Comput Stat
DOI 10.1007/s00180-013-0434-5

ORIGINAL PAPER

From simple structure to sparse components: a review

Nickolay T. Trendafilov

Received: 4 September 2012 / Accepted: 5 July 2013

Abstract The article begins with a review of the main approaches for interpretation the results from principal component analysis (PCA) during the last 50–60 years. The simple structure approach is compared to the modern approach of sparse PCA where interpretable solutions are directly obtained. It is shown that their goals are identical but they differ by the way they are realized. Next, the most popular and influential methods for sparse PCA are briefly reviewed. In the remaining part of the paper, a new approach to define sparse PCA is introduced. Several alternative definitions are considered and illustrated on a well-known data set. Finally, it is demonstrated, how one of these possible versions of sparse PCA can be used as a sparse alternative to the classical rotation methods.

Keywords Simple structure loadings · Orthogonal and oblique rotations · Factor analysis · Sparse component loadings · Sparseness inducing constraints · LASSO · Constrained optimization on matrix manifolds · Projected gradients

1 Introduction

The article begins with a review of the approaches for interpretation the results from principal component analysis (PCA) during the last 50–60 years. It will become clear that the modern trend to look for sparse component loadings is not entirely new. For this reason, the simple structure approach and its features are discussed with respect to the modern methods for sparse PCA. The main types of simple structure rotations are

N. T. Trendafilov (✉)
Department of Mathematics and Statistics, The Open University,
Walton Hall, Milton Keynes MK7 6AA, UK
e-mail: N.Trendafilov@open.ac.uk

Published online: 16 July 2013

briefly outlined. It is demonstrated that the simple structure approach relies on sparse component loadings too, which are, unfortunately, found in a suboptimal (loose) way.

Next, the most popular and influential methods for sparse PCA are briefly reviewed. Then, they are classified according to the form of the objective function and constraints involved in their definitions. It will be seen that there is one way to define sparse PCA, known as *function-constrained form*, which is not explored so far. The remaining part of the paper follows this way. Several possible definitions of sparse PCA are considered and illustrated on the well-known Pitprop data (Jeffers 1967). It is be demonstrated that one of them can serve as an alternative to the simple structure rotation methods in PCA, in a sense that it possesses most of their features but produces sparse loadings, and thus, unifies the different trends in the PCA interpretation.

Detailed study of the new function-constrained form of sparse PCA will be considered elsewhere, because of the review nature of the paper. The numerical examples in the paper are solved by projected gradient method.

2 PCA revisited

2.1 Origins and background

Before considering the interpretation problem in PCA, let us summarize its main features. PCA was first defined in the form that is used nowadays by Pearson (1901). He found the best-fitting line in the least squares sense to the data points, which is known today as the first principal component. Hotelling (1933) showed that the loadings for the components are the eigenvectors of the sample covariance matrix. We should note, that PCA as a scientific tool appeared first in psychological journals. This should not be very surprising if we recall that the first paper on singular value decomposition (SVD) was published in Psychometrika (Eckart and Young 1936). This simply reflects the changing nature of the human practical and intellectual priorities. The quantitative measurement of the human intelligence and personality in the 30's of the last century should have been of the same challenging importance as the contemporary interest in gene and tissue engineering, signal processing, or analyzing huge climate, financial or internet data.

Now, PCA is well known and efficient technique for reducing the dimension of high-dimensional data (Jolliffe 2002). Suppose that a vector of p random variables x is observed, and let R denote the sample correlation matrix of x. PCA forms p new 'variables', called principal components (PCs), which are linear combinations of the original ones with the following *unique* properties: they are ordered according to their variances magnitudes, are uncorrelated and the vectors of their coefficients, called component loadings, are orthogonal.

Formally, the ith PC ($i = 1, \ldots, p$) is the linear combination $y_i = a_i^\top x$ with maximum variance $a_i^\top R a_i$ subject to $a_i^\top a_i = 1$ and $a_i^\top a_j = 0$ for $i > j$. This is variational definition of the eigenvalue decomposition (EVD) of symmetric R. The vector a_i contains the component loadings of the ith PC and is the ith eigenvector corresponding to the ith largest eigenvalue of R. In practice, the first r ($\ll p$) PCs,

accounting for the majority of the variation in the original variables, are used for further analysis which is known as *dimension reduction*.

Let a vector of p random variables be observed on each of n individuals, and the measurements be collected in an $n \times p$ data matrix X. From now on we assume that X is centered, i.e. $X^\top 1_n = 0_p$, and the columns (variables) have unit length, i.e. $\mathrm{diag}(X^\top X) = 1_p$ and diag() is the main diagonal of the matrix argument. The EVD of R in the PCA definition given above, can be replaced by the SVD of X, i.e. $X = FDA^\top$, where the diagonal elements of D are in decreasing order. For any $r(\ll p)$, let $X_r = F_r D_r A_r^\top$ be the truncated SVD of X. As usual, F_r and A_r denote the first r columns of F and A respectively ($F_r^\top F_r = A_r^\top A_r = I_r$), and D_r is diagonal matrix with the first r singular values of X. The matrices $Y_r = F_r D_r$ and A_r contain the component scores and loadings respectively, and D_r^2—the variances of the first r PCs.

A right multiplication of $X = FDA^\top$ by A_r gives:

$$XA_r = FDA^\top A_r = FD\begin{bmatrix} I_r \\ 0_{(p-r)\times r} \end{bmatrix} = F\begin{bmatrix} D_r \\ 0_{(p-r)\times r} \end{bmatrix} = F_r D_r = Y_r, \qquad (1)$$

i.e. the principal components Y_r are linear combination of the original variables X with coefficients A_r. Right multiplication of (1) by $A_r^\top$ gives $X_r = XA_rA_r^\top$, i.e. the best approximation X_r to X of rank r is given by the orthogonal projection of X onto the r-dimensional subspace in $\mathbb{R}^p$ spanned by the columns of A_r. That is, X_r is low-dimensional representation of the original data X.

The variances of the first r PCs are given by

$$Y_r^\top Y_r = A_r^\top X_r^\top X_r A_r = D_r F_r^\top F_r D_r = D_r^2, \qquad (2)$$

which also shows that the principal components are uncorrelated, as $Y_r^\top Y_r$ is diagonal. Then, the total variance explained by the first r PCs is given by $\mathrm{trace} D_r^2 = \mathrm{trace}(A_r^\top X_r^\top X_r A_r)$. In the sequel, the index r is omitted when it is clear from the context whether the original or the truncated matrix is in use.

2.2 PCA interpretation

The PCA is interpreted by considering the magnitudes of the component loadings, which indicate how strongly each of the original variables contribute to the PC. PCs are really useful if they can be simply interpreted. However, this is not often the case. Even if a reduced dimension of r PCs is considered for further analysis, each PC is still a linear combination of *all* original variables. This complicates the PCs interpretation, especially when p is large. The usual *ad hoc* practice in the PC interpretation is to ignore the variables with small absolute loadings or set to zero loadings smaller than some threshold value. This makes the PCs sparse artificially as no care is taken on how well they fit the data. Not surprisingly, such a practice is found to be misleading especially for PCs computed from a covariance matrix (Cadima and Jolliffe 1995).

The oldest approach to solve the PCA interpretation problem is the simple structure rotation, initially designed in factor analysis (FA) (Thurstone 1935) and later adapted to PCA to make the components as interpretable as possible (Jolliffe 2002, Chap. 11). It is based on the following simple identity:

$$X = FDA^\top = FQQ^{-1}DA^\top, \tag{3}$$

where Q can be any non-singular transformation matrix. Two types of matrices Q are used in PCA and FA: orthogonal and oblique. A non-singular $r \times r$ matrix Q is called oblique if $Q^\top Q$ is correlation matrix. As the orthogonal matrix Q can be viewed as a rotation, these methods are known as the rotation methods.

Here, we stress that the traditional PCA employs (3) and refers to AD as a loadings matrix, which differs from the notations in Sect. 2.1. This is a consequence from the fact that the rotation methods were first introduced in FA and then, adopted in PCA, which introduces a number of problems (Jolliffe 2002, p.272). The PC interpretation relies on either AD or its rotated version $ADQ^{-\top}$. In contrast, in sparse PCA the notations from Sect. 2.1 are followed, and the sparsification and interpretation are based on A.

The simple structure rotation approach proceeds as follows. First, the dimension reduction is performed, i.e. the appropriate r is chosen, and A and D are obtained. Then, a rotation Q is found by optimizing certain simple structure criterion, such that the resulting $ADQ^{-\top}$ has simple structure. The belief, that the rotated component has its absolute loadings near 1 or 0 while avoiding intermediate values, is usually false and makes the approach ambiguous. Its application is additionally complicated by the huge number of simplicity criteria to choose from. Another important shortcoming is that the rotated components lack the nice PCs properties to be uncorrelated and explain successively decreasing amount of variance.

