Information Maximization in Sparse Principal Components Analysis: An Exact Approach for Sparse and Interpretable Dimension Reduction

Alessio Farcomeni*

Università di Roma "La Sapienza"

Abstract

We show a branch and bound approach to exactly find the best sparse dimension reduction of a matrix. We call our approach IMS-PCA: Information Maximization in Sparse Principal Components Analysis. We can choose between enforcing orthogonality of the coefficients and uncorrelation of the components, and can explicitly set the degree of sparsity as the number of variables used by each derived component. We suggest methods to choose the number of non-zero loadings for each component; and illustrate and compare IMS-PCA with existing methods through a benchmark data set.

Keywords: branch and bound, dimension reduction, feature selection, feature extraction, interleaving eigenvalues theorem, sparse principal components

1 Introduction

Principal Component Analysis (PCA) is a popular dimension reduction and descriptive multivariate technique (Jolliffe, 2002; Chatfield and Collins, 1980).

Given an n by m matrix X, a new n by p matrix Y is built. The columns of Y are functions of the columns of the original data matrix, and $p << m$. It is well known that the highest possible information/variability is retained in Y if its columns (the principal components) are an affine linear combination of the original columns, with weights given by unit length eigenvectors of $X'X$ (the loadings). This is the well-known standard PCA. Furthermore (i) the columns of the derived matrix Y are uncorrelated and (ii) the loadings are orthogonal; so that the information is well separated.

There is an impressive number of applications of PCA in biology, medicine, psychology, financial econometrics, engineering, etc.

*The author is grateful to Nicola Apollonio for interesting discussions about the mathematical background, to Luca Tardella for suggestions that lead to improvement in the presentation, and to Paolo Giordani for careful reading of a first draft.

In such applications usually X is a two-mode data matrix, in which the n rows represent subjects and the m columns represent numeric variables.

The main drawback of dimensionality reduction through PCA is that each principal component (PC) is in general a linear combination of *all* the m variables used in input. This complicates the interpretation of the information contained in the derived PCs, and do not help the user in discarding less important variables. It is in general believed that sparseness of the loadings would be of great relevance in aiding in the interpretation of the derived variables. If principal components were linear combinations of only a small number of original variables, with different variables being used by different components, the subjective interpretation step would be much easier. For this reason, principal components extraction is often followed by some kind of transformation which aims at making the interpretation easier. A common approach then is to discard the smallest coefficients (hard thresholding) of the ordinary or rotated principal components. Cadima and Jolliffe (1995) note that such "simple thresholding" of the loadings, even after a rotation (Jolliffe, 1995), can be misleading; and in general does not produce an optimal solution. Jolliffe (1982) notes that simple thresholding can produce substantial problems when using the new variables to do multivariate regression. It is also not uncommon that the principal components are not easily interpreted even after rotation and thresholding (as it is known in the literature about the example of Section 5). Rotation and thresholding do not guarantee interpretability; while a general idea is that if enough loadings are zero, the derived axes will be interpretable. Note that the interpretability would not be the only advantage: sparse matrices are preferrable for lossy information compression since they are better stored and handled, variables which receive a zero loading in all p components will be discarded, thus performing an automatic feature selection, etc.

Jolliffe and Uddin (2000) advocate techniques that combine extraction and interpretation, so as to secure interpretability. A possible solution is then to extract components in which a certain number of loadings are directly set to zero.

To tackle this problem, Jolliffe *et al.* (2003) introduce SCoTLASS to get modified principal components with possible zero loadings, with orthogonal loadings of different principal components. SCoTLASS solves a non-convex constrained optimization problem, and usually it is of high computational cost. An efficient algorithm is proposed in Trendafilov and Jolliffe (2006). Zou *et al.* (2004) put the PCA problem under a regression-type optimization framework, and use the elastic net of Zou and Hastie (2003) to find sparse approximations of the PCs. In their approach the computational cost is much lower and there is the possibility to retain the same amount of information with more sparsity, main drawback is that orthogonality of the loadings is not guaranteed. In both approaches the degree of sparsity is controlled via a penalization parameter, and the choice of such parameter is an open problem.

Further and more importantly, neither approach can produce uncorrelated sparse components, and another open problem is that the user does not know in advance if and how sparse the loadings will be.

Our main goal in this paper is to give non-zero weight to a pre-specified number of variables, with as little information loss as possible with respect to ordinary PCA. We view the sparse principal component analysis (sPCA) problem as a variability/information maximization problem, and show how to derive an exact solution for prescribed degree of sparsity.

We will also be able to choose whether to enforce orthogonality of the loadings or uncorrelation of the new variables. We will see that our approach allows to find good solutions in terms of information compression, and interpretability. Information loss with respect to PCA, in practice, will often be negligible; but with the advantage of sparsity of the loadings and one property between loading orthogonality and uncorrelation of the new variables. If orthogonality of the loadings is enforced, solutions in which different variables are used by different components are favored. If uncorrelation of the new variables is enforced, the information is perfectly separated and multicollinearity problems are solved.

The main drawback, partly shared with the other methods, is that such exact approach will be suitable only for small dimensional situations (say up to 100 variables). We are currently working on a solution for high-dimensional problems.

The remainder of the paper is organized as follows: in Section 2 we formalize the sPCA problem for the first sparse principal component (sPC). In Section 3 we show how to derive the other sparse components. Section 4 will propose heuristic and formal strategies to choose the degree of sparsity of each sPC, and Section 5 will illustrate the method in a real data application. Finally, we give a short discussion in Section 6.

