Sparse Principal Component Analysis in Medical Shape Modeling

Karl Sjöstrand, Mikkel B. Stegmann and Rasmus Larsen

Informatics and Mathematical Modelling, Technical University of Denmark,
Richard Petersens Plads, DK-2800 Kgs. Lyngby, Denmark

ABSTRACT

Principal component analysis (PCA) is a widely used tool in medical image analysis for data reduction, model building, and data understanding and exploration. While PCA is a holistic approach where each new variable is a linear combination of all original variables, *sparse PCA* (SPCA) aims at producing easily interpreted models through sparse loadings, i.e. each new variable is a linear combination of a subset of the original variables. One of the aims of using SPCA is the possible separation of the results into isolated and easily identifiable effects.

This article introduces SPCA for shape analysis in medicine. Results for three different data sets are given in relation to standard PCA and sparse PCA by simple thresholding of small loadings. Focus is on a recent algorithm for computing sparse principal components, but a review of other approaches is supplied as well. The SPCA algorithm has been implemented using Matlab and is available for download.

The general behavior of the algorithm is investigated, and strengths and weaknesses are discussed. The original report on the SPCA algorithm argues that the ordering of modes is not an issue. We disagree on this point and propose several approaches to establish sensible orderings. A method that orders modes by decreasing variance and maximizes the sum of variances for all modes is presented and investigated in detail.

Keywords: Sparse Principal Component Analysis, PCA, Evaluation of Principal Components, Shape Modeling, Corpus Callosum

1. INTRODUCTION

Few computational methods for data understanding, exploration and reduction has found more use than principal component analysis (PCA). PCA takes an $(n \times p)$ data matrix $\mathbf{X}$, n being the number of observations and p being the number of variables, and transforms it by $\mathbf{Z} = \mathbf{XB}$ such that the derived variables (the columns of $\mathbf{Z}$) are uncorrelated and correspond to directions of maximal variance in the data. The derived coordinate axes are the columns of $\mathbf{B}$, called *loading vectors* with individual elements known as *loadings*. These are at right angles with each other; PCA is simply a rotation of the original coordinate system, and the $(p \times p)$ loading matrix $\mathbf{B}$ is the rotation matrix. The new variables (the columns of $\mathbf{Z}$) are known as *principal components* (PCs). Usually only the first k components, $k < p$, are retained since these explain the majority of the sample set variance. This makes $\mathbf{Z}$ $(n \times k)$ and $\mathbf{B}$ $(p \times k)$. The loading matrix can be calculated using a singular value decomposition of the data matrix $\mathbf{X}$ or through an eigenanalysis of the corresponding covariance or correlation matrix.

Another way of viewing PCA is by treating each new variable as a linear combination of the original variables. The loadings then translate to coefficients and may be investigated in detail to determine the important factors behind each PC. The problem is that each new variable is a linear combination of *all* variables, and the loadings are typically non-zero. This makes interpretation difficult. *Sparse principal component analysis* aims at approximating the properties of regular PCA while keeping the number of non-zero loadings small.

The most straight-forward way of obtaining sparse loadings is by simple thresholding, where sufficiently small loadings are truncated to zero. The threshold can be chosen using e.g. Jeffers' criterion[1] of excluding,

Further author information:
Karl Sjöstrand: E-mail: kas@imm.dtu.dk, Telephone: +45 4525 3423
Mikkel B. Stegmann: E-mail: mbs@imm.dtu.dk, Telephone: +45 4525 3422
Rasmus Larsen: E-mail: rl@imm.dtu.dk, Telephone: +45 4525 3415

disregarding signs, loadings below 70% of the largest loading for each PC. Thresholding can be misleading in several respects, as discussed by Jolliffe.[2] The influence of a variable on a specific PC is not dependent on the magnitude of the corresponding loading only, but is governed by a series of relationships, such as variable size, or analogously, variance.

Among the earliest methods for obtaining a *simple structure* of the loadings of the original variables is the class of orthomax rotations,[3] where an initial basis is rotated due to some objective criterion. The basis can for example be provided by a PCA. Let $\mathbf{B}$ be a $p \times k$ orthonormal matrix (of column eigenvectors) and $\mathbf{\Omega}$ be an orthonormal rotation matrix in $\mathbb{R}^k$, i.e. $\mathbf{\Omega}^\mathsf{T}\mathbf{\Omega} = \mathbf{I}$. Then, the class of orthomax rotations can be defined as

$$\mathbf{\Omega}_o = \arg\max_{\mathbf{\Omega}} \left(\sum_{j=1}^{k} \sum_{i=1}^{p} (\mathbf{B\Omega})_{ij}^4 - \frac{\gamma}{p} \sum_{j=1}^{k} \left(\sum_{i=1}^{p} (\mathbf{B\Omega})_{ij}^2 \right)^2 \right), \tag{1}$$

where $\mathbf{\Omega}_o$ denotes the resulting rotation and γ denotes the type. In the orthomax class we find the Varimax[4] case where $\gamma = 1$. Here, Equation 1 simplifies to a sum of variances. The variances are calculated for each loading vector where the individual loadings are squared. This emphasizes sparsity within each loading vector by clustering loadings into an approximate bimodal distribution of large and very small loadings. Although the resulting components may not be strictly sparse, one benefit of the Varimax method is that it is computationally feasible in high-dimensional cases, see e.g.[5]

Chennubhotla and Jepson present another criterion for finding a suitable rotation matrix based on the entropy of the loading matrix.[6] A cost function,

$$C = C_1 + \lambda C_2,$$

is minimized where $C_1 = \sum_{j=1}^{k} -d_j \log d_j$ and d_j is the relative variance of the jth principal component. Next, $C_2 = \sum_{i=1}^{p} \sum_{j=1}^{k} -b_{i,j}^2 \log b_{i,j}^2$, were $b_{i,j}$ denotes the elements of the $(p \times k)$ loading matrix. Optimizing C_1 alone gives the standard PCA solution, while C_2 is minimal for the identity matrix, thus promoting sparsity. Similarly to the Varimax criterion, suppressed loadings will be small but non-zero. To achieve strict sparsity, thresholding of small loadings is performed as discussed above. The resulting loading vectors will, contrary to those constructed using the Varimax criterion, explain a decreasing amount of variance of the original data set; a feature it has in common with regular PCA. Additionally, the number of non-zero loadings also decrease, making a multi-scale interpretation possible.

