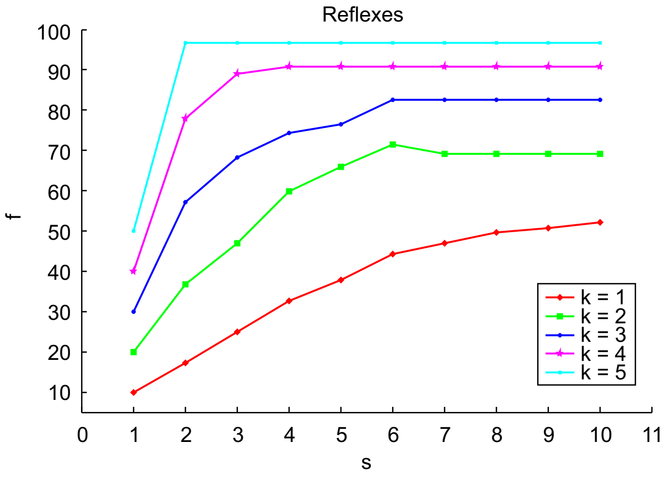


Fig. 3. f varying s and k for Reflexes dataset.

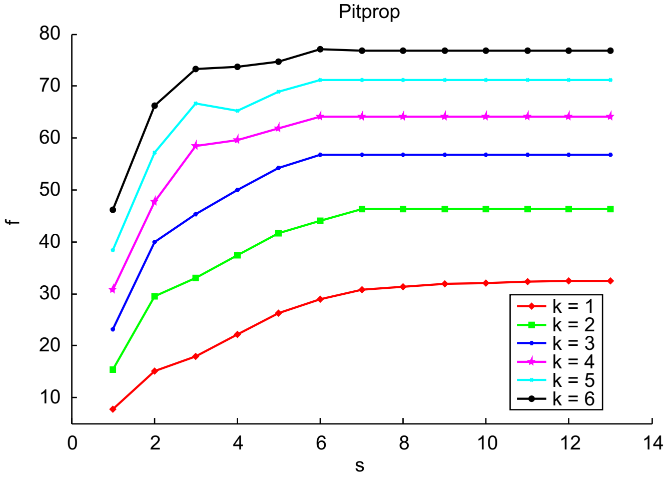


Fig. 4. f varying s and k for Pitprop dataset.

A second analysis has been done to study how sensitive our procedure is with respect to the choice of k and s. The performance of our method on the different test sets is depicted in Figs. 2–6, which show f for $r=1$ and varying s and k. These figures correspond to the case $r=1$, as analyzed in Section 3.

For the case of an arbitrary r, as discussed in Section 4, the local search routine may fail to find a feasible solution. Table 4 shows the percentages of feasible solutions obtained for the different datasets, when r is varying between 1 and k, and s varies between 1 and n. We conclude that failing to obtain feasibility is rare in small-dimensional datasets, while for the larger set ($n=51$), a feasible solution was obtained in less than 40% of the instances.

6. Conclusions

In this paper we have introduced a new dimensionality reduction method which ensures sparsity of the procedure. We have modeled the problem as a MINLP, heuristically solved by using VNS. The numerical experience reported shows that our procedure outperforms, in terms of error minimization and sparsity, benchmark methods in the literature. Several research lines remain open. Although for high-dimensional problems heuristics such as VNS seem to be the only feasible strategy, it would be interesting to be able to solve the MINLPs obtained with exact methods. Plugging directly the problem into state-of-the-art MINLP branch-

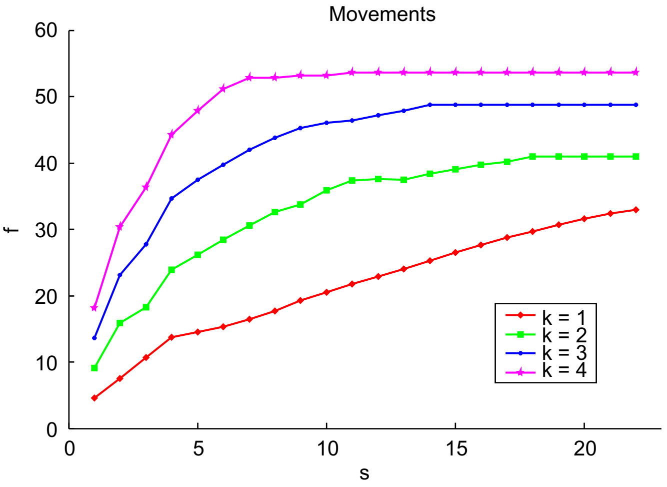


Fig. 5. *f* varying *s* and *k* for Movements dataset.