2 The sPCA problem

Let $\Sigma = \{\sigma_{ij}; i = 1, \ldots, m, j = 1, \ldots, m\}$ be the covariance matrix computed from data X, with n observations and m variables. The sparse principal component problem is:

$$\begin{cases} \max_x x'\Sigma x \\ x'x = 1 \\ \mathrm{sign}(x)'\mathrm{sign}(x) \leq k_1, \end{cases} \tag{1}$$

where sign(x) is the sign function, which is zero when x is zero. The solution β_1 to such problem is well known to be the first eigenvector of Σ whenever $k_1 \geq m$, and the achieved maximum is the corrisponding eigenvector (the amount of information retained by the first principal component). If $k_1 < m$ there is a genuine sparsity constraint on the loadings vector x: at least $m - k_1$ variables will receive zero weight. It can be shown that the maximum is not decreasing in k_1, but for appropriately chosen k_1 very little information loss can happen while discarding a few variables. The main difference with the other approaches to sPCA is that we express the problem in terms of maximization of information retained in the "new" variable $X\beta_1$, rather than try and approximate the ordinary principal components. For this reason we call our approach Information

Maximization for Sparse Principal Components Analysis (IMS-PCA).

2.1 The First Sparse Principal Component

We propose now a branch and bound algorithm to exactly solve the problem in (1). Branch and bound algorithms are a clever way to enumerate the possible solutions to a given (difficult) problem. The set of possible solutions is split into subsets by *branching*, and a criterion is used to *bound* the solutions into each subset. Finally, only subsets for which the bound is bigger than the current maximum are explored. A nice review of branch and bound in statistical analysis is given in Hand (1981).

We will use the interleaving eigenvalues theorem (Wilkinson, 1965) for bounding:

Theorem 1. *Let A be an m by m real symmetric matrix and denote by $A^{(-j)}$ the matrix obtained by removing the j-th row and j-th column of A. Let ρ_i be the i-th eigenvalue of A and μ_i the i-th eigenvalue of $A^{(-j)}$. Then,*

$$\rho_1 \geq \mu_1 \geq \rho_2 \geq \mu_2 \geq \cdots \geq \mu_{m-1} \geq \rho_m.$$

Note that removing the j-th row and column of a covariance matrix is equivalent to removing the j-th variable from the data matrix. Theorem 1 can be used to claim that when in a set of $k \geq k_1$ variables the largest eigenvalue of the covariance matrix is lower than the current maximum, all the subsets of size k_1 have a lower largest eigenvalue and can be discarded. Note that the theorem also implies that the third constraint in problem (1) is changed into an equality.

The branching step can be the same as the algorithms for best subset selection (Miller, 1990): split the current set into subsets in which variables are removed one at a time.

Along the lines of Narendra and Fukunaga (1977) and Ridout (1988), we also suggest to accelerate the algorithm by sorting the variables with respect to $\sigma_{ii} + \sum_i |\sigma_{ij}|$, $i = 1, \ldots, m$, in light of the following theorem:

Theorem 2. *Let A be any real square symmetric matrix. Let λ_i be the i-th eigenvalue of A. Then,* $\min_i \sum_j |A_{ij}| \leq \lambda_i \leq \max_i \sum_j |A_{ij}|$.

Proof. Let $x \in \mathcal{R}^n - \{\underline{0}\}$ be the eigenvector associated to the i-th eigenvalue of A. Let j be the index of the element of the vector farthest from 0. By definition, $Ax = \lambda_i x$. We have:

$$\begin{aligned} \lambda_i |x_j| &= |\sum_k A_{jk} x_k| \\ &\leq \sum_k |A_{jk}||x_k| \\ &\leq |x_j| \sum_k |A_{jk}| \\ &\leq |x_j| \max_j \sum_k |A_{jk}|. \end{aligned}$$

The other inequality follows from the same reasoning. □

Subsets of variables with small $\sigma_{ii}+\sum_i |\sigma_{ij}|$ do not give strong contributes to the overall information, and they will likely be soon discarded by the algorithm.

In what follows, we denote by $\lambda_{\max}(\cdot)$ the operator that computes the largest eigenvalue of a square symmetric matrix, and by Σ_S the covariance matrix obtained from a subset S, with cardinality $\text{card}(S)$, of the variables.

The proposed branch and bound algorithm is briefly described below: suppose we are examining a subset of variables S, and $\lambda_{\max}(\Sigma_S)$ is bigger than the current maximum λ_0. At the beginning, we suggest to set λ_0 as the first eigenvector of the first k_1 variables after ordering, since they are the best candidates for final optimality.

(i) Split the set S into $\text{card}(S)$ subsets $S_1, \ldots, S_{\text{card}(S)}$, each of which is composed by the elements of S to which one component is removed.

(ii) Set $i := 1$

(iii) Let $\lambda_i = \lambda_{\max}(\Sigma_{S_i})$.

(iv) If $\lambda_i \leq \lambda_0$, go to (v). Otherwise go to (vii).

(v) We can reject S_i. If $i = \text{card}(S)$ go to (vi), otherwise set $i := i + 1$ and go to (iii).

(vi) All subsets of S have been evaluated.

(vii) We cannot reject S_i. If $\text{card}(S_i) = k_1$, go to (viii), otherwise go to (ix).

(viii) We have found a better solution than the current maximum. Set $\lambda_0 = \lambda_{\max}(S_i)$. If $i = \text{card}(S)$ go to (vi) and otherwise set $i := i + 1$ and go to (iii).