Simple principal components[7] is a technique for producing particularly simple, and possibly sparse, loading vectors. It uses a series of in-plane rotations affecting two loading vectors at a time such that the resulting directions explain maximal variance subject to being represented by integers. The end result is a set of orthogonal loading vectors represented by (primarily small) integers. Empirical evidence shows that the correlations between the resulting PCs are low. Small loadings will typically be translated to zeros, resulting in a sparse loading matrix structure. Similar ideas have been put forth by Hausman[8] and Rousson and Gasser.[9]

D'Aspremont et al. takes a variational approach to sparse PCA.[10] The PCs are estimated separately by approximating a positive semidefinite symmetric matrix (the covariance or correlation matrix) by a rank-one matrix, $\mathbf{b}\mathbf{b}^T$. To impose sparsity, a constraint is added on the maximum number of non-zero elements of $\mathbf{b}$, known as the *cardinality* of $\mathbf{b}$. This direct formulation results in a non-convex optimization problem that is difficult to solve. The problem is therefore relaxed by replacing the cardinality constraint with a convex one, making the computation feasible. The resulting PCs are reported to explain a larger proportion of variance than competing algorithms, but the complexity of the formulation grows quickly with the number of variables.

This article focuses on a method for computing sparse loading vectors using concepts from variable selection in regression. A method coined SCoTLASS[11] (Simplified Component Technique-LASSO) predates this method and is based on similar ideas. Maximizing the expression

$$\mathbf{b}_i^T \mathbf{R} \mathbf{b}_i,$$

where $\mathbf{R}$ denotes the covariance matrix of $\mathbf{X}$, subject to

$$\mathbf{b}_i^T \mathbf{b}_i = 1 \quad \text{and} \quad \mathbf{b}_j^T \mathbf{b}_i = 0, \;\; j \neq i,$$

renders the solution of a regular PCA. The authors propose to add the constraint

$$\|\mathbf{b}_i\|_1 = \sum_{j=1}^{p} |b_{ij}| \leqslant t, \quad t \in \mathbb{R}^+, \quad \forall i.$$

The parameter t controls the sparsity of the loading vectors $\mathbf{b}_k$. The addition of this constraint was inspired by the LASSO[12] regression method described below. However, this necessitates the use of a numerical optimization method. The problem formulation contains p parameters which is a potentially large number, and the cost function contains several local minima. The authors use a simulated annealing approach for optimization, which adds a number of tuning parameters in itself.

The following section presents the theory of the present method of sparse PCA, hereafter simply denoted SPCA. Section 3 shows results on shape data from three different data sets along with results on the general properties of SPCA. Section 4 discusses the obtained results, debates the advantages and drawbacks of SPCA and proposes a range of different possibilities for ordering of modes. Section 5 concludes the paper.

2. METHODS

This section gives a brief description of the SPCA algorithm and discusses its relation to variable selection methods in regression. For a complete treatment, consult Ref. 13 and the preliminary papers 12, 14 and 15.

2.1. Regression Techniques

The regression methods presented here all originate from ordinary least squares (OLS) approximations. The *response* variable $\mathbf{y}$ is approximated by the *predictors* in $\mathbf{X}$. The coefficients for each variable (column) of $\mathbf{X}$ are contained in $\mathbf{b}$,

$$\mathbf{b}_{\text{OLS}} = \arg\min_{\mathbf{b}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2, \tag{2}$$

where $\|\cdot\|$ represents the L2-norm. This is the best linear unbiased estimator given a number of assumptions, such as independent and identically distributed (i.i.d.) residuals. However, if some bias is allowed, estimators can be found with lower mean square error than OLS when tested on an unseen set of observations. A common way of implementing this is by introducing some constraint on the coefficients in $\mathbf{b}$. The methods described here use constraints on either the L1-norm or the L2-norm of $\mathbf{b}$, or both. Adding the L2 constraint gives

$$\mathbf{b}_{\text{ridge}} = \arg\min_{\mathbf{b}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \lambda\|\mathbf{b}\|^2. \tag{3}$$

This is known as ridge regression. Any positive λ will shrink the coefficients of $\mathbf{b}$; if λ is chosen carefully, this may lead to improved prediction accuracy and better numerical properties. Replacing the L2-norm in the constraint with the L1-norm gives

$$\mathbf{b}_{\text{LASSO}} = \arg\min_{\mathbf{b}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \delta\|\mathbf{b}\|_1, \tag{4}$$

where $\|\mathbf{b}\|_1 = \sum_{i=1}^{p} |b_i|$. This method is coined LASSO,[12] the least absolute shrinkage and selection operator. As the name implies, using the L1-norm not only shrinks the coefficients, but drives them one by one to exactly zero as δ grows. This implements a form of variable selection, as minor coefficients will be set to zero in a controllable fashion, while the remaining coefficients will be altered to mimic the response in the best possible way. The relation to the problem of setting small PCA loadings to zero is already evident, but some more theory is needed before this can be properly handled.