The main types of rotation methods are briefly outlined in the next Section. Browne (2001) gives a comprehensive overview of the field. Some additional aspects are discussed in Jennrich (2007). Details can be found in the cited papers there and in the standard texts on FA (Harman 1976; Mulaik 2010). To save space, no reference to original papers on rotation methods are given.

3 Why sparse PCA?

In this Section we show that the classical approach to PCA interpretation—the simple structure rotation—implicitly relies on the sparse representation of the component loadings.

3.1 Simple structure concept in PCA (and FA)

It was mentioned in Sect. 2.2 that the goal of the simple structure approach to the PCA interpretation is to find orthogonal or oblique matrix Q such that the resulting loadings $ADQ^{-\top}$ have simple structure.

The original simple structure concept was introduced in FA as three rules (Thurstone 1935, p.156), which were later extended and elaborated in (Thurstone 1947, p.335). In more contemporary language they look as follows (Harman 1976, p.98), where the term "factor matrix" used in FA context should be understood as component loadings matrix for the PCA case:

1. Each row of the factor matrix should have at least one *zero*,
2. If there are r common factors each column of the factor matrix should have at least r *zeros*,
3. For every pair of columns of the factor matrix there should be several variables whose entries *vanish* in one column but not in the other,
4. For every pair of columns of the factor matrix, a large proportion of the variables should have *vanishing* entries in both columns when there are four or more factors,
5. For every pair of columns of the factor matrix there should be only a small number of variables with *non-vanishing* entries in both columns.

The words in italic are made by me to stress that the original simple structure concept requires, in fact, *sparse* loadings. Unfortunately, this sparseness has never been achieved by the classical rotation methods.

There are situations when the first PC is a measure of "overall size" with non-trivial loadings for all variables. The same phenomenon occurs in the so-called Bi-factor analysis (Harman 1976, pp.120–7). In such PCA/FA solutions, the five rules are applied to the remaining PCs/factors.

Let us take a closer look at how the simple structure concept is implemented in a number of rotation methods. It seems that the oldest rotation method was graphical. Its main conceptual drawback is that it is subjective. In order to make the interpretation process objective, analytic rotation methods were introduced. The heart of these methods is the simplicity criterion, which is supposed to express the Thurstone's five rules as mathematical formula. Unfortunately, it turned out that this is a very difficult task. As a results there were produced a huge number of criteria, because none of them is capable to produce satisfying solution for any problem. To get some flavor of what is this simple structure criterion, consider VARIMAX (VARiance MAXimization) (Mulaik 2010, p.310). This is arguably the most popular (but not the best) criterion, because it produces satisfactory results in many applications and is based on a simple to explain concept. Let $C = ADQ$ be the $p \times r$ matrix of orthogonally rotated loadings. The variance of the squared loadings of the jth rotated column c_j is:

$$\mathcal{V}_j = \mathcal{V}(c_j) = \sum_{i=1}^{p} c_{ij}^4 - \frac{1}{p}\left(\sum_{i=1}^{p} c_{ij}^2\right)^2. \tag{4}$$

The variance $\mathcal{V}_j$ will be large if there are few large squared loadings and all the rest are near zero. The variance $\mathcal{V}_j$ will be small when all squared loadings have similar magnitudes. The VARIMAX rotation problem is to find an $r \times r$ orthogonal matrix Q such that the total variance $\mathcal{V} = \sum_{j=1}^{r} \mathcal{V}_j$ is maximized.

The simple structure rotation problems are optimization problems subject to orthogonality or oblique constraints. The original VARIMAX and the other early rotation

algorithms are based on successive planar rotations of all possible $r(r-1)/2$ pairs of factors to find the optimal Q. They can be solved in a unified way by differential geometric methods and algorithms, e.g. (Edelman et al. 1998; Diele et al. 1998; Trendafilov and Lippert 2002). Chu and Trendafilov (1998), Trendafilov (1999) employed projected gradient approach to construct a general matrix algorithm for solving arbitrary rotation problems. Now, this approach is routinely used in modern FA (Mulaik 2010).

There are a number of alternative rotation methods which do not rely on any particular simplicity criterion. In contrast, they search for well defined configurations (of variables), e.g. hyperplanes. They can be divided into two categories: hyperplane fitting rotations and hyperplane counting methods. The hyperplane fitting rotations "refine" the solution C from certain analytic rotation (Mulaik 2010, p.342–347). They construct a target matrix T reflecting the simple structure of the problem, based on the already obtained C. Then the new solution CP_∞ is obtained by Procrustes fitting of C to the target T, where $P_\infty = \operatorname{argmin}_P \|T - CP\|_F$ and $\| \ \|_F$ denotes the Frobenius norm of the argument. It is interesting to note that one of these methods, PROMAJ (PROcrustes MAJorization) (Mulaik (2010), p.343), constructs a simple structure target matrix by a procedure identical to what is known now as *soft-thresholding* (Donoho and Johnstone 1994). The soft thresholding operator is defined as $\eta_S(\alpha, \delta) = \operatorname{sgn}(\alpha) \max\{|\alpha| - \delta, 0\}$, for any scalar α and threshold δ. In both cases, the expected result is the same: the soft-thresholding operator gives the best *sparse* linear regressor in least squares sense subject to LASSO (Least Absolute Shrinkage and Selection Operator) constraint, while PROMAJ produces a *sparse* target matrix majorizing the relative contributions of each component (column) (Marshall and Olkin (1979), Ch.5,B).

The hyperplane counting methods are also introduced by Thurstone (1947). As suggested by their name, they count the variables, close to certain hyperplane, and then try to maximize it. Recently this rationale inspired the introduction of a class of rotation criteria called *component loss functions* (CLF) (Mulaik 2010, p.360–66). The most intriguing of them is given by the ℓ_1 matrix norm of the rotated loadings $\|C\|_1 = \sum_i^p \sum_j^r |c_{ij}| = \operatorname{trace}[C^\top \operatorname{sign}(C)]$ which should be minimized over a set of admissible rotations.

3.2 PCA interpretation via rotation methods

Traditionally, PCs are considered easily interpretable if there are plenty of small component loadings indicating the negligible importance of the corresponding variables. The rotation methods are incapable to produce vanishing (exactly zero) loadings. As a consequence, the "classic" way to apply the simple structure concept for interpretation is to ignore (effectively set to zero) component loadings whose absolute values fall below some threshold (Jolliffe 2002, p. 269). Thus, the PCs simplicity and interpretability are implicitly associated with their *sparseness*. This will be illustrated by the following example, on a data set, first used by Jolliffe et al. (2003), and then became a standard example in any work on sparse approximation of PCA.

Jeffers's Pitprop data example: The Pitprop data contains 13 variables measured for 180 pitprops cut from Corsican pine timber (Jeffers 1967). Denote by $x_1, x_2, \ldots, x_{13}$ the variables in the order they appear in the cited paper. Unfortunately,

Table 1 Jeffers's Pitprop initial loadings and their interpretation

Vars	Component loadings (AD)						Jeffers's interpretation after normalization					
	1	2	3	4	5	6	1	2	3	4	5	6
x_1	0.83	0.34	−0.28	−0.10	0.08	0.11	1.0					
x_2	0.83	0.29	−0.32	−0.11	0.11	0.15	1.0					
x_3	0.26	0.83	0.19	0.08	−0.33	−0.25		1.0				
x_4	0.36	0.70	0.48	0.06	−0.34	−0.05		0.84	0.73			
x_5	0.12	−0.26	0.66	0.05	−0.17	0.56			1.0			1.0
x_6	0.58	−0.02	0.65	−0.07	0.30	0.05	0.70		0.99			
x_7	0.82	−0.29	0.35	−0.07	0.21	0.00	0.99					
x_8	0.60	−0.29	−0.33	0.30	−0.18	−0.05	0.72					
x_9	0.73	0.03	−0.28	0.10	0.10	0.03	0.88					
x_{10}	0.78	−0.38	−0.16	−0.22	−0.15	−0.16	0.93					
x_{11}	−0.02	0.32	−0.10	0.85	0.33	0.16				1.0		
x_{12}	−0.24	0.53	0.13	−0.32	0.57	−0.15					1.0	
x_{13}	−0.23	0.48	−0.45	−0.32	−0.08	0.57						1.0

the raw Pitprop data seem lost, and their correlation matrix is available only (Jeffers 1967, Table 2). Following Jeffers (1967), six PCs are to be considered. The first six eigenvalues of the correlation matrix are: 4.2186, 2.3781, 1.8782, 1.1094, 0.9100 and 0.8154. Their sum is 11.3098, i.e. the total variance explained by the first six PCs is 86.9 %. The component loadings are given in the first six columns of Table 1.