(ix) It is possible that a subset of elements of S_i contain a better solution than the current maximum. Set $S = S_i$ and go to (i).

It shall be noted that a single run of the algorithm can produce all possible solutions from a chosen k_1 to m; and that the algorithm *exactly* solves the optimization problem.

3 The Other Principal Components

It is straightforward to check that only the loadings resulting from ordinary PCA can be orthogonal and simultaneously yield uncorrelated components. We will have to choose then between enforcing orthogonality of the loadings and uncorrelation of the components, by adding a specific constraint into problem in (1). Let β_i be the loadings of the i-th sPC. It is straightforward to check that uncorrelation between the i-th and j-th sPC will be given by the constraint

$\beta_i'\Sigma\beta_j = 0$, while orthogonality of the loadings is obviously given by $\beta_i'\beta_j = 0$. There is no sparse solution that satisfies both constraints.

Call now $J(\Sigma, x)$ the objective function, where maximization is in x. Note that now the maximum will not necessarily be an eigenvalue of a submatrix. An important choice for the objective function is between maximization of the variance explained by each component, and the *adjusted variance*, an index of variability introduced by Zou *et al.* (2004) to cope with correlated components.

If we enforce orthogonality, uncorrelation of the components is sacrificed and the information added by the i-th component, with $i > 1$, is in general lower than its variance.

When the extracted features are used separately, for instance for descriptive purposes, it may be desirable to maximize the variance of each component ($J(\Sigma, x) = x'\Sigma x$). In the other cases, Zou *et al.* (2004) devised how to measure the additional variability under correlation, and showed that the adjusted variance of the j-th component is given by the square of the j-th diagonal element R_{jj} of the upper triangular matrix in the QR decomposition of the new matrix Y ($J(\Sigma, x) = R_{jj}^2$). Maximization of the adjusted variance may be more sensible in cases in which the extracted features are used jointly. If uncorrelation is enforced, adjusted variance reduces to the variance of the new variable. See Zou *et al.* (2004) for other computational strategies and further comments.

The sPCA problem for the second sPC is then:

$$\begin{cases} \max_x J(\Sigma, x) \\ x'x = 1 \\ \text{sign}(x)'\text{sign}(x) \leq k_2 \\ C(x, \beta_1) = 0, \end{cases} \tag{2}$$

where $J(\Sigma, x)$ is either the variance or the adjusted variance of the new PC and $C(x, \beta_1)$ is either $x'\beta_1$ or $x'\Sigma\beta_1$. The solution to problem (2) will be an sPC with k_2 non-zero elements.

As before, a branch and bound algorithm can be used to solve the problem, just by substituting the $\lambda_{\max}(\cdot)$ operator with the operator that computes the solution to problem:

$$\begin{cases} J(\Sigma_{S_i}, x) \\ x'x = 1 \\ C(x, \beta_{1,S_i}) = 0, \end{cases} \tag{3}$$

where β_{1,S_i} are the loadings of the variables in subset S_i.

The bounding is the same in light of the following straightforward generalization of the interleaving eigenvalues theorem:

Theorem 3. *Let λ be the maximum for problem (3) for a given S_i and β_1. Let $\Sigma_{S_i^{(-j)}}$ be the matrix S_i in which the j-th row and column are removed, and $\beta_1^{(-j)}$ the vector β_1 in which the j-th element is removed. Let μ be the maximum for problem (3) in which Σ_{S_i} is substituted with $\Sigma_{S_i^{(-j)}}$ and β_1 with $\beta_1^{(-j)}$. Then, $\lambda \geq \mu$.*

Proof. Without loss of generality suppose we remove the last row and column from Σ_{S_i} and correspondingly the last element from β_1. Call y the solution of the reduced problem, with maximum μ. Let $x = \begin{pmatrix} y \\ 0 \end{pmatrix}$. It suffices to show that x satisfies the constraints of the enlarged problem, which is straightforward. □

Maximization in (3) can be easily and quickly solved by quadratically constrained convex optimization (Gill *et al.*, 1981). Fast routines are available even for large scale problems (Coleman and Li, 1994).

If $C(x, \beta_1) = x'\beta_1$, note that whenever the set S_i at a given node is made of variables receiving a zero loading in the first sPC, the solution is the first eigenvector of the reduced matrix. A possible acceleration of the algorithm is given by first solving problem (1) for the variables not used by the first PC, and then by using the optimum as a starting solution for problem (2). If the variances are more or less homogeneous, it is reasonable to expect that the second sPC will be a linear combination of variables which receive a zero loading in the first sPC and a very early harvesting of the tree will happen.

After finding the second sPC, it is straightforward to add a further constraint to find a *third* sPC in which only k_3 variables are used, and so on.

A sufficient condition for the existence of a solution to problem (2) is $k_i \geq i$, $i = 2, \ldots, m$. The solution may exist anyway also for $k_i < i$, for instance if $J(x, \beta_i) = x'\beta_i$ and at least k_i variables have not been used by the previous components. Nevertheless, so much sparsity in the first sPCs is rarely needed.

3.1 Different Criteria

It is straightforward to modify the algorithm in order to solve other kind of problems. So far each component was allowed to use any of the variables. We can very easily further constraint the p sparse principal components to use the *same* k variables. This would provide p linear combinations in which just k variables are used (with $p \leq k \leq m$), performing simultaneous variable selection and dimension reduction. It is straightforward to see that it suffices to use a single branch and bound algorithm, in which the objective function is given by the sum of the first p eigenvalues of the covariance matrix. In this case we will have finally applied ordinary PCA on a subset of the variables. Uncorrelation of sparse components and orthogonality of the loadings follow by construction.