LASSO has proven to be a very powerful regression and variable selection technique, but it has a few limitations. If $p > n$, i.e. there are more variables than observations, LASSO chooses a maximum of n variables. If there is a group of strongly correlated predictors, LASSO tends to choose a single predictor from that group only. The elastic net regression method[15] was developed to address these shortcomings. It uses a combination of the constraints from ridge regression and LASSO,

$$\mathbf{b}_{\text{nEN}} = \arg\min_{\mathbf{b}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \lambda\|\mathbf{b}\|^2 + \delta\|\mathbf{b}\|_1, \tag{5}$$

where nEN is short for naive elastic net for reasons described below. The elastic net can be formulated as a LASSO problem on augmented variables,

$$\mathbf{b}^*_{\text{nEN}} = \arg\min_{\mathbf{b}^*} \|\mathbf{y}^* - \mathbf{X}^*\mathbf{b}^*\|^2 + \frac{\delta}{\sqrt{1+\lambda}}\|\mathbf{b}^*\|_1, \tag{6}$$

where

$$\mathbf{X}^*_{(n+p)\times p} = \frac{1}{\sqrt{1+\lambda}}\begin{bmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{I}_p\end{bmatrix}, \quad \mathbf{y}^*_{n+p} = \begin{bmatrix}\mathbf{y}\\ \mathbf{0}_p\end{bmatrix}, \quad \mathbf{b}^* = \sqrt{1+\lambda}\mathbf{b}.$$

The authors argue that the formulation in Equation 6 incurs a double amount of coefficient shrinkage, which is why the solution of Equation 5 is referred to as naive. The excessive shrinkage is compensated for in the final solution for $\mathbf{b}_{\text{EN}}$ which is

$$\mathbf{b}_{\text{EN}} = \sqrt{1+\lambda}\mathbf{b}^*_{\text{nEN}} = (1+\lambda)\mathbf{b}_{\text{nEN}}. \tag{7}$$

The resulting LASSO problem has more observations $(p+n)$ than variables (p), which is why cases where $p > n$ are handled gracefully. If $\lambda > 0$, the elastic net constraint function $\lambda\|\mathbf{b}\|^2 + \delta\|\mathbf{b}\|_1$ is strictly convex. It can be shown[15] that the difference between coefficients of highly correlated variables in such a system is very small. The elastic net therefore has a tendency of grouping variables, contrary to the LASSO. These are two properties that are desirable in a PCA framework. Problems where there are more variables than observations are common, and principal components built from highly correlated and significant variables are easier to interpret.

Ordinary least squares and ridge regression have closed-form solutions, that is, $\mathbf{b}_{\text{OLS}}$ and $\mathbf{b}_{\text{ridge}}$ can be expressed as simple functions of $\mathbf{X}$, $\mathbf{y}$ and λ. This is not true for the LASSO and elastic net methods. For many years, LASSO solutions were found using standard optimization techniques, which made for long computation times. In 2002, Efron et al. published a report on a new regression method called least angle regression[14] (LARS). The terminal S in LARS refers to its close relation to stagewise regression and LASSO. Although conceptually different, the method is shown to be very similar to LASSO, and through a small modification, the exact LASSO solution can be computed. The method is built on a powerful geometric framework, through which a computationally thrifty algorithm is conceived. The algorithm starts with all coefficients at zero, and successively adds predictors until all variables are active and the ordinary least squares solution is reached. In other words, LARS returns the solutions for all possible values of δ. What remains is to pick a suitable solution, a proper value of δ. This can for instance be done using cross-validation or prior knowledge of the desired number of non-zero coefficients. In the elastic net setting, LARS returns the solutions corresponding to all possible values of δ given a value of λ. Ref. 15 describes a further development of the LARS algorithm tailor made to suit the elastic net framework. This extension is called LARS-EN.

In summary, a regression approach has been presented through which a relevant subset of variables can be selected, which handles the case of more variables than observations gracefully, and which can be computed efficiently. We now turn to the problem of calculating sparse PCs. Note that "sparse PCs" refers to principal components formed by linear combinations of sparse sets of variables.

2.2. Sparse Principal Component Analysis (SPCA)

The simplest approach to SPCA using regression is by treating each principal component as a response vector and regressing this on the p variables. Denoting the ith PC and loading vector by $\mathbf{z}_i$ and $\mathbf{b}_i$ respectively, and inserting this into the elastic net framework gives

$$\hat{\mathbf{b}}_i = \arg\min_{\mathbf{b}_i} \|\mathbf{z}_i - \mathbf{X}\mathbf{b}_i\|^2 + \lambda\|\mathbf{b}_i\|^2 + \delta\|\mathbf{b}_i\|_1. \tag{8}$$

The principal component $\mathbf{z}_i$ is calculated using regular PCA. The regression procedure will calculate a loading vector $\mathbf{b}_i$ such that the resulting PC is close to $\mathbf{z}_i$ while being sparse. The weakness of this approach is that all solutions are constrained to the immediate vicinity of a regular PCA. A better approach would be to approximate the *properties* of PCA, rather than its exact results. Specifically, the loading matrix $\mathbf{B}$ should be near orthogonal,

and the correlations between the PCs of the scores matrix $\mathbf{Z}$ should be kept low. Zou and Hastie propose a problem formulation called the *SPCA criterion*[13] to address this.

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i\|^2 + \lambda \sum_{j=1}^{k} \|\mathbf{b}_j\|^2 + \sum_{j=1}^{k} \delta_j \|\mathbf{b}_j\|_1 \quad \text{subject to} \quad \mathbf{A}^T\mathbf{A} = \mathbf{I}_k \tag{9}$$

To clarify this expression, it will be broken down into components. First, $\mathbf{B}^T\mathbf{x}_i$ takes the variables of observation i and projects them onto the principal axes (loading vectors) of $\mathbf{B}$. Note that $\mathbf{x}_i$ denotes the ith column of $\mathbf{X}^T$. Only k PCs are retained, meaning that some information is lost in this transformation. Further, $\mathbf{A}\mathbf{B}^T\mathbf{x}_i$ takes the scores of $\mathbf{B}^T\mathbf{x}_i$ and transforms them back into the original space. The orthogonality constraint on $\mathbf{A}$ makes sure $\mathbf{B}$ is near orthogonal. The whole term $\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^T\mathbf{x}_i\|^2$ measures the reconstruction error. The remaining constraints are the same as for elastic net regression, driving the columns of $\mathbf{B}$ towards sparsity. The constraint weight λ must be chosen beforehand, and has the same value for all PCs, while δ may be set to different values for each PC, offering good flexibility. It can be shown[13] that for $\delta_j = 0 \ \forall j$, the SPCA criterion is minimized by setting $\mathbf{A}$ and $\mathbf{B}$ equal to the loading matrix of ordinary PCA. Hence, the solutions of the present formulation of SPCA conveniently range from ordinary PCA on one end, to the (maximally sparse) zero matrix on the other.