The loadings are interpreted by normalizing first each column to have maximal magnitude one, and then, considering only the loadings greater than 0.7 (Jeffers 1967, pp.229–230). Effectively, the interpretation is based on the sparse loadings matrix given in the last six columns of Table 1. We can check that they fulfill the five rules of the simple structure. One may not be completely satisfied, because the first column has exactly six zeros. Indeed, this complies with the second rule for at least six zeros, but one feels tempted to improve it.

Can we obtain a clearer interpretation with the help of simple structure rotation? We apply VARIMAX from MATLAB (2011) to rotate the first six component loadings of the Jeffers's Pitprop data. The rotated loadings are depicted in the first six columns of Table 2. Indeed, the VARIMAX loadings look better. This conclusion is mainly based on the observation that the last three columns have now clearly dominating loadings. As Jeffers, to interpret the loadings, we normalize them first. Then we have to choose a threshold to discard the "non contributing" variables. If we take 0.7 as in the previous example, we will end up with three empty (zero) rows, i.e. three variables do not contribute at all, which seems unreasonable. By taking 0.59 as a threshold value we find quite clear simple structure which is depicted in the last six columns of Table 2. Each variable contribute to only one component and the first column has now eight zeros. However, the weakness of both interpretations is that the choice of the threshold is completely subjective. The "sparse" matrices obtained after normalization

Table 2 Jeffers's Pitprop rotated loadings and their interpretation

Vars	VARIMAX loadings						Normalized loadings greater than 0.59					
	1	2	3	4	5	6	1	2	3	4	5	6
x_1	0.91	0.26	−0.01	0.03	0.01	0.08	0.97					
x_2	0.94	0.19	−0.00	0.03	0.00	0.10	1.0					
x_3	0.13	0.96	−0.14	0.08	0.08	0.04		1.0				
x_4	0.13	0.95	0.24	0.03	0.06	−0.03		0.98				
x_5	−0.14	0.03	0.90	−0.03	−0.18	−0.03			1.0			
x_6	0.36	0.19	0.61	−0.03	0.28	−0.49			0.68			
x_7	0.62	−0.02	0.47	−0.13	−0.01	−0.55	0.66					
x_8	0.54	−0.10	−0.10	0.11	−0.56	−0.23					−0.64	
x_9	0.77	0.03	−0.03	0.12	−0.16	−0.12	0.82					
x_{10}	0.68	−0.10	0.02	−0.40	−0.35	−0.34	0.73					
x_{11}	0.03	0.08	−0.04	0.97	0.00	−0.00				1.0		
x_{12}	−0.06	0.14	−0.14	0.04	0.87	0.09					1.0	
x_{13}	0.10	0.04	−0.07	−0.01	0.15	0.93						1.0

and thresholding quite loosely reflect the fitting power of the loadings they are obtained from.

4 Analyzing high-dimensional multivariate data

The modern data analysis faces analyzing high-dimensional multivariate data. As we mentioned in the previous Section, traditionally the first step in data analysis is some kind of low-dimensional data representation, e.g. performing PCA (or FA), and then, followed by interpretation.

However, the main problems in modern applications is that PCA might be too slow, and that the results from PCA involve all input variables which complicates the interpretation. For problems involving thousands of variables it is natural to look for solutions involving fairly limited part of them, which, in other words, calls for sparseness. Let $\|a\|_0$ denote the cardinality of $a \in \mathbb{R}^p$, i.e. the number of its non-zero entries. Then a is called sparse if $\|a\|_0 \ll p$. Usually it is more convenient to work with $\|a\|_1$, which also promotes sparsity in a, but avoids using discrete variables.

4.1 Abandoning the rotation methods

A possible mode of attack is to modify PCA to explicitly produce simple PCs in a sense that the values of the component loadings are restricted. The first such method was proposed by Hausman (1982). It finds PC loadings from a prescribed subset of values, say $S = \{-1, 0, 1\}$. Later this idea was extended to arbitrary integers (Vines 2000).

The rotation methods cannot produce sparse loading because they are designed to simplify the loadings while preserving the already chosen percentage of variance. Therefore it seems reasonable to relax this requirement by putting more emphasis on simplicity than on variance maximization. Jolliffe and Uddin (2000) are the first to modify the original PCs to additionally satisfy the VARIMAX criterion still explaining successively decreasing portion of the variance. They solve the following problem:

$$\max_{a^\top a=1} a^\top Ra + \tau \mathcal{V}(a),$$

which is the first penalized formulation of PCA designed to directly produce interpretable component loadings A. The method is called SCoT (Simplified Component Technique) and the VARIMAX criterion function $\mathcal{V}(a)$ from (4) is used as a sparseness inducing constraint. However, the resulting loadings contain plenty of small nonzero values, i.e. they are still not sparse. Then, Jolliffe et al. (2003) modify the original PCs to additionally satisfy the LASSO constraint, which drives many loadings to exact zeros. This method, called SCoTLASS (Simplified Component Technique subject to LASSo), became the first one to produce genuine sparse component loadings.

4.2 Sparse components: definitions and algorithms

In this Section several most influential sparse PCA definitions and methods for their solution are listed. The first major group of methods explore some kind of penalized reformulation of PCA. Let X and R denote the data and the sample correlation matrices.

- SCoTLASS: Jolliffe et al. (2003), Trendafilov and Jolliffe (2006) find the the ith vector of sparse loadings as solution of the following problem:

$$\max_{\substack{\|a\|_2=1 \text{and } \|a\|_1\le\tau \\ a\perp\{a_1,a_2,\ldots,a_{i-1}\}}} a^\top Ra,$$

  where $a_1, \ldots, a_{i-1}$ are the already found vectors of sparse component loadings. The requirement $\|a\|_1 \le \tau$ is known as the LASSO constraint.
- SPCA: Zou et al. (2006) transform the standard PCA into a regression form and solve the following problem, for an auxiliary vector $b \in \mathbb{R}^p$:

$$\min_{a,b} \|X - Xab^\top\|_F^2 + \lambda_1\|a\|_2^2 + \lambda_2\|a\|_1,$$
$$\text{subject to } \|b\|_2 = 1,$$

  The solution a gives the ith vector of sparse component loadings. Witten et al. (2009) rewritten SPCA in the following equivalent *bound* form:

$$\min_{a,b} \|X - Xab^\top\|_F^2 + \lambda_1\|a\|_2^2 + \lambda_2\|a\|_1,$$
$$\text{subject to } \|a\|_2^2 \le 1, \quad \|a\|_1 \le \tau, \quad \|b\|_2 = 1.$$

– d'Aspremont et al. (2007) apply semidefinite-programming (SDP) relaxation of the original sparse PCA formulation which leads to finding a symmetric and SD matrix A which

$$\begin{aligned} &\max \ \operatorname{trace}(RA) \\ &\text{subject to } \operatorname{trace}(A) = 1, \quad \|A\|_1 \le \kappa, \end{aligned}$$

where κ is a tuning parameter controlling the sparsity. The first sparse component loadings are given by the leading eigenvector a of A. Then, after the deflating $R := R - (a^\top Ra)aa^\top$, the same problem is solved again to find the next vector a of sparse loadings.
– Another approach to obtain sparse components is based on the spectral bounds of submatrices of the sample correlation matrix R. The idea is to identify the subset of m variables explaining the maximum variance among all possible subsets of size m and replace the loadings of the rest $p - m$ variables by 0s. Cadima and Jolliffe (2001) employed this approach for variable selection in Fisher's linear discriminant analysis (LDA). Moghaddam et al. (2006) further developed the idea and used it to construct a greedy algorithm for sparse PCA and LDA subject to cardinality constraints.
– d'Aspremont et al. (2008) proposed another SDP relaxation to construct a more efficient greedy algorithm than d'Aspremont et al. (2007), Moghaddam et al. (2006). They find a symmetric and SD matrix A, which

$$\begin{aligned} &\max_A \sum_{i=1}^{p} \max\{w_i^\top A w_i - \rho, 0\}, \\ &\text{subject to } \operatorname{trace}(A) = 1, \end{aligned}$$

where ρ is a tuning parameter controlling the sparsity and w_i is the ith column of W defined as the square root of $R = W^\top W$.
– sPCA–rSVD: Shen and Huang (2008) solve the following problem,

$$\min_{a,b,\|b\|_2=1} \|X - ba^\top\|_F^2 + \mathcal{P}_\lambda(a),$$

where $a \in \mathbb{R}^p$ and $b \in \mathbb{R}^n$, and $\mathcal{P}_\lambda(a)$ is a particular penalty term. The solution a gives the ith vector of sparse component loadings.
– Witten et al. (2009) proposed a general algorithm for sparse SVD:

$$\begin{aligned} &\min_{a,b} \ b^\top Xa, \\ &\text{subject to } \|a\|_2^2 \le 1, \quad \|b\|_2^2 \le 1, \quad \mathcal{P}_1(a) \le \tau_1, \quad \mathcal{P}_2(b) \le \tau_2. \end{aligned}$$

This formulation generalizes and simplifies the previous sparse PCA definitions given by SCoTLASS, SPCA and sPCA–rSVD, and can be applied for sparse canonical correlation analysis (CCA).