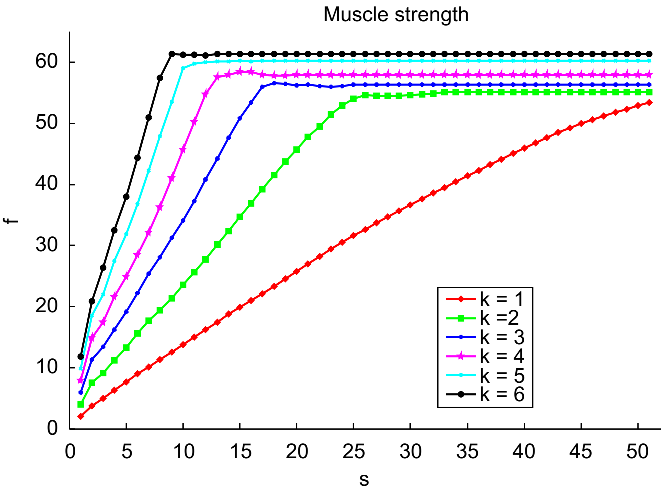


Fig. 6. *f* varying *s* and *k* for Musclestrength dataset.

Table 4
Percentage of feasibility in *rs*-SPCA general problem.

Datasets	Feasibility
Hearing Loss	100%
Reflexes	100%
Pitprop	94.87%
Movements	92.04%
Musclestregth	38.89%

and-bound methods such as Couenne, [2], was unsuccessful, since no solution was provided in reasonable times. Our method, as well as its competitors, seeks orthogonal vectors, but the output is not uncorrelated. Seeking a basis (not necessarily orthogonal) yielding uncorrelated projections should lead to more challenging optimization problems, which are now under study.

Since both uncorrelation and orthogonality are of interest, addressing a biobjective problem trading-off components correlation and orthogonality, but maintaining a high variance explained, is also a challenging problem which deserves a deeper analysis.

Acknowledgment

We are grateful to Professor Valentin Rousson for providing us the muscle strength correlation matrix.

References

[1] Anaya-Izquierdo K, Critchley F, Vines K. Orthogonal simple component analysis: a new, exploratory approach. The Annals of Applied Statistics 2011;5(1):486–522.
[2] Belotti P. Couenne: a user's manual. Technical Report, Lehigh University; 2009.
[3] Brusco M, Singh R, Steinley D. Variable neighborhood search heuristics for selecting a subset of variables in principal component analysis. Psycometrika 2009;74(4):705–26.
[4] Cadima J, Jolliffe T. Loading and correlation in the interpretation of principal components. Journal of Applied Statistics 1995;22:203–14.
[5] d'Aspremont A, Ghaoui L, Jordan M, G. L. Direct formulation of sparse PCA using semidefinite programming. SIAM Review 2007;49(3):434–48.
[6] Farcomeni A. An exact approach to sparse principal component analysis. Computational Statistics 2009;24(583–604).
[7] Gervini D, Rousson V. Criteria for evaluating dimension-reducing components for multivariate data. The American Statitician 2004;58:72–6.
[8] Hansen P, Mladenović N. Variable neighbourhood search: methods and applications. Annals of Operations Research 2010;175:35–45.
[9] Jackson JE. A user's guide to principal components. New York: Wiley; 1991.
[10] Jeffers J. Two case studies in the application of principal component analysis. Applied Statistics 1967;16(3):225–36.
[11] Jollife IT, Trendafilov NT, Uddin M. A modified principal component technique based on the Lasso. Journal of Computational and Graphical Statistics 2003;12:531–47.
[12] Jolliffe IT. Discarding variables in a principal component analysis I: artificial data. Applied Statistics 1972;21(2):160–73.
[13] Jolliffe IT. Discarding variables in a principal component analysis II: artificial data. Applied Statistics 1973;22(1):21–31.
[14] Jolliffe IT. Principal component analysis. 2nd ed. New York: Springer; 2002.
[15] Kaiser HF. The varimax criterion for analytic rotation in factor analysis. Psychometrica 1958;23(3):187–200.
[16] Largo R, Caflisch J, Hug F, Muggli K, Molnar A, Molinari L. Neuromotor development from 5 to 18 years. Part 2: associated movements. Developmental Medicine on Child Neurology 2001;43:444–53.
[17] McCabe GP. Principal variables. Technometrics 1984;26(2):137–44.
[18] Mladenović N, Hansen P. Variable neighborhood search. Computers and Operations Research 1997;24:1097–100.
[19] Pacheco J, Casado S, Porras S. Exact methods for variable selection in principal component analysis: guide functions and pre-selection. Computational Statistics & Data Analysis 2013;57:95–111.
[20] Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 1901;2:559–72.
[21] Qi X, Luo R, Zhao H. Sparse principal component analysis by choice of norm. Journal of Multivariate Analysis 2013;114:127–60.
[22] Rousson V, Gasser T. Some case studies of simple component analysis, ⟨http://www.biostat.uzh.ch/research/manuscripts/scacases.pdf⟩; 2003.
[23] Rousson V, Gasser T. Simple component analysis. Applied Statistics 2004;53(4):539–55.
[24] Sabatier R, Reynès C. Extensions of simple component analysis and simple linear discriminant analysis using genetic algorithms. Computational Statistics & Data Analysis 2008;52:4779–89.
[25] Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 1996;58:267–88.
[26] Vines S. Simple principal components. Applied Statistics 2000;49(4):441–51.
[27] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B 2005;(67):301–20.
[28] Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. Journal of Computational and Graphical Statistics 2006;15(2):265–86.