4 Choice of the Degree of Sparsity

The choice of the number of principal components is a problem shared with classical PCA, and it may be solved with classical methods, some of which we briefly summarize:

- Cattell (1966) proposes the scree-test, in which the proportion of explained variance is plotted against the number of components, and an “elbow” is looked for in the graph.

- Kaiser (1960) and Horn (1965) propose to retain only components who explain more than the average variance of the original variables.

- Bartlett (1950, 1951) propose a chi-squared based test to sequentially verify, in the Cattell (1966) spirit, that the last components all explain the same amount of variability.

- Velicer (1976) proposes the minimum average partial rule, which exploits a matrix of partial correlations.

A comparison and complete overview of the methods is done in Zwick and Velicer (1986).

On the other hand, a different problem is posed by the choice of the number of variables to use in each sPC (k_i). In this section we will propose heuristic and formal methods to choose the degree of sparsity of each sPC; which we will explore in the example below.

Unlike other methods, we can explicitly choose the number of variables to be used in computing the sparse principal component (sPC), namely, k_i.

A possibility is to exploit the same idea of Cattell (1966): a *scree plot* can be made for the single component, in which for each value of k_i it is reported the objective function for the sPC. An example is given in Figure 1. The variance of the sPC increases with the number of non-zero loadings k_i, but from some k_i on the growth may flatten markedly. This elbow phenomenon can be used to choose k_i as the maximum number of non-zero loadings for which adding one variable does not give a significant contribution.

Zou *et al.* (2004) suggest to choose their penalty parameter as the one giving best approximation to the ordinary PC. Along those lines, the Euclidean distance between each sPC and the PC can be plot in function of k_i, and the scree-test applied.

Following the idea of information maximization proposed in this paper, we can also suggest other (formal) criteria.

We suggest in fact to choose k as the maximizer in q of $J(\Sigma, \beta_{(q),i}) - \rho(i) f(q)$, where $\beta_{(q),i}$ are the loadings of the i-th sPC (given the previous) with degree of sparsity q, $\rho(i)$ is a penalty parameter, and $f(\cdot)$ is a strictly monotone function (for instance, the identity or the logarithmic function). The expression at the first term is the proportion of variability contained in the component, which is penalized by a function of the proportion of variables used by it. This approach will favor sparser and more interpretable results for reasonable choices of the penalty parameter $\rho(i)$, which may be proportional to the average variance $\bar{\sigma}_{ii} = 1/m \sum_i \sigma_{ii}$. If $\rho(i)$ is taken to be constant with respect to i, the first principal components will be less sparse than the others. For this reason, another possibility is given by decreasing the penalty parameter as i increases, for instance by a constant quantity, like $\rho(i) = \bar{\sigma}_{ii}/(i+1)$.

In this paper we will always use the criterion:

$$\max_q \; J(\Sigma, \beta_{(q),i}) - \frac{\log(q)\bar{\sigma}_{ii}}{i+1} \tag{4}$$

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	0.546	0.047	-0.087	0.066	-0.046	0.000
length	0.568	0.000	-0.076	0.117	-0.081	0.000
moist	0.000	0.641	-0.187	-0.127	0.009	0.017
testsg	0.000	0.641	0.000	-0.139	0.000	0.000
ovensg	0.000	0.000	0.457	0.000	-0.614	-0.562
ringtop	0.000	0.356	0.348	0.000	0.000	-0.045
ringbut	0.279	0.000	0.325	0.000	0.000	0.000
bowmax	0.132	-0.007	0.000	-0.589	0.000	0.000
bowdist	0.376	0.000	0.000	0.000	0.000	0.065
whorls	0.376	-0.065	0.000	-0.067	0.189	-0.065
clear	0.000	0.000	0.000	0.000	-0.659	0.725
knots	0.000	0.206	0.000	0.771	0.040	0.003
diaknot	0.000	0.000	-0.718	0.013	-0.379	-0.384
Number of nonzero loadings	6	7	7	8	8	8
Variance (%)	27.2	16.4	14.8	9.4	7.1	7.9
Adjusted Variance (%)	27.2	15.3	14.4	7.1	6.7	7.5
Cumulative Adjusted Variance (%)	27.2	42.5	56.9	64.0	70.7	78.2
Variance of PCA solution (%)	32.4	18.3	14.4	8.5	7.0	6.3
Cumul. Variance of PCA solution (%)	32.4	50.7	65.2	73.7	80.7	87.0

Table 1: sPCA of Pitprops data, SCoTLASS

Note that the choice of the degree of sparsity is to be made sequentially, following the natural ordering of the sparse components: for a different choice of k_1 the second sPC will be different, possibly leading to a different "optimal" k_2.

5 Application to Pitprops data

The Pitprops data was first used by Jeffers (1967) as an example of the difficulty of interpreting principal components. The data set has 180 observations and 13 standardized variables, so that Σ reduces to the correlation matrix. Jeffers (1967) performed a PCA and suggested using the first 6 principal components.

sPCA can be applied to this data to enhance interpretability. SCoTLASS produced results in Table 1, while Zou *et al.* (2004) method produced results in Table 2. The variance and cumulative variance of the ordinary PCA solution is reported only in Table 1 for reasons of space.