Equation 9 resembles the elastic net formulation but there is a significant difference. Instead of estimating a single coefficient vector, this problem has two matrices of coefficients, $\mathbf{A}$ and $\mathbf{B}$. A reasonably efficient optimization method for minimizing the SPCA criterion is presented in Ref. 13. First, assume that $\mathbf{A}$ is known. By expanding and rearranging Equation 9, it is shown that $\mathbf{B}$ can be estimated by solving k independent naive elastic net problems, one for each column of $\mathbf{B}$. Referring to Equation 5, the data matrix is $\mathbf{X}$ as usual while $\mathbf{y} = \mathbf{X}\mathbf{a}_i$ for the ith loading vector. On the other hand, if $\mathbf{B}$ is known, $\mathbf{A}$ can be calculated using a singular value decomposition; if $\mathbf{X}^T\mathbf{X}\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, then $\mathbf{A} = \mathbf{U}\mathbf{V}^T$. Since both matrices are unknown, an initial guess is made and $\mathbf{A}$ and $\mathbf{B}$ are estimated alternately until convergence. Ref. 13 suggests initializing $\mathbf{A}$ to the loadings of k first ordinary principal components.

2.3. Ordering of principal components

One goal of PCA is to recover latent variables that are as descriptive as possible. This is done by maximizing the variance of each PC subject to being orthogonal to higher order PCs. The performance of PCA methods is commonly measured by the amount of variance explained by each PC, and the total amount of variance for k modes. Regular PCA is the only linear transformation that produces both orthogonal loadings and uncorrelated scores.[16] For methods that produce correlated scores, variances cannot be calculated directly, as some of the variance explained by one PC will be present in others. This calls for a fair evaluation method. Several such methods are presented in Ref. 17, where it is concluded that the most powerful method is to measure *adjusted variance*, a term used by Zou and Hastie who suggest the same method in Ref. 13. The idea is that the variance of each PC should be adjusted for the variance already explained by higher order components. For mean centered variables, such as those derived by PCA, correlation is equivalent to the cosine of the angle between vectors. Zero correlation corresponds to a 90° angle between vectors while fully correlated variables are parallel. Adjustment of a PC therefore amounts to a transformation such that the resulting vector is at right angles with all higher order PCs. This is also known as Gram-Schmidt orthogonalization.

The variance of the jth PC is proportional to its squared length, $\operatorname{var}\mathbf{z}_j = \frac{\mathbf{z}_j^T\mathbf{z}_j}{n} \propto \mathbf{z}_j^T\mathbf{z}_j$. Any ordering that maximizes the total variance therefore also maximizes the sum of squared lengths. For ease of notation, squared lengths are considered in the following equations.

A vector $\mathbf{z}_j$ may be orthogonalized, or *adjusted* with respect to another vector $\mathbf{z}$ using orthogonal projection by

$$\hat{\mathbf{z}}_j = \mathbf{z}_j - \mathbf{z}(\mathbf{z}^T\mathbf{z})^{-1}\mathbf{z}^T\mathbf{z}_j.$$

The jth PC should be adjusted for all higher order PCs. Assume that the variables, the columns of $\mathbf{Z}$, have been sorted according to decreasing order. The adjustment can then be carried out for all higher order PCs simultaneously by

$$\hat{\mathbf{z}}_j = \mathbf{z}_j - \mathbf{Z}_{(j-1)}(\mathbf{Z}_{(j-1)}^T\mathbf{Z}_{(j-1)})^{-1}\mathbf{Z}_{(j-1)}^T\mathbf{z}_j,$$

where $\mathbf{Z}_{(j)} = [\mathbf{z}_1 \dots \mathbf{z}_j]$.[17]

The SPCA criterion (9) keeps the loading matrix near orthogonal by forcing $\mathbf{A}$ to be orthogonal, but does nothing to encourage uncorrelated scores. This makes an orthogonalization process central to SPCA. Zou and Hastie argue that the order of the components is not an issue; the order is left unaltered, making it possible for lower order modes to explain more variance than higher order modes. Furthermore, the amount of total adjusted variance is dependent on the ordering of the PCs, and may not be maximal in this case.

Formally, the variable ordering that maximizes the total variance can be established by maximizing $\sum_j \hat{\mathbf{z}}_j^T \hat{\mathbf{z}}_j$ and allowing for permutations,

$$\arg\max_{\mathbf{P}\in\mathcal{P}_k} \tilde{\mathbf{z}}_1^T\tilde{\mathbf{z}}_1 + \sum_{j=2}^{k} \tilde{\mathbf{z}}_j^T\tilde{\mathbf{z}}_j - \tilde{\mathbf{z}}_j^T\tilde{\mathbf{Z}}_{(j-1)}(\tilde{\mathbf{Z}}_{(j-1)}^T\tilde{\mathbf{Z}}_{(j-1)})^{-1}\tilde{\mathbf{Z}}_{(j-1)}^T\tilde{\mathbf{z}}_j, \tag{10}$$

where $\tilde{\mathbf{Z}} = \mathbf{ZP}$ is the permuted scores matrix and $\mathcal{P}_k$ is the set of permutation matrices of size k. Note that the supremum of Equation 10 is the sum of unadjusted variances which is equal to the maximum iff $\mathbf{Z}$ is orthogonal. The simplest way of finding the optimal permutation is by trying all $k!$ possible permutations, which is feasible for a low number of PCs. This paper proposes a forward selection-type rule for picking an ordering with two properties; the variance of a PC is less than or equal to the variances of higher order PCs, and the expression in Equation 10 is maximized in most cases. The rule is simple. Treat one PC at a time. At each step, choose the PC with largest (adjusted) variance, and adjust the scores matrix for this PC. This means that in the first step, we calculate the variances of all (unadjusted) PCs and choose the one with greatest variance. All PCs ($\mathbf{Z}$) are then adjusted with respect to the chosen PC ($\mathbf{z}_j$) using

$$\hat{\mathbf{Z}} = \mathbf{Z} - \mathbf{z}_j(\mathbf{z}_j^T\mathbf{z}_j)^{-1}\mathbf{z}_j^T\mathbf{Z}. \tag{11}$$