– Journée et al. (2010) treat sparse PCA with either LASSO or cardinality constraints in a very elegant unified way. Their algorithm is probably the most efficient algorithm for sparse PCA available now.
– Sriperumbudur et al. (2011) consider sparse version of the generalized EVD which can be then used for PCA, CCA and Fisher's LDA (when $n > p$). Instead of constraining the cardinality $\|a\|_0$ of the loadings $a \in \mathbb{R}^p$ or their ℓ_1 norm, as in the LASSO, the authors choose a small ε and work with:

$$\|a\|_\varepsilon = \sum_i^p \frac{\log(1 + |a_i|/\varepsilon)}{\log(1 + 1/\varepsilon)},$$

which is known to induce sparseness even more strongly than the ℓ_1 norm Candès et al. (2008). They propose very interesting and efficient solution by first rewriting the sparse PCA problem as a difference of two convex functions, and then solving it by the majorization-minimization approach.
– Qi et al. (2013) construct sparse components making use of a weighted sum of ℓ_1 and ℓ_2 norms as follows: $\|a\|_\lambda^2 = (1-\lambda)\|a\|_2^2 + \lambda\|a\|_1^2$. Then, similar to SCoTLASS, the ith vector of sparse component loadings is found as solution of the following optimization problem:

$$\max_{\substack{\|a\|_2=1 \\ a \perp \{a_1, a_2, \ldots, a_{i-1}\}}} \frac{a^\top R a}{\|a\|_{\lambda_i}^2},$$

where $a_1, \ldots, a_{i-1}$ are already found vectors of sparse component loadings and the (tuning) parameter λ_i is specific for the ith column. In contrast to SCoTLASS and most of the other methods, the constraint set is strictly convex for $0 \leq \lambda < 1$, which makes the optimization problem convex, and thus, having *unique* solution. The solution of this problem is found by a new type of thresholding. The sparse component loadings $A = \{a_1, \ldots, a_r\}$ are orthonormal, i.e. $A^\top A = I_r$, but the components are correlated, i.e. $A^\top RA$ is not diagonal. By considering the following modified problem:

$$\max_{\substack{\|a\|_2=1 \\ Ra \perp \{a_1, a_2, \ldots, a_{i-1}\}}} \frac{a^\top R a}{\|a\|_{\lambda_i}^2},$$

the authors obtain uncorrelated sparse components, i.e. $A^\top RA$ is diagonal, but the sparse component loadings are *not* orthonormal, i.e. $A^\top A \neq I_r$.
Note, that penalty as a mixture of ℓ_1 and ℓ_2 norms has been already considered for regression type of problems by Friedlander and Tseng (2007).

Sparse PCA can also be attacked through Bayesian approach. In general, the strategy is the same as in the methods listed above. The standard PCA formulation is replaced by its probabilistic version. The LASSO (or ℓ_1) regularization term of the loadings in the sparse PCA is replaced by independent Laplace (double-exponential) priors for

each of the loadings. For example, Guan and Dy (2009) find sparse PCA solution (Bayesian posterior mode estimates) by considering also another two types of priors: Inverse-Gaussian Prior and Jeffrey's Prior, which resemble ℓ_γ regularization terms with $0 < \gamma \leq 2$. A hierarchical Bayesian model is considered in (Ding et al. 2011) for decomposing a noisy matrix into low-rank and sparse components. The model is robust to a broad range of noise levels, possibly non-stationary. The Bayesian model is compared to optimization-based implementation of robust PCA and demonstrates competitive performance.

The main conceptual difficulty related to the methods employing LASSO or cardinality constraint or regularization term is the right choice of the LASSO threshold τ, or the number of zero loadings per component (κ, ρ, etc) or respectively, the weight of the regularization term λ that compromise between sparseness and explained variance. The first sparse PCA method, SCoTLASS, solves this problem by running the algorithm for a decreasing sequence of values of the tuning parameter $\tau \in [\sqrt{p} \searrow 1]$ (Jolliffe et al. 2003; Trendafilov and Jolliffe 2006). Each solution is used as a starting value for the next run. Thus, the result is a sequence of loadings matrices with increasing sparseness. The user is supposed to choose the most appropriate one according to specific requirements for explained variance and sparseness. The second sparse PCA method (Zou et al. 2006) considerers the explained variance for each vector of sparse loadings in turn as a function of the tuning parameter λ_2 (Fig.2) in order to decide on its value. In general, cross-validation is the most popular method for choosing such tuning parameters with different implementations employed. Other possible, but less explored, options are employing information criteria as AIC, BIC, etc (Guo et al. 2010; Qi et al. 2013; Zou et al. 2007) and bootstrap methods (Bach 2008).

Tuning the penalty parameters in the methods involving LASSO or cardinality constraint is time consuming for large data applications. That is why, there exists another large group of methods trying to avoid them.

- Rousson and Gasser (2004) proposed to find block sparse components by classifying first the original variables into disjoint groups, and then finding components from regressing the original variables on the preceding sparse components. Unfortunately, the method neither produces sparse enough loadings, nor they explain enough variance. Several procedures for sparse PCA related to clustering were proposed recently, e.g. see (Enki and Trendafilov 2012; Enki et al. 2013; Vichi and Saporta 2009) and the works cited there.
- Chipman and Gu (2005) introduced three types of "interpretable" components. Then, the best sparse component to be chosen is the "interpretable" one that has the least angle with the original PC.
- Johnstone and Lu (2009) proposed a kind of a thresholding method which also does not involve particular penalties. However, it requires the selection of a wavelet basis E, as well as the choice of a number of parameters. The method has the following steps (Johnstone and Lu 2009, p.686):
  1. let X_0 be a $n \times p$ data matrix and E be a basis in $\mathbb{R}^p$; find the coordinates of the data in this basis, i.e. form an $n \times p$ matrix $X = X_0 E$;
  2. calculate the sample variances of X and let $I \subset \{1, \ldots, p\}$ be the subset of indexes corresponding the largest r variances;

3. apply PCA to the reduced data X_I of size $n \times r$ and store the estimated eigenvectors in A of size $r \times r$;
4. apply hard thresholding elementwise to the estimated eigenvectors in A (for a scalar α and threshold δ the hard thresholding is defined as $\eta_H(\alpha, \delta) = \alpha \mathcal{I}(|\alpha| > \delta)$, where $\mathcal{I}$ is indicator function of the argument);
5. return to the original domain by $A^* = E_I \eta_H(A, \Delta)$, where E_I is the reduced basis of size $p \times r$.

In contrast to most works on sparse PCA, Johnstone and Lu (2009) assumed observations with Gaussian distribution following a single component (rank-one) model. Under these assumptions they show inconsistency of the principal components when $p \gg n$ and their sparse components remedy. Paul and Johnstone (2007) extended this analysis to a multiple components (finite rank) model. Moreover, they suggested using the whole covariance structure of the data (variances and covariances) which is implemented in the augmented sparse PCA. Cai et al. (2012) also employ a multiple components model (and normality assumptions) but their analysis is in terms of principal subspaces rather than PCs. They find the minimax risk for estimating the principal subspace and construct an explicit aggregated estimator, which *attains* the same rates of convergence as those of the minimax lower bounds. Unfortunately, it is impractical for large p and is replaced by an adaptive estimation reducing the sparse PCA problem to a regression one.

A weakness of most of the listed methods is that they produce sparse loadings that are not completely orthonormal and the corresponding components are correlated. Only SCoTLASS (Jolliffe et al. 2003; Trendafilov and Jolliffe 2006) and the method recently proposed by Qi et al. (2013) are capable to produce either orthonormal loadings or uncorrelated sparse components. Recently, Lu and Zhang (2012) consider a novel type of sparse PCA explicitly controlling the orthonormality of the loadings and the correlations among components. The presumption is that such sparse PCA preserves better the optimal features of PCA. Further research is needed to make clear the benefits from such solutions, e.g. with respect to consistency in high dimensions, etc.