In Table 3 we show IMS-PCA for the same degree of sparsity of the first two methods, with orthogonal coefficients and maximization of the variance of each component. When we set $k_5 = k_6 = 1$, there is no feasible solution so we turn to the best approximation. In all cases it can be seen that the information loss in using a sparse solution is minimal, while at most 8 variables out of 13 are used by sparse principal components.

In Table 4 we maximize the adjusted variance. It can be seen that in all

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.477	0.000	0.000	0	0	0
length	-0.476	0.000	0.000	0	0	0
moist	0.000	0.785	0.000	0	0	0
testsg	0.000	0.620	0.000	0	0	0
ovensg	0.177	0.000	0.640	0	0	0
ringtop	0.000	0.000	0.589	0	0	0
ringbut	-0.250	0.000	0.492	0	0	0
bowmax	-0.344	-0.021	0.000	0	0	0
bowdist	-0.416	0.000	0.000	0	0	0
whorls	-0.400	0.000	0.000	0	0	0
clear	0.000	0.000	0.000	-1	0	0
knots	0.000	0.013	0.000	0	-1	0
diaknot	0.000	0.000	-0.015	0	0	1
Number of nonzero loadings	7	4	4	1	1	1
Variance (%)	28.0	14.4	15.0	7.7	7.7	7.7
Adjusted Variance (%)	28.0	14.0	13.3	7.4	6.8	6.2
Cumulative Adjusted Variance (%)	28.0	42.0	55.3	62.7	69.5	75.8

Table 2: sPCA of Pitprops data, Zou *et al.* (2004) method

but one case (the sixth sPC with $k_6 = 8$) we manage to achieve higher objective functions than SCoTLASS. When compared with Zou *et al.* (2004), it can be seen that IMS-PCA achieves higher objective functions in the first three sPC, even in presence of the further orthogonality constraint. It can then be said that in this example our approach to sPCA leads to more compression of the information in the first axes, and that for the first 5 axes the total variance/adjusted variance is always higher than the other methods, for the same degree of sparsity. It is worth also noticing that in all cases both the variance and adjusted variance follow the right not-decreasing order. This happens only for the adjusted variance in the other methods.

When coming to the interpretation of the axes, it can be seen that there is the same or less overlap among components (especially when maximizing the adjusted variance). For instance, the second sPC with $k_2 = 4$ and the fourth with $k_4 = 1$ in Table 3 put zero weight to all of the previously used variables.

Even more encouraging results can be seen in Table 5, where we change the constraint and ask for uncorrelated components. Even if the new variables are now uncorrelated, the objective functions are very close to the results of SCoTLASS and Zou *et al.* (2004) and even higher in the first components. Once again we turn to the best approximation in quadratic loss when $k_4 = k_5 = k_6 = 1$. Loadings are seldom orthogonal, but always very close to orthogonality.

To give an idea of the efficiency of the branch and bound algorithm we note that at most only 27% of the possible groupings of k variables were actually explored by the algorithm, in these examples. Even less are expected when the

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.444	0.226	0.205	0.000	-0.093	0.000
length	-0.453	0.000	-0.146	0.288	-0.235	0.257
moist	0.000	0.604	0.000	0.167	0.222	0.309
testsg	0.000	0.623	0.000	0.000	0.276	-0.050
ovensg	0.000	0.000	-0.597	0.000	0.000	0.000
ringtop	0.000	0.290	-0.182	-0.440	0.000	-0.047
ringbut	-0.379	0.000	-0.088	-0.336	0.000	-0.798
bowmax	-0.341	-0.154	0.000	0.000	0.294	0.000
bowdist	-0.403	0.000	0.000	0.234	0.000	0.000
whorls	-0.418	-0.114	0.000	-0.234	0.113	0.043
clear	0.000	0.000	0.000	0.681	0.084	-0.103
knots	0.000	0.271	0.018	0.000	-0.839	0.000
diaknot	0.000	0.000	0.734	-0.092	0.000	-0.429
Number of nonzero loadings	6	7	7	8	8	8
Variance (%)	29.0	17.3	15.8	9.9	7.8	7.0
Adjusted Variance (%)	29.0	16.7	10.6	7.5	6.5	2.1
Cumulative Adjusted Variance (%)	29.0	45.7	56.4	63.9	70.3	72.4

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.423	0.000	0.000	0	0	0
length	-0.430	0.000	0.000	0	0	0
moist	0.000	0.676	0.000	0	0	0
testsg	0.000	0.662	0.000	0	0	0
ovensg	0.000	0.000	0.000	1	0	0
ringtop	-0.268	0.000	0.000	0	0	0
ringbut	-0.403	0.000	0.000	0	0	0
bowmax	-0.313	0.000	0.558	0	0	0
bowdist	-0.379	0.000	0.000	0	0	0
whorls	-0.400	0.000	0.187	0	0	0
clear	0.000	0.182	0.000	0	1	0
knots	0.000	0.267	-0.679	0	0	0
diaknot	0.000	0.000	-0.438	0	0	1
Number of nonzero loadings	7	4	4	1	1	1
Variance (%)	30.7	15.3	13.4	7.7	7.7	7.7
Adjusted Variance (%)	30.7	15.0	7.5	7.5	7.0	4.9
Cumulative Adjusted Variance (%)	30.7	45.8	53.3	60.7	67.7	72.6