In the second step, the adjusted variances from the first step are considered. Again, the PC with greatest variance is chosen, and all PCs are updated using Equation 11. This process is repeated for all PCs. This results in a zero $\mathbf{Z}$ matrix, however, a sensible ordering has been established. This ordering is finally applied to the loading vectors and the original scores matrix.

The first property of this rule, which states that variances are decreasing, is easily realized since the longest vector is chosen in each step and since the squared length cannot grow as the vector is adjusted for some other vector. The second property of maximal total variance is empirically shown below to be fulfilled to a large extent, but as shall be seen, there are counter examples, e.g. $\mathbf{Z} = [[0\ 1.5]^T\ [1\ 1]^T\ [1\ \ -1]^T]$.

3. RESULTS

The SPCA algorithm has been applied to medical shape analysis. The shape data is contained in a data matrix $\mathbf{X}$ $(n \times p)$ where each shape corresponds to one row (observation) and the variables consist of the different landmark positions. Landmarks are defined by two coordinates (2D data); these are treated separately such that one coordinate is one variable. This project is concerned with 2D data only, although the techniques described herein are directly applicable to data of any dimensionality.

Three data sets were used in this study. The first consists of 37 annotations of the human face. Each face is represented by 58 landmarks. The second data set is a shape model of the lungs, the heart and the clavicles. The set contains 247 observations, each with 166 landmarks. The final data set represents the *corpus callosum* brain structure. This is the bundle of nerve fibers connecting the two cerebral hemispheres of the brain. The structure is well defined in the *mid-sagittal plane*, the plane that separates the left hemisphere from the right.[18] Further away from this plane, the structure dissolves into separate fibers, which is why it is best analyzed in 2D. The set has 62 observations, each with 78 landmarks.

Figure 1 shows regular and sparse decompositions of the face data set. Each set of figures shows the first 12 *modes of variation**, ordered by the method described above. It is evident that regular PCA produces holistic

*Modes of variation is a commonly used term where the jth mode denotes movements along the axis defined by the jth loading vector. The mean shape defines the origin and perturbations are measured offset to this.

modes of variation, each describing a series of effects at once, making interpretation difficult. SPCA, on the other hand, manages to display more or less separate effects for each mode. SPCA modes 2, 8 and 12 correspond to mouth opening/closing, upper lip thickness and smile/frown respectively. SPCA modes 4, 6 and 7 show differing eyebrow configurations. Figure 2 shows corresponding images for the lungs, heart and clavicles data set. SPCA mode 2 depicts the length of the clavicles, while most other modes are concerned with either lung or heart geometry, or both (e.g. SPCA mode 5). Figure 3 presents results for the corpus callosum data set.

Table 1 shows variance proportions for ten modes of variation of the corpus callosum data set. The top row contains results for a regular PCA, while the second row represents sparse PCA using thresholding. The third row presents the adjusted variances of SPCA with no reordering of modes, and the results in the bottom row are for SPCA using the proposed forward selection-type rule for mode ordering. It is seen that reordering the modes increases the total explained variance and ensures that the variances are decreasing.

Variance (%)	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10	$\sum$
PCA	43.27	18.55	13.74	7.71	4.93	2.01	1.58	1.36	0.98	0.78	94.92
threshold PCA	14.36	8.86	4.89	2.46	3.10	0.90	0.66	0.57	0.32	0.24	36.37
SPCA	13.21	7.51	5.18	3.44	2.27	0.60	0.26	0.90	0.11	0.03	33.50
reordered SPCA	15.12	7.88	7.10	3.37	1.35	1.06	0.42	0.32	0.11	0.03	36.77

Table 1. Explained proportion of variance for each mode and method for the corpus callosum data set. The last column shows the cumulative variance for all ten modes. Each sparse mode is set to affect 20 coordinates exactly (total 78), explaining the low proportions of variation.

The simplest alternative to SPCA is straight-forward truncation of loadings as described in the introduction. Some results of this scheme is found in Figure 4. The difficulties of this method are clear from these images; the modes of variation are merely pruned versions of those of regular PCA. Hence, the effects are scattered and hard to interpret.

Figure 5 shows an important property of SPCA. The results vary slowly with values of λ, the weighting term on the L2 norm of the loadings. Here, vastly different values are chosen, but with similar results.

The relatively strong correlations among the PCs produced by SPCA are evident in Figure 6 where correlations are plotted for the PCs, next to the angles between the loading vectors. The correlations become considerable, while most angles are in the vicinity of 90°, although with a few clear exceptions. These properties follow from the definition of the SPCA criterion as discussed earlier. The implication of the high correlations is that it becomes impossible to refer to one PC without referring to others. This is what motivates the discussion on ordering of modes.

The proposed method for ordering the principal components and the corresponding loading vectors proved successful in the majority of cases. To test the performance of the method, 100 random scores matrices were used as input and the average total amount of adjusted variance was measured in three different ways; using no reordering, the proposed method, and, by trying all possible combinations, the average maximal adjusted variance. This test was carried out for a number of combinations of the number of observations n, and the number of PCs k. Table 2 shows the complete set of results. The test matrices were all random, but to produce relevant scores matrices, a predefined covariance structure was used, and all variables had zero mean. The covariance structure from the SPCA calculations on the face data set was used. Similar results were obtained using other SPCA covariance matrices.