The Section will not be complete without reminding how to find the total variance explained by the sparse components. The standard PCs are linear combinations of orthonormal vectors collected as columns of F_r, see (1). However, (2) does not hold for sparse A and $Y^\top Y$ is not diagonal, i.e. the sparse componets are correlated. Then $\text{trace}(A^\top X^\top X A)$ is no longer suitable measure for the total variance explained by them. To take into account the correlations among the sparse components, Zou et al. (2006) proposed the following. Assuming Y has rank r, its QR decomposition $Y = QU$ finds a basis of r orthonormal vectors collected in Q, such that they span the same subspace in $\mathbb{R}^p$ as Y and replace the original F_r. Let Y_i and Q_i denote the ith column of Y and Q. Then, the diagonal entries of U are given by $u_{ii} = \|Y_i - \sum_{j=1}^{i-1}(Q_j^\top Y_i) Q_j\|_F$, or $u_{ii}^2 = Y_i^\top Y_i - \sum_{j=1}^{i-1}(Q_j^\top Y_i)^2 = \|Y_i\|^2(1 - \sum_{j=1}^{i-1} \cos^2(\angle(Q_j, Y_i))$. Thus, u_{ii}^2 gives the variance of Y_i reduced by a term accounting for its correlations with the preceding sparse components. Then, the adjusted variance is defined as: $\text{AdjV}(Y) = \sum_i^r u_{ii}^2 = \text{trace}[\text{diag}(U)^2] \leq \text{trace}(U^\top U) = \text{trace}(Y^\top Y)$. The adjusted variance as defined in (Shen and Huang 2008, p.1021) seems misleading, because for orthonormal sparse loadings it coincides with the unadjusted.

4.3 Taxonomy of PCA subject to ℓ_1 (LASSO) constraint

Wright (2011) proposed the following taxonomy of problems seeking for sparse minimizers x of $f(x)$ through the ℓ_1 norm:

- Weighted form: $\min f(x) + \tau\|x\|_1$, for some $\tau > 0$;
- ℓ_1-constrained form (variable selection): $\min f(x)$ subject to $\|x\|_1 \leq \tau$;
- Function-constrained form: $\min \|x\|_1$ subject to $f(x) \leq \bar{f}$.

All three options were explored for regression type of problems, i.e. $f(x) = Ax$ (Candès and Tao 2007; Tibshirani 1996). This taxonomy can be restated accordingly for sparse PCA. For a given $p \times p$ correlation matrix R find vector of loadings a, $(\|a\|_2 = 1)$, by solving one of the following:

- Weighted form: $\max a^\top Ra + \tau\|a\|_1$, for some $\tau > 0$.
- ℓ_1-constrained form (variable selection): $\max a^\top Ra$ subject to $\|a\|_1 \leq \tau$, $\tau \in [1, \sqrt{p}]$.
- Function-constrained form: $\min \|a\|_1$ subject to $a^\top Ra \nearrow \lambda$, where λ is eigenvalue of R and $a^\top Ra \leq \lambda$ for any sparse a.

The first two forms were explored in a number of papers. For example, SCoTLASS is in the ℓ_1-constrained form, while SPCA (Zou et al. 2006) utilizes the weighted form to define the sparsification problem. It is interesting that the function-constrained form has never been used to attack the PCA sparsification. In the remaining part of the article sparse component loadings will be constructed by considering the function-constrained form of PCA. Several possible definitions of sparse PCA will be listed aiming to approximate different features of the original PCs and serve different applied scenarios. They will be illustrated on the Pitprop data. Also, it will be demonstrated that one of them can be used as a sparse alternative of the old analytic rotation methods.

5 Function-constrained sparse components

5.1 Orthonormal sparse component loadings

First we define problems involving orthonormal sparse component loadings A, i.e. $A^\top A = I_r$, which necessarily produce correlated sparse components.

5.1.1 Weakly correlated sparse components

The following version of sparse PCA:

$$\min_{A^\top A = I_r} \|A\|_1 + \mu\|A^\top RA - D^2\|_F^2, \tag{5}$$

is seeking for sparse loadings A which additionally diagonalize R, i.e. they are supposed to produce sparse components which are as weakly correlated as possible. D^2 is diagonal matrix of the original PCs' variances. We solve (5) for a range of values

of μ and consider the *index of sparseness*:

$$\text{IS} = \frac{V_a V_s}{V_o^2} \times \frac{\#_0}{pr}, \tag{6}$$

where V_a, V_s and V_o are the adjusted, unadjusted and ordinary total variances for the problem, and $\#_0$ is the number of zero loadings in A. IS in (6) increases with the goodness-of-fit (V_s/V_o), the higher adjusted variance (V_a/V_o) and the sparseness. The IS values help to choose μ with respect to sparseness and fit.

Jeffers's Pitprop data example (continued): We solve (5) with $\mu = 7$, a local IS maximum in [1, 14]. The sparse loadings A are given in the first six columns of Table 3. The diagonal elements of $A^\top RA$ approximating the original eigenvalues are: 4.0290, 1.8113, 1.6150, 0.9997, 0.7960 and 0.8069 (compare with their original values from Sect. 3.2). The sparse components are ordered according to the magnitude of their variances (except the smallest two). They explain 77.37 % of the total variance, which is quite close to their total adjusted variance 76.53 %. There are 48 zero loadings (out of 78), i.e. the loadings are moderately sparse. We note that most of the correlations among these sparse components are quite weak as seen from their correlation matrix, which departure from identity I_6 is 0.4764. They can be weaken further by increasing μ. In summary, these sparse components reasonably well approximate the desired features of the PCs: they have orthonormal loadings, are weakly correlated, and ordered according to the magnitudes of their variances.

5.1.2 Sparse components approximating the PCs variances

The next version of sparse PCA is:

$$\min_{A^\top A = I_r} \|A\|_1 + \mu \|\text{diag}(A^\top RA) - \text{diag}(D^2)\|_2^2 , \tag{7}$$

in which the variances of the sparse components should fit better the initial variances D^2, without paying attention to the off-diagonal elements of R. As a result $A^\top RA$ is expected to be less similar to a diagonal matrix than in (5) and the resulting sparse components—more correlated.

Jeffers's Pitprop data example (continued): We solve (7) with $\mu = 6$ (maximal IS). The sparse loadings A are given in the last six columns of Table 3. The diagonal elements of $A^\top RA$ approximating the original eigenvalues are: 3.7504, 1.8832, 1.5511, 0.9999, 0.9999 and 0.9990. The sparse components are ordered according to the magnitudes of their variances (except the last two), and explain 78.33 % of the total variance. This is a bit more than the variance explained by the components obtained from (5) in the previous example. However, the drop in their total adjusted variance 74.44 % is bigger than (5), because, as expected, the correlations among these sparse components are not as weak as those from (5). The deviation of the correlation matrix from I_6 is 0.9335. The loadings are very sparse, 63 zeros out of 78. With respect to the desired features of the PCs, these sparse components have orthonormal loadings

Table 3 Sparse component loadings for Jeffers's Pitprop data

Var	Sparse A by solving (5) with $\mu = 7$						Sparse A by solving (7) with $\mu = 6$					
	1	2	3	4	5	6	1	2	3	4	5	6
x_1	0.392	0.291			0.001		−0.468	−0.001				
x_2	0.419	0.311			0.024		−0.479					
x_3	0.123	0.001	0.792			−0.001		−0.707				
x_4			0.518					−0.707				
x_5		−0.479			0.774		0.059		0.677			
x_6	0.305	−0.469							0.726			
x_7	0.458	−0.200					−0.327		0.122			
x_8	0.361						−0.318					
x_9	0.300	0.001	−0.322		0.001		−0.406					
x_{10}	0.349	0.001				0.272	−0.419					
x_{11}				−1.00							1.00	
x_{12}	−0.099		0.001			0.962				−1.00		
x_{13}	−0.017	0.573			0.633							1.00
0s	3	4	9	12	9	11	6	11	10	12	12	12
V	31.0	13.9	12.4	7.7	6.1	6.2	28.9	14.5	11.9	7.7	7.7	7.7
C	31.0	44.9	57.4	65.0	71.2	77.4	28.9	43.3	55.3	63.0	70.7	78.3
A	31.0	44.9	57.2	64.7	70.7	76.5	28.9	42.9	53.7	60.5	68.0	74.4
#	Correlations among sparse components						Correlations among sparse components					
1	1.0	−0.05	0.05	0.02	−0.05	−0.02	1.0	0.18	−0.28	−0.21	0.02	0.11
2	−0.05	1.0	0.06	−0.10	−0.05	0.13	0.18	1.0	−0.19	0.20	−0.13	−0.07
3	0.05	0.06	1.0	−0.12	0.08	0.20	−0.28	−0.19	1.0	0.09	−0.08	−0.35
4	0.02	−0.10	−0.12	1.0	0.07	0.07	−0.21	0.20	0.09	1.0	−0.03	−0.18
5	−0.05	−0.05	0.08	0.07	1.0	−0.04	0.02	−0.13	−0.08	−0.03	1.0	0.01
6	−0.02	0.13	0.20	0.07	−0.04	1.0	0.11	−0.07	−0.35	−0.18	0.01	1.0

and are ordered according to the magnitudes of their variances. They are quite sparse but are nearly correlated as most of the available sparse solutions.