Table 3: IMS-PCA of Pitprops data, maximization of variance

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.444	0.235	-0.179	0.000	0.000	0.000
length	-0.453	0.000	-0.179	0.000	-0.076	0.121
moist	0.000	0.602	0.000	-0.118	0.304	0.148
testsg	0.000	0.617	0.000	0.000	0.264	-0.070
ovensg	0.000	0.000	0.477	0.000	0.000	0.000
ringtop	0.000	0.268	0.439	0.000	-0.304	0.718
ringbut	-0.379	0.000	0.424	0.139	-0.157	0.000
bowmax	-0.341	-0.160	0.000	-0.265	0.274	-0.158
bowdist	-0.403	0.000	0.000	-0.172	0.000	0.000
whorls	-0.418	-0.118	0.000	0.256	0.000	-0.160
clear	0.000	0.000	0.000	-0.874	-0.313	0.000
knots	0.000	0.299	-0.254	0.195	-0.736	-0.103
diaknot	0.000	0.000	-0.519	0.018	0.000	0.618
Number of nonzero loadings	6	7	7	8	8	8
Variance (%)	29.0	17.3	15.7	8.9	7.0	6.5
Adjusted Variance (%)	29.0	17.2	15.3	8.8	6.8	6.4
Cumulative Adjusted Variance (%)	29.0	46.2	61.5	70.3	77.1	83.5

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.423	0.000	0.000	0	0	0
length	-0.430	0.000	0.000	0	0	0
moist	0.000	0.676	0.000	0	0	0
testsg	0.000	0.662	0.000	0	0	0
ovensg	0.000	0.000	-0.659	0	0	0
ringtop	-0.268	0.000	-0.435	0	0	0
ringbut	-0.403	0.000	0.000	0	0	0
bowmax	-0.313	0.000	0.000	0	0	0
bowdist	-0.379	0.000	0.308	0	0	0
whorls	-0.400	0.000	0.000	0	0	0
clear	0.000	0.182	0.000	1	0	0
knots	0.000	0.267	0.000	0	1	0
diaknot	0.000	0.000	0.530	0	0	1
Number of nonzero loadings	7	4	4	1	1	1
Variance (%)	30.7	15.3	11.1	7.7	7.7	7.7
Adjusted Variance (%)	30.7	15.0	10.9	7.1	5.9	4.1
Cumulative Adjusted Variance (%)	30.7	45.8	56.7	63.8	69.7	73.8

Table 4: IMS-PCA of Pitprops data, Maximization of Adjusted Variance

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.423	0.000	0.000	0	0	0
length	-0.430	0.000	0.431	0	0	0
moist	0.000	0.654	0.000	0	0	0
testsg	0.000	0.635	0.000	0	0	0
ovensg	0.000	0.000	-0.682	0	0	0
ringtop	-0.268	0.000	-0.551	0	0	0
ringbut	-0.403	0.000	0.000	0	0	0
bowmax	-0.313	0.000	0.000	0	0	0
bowdist	-0.379	0.000	0.000	0	0	0
whorls	-0.400	-0.222	0.000	0	0	0
clear	0.000	0.000	0.000	1	0	0
knots	0.000	0.345	0.000	0	1	0
diaknot	0.000	0.000	0.214	0	0	1
Number of nonzero loadings	7	4	4	1	1	1
Variance (%)	30.7	15.3	10.5	7.7	7.7	7.7
Adjusted Variance (%)	30.7	15.3	10.5	7.4	5.5	5.2
Cumulative Adjusted Variance (%)	30.7	46.0	56.6	64.0	69.4	74.6

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.444	0.097	-0.285	0.000	0.000	-0.120
length	-0.453	0.000	-0.301	0.000	0.000	-0.166
moist	0.000	0.585	0.000	-0.078	0.272	0.373
testsg	0.000	0.572	0.210	0.000	0.321	0.000
ovensg	0.000	0.000	0.526	0.000	0.323	-0.591
ringtop	0.000	0.129	0.456	0.000	-0.364	-0.116
ringbut	-0.378	0.000	0.300	0.076	-0.275	0.000
bowmax	-0.341	-0.328	0.000	-0.222	0.146	0.000
bowdist	-0.403	0.000	0.000	-0.101	0.000	0.000
whorls	-0.418	0.000	0.000	0.279	0.000	0.194
clear	0.000	0.000	0.000	-0.880	-0.238	-0.145
knots	0.000	0.394	0.000	0.169	-0.658	0.000
diaknot	0.000	0.202	-0.456	0.220	0.000	-0.630
Number of nonzero loadings	6	7	7	8	8	8
Variance (%)	29.0	16.3	14.5	8.6	6.7	6.2
Adjusted Variance (%)	29.0	16.3	14.5	8.6	6.7	6.2
Cumulative Adjusted Variance (%)	29.0	45.3	59.8	68.4	75.1	81.3

Table 5: IMS-PCA of Pitprops data, uncorrelated components (same sparsity as Zou *et al.* (2004) and SCoTLASS).

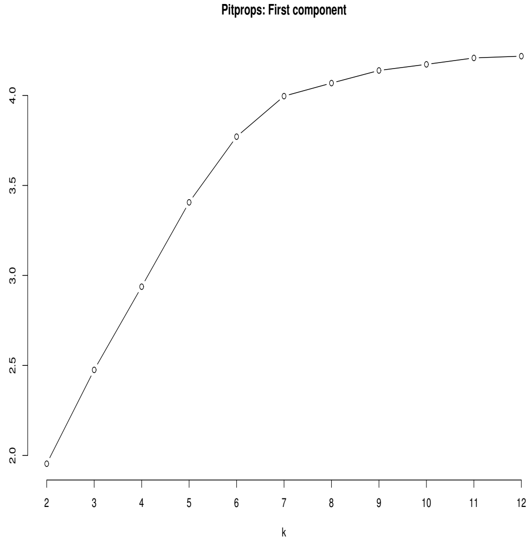

Figure 1: Scree Test for the first sPC, Pitprops data

covariance matrix is used instead of the correlation matrix.