The $\mathbf{A}$ matrix is initialized to the first k loading vectors of a regular PCA. In the first SPCA iteration, the values of $\mathbf{B}$ will be influenced by this. However, as Figure 7 shows, as the iterations progress, the values of $\mathbf{B}$ converge to very different values; the resulting $\mathbf{B}$ seems to be independent of regular PCA. Tests with initialization of $\mathbf{A}$ to the identity matrix gives slightly different, but acceptable results.

	no reordering			forward selection reordering		
	$n = 10$	$n = 100$	$n = 1000$	$n = 10$	$n = 100$	$n = 1000$
$k = 3$	14.0 (98.8)	0	0	1.0 (98.8)	0	0
$k = 4$	46.0 (97.9)	16.0 (99.8)	5.0 (100.0)	0	0	0
$k = 5$	57.0 (97.6)	21.0 (99.8)	2.0 (100.0)	0	0	0
$k = 6$	76.0 (96.4)	39.0 (99.8)	4.0 (100.0)	6.0 (98.7)	4.0 (99.8)	0
$k = 7$	87.0 (96.8)	46.0 (99.8)	5.0 (100.0)	9.0 (98.3)	6.0 (99.9)	0
$k = 8$	95.0 (96.2)	50.0 (99.8)	2.0 (100.0)	6.0 (99.1)	16.0 (99.5)	0

Table 2. Results of the proposed ordering method (right) versus no reordering (left) for k n-dimensional random PCs with a static covariance structure. Numbers represent the average proportion (%) over 100 trials where the optimal ordering was not found. The optimal ordering was established by an all-subsets calculation in each case. The parenthesized numbers denote the average proportion of maximal variance reached in cases of failure. Note that the average proportion of maximal variance over all trials is higher.

4. DISCUSSION

The results presented in this article provide evidence that the presented SPCA algorithm is able to produce separate and easily identifiable modes of variation. We anticipate that SPCA will find good use in many clinical applications. In particular, the ability of SPCA to extract latent variables that are easily interpreted and visualized may help to understand the present variability. For instance, studies of atrophic processes in the human brain due to aging, dementia, Alzheimer's disease etc. may benefit from this treatment.

The algorithm requires $k + 1$ parameters, λ and one δ_i for each PC. From the results, it can be seen that the resulting loading vectors vary slowly with λ. The values of δ_i are, however, crucial. In this project, δ is set such that precisely 20 coordinates are affected in each mode, but any other choice is equally valid, and results would differ greatly. This makes the algorithm flexible, but parameter tuning requires knowledge of the problem at hand.

The computational complexity of SPCA for $n > p$ is at most $np^2+mO(p^3)$ where m is the number of iterations before the algorithm converges. If $p > n$, the complexity is of order $mkO(pln + l^3)$ where l is the number of non-zero loadings, see Ref. 13 for a more thorough discussion. Typical computation times for the examples in this article are less than one minute on a standard laptop computer. However, the number of iterations grows rapidly with the number of PCs, and computation times for each elastic net problem grow with the number of non-zero loadings. Memory consumption depends mostly on the number of non-zero loadings, as the algorithm creates an $(l \times l)$ matrix in each iteration. This makes it difficult to handle e.g. texture data in this setting, where thousands of non-zero loadings may be of interest. The SPCA article[13] presents a designated SPCA algorithm where λ is set to infinity. Each elastic net computation is replaced by a single matrix multiplication, allowing for much lower memory consumption and computational complexity. Results on this extension are, however, yet to come.

This article presents one simple way of ordering principal components. This method sorts the PCs according to descending variance and maximizes the total explained (adjusted) variance in most cases. Table 2 shows that the fail rate increases dramatically with increasing k if no reordering is performed, especially for a low number of observations. With reordering, this effect is considerably lower. It is also apparent that the negative impact of a failure drops with the number of dimensions, n. The shape data used in this article has approximately $k = 12$ and $37 \leqslant n \leqslant 247$. Without reordering the PCs, there is a considerable risk that the resulting total adjusted variance is sub-maximal, and, as shown in Table 1, the individual variances may not be sorted in descending order.

The proposed method of measuring the performance of SPCA is convenient, as it resembles the results of a regular PCA. However, other ways of ordering modes can be beneficial.

Sparsity Modes can be ordered according to the amount of sparsity of the corresponding loading vectors. Several SPCA calculations may be carried out, each with different sparsity constraints. The modes are then ordered according to sparsity, e.g. from highly local modes to more global effects.

Spatially Modes may also be ordered according to spatial locality. The center of attention is calculated for each mode. These are then ordered along the contour of the object.

Entropy Although the resulting loading vectors are sparse, each mode may describe more than one effect. Using results from information theory, the *entropy* of each mode can be calculated, effectively giving a measure on the amount of clustering. Modes can be ordered accordingly, for instance going from low entropy, where a mode describes a single effect and has limited spatial extent, to high entropy, where effects are scattered and/or affect a larger proportion of the contour.

Combinations To obtain a more thorough library of modes, they may be ordered according to two criteria simultaneously and put in a two-dimensional grid. For instance, a combination of sparsity and a spatial ordering may be useful, especially in an exploratory setting. It is plausible that an examiner has some idea of the spatial location and extent of the relevant effect. The search may then be constrained by isolating relevant modes by defining for instance a rectangle in the two-dimensional grid.

5. CONCLUSION

This article has introduced sparse principal component analysis (SPCA) to medical shape modeling. Results, shown on three different data sets, provide some evidence that SPCA manages to isolate relevant sparse effects in each mode of variation. The inherent design of SPCA keeps loading vectors near orthogonal, while correlations between principal components are typically high. This motivates a discussion on the ordering of PCs. A method that orders the modes according to descending variance was discussed in detail and shown to improve the estimates of adjusted variances notably, while a few other possibilities where mentioned briefly. The convergence of SPCA was shown to be irregular and slow at times, but results are superior to those of the more straightforward approaches, such as thresholding of loading vectors.