For comparison, the correlation matrices of the first six sparse components obtained in Zou et al. (2006) and Shen and Huang (2008) are reproduced respectively in the upper and the lower triangular parts of the following matrix:

$$\llcorner R_{SH} \backslash R_{ZHT} \urcorner = \begin{pmatrix} 1.0 & -0.17 & -0.33 & 0.00 & -0.20 & 0.08 \\ 0.20 & 1.0 & 0.13 & -0.14 & -0.22 & 0.08 \\ -0.46 & -0.11 & 1.0 & 0.10 & 0.14 & -0.40 \\ -0.33 & 0.27 & 0.26 & 1.0 & 0.03 & -0.01 \\ -0.20 & 0.13 & 0.16 & 0.20 & 1.0 & -0.18 \\ -0.04 & 0.05 & -0.10 & 0.07 & -0.05 & 1.0 \end{pmatrix}.$$

One can check that $\|R_{ZHT} - I_6\|_F = 0.9985$ and $\|R_{SH} - I_6\|_F = 1.1464$. i.e. their components are more correlated. They explain 75.8 and 80.3 % adjusted variance, with less sparse (60 and 53 zeros) and not orthonormal loadings.

5.1.3 Sparse components approximating the total variance

Instead of fitting the variances of individual sparse components to the initial ones as in (7), one can consider sparse components which total variance fits the total variance of the first r PCs. This is expressed as the following problem:

$$\min_{A^\top A = I_r} \|A\|_1 + \mu[\text{trace}(A^\top RA) - \text{trace}(D^2)]^2 \,. \tag{8}$$

The obvious drawback of this sparse PCA formulation is that the resulting sparse components will not be ordered according to the magnitudes of their variances. As the total variance is fitted only, one expects that the explained variance will be higher than with the previous formulations (5) and (7).

Jeffers's Pitprop data example (continued): We solve (8) with $\mu = 3$, for which IS has its highest value. The sparse loadings A are given in the first six columns of Table 4. The diagonal elements of $A^\top RA$ are: 3.4647, 1.0005, 2.0202, 1.8827, 1.0004 and 1.0006. The sparse components are not ordered according to the magnitude of their variances. They explain 79.77 % of the total variance, a pretty high value for sparse components with 62 zero loadings (out of 78). However, this value is far from their total adjusted variance (73.19 %), i.e. these sparse components are not quite valuable. In fact, this is a common feature for all solutions of (8). The correlations among them are of the same magnitudes as the ones obtained by (7). To summarize, these sparse components have orthonormal loadings, are reasonably correlated, but are not ordered according to the magnitudes of their variances. This makes them not very attractive candidate for a sparse dimension reduction. However, their features remind for the old rotation methods in PCA where the rotated components are also not ordered according to the magnitudes of their variances, but preserve exactly the total variance of the initial PCs. This relation will be further elaborated in Sect. 6.2

5.1.4 Sequential sparse components approximating the PCs variance

Finally, problem (7) can be rewritten in the following vectorial form:

$$\min_{\substack{a^\top a = 1 \\ a \perp A_{i-1}}} \|a\|_1 + \mu(a^\top Ra - d_i^2)^2 \,, \tag{9}$$

where $A_0 := 0$ and $A_{i-1} = [a_1, a_2, \ldots, a_{i-1}]$.

Jeffers's Pitprop data example (continued): We solve (9) with $\mu = 4$ (maximal IS). The sparse loadings A are orthonormal and are given in the last six columns of Table 4. The diagonal elements of $A^\top RA$ are: 3.8626, 1.8825, 1.0005, 0.9997, 0.9998 and 0.9996. The sparse components are ordered according to the magnitude of their variances, and explain 74.96 % of the total variance. The adjusted variance (72.51 %)

Table 4 Sparse component loadings for Jeffers's Pitprop data

Var	Sparse A by solving (8) with $\mu = 3$						Sparse A by solving (9) with $\mu = 4$					
	1	2	3	4	5	6	1	2	3	4	5	6
x_1	0.500						−0.482					
x_2	0.513						−0.499					
x_3				−0.707				0.708				
x_4				−0.707				0.710				
x_5	−0.111		−0.432								−1.00	
x_6			−0.681				−0.102					
x_7	0.081		−0.591				−0.370					
x_8	0.334						−0.244					
x_9	0.429						−0.381					
x_{10}	0.415				−0.001		−0.410					
x_{11}					1.00				1.00			
x_{12}						−1.00				1.00		
x_{13}		1.00										−1.00
0s	6	12	10	11	11	12	6	10	11	12	12	12
V	26.7	7.7	15.5	14.5	7.7	7.7	29.7	14.5	7.7	7.7	7.7	7.7
C	26.7	34.4	49.9	64.4	72.1	79.8	29.7	44.2	51.9	59.6	67.3	75.0
A	26.7	34.4	45.5	59.0	66.4	73.2	29.7	43.6	51.1	58.1	65.5	72.5
#	Correlations among sparse components						Correlations among sparse components					
1	1.0	−0.05	−0.37	−0.19	0.00	0.19	1.00	−0.20	0.03	0.20	0.01	−0.13
2	−0.05	1.0	0.40	−0.07	0.01	−0.18	−0.20	1.0	0.13	0.20	−0.04	−0.07
3	−0.37	0.40	1.0	0.17	0.09	−0.13	0.03	0.13	1.0	0.03	0.09	−0.01
4	−0.19	−0.07	0.17	1.0	−0.13	0.20	0.20	0.20	0.03	1.0	0.15	−0.18
5	0.00	0.01	0.09	−0.13	1.0	−0.03	0.01	−0.04	0.09	0.15	1.0	−0.21
6	0.19	−0.18	−0.13	0.20	−0.03	1.0	−0.13	−0.07	−0.01	−0.18	−0.21	1.0

is close to this value, which suggests for rather weakly correlated sparse components. There are 63 zero loadings out of 78. As expected, these results are comparable to the one obtained by solving (7).

5.2 Uncorrelated sparse components

5.2.1 Sequential sparse components approximating the PCs variances

The problem (9) can be modified for obtaining uncorrelated sparse components (however, loosing $A^\top A = I_r$) as follows:

$$\min_{\substack{a^\top a=1 \\ Ra \perp A_{i-1}}} \|a\|_1 + \mu(a^\top Ra - d_i^2)^2 \,, \tag{10}$$

where $A_0 := 0$ and $A_{i-1} = [a_1, a_2, \ldots, a_{i-1}]$.

Table 5 More sparse component loadings for Jeffers's Pitprop data

Var	Sparse A by solving (10) with $\mu = 5$						Sparse A by solving (11) with $\mu = 4$					
	1	2	3	4	5	6	1	2	3	4	5	6
x_1	−0.472	−0.234	−0.001				0.451					
x_2	−0.485	−0.246	−0.146		0.258		0.483	0.001				
x_3			0.542					−0.703			−0.001	0.091
x_4			0.808	0.154	−0.288	0.071		−0.691				
x_5		0.562			0.357	−0.687				1.00		
x_6	−0.132	0.375			0.001		0.016					
x_7	−0.382	0.186		−0.059	0.001		0.325				0.087	
x_8	−0.253						0.328	0.156			0.001	
x_9	−0.383						0.405	0.001				
x_{10}	−0.410		−0.178		0.001	−0.063	0.431	0.059			0.001	−0.230
x_{11}				−0.981	0.031	−0.153			−1.00			
x_{12}					0.850	0.220					−0.001	−0.969
x_{13}		−0.627		0.105		−0.669					0.996	
0s	6	7	8	9	5	7	6	8	12	12	10	10
V	30.1	13.8	13.1	7.4	5.4	5.7	29.1	14.6	7.7	7.7	7.1	6.2
C	30.1	43.9	57.00	64.4	69.8	75.5	29.1	43.7	51.4	59.1	66.2	72.4
#	$A^\top A$						$A^\top A \llcorner \backslash \urcorner A^\top RA$					
1	1.0	0.11	0.14	0.02	−0.13	0.03	1.0	−0.07	0.02	−0.01	−0.06	0.05
2	0.11	1.0	0.04	−0.08	0.14	0.03	0.08	1.0	0.14	−0.03	−0.11	0.14
3	0.14	0.04	1.0	0.12	−0.27	0.07	−0.00	−0.00	1.0	0.09	0.00	−0.07
4	0.02	−0.08	0.12	1.0	−0.08	0.09	−0.00	−0.00	−0.00	1.0	−0.19	0.13
5	−0.13	0.14	−0.27	−0.08	1.0	−0.08	0.03	0.00	0.00	0.00	1.0	−0.11
6	0.03	0.03	0.07	0.09	−0.08	1.0	−0.10	−0.08	−0.00	−0.00	0.00	1.0

Jeffers's Pitprop data example (continued): We solve (10) with $\mu = 5$ (maximal IS). The loadings A of the sparse components are given in the first six columns of Table 5. They are not orthonormal, but not much. $A^\top RA$ is diagonal matrix with elements: 3.9090, 1.7924, 1.7087, 0.9587, 0.7030 and 0.7421. The sparse components are ordered according to the magnitude of their variances, except the last two. They explain 75.5 % of the total variance, and because are uncorrelated, their adjusted variance is the same. The loadings are not very sparse, 42 zeros of 78, a common feature for all solutions of (10).