We also illustrate our methods to choose the degree of sparsity. Figure 1 shows the scree-test in which the variability of the first component is plot as a function of k. Such scree test would probably lead to agreement with the previous methods by choosing k as 6 or 7.

Finally, in Tables 6 and 7 results for degree of sparsity of each component chosen as the maximizer of (4) are shown. Note that the first three components can explain almost the same amount of variance as the first three of Zou *et al.* (2004), using the same number of variables in the first component and more sparsity in the second and third component. Not surprisingly all the non-zero loadings are far from zero, and additionally our loadings are orthogonal so that interpretation is easier. Maximization of adjusted variance under orthogonality of the components is the strategy we finally advocate as the most sensible in this application.

Table 8 shows loadings for uncorrelated components with automatically chosen sparsity, arising the same comments.

Finally, suppose we want to perform a simultaneous variable selection and dimension reduction. For each $k = 6, \ldots, 12$ we select the k variables whose first 6 principal components explain the maximum amount of information. We note that the first 6 ordinary principal components explain 87% of the variance. If we use only 10 variables, we can still explain 79% of the variance, and 61% can be explained by using only 8 variables. This yields sparse principal components that are both uncorrelated and whose loadings are orthogonal.

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.423	0.000	0.000	0.000	0	0
length	-0.430	0.000	0.000	0.000	0	0
moist	0.000	0.707	0.000	0.000	0	0
testsg	0.000	0.707	0.000	0.000	0	0
ovensg	0.000	0.000	0.000	0.707	0	0
ringtop	-0.268	0.000	0.488	0.000	0	0
ringbut	-0.403	0.000	0.000	0.000	0	0
bowmax	-0.313	0.000	-0.417	0.000	0	0
bowdist	-0.379	0.000	0.000	0.000	0	1
whorls	-0.400	0.000	0.000	0.000	0	0
clear	0.000	0.000	0.000	0.000	1	0
knots	0.000	0.000	0.766	0.000	0	0
diaknot	0.000	0.000	0.000	-0.707	0	0
Number of nonzero loadings	7	2	3	2	1	1
Variance (%)	30.7	14.5	9.3	9.3	7.7	7.7
Adjusted Variance (%)	30.7	13.9	8.2	9.0	7.5	3.1
Cumulative Adjusted Variance (%)	30.7	44.6	52.8	61.8	69.3	72.4

Table 6: IMS-PCA of Pitprops data, maximization of variance, automatically chosen sparsity

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.423	0.000	0.000	0	0	0.000
length	-0.430	0.000	-0.283	0	0	0.000
moist	0.000	0.707	0.000	0	0	0.000
testsg	0.000	0.707	0.000	0	0	0.000
ovensg	0.000	0.000	0.600	0	0	0.704
ringtop	-0.268	0.000	0.455	0	0	0.000
ringbut	-0.403	0.000	0.000	0	0	0.000
bowmax	-0.313	0.000	0.000	0	0	0.000
bowdist	-0.379	0.000	0.000	0	0	0.000
whorls	-0.400	0.000	0.000	0	0	0.000
clear	0.000	0.000	0.000	1	0	0.000
knots	0.000	0.000	0.000	0	1	0.000
diaknot	0.000	0.000	-0.594	0	0	0.710
Number of nonzero loadings	7	2	4	1	1	2
Variance (%)	30.7	14.5	11.8	7.7	7.7	6.1
Adjusted Variance (%)	30.7	13.9	11.7	7.5	6.8	5.9
Cumulative Adjusted Variance (%)	30.7	44.6	56.3	63.8	70.6	76.5

Table 7: IMS-PCA of Pitprops data, maximization of adjusted variance, automatically chosen sparsity

Variable	PC1	PC2	PC3	PC4	PC5	PC6
topdiam	-0.423	0.000	0.306	0.000	0.000	0.000
length	-0.430	0.000	0.000	0.000	-0.202	0.000
moist	0.000	-0.654	0.000	0.000	0.221	0.260
testsg	0.000	-0.635	0.000	0.000	0.000	0.271
ovensg	0.000	0.000	-0.867	0.000	-0.331	0.219
ringtop	-0.268	0.000	-0.392	0.000	0.000	-0.374
ringbut	-0.403	0.000	0.000	0.000	0.000	-0.255
bowmax	-0.313	0.000	0.000	0.000	0.000	0.219
bowdist	-0.379	0.000	0.000	0.000	0.000	0.000
whorls	-0.400	0.222	0.000	0.093	0.000	0.000
clear	0.000	0.000	0.000	-0.918	-0.223	-0.289
knots	0.000	-0.345	0.000	0.328	0.000	-0.689
diaknot	0.000	0.000	0.000	0.200	-0.867	0.000
Number of nonzero loadings	7	4	3	4	5	8
Variance (%)	30.7	15.3	9.5	7.9	6.5	6.9
Adjusted Variance (%)	30.7	15.3	9.5	7.9	6.5	6.9
Cumulative Adjusted Variance (%)	30.7	46.0	55.6	63.5	70.0	76.9

Table 8: IMS-PCA of Pitprops data, uncorrelated components, automatically chosen sparsity

6 Conclusions

PCA falls into the category of *feature extraction* techniques, trying to build new variables that carry a large part of the global information. On the other hand, *feature selection* techniques (Guyon and Elisseeff (2003), Miller (1990), etc.) find an appropriate subset of the original variables to represent the data. It may be desirable to combine feature selection and extraction, and a possibility is given by sparseness of the loadings in PCA. We have seen for instance how to select the best subset of $k \leq m$ variables in order to extract $p \leq k$ linear combinations of those that carry the largest possible proportion of global information.