Future work includes using SPCA for other applications, such as exploratory analysis of fMRI data. The main obstacle in such analyses is the large number of variables. An examination of the discriminative power of SPCA calculations in medical shape modeling is also planned.

Source code for the statistics software S-Plus and its freeware sibling R has been written and made available by Hui Zou and Trevor Hastie, see `www.r-project.org`. The first author of this article has made a corresponding implementation for Matlab, available on `www.imm.dtu.dk/~kas/software/spca/`.

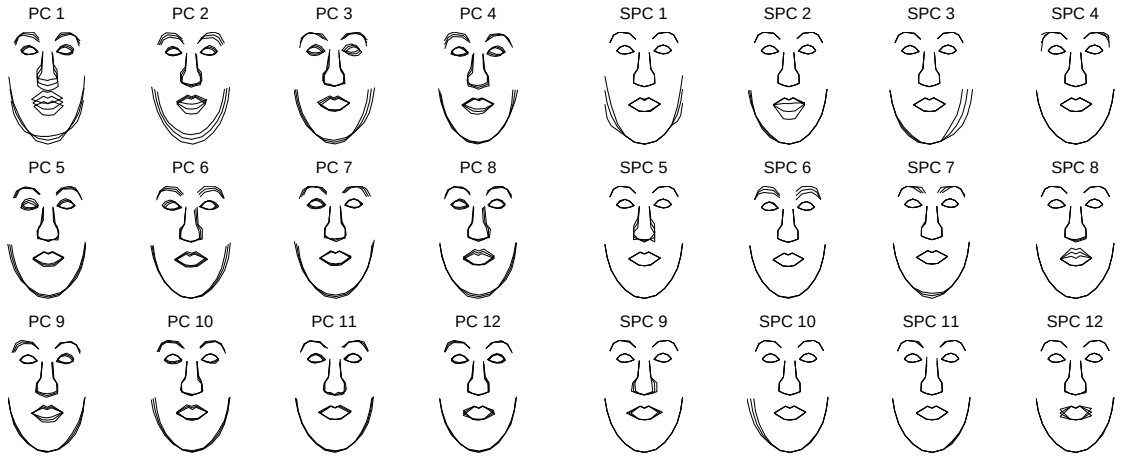
ACKNOWLEDGMENTS

Dr. Bram van Ginneken, Image Sciences Institute, University Medical Center Utrecht, kindly provided the lung annotations used in this study. Dr. Ginneken also assisted with the anatomical interpretation. Charlotte Ryberg and Egill Rostrup, The Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital, Hvidovre, kindly provided the MRIs used to produce the corpus callosum annotations. Hui Zou and Trevor Hastie are acknowledged for making the source code of the SPCA algorithm publicly available. K. Sjöstrand was supported by The Technical University of Denmark, DTU. M. B. Stegmann was supported by The Danish Research Agency, grant no. 2059-03-0032.

REFERENCES

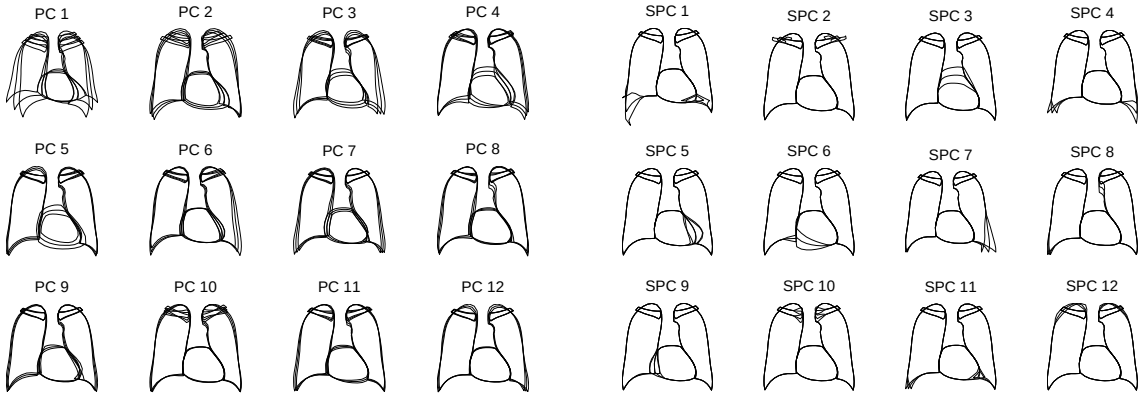
1. J. N. R. Jeffers, "Two case studies in the application of principal component analysis," *Applied Statistics* **16**(3), pp. 225–236, 1967.
2. J. Cadima and I. T. Jolliffe, "Loadings and correlations in the interpretation of principal components," *Journal of Applied Statistics* **22**(2), pp. 203–214, 1995.
3. H. H. Harman, *Modern Factor Analysis*, The University of Chicago Press, 1967.