5.2.2 Weakly correlated sparse components with oblique loadings

It was mentioned that Lu and Zhang (2012) proposed a sparse PCA procedure which explicitly controls the orthonormality of the sparse loadings and the correlations among the components. Similar results can also be obtained with the function-constraint PCA

approach as follows. Along with problem (10), one can consider solving the following matrix optimization problem:

$$\min_{\text{diag}(A^\top A)=1_r} \|A\|_1 + \mu \|A^\top RA - D^2\|_F \,. \tag{11}$$

Clearly, the components obtained from (11) can be uncorrelated only approximately, as in (5). This sparse PCA formulation is very interesting, because (for some reason not completely understood yet) the resulting loadings A stay nearly orthonormal while diagonalizing $A^\top RA$. Such a feature may help to design algorithms that avoid the explicit preserving of orthonormality.

Jeffers's Pitprop data example (continued): We solve (11) with $\mu = 4$, for which IS has its highest value. The sparse loadings A are given in the last six columns of Table 5. They are not orthonormal, but not much (the entries below the main diagonal under the loadings). Indeed, $\|A^\top A - I_6\| = 0.2135$. The diagonal elements of $A^\top RA$ are: 3.7822, 1.8994, 1.0003, 0.9997, 0.9257 and 0.8001. The sparse components are ordered according to the magnitudes of their variances. They explain 72.36 % of the total variance. The adjusted variance (71.34 %) is very close to this value, because the sparse components are weakly correlated (the entries above the main diagonal under the loadings). Indeed, $\|\text{corr}(A^\top RA) - I_6\| = 0.5325$. The number of zero loadings is 58 (out of 78). These results can be compared to the results obtained by solving (5). Simple simulations show that solving (11) does not provide any obvious advantage over solutions produced by (5) in terms of sparseness or fit.

6 Applications

6.1 Simple structure rotation versus sparse components

There are very few situations where classic rotation methods are applied to large data, especially with modern format $p \gg n$. Some exceptions are available in climate data analysis, where classic rotation methods are used to interpret the empirical orthogonal functions (EOFs), corresponding to the component loadings in PCA. Hannachi et al. (2006) analyze a "modern" kind of data, a $174 \times 4{,}176$ matrix of winter sea level pressure. First, VARIMAX is applied to interpret the EOFs, followed by sparse PCA (SCoTLASS). It is demonstrated that much clearer interpretation is possible based on the sparse EOFs.

6.2 Application to simple structure rotation

The definition of sparse PCA with sparse components approximating the total variance (8) seems least appealing among all possible definitions: it may do best with respect to (unadjusted) variance, but does worst for adjusted variance. However, it can be used as a sparse alternative to the simple structure rotation methods, where the total variance of the already chosen r PCs collected in A is preserved. The solution of (8) will preserve this total variance approximately, but in reward will produce sparse loadings A.

This is illustrated by solving (8) for a well known and challenging data set which is always used when a new method for simple structure rotation is proposed. The optimal μ is located by comparing IS (6) for $\mu = 2, \ldots, 30$.

Thurstone's 26 Box Problem: L. L. Thurstone constructed an artificial data set with 30 observations and 26 variables in the following way. He collected at random 30 boxes and measured their three dimensions x_1 (length), x_2 (width) and x_3 (height). The variables of the data set are twenty-six functions of these dimensions listed in Table 6. Unfortunately, these 30 boxes and their dimensions are lost, but their correlation matrix is available from (Thurstone 1947, p.370). Its first three eigenvalues are considerably greater than the rest, and thus, three components are considered. The problem is to identify the three "latent" dimensions based on the component loadings. This problem is notorious as most of the rotation methods fail to reveal them, e.g. VARIMAX. The desired simple structure is depicted in the first three columns of Table 6.

The first solution is obtained in (Thurstone 1947, p.371) by graphical oblique rotation of initial centroid solution. For convenience, this solution is given in the middle three columns of Table 6. The matrix of oblique graphical rotation is found approximately by Procrustes fitting of the initial centroid solution to the graphically rotated loadings and is

$$Q = \begin{pmatrix} 0.5539 & 0.7328 & 0.7021 \\ -0.6947 & 0.5768 & 0.0093 \\ 0.4596 & 0.3619 & -0.7120 \end{pmatrix}.$$

The (approximate) correlations among the graphically rotated components are $Q^\top Q$ (Harman 1976, p.265), and are reproduced under the loadings.

Next, the sparse PCA problem (8) is solved with $\mu = 22$, for which IS is maximal among the solutions with 27 zeros in the sparse orthonormal A. The last three columns of Table 6 depict $A\sqrt{\mathrm{diag}(A^\top RA)}$, with "zero" entries ≤ 0.002. The desired simple structure is clearly and correctly identified.

7 Conclusion

During the last ten years, since the first paper on sparse PCA appeared, enormous number of works on the topic was published. Only 31 relevant papers are listed in this review. We mentioned, that one of the problems when working with the old rotation methods is that there are about 50 different of them to choose from. However, the situation with sparse PCA is getting worse. The existing methods to obtain sparse component loadings are so many and diverse, and their number constantly increases.

On the basis of a number of mainly empirical comparisons (e.g. Journée et al. 2010; Lu and Zhang 2012; Richtárik et al. 2012; Sriperumbudur et al. 2011) one can conclude that the generalized power method (GPower) of Journée et al. (2010) is probably the fastest and most versatile available method for sparse PCA. Their algorithm adopting ℓ_0 penalty seems more efficient than those with ℓ_1 penalty thanks to involving exact

Table 6 The desired simple structure and two loadings matrices for 26 Box data

Var	Simple str.			Thurstone's			Sparse ($\mu = 22$)		
	1	2	3	1	2	3	1	2	3
x_1	×			0.95	0.01	0.01	1.03		
x_2		×		0.02	0.92	0.01		1.01	
x_3			×	0.02	0.05	0.91			1.02
x_1x_2	×	×		0.59	0.64	−0.03	0.554	0.615	
x_1x_3	×		×	0.60	0.00	0.62	0.540		0.627
x_2x_3		×	×	−0.04	0.60	0.58		0.564	0.577
$x_1^2x_2$	×	×		0.81	0.38	0.01	0.852	0.281	
$x_1x_2^2$	×	×		0.35	0.79	0.01	0.225	0.872	
$x_1^2x_3$	×		×	0.79	−0.01	0.41	0.809		0.322
$x_1x_3^2$	×		×	0.40	−0.02	0.79	0.275		0.845
$x_2^2x_3$		×	×	−0.04	0.74	0.40		0.789	0.312
$x_2x_3^2$		×	×	−0.02	0.41	0.74		0.328	0.784
x_1/x_2	×	×		0.74	−0.77	0.06	0.887	−0.922	
x_2/x_1	×	×		−0.74	0.77	−0.06	−0.887	0.922	
x_1/x_3	×		×	0.74	0.02	−0.73	0.929		−0.914
x_3/x_1	×		×	−0.74	−0.02	−0.73	−0.929		0.914
x_2/x_3		×	×	−0.07	0.80	−0.76		0.981	−0.953
x_3/x_2		×	×	0.07	−0.80	0.76		−0.981	0.953
$2x_1 + 2x_2$	×	×		0.51	0.70	−0.03	0.429	0.710	
$2x_1 + 2x_3$	×		×	0.56	−0.04	0.69	0.481		0.678
$2x_2 + 2x_3$		×	×	−0.02	0.60	0.58		0.576	0.566
$\sqrt{x_1^2 + x_2^2}$	×	×		0.50	0.69	−0.03	0.421	0.704	
$\sqrt{x_1^2 + x_3^2}$	×		×	0.52	−0.01	0.68	0.446		0.681
$\sqrt{x_2^2 + x_3^2}$		×	×	−0.01	0.60	0.55		0.588	0.534
$x_1x_2x_3$	×	×	×	0.43	0.46	0.45	0.360	0.439	0.415
$\sqrt{x_1^2 + x_2^2 + x_3^2}$	×	×	×	0.31	0.51	0.46	0.209	0.511	0.446
# zeros	9	9	9	0	0	0	9	9	9
s.s.				6.68	8.82	8.23	7.44	9.03	8.68
Comps				$Q^\top Q$			$A^\top RA$		
1				1.0	0.17	0.06	1.0	0.34	0.40
2				0.17	1.0	0.26	0.34	1.0	0.38
3				0.06	0.26	1.0	0.40	0.38	1.0

rather than approximate update (Journée et al. 2010, p. 543). The GPower performance is followed closely by the method proposed by Zou et al. (2006).