We showed an information maximization approach to the sparse principal components problem. Maximizing the information of each sPC instead of achieving a sparse approximation of the ordinary PC avoids taking into account the (maybe small) weight given by the ordinary PC to less important variables. In this sense, sparse principal component analysis is a valid alternative to the common practice of rotating principal components to enhance interpretability.

Our approach is flexible, as both the constraints and objective function can be chosen, together with the degree of sparsity; and in the example we saw it leads to a good compression with large sparsity. We gave some guidelines to such choices, and furthermore in real data applications different combinations can be tried and the sparse solution achieving exact orthogonality/uncorrelation and closest to uncorrelation/orthogonality adopted.

The most pressing refinement to our approach is an efficient extension to high dimensional situations. For moderate dimensional situations the algorithm is surprisingly fast, because of the availability of efficient maximization procedures for the simple problem at each node of the branch and bound algorithm. Still, even if our approach can efficiently handle large n data matrices, unfortunately when m is big (say $m > 100$) the number of nodes can get too large to allow for an exploration of the entire tree. Due to the nature of the problem, we actually expect the method to be applicable also in high-dimensional situations when the covariance matrix is used and few variables have much higher variance than the remaining. If this is not the case, it still happens in few high-dimensional problems (like in DNA Microarray data analysis) that only less than 1% of the variables are expected to finally enter into the model, so we can moreover suggest a preliminary variable selection. The algorithm will be finally applied only to a subset of prospective relevant variables. Another possibility is to split the variables into groups, perform an sPCA on the groups; and finally aggregate and perform a final sPCA on the extracted variables. The results are not guaranteed to be optimal but certainly a genuine sparse principal components analysis. It is worth noticing that, among the competing approaches, Zou *et al.* (2004) have managed to apply their method in high-dimensional situations.

References

M.S. Bartlett (1950). Tests of significance in factor analysis. *British Journal of Psychology*, **3**, 77–85.

M.S. Bartlett (1951). A further note on tests of significance in factor analysis. *British Journal of Psychology*, **4**, 1–2.

J. Cadima and I.T. Jolliffe (1995). Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, **22**.

R.B. Cattell (1966). The meaning and strategic use of factor analysis. In: *Handbook of multivariate experimental psychology*. Cattell, R.B.

C. Chatfield and A.J. Collins (1980). *Introduction to Multivariate Analysis*. Chapman and Hall.

T.F. Coleman and Y. Li (1994). On the convergence of reflective newton methods for large-scale nonlinear minimization subject to bounds. *Mathematical Programming*, **67**, 189–224.

P.E. Gill, W. Murray, and M.H. Wright (1981). *Practical Optimization*. Academic Press.

I. Guyon and A. Elisseeff (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157–1182.

D.J. Hand (1981). Branch and bound in statistical data analysis. *The Statistician*, **30**.

J.L. Horn (1965). A rationale and test for the number of factoris in factor analysis. *Psychometrika*, **30**, 179–185.

J. Jeffers (1967). Two case studies in the application of principal components. *Applied Statistics*, **16**, 225–236.

I. Jolliffe (2002). *Principal Component Analysis*. Springer-Verlag.

I.T. Jolliffe (1982). A note on the use of principal components in regression. *Appl. Statist.*, **31**, 300–303.

I.T. Jolliffe (1995). Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, **22**, 29–35.

I.T. Jolliffe, N.T. Trendafilov, and M. Uddin (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, **12**, 531–547.

I.T. Jolliffe and M. Uddin (2000). The simplified component technique - an alternative to rotated principal components. *Journal of Computational and Graphical Statistics*, **9**, 689–710.

H.F. Kaiser (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, **20**, 141–151.

A. Miller (1990). *Subset Selection in Regression*. Chapman & Hall.

P.M. Narendra and K. Fukunaga (1977). A branch and bound algorithm for feature subset selection. *IEEE trans. Comput. C*, **26**, 917–922.

M.S. Ridout (1988). Algorithm AS 233: An improved branch and bound algorithm for feature subset selection. *Applied Statistics*, **37**, 139–147.

N.T. Trendafilov and I.T. Jolliffe (2006). Projected gradient approach to the numerical solution of the SCoTLASS. *Computational Statistics and Data Analysis*, **50**, 242–253.

W.F. Velicer (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, **41**.

J.H. Wilkinson (1965). *The algebraic eigenvalue problem*. Oxford University Press.

H. Zou and T. Hastie (2003). Regression shrinkage and selection via the elastic net, with applications to microarrays. *Tech. rep.*, Department of Statistics, Stanford University.

H. Zou, T. Hastie, and R. Tibshirani (2004). Sparse principal components analysis. *Tech. rep.*, Department of Statistics, Stanford University.

W.R. Zwick and W.F. Velicer (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, **99**, 432–442.