Nevertheless, even if particular method for sparse PCA proves its superiority, further research is needed to reveal the statistical properties of sparse PCA and, more generally,

sparse data analysis. The situation reminds the classic PCA: fast and reliable methods for SVD/EVD do exist already for many years, but PCA, as a tool for data analysis, remains a central research topic.

Acknowledgments I thank the Editor, the Associate Editor, and the anonymous reviewers for their careful work and for the many helpful comments and suggestions.

References

Bach F (2008) BOLASSO: model consistent LASSO estimation through the bootstrap. In: ICML '08 proceedings of the 25th international conference on machine learning. ACM Press, New York, pp 33–40

Browne MW (2001) An overview of analytic rotation in exploratory factor analysis. Multivar Behav Res 36:111–150

Cadima J, Jolliffe IT (1995) Loadings and correlations in the interpretations of principal components. J Appl Stat 22:203–214

Cadima J, Jolliffe IT (2001) Variable selection and the interpretation of principal subspaces. J Agric Biol Environ Stat 6:62–79

Cai T, Ma Z, Wu Y (2012) Sparse PCA: optimal rates and adaptive estimation. http://arxiv.org/abs/1211.1309

Candès EJ, Tao T (2007) The Dantzig selector: statistical estimation when p is much larger than n. Ann Stat 35:2313–2351

Candès EJ, Wakin M, Boyd SP (2008) Enhancing sparsity by reweighted ℓ_1 minimization. J Fourier Anal Appl 14:877–905

Chipman HA, Gu H (2005) Interpretable dimension reduction. J Appl Stat 32:969–987

Chu MT, Trendafilov NT (1998) ORTHOMAX rotation problem. A differential equation approach. Behaviormetrika 25:13–23

d'Aspremont A, Ghaoui L, Jordan M, Lanckriet G (2007) A direct formulation for sparse PCA using semidefinite programming. SIAM Rev 49:434–448

d'Aspremont A, Bach F, Ghaoui L (2008) Optimal solutions for sparse principal component analysis. J Mach Learn Res 9:1269–1294

Diele F, Lopez L, Peluso R (1998) The Cayley transform in the numerical solution of unitary differential systems. Adv Comput Math 8:317–334

Ding X, He L, Carin L (2011) Bayesian robust principal component analysis. IEEE Trans Image Process 20:3419–3430

Donoho DL, Johnstone IM (1994) Ideal spatial adaptation via wavelet shrinkage. Biometrika 81:425–455

Eckart C, Young G (1936) The approximation of one matrix by another of lower rank. Psychometrika 1:211–218

Edelman A, Arias TA, Smith ST (1998) The geometry of algorithms with orthogonality constraints. SIAM J Matrix Anal Appl 20:303–353

Enki D, Trendafilov NT (2012) Sparse principal components by semi-partition clustering. Comput Stat 4:605–626

Enki D, Trendafilov NT, Jolliffe IT (2013) A clustering approach to interpretable principal components. J Appl Stat 3:583–599

Friedlander M, Tseng P (2007) Exact regularization of convex programs. SIAM J Optim 4:1326–1350

Guan Y, Dy J (2009) Sparse probabilistic principal component analysis. Proc Twelfth Int Conf Artif Intell Stat 5:185–192

Guo F, Gareth J, Levina E, Michailidis G, Zhu J (2010) Principal component analysis with sparse fused loadings. J Comput Graph Stat 19:947–962

Hannachi A, Jolliffe IT, Stephenson DB, Trendafilov NT (2006) In search of simple structures in climate: simplifying EOFs. Int J Climatol 26:7–28

Harman HH (1976) Modern factor analysis, 3rd edn. University of Chicago Press, Chicago

Hausman RE (1982) Constrained multivariate analysis. In: Zanakis SH, Rustagi JS (eds) Optimization in statistics. North-Holland, Amsterdam, pp 137–151

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24(417–441):498–520

Jeffers JNR (1967) Two case studies in the application of principal component analysis. Appl Stat 16:225–236
Jennrich RI (2007) Rotation methods, algorithms, and standard errors. In: Cudeck R, MacCallum RC (eds) Factor analysis at 100. Lawrens Erlbaum Associates, Mahwah, NJ, pp 315–335
Johnstone IM, Lu AY (2009) On consistency and sparsity for principal components analysis in high dimensions. J Am Stat Assoc 104:682–693
Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York
Jolliffe IT, Uddin M (2000) The simplified component technique: An alternative to rotated principal components. J Comput Graph Stat 9:689–710
Jolliffe IT, Trendafilov NT, Uddin M (2003) A modified principal component technique based on the LASSO. J Comput Graph Stat 12:531–547
Journée M, Nesterov Y, Richtárik P, Sepulchre R (2010) Generalized power method for sparse principal component analysis. J Mach Learn Res 11:517–553
Lu Z, Zhang Y (2012) An augmented Lagrangian approach for sparse principal component analysis. Math Program Ser A 135:149–193
Marshall A, Olkin I (1979) Inequalities: theory of majorization and its applications. Academic Press, London
MATLAB (2011) MATLAB R2011a. The MathWorks, Inc., New York
Moghaddam B, Weiss Y, Avidan S (2006) Spectral bounds for sparse PCA: exact and greedy algorithms. Adv Neural Inf Process Syst 18:915–922
Mulaik SA (2010) The foundations of factor analysis, 2nd edn. Chapman and Hall/CRC, Boca Raton, FL
Paul D, Johnstone IM (2007) Augmented sparse principal component analysis for high dimensional data. http://arxiv.org/abs/1202.1242
Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philos Mag 2:559–572
Qi X, Luo R, Zhao H (2013) Sparse principal component analysis by choice of norm. J Multivar Anal 114:127–160
Richtárik P, Takáč M, Ahipaşaoğlu SD (2012) Alternating maximization: unifying framework for 8 sparse PCA formulations and efficient parallel codes. http://www.maths.ed.ac.uk/~richtarik/24AM.pdf
Rousson V, Gasser T (2004) Simple component analysis. Appl Stat 53:539–555
Shen H, Huang JZ (2008) Sparse principal component analysis via regularized low-rank matrix approximation. J Multivar Anal 99:1015–1034
Sriperumbudur BK, Torres DA, Lanckriet GRG (2011) A majorization-minimization approach to the sparse generalized eigenvalue problem. Mach Learn 85:3–39
Thurstone LL (1935) The vectors of mind. University of Chicago Press, Chicago, IL
Thurstone LL (1947) Multiple factor analysis. University of Chicago Press, Chicago, IL
Tibshirani R (1996) Regression shrinkage and selection via the LASSO. J R Stat Soc 58:267–288
Trendafilov NT (1999) A continuous-time approach to the oblique Procrustes problem. Behaviormetrika 26:167–181
Trendafilov NT, Jolliffe IT (2006) Projected gradient approach to the numerical solution of the SCoTLASS. Comput Stat Data Anal 50:242–253
Trendafilov NT, Lippert RA (2002) The multimode Procrustes problem. Linear Algebra Appl 349(1–3):245–264
Vichi M, Saporta G (2009) Clustering and disjoint principal component analysis. Comput Stat Data Anal 53:3194–3208
Vines SK (2000) Simple principal components. Appl Stat 49:441–451
Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation. Biostatistics 10:515–534
Wright S (2011) Gradient algorithms for regularized optimization. SPARS11, Edinburgh, Scotland. http://pages.cs.wisc.edu/~swright
Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. J Comput Graph Stat 15:265–286
Zou H, Hastie T, Tibshirani R (2007) On the "degrees of freedom" of the LASSO. Ann Stat 35:2173–2192