4. H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis," *Psychometrika* **23**, pp. 187–200, 1958.
5. M. B. Stegmann and K. Sjöstrand, "Sparse modeling of landmark and texture variability using the orthomax criterion," International Symposium on Medical Imaging 2006, San Diego, CA (to appear), feb 2006.
6. C. Chennubhotla and A. Jepson, "Sparse PCA extracting multi-scale structure from data," *Proceedings of the IEEE International Conference on Computer Vision* **1**, pp. 641–647, 2001.
7. S. Vines, "Simple principal components," *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **49**(4), pp. 441–451, 2000.
8. R. Hausman, "Constrained multivariate analysis," in *Optimization in Statistics*, S. Zanakis and J. Rustagi, eds., *Studies in the management sciences*(19), pp. 137–151, North-Holland, Amsterdam, 1982.
9. V. Rousson and T. Gasser, "Simple component analysis," *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **53**(4), pp. 539–555, 2004.
10. A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming,"
11. I. T. Jolliffe, N. T. Trendafilov, and M. Uddin, "A modified principal component technique based on the LASSO," *Journal of Computational and Graphical Statistics* **12**(3), pp. 531–547, 2003.
12. R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society - Series B Methodological* **58**(1), pp. 267–288, 1996.
13. H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," tech. rep., Statistics Department, Stanford University, 2004.
14. B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics* **32**(2), pp. 407–451, 2004.
15. H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), pp. 301–320, 2005.
16. I. T. Jolliffe, "Rotation of principal components: Choice of normalization constraints," *Journal of Applied Statistics* **22**(1), pp. 29–36, 1995.
17. D. Gervini and V. Rousson, "Criteria for evaluating dimension-reducing components for multivariate data," *American Statistician* **58**(1), pp. 72–76, 2004.
18. M. B. Stegmann, K. Skoglund, and C. Ryberg, "Mid-sagittal plane and mid-sagittal surface optimization in brain MRI using a local symmetry measure," in *International Symposium on Medical Imaging 2005, San Diego, CA, Proc. of SPIE vol. 5747*, SPIE, feb 2005.



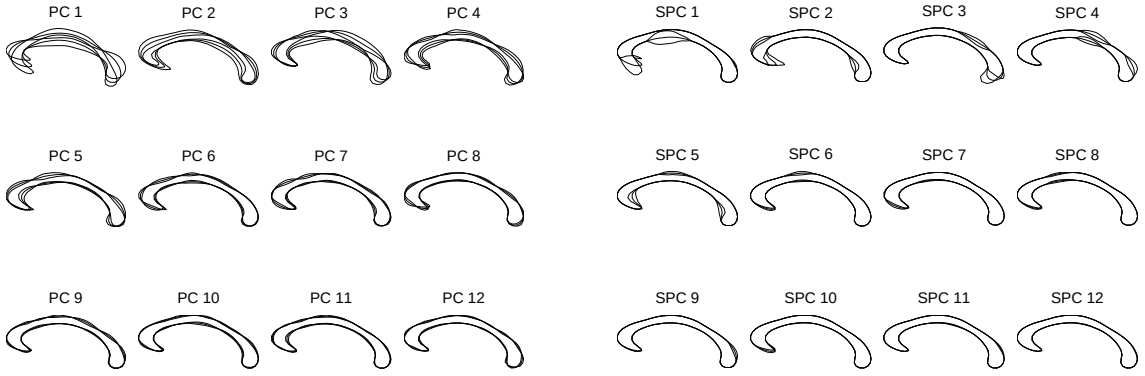
(a) PCA modes (0, ±2.5 std. dev. overlaid) (b) SPCA modes (0, ±2.5 std. dev. overlaid)

Figure 1. PCA (left) versus SPCA (right) shape models of the human face. Each mode describes an identifiable effect, such as smile/frown, nose size and shape, and eyebrow configurations.



(a) PCA modes (0, ±2.5 std. dev. overlaid) (b) SPCA modes (0, ±2.5 std. dev. overlaid)

Figure 2. Lungs, heart and clavicles. Mode 3, 5, 6 and 9 depict the heart geometry while mode 8 describes the position of the aortic arch.



(a) PCA modes (0, ±2.5 std. dev. overlaid) (b) SPCA modes (0, ±2.5 std. dev. overlaid)

Figure 3. PCA (left) and Sparse PCA (right) models of the corpus callosum brain structure.

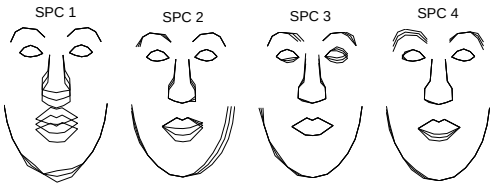


Figure 4. SPCA using simple thresholding. Although the same L1 constraint has been used, these images do not show the same amount of separation as those in Figure 1(b).

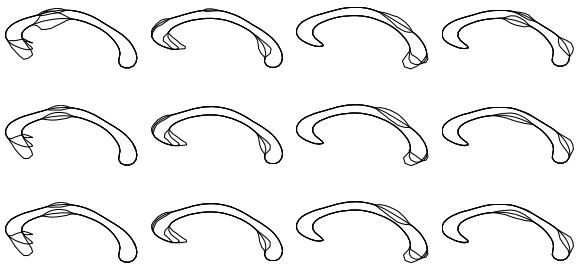
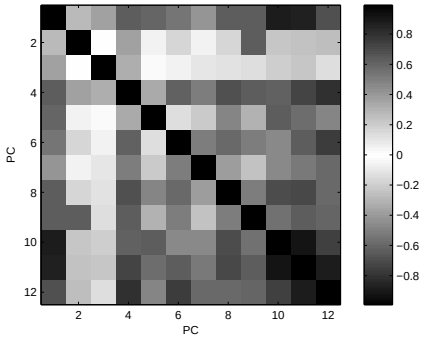
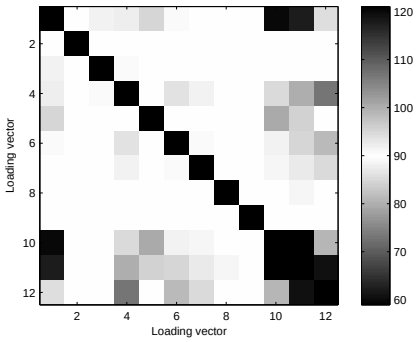


Figure 5. The first four modes of variation for the corpus callosum data set. Rows correspond to λ-values 0.001, 1, and 1000 (top to bottom). Note the insensitivity to values of λ.



(a) Correlations among the sparse principal components.



(b) Angles (degrees) between the sparse loading vectors.

Figure 6. Correlations of PCs and angles between loading vectors for the lungs data set, using the SPCA method. Regular PCA produces an orthonormal loading matrix and uncorrelated principal components. SPCA typically results in significant correlations, while angles are relatively close to 90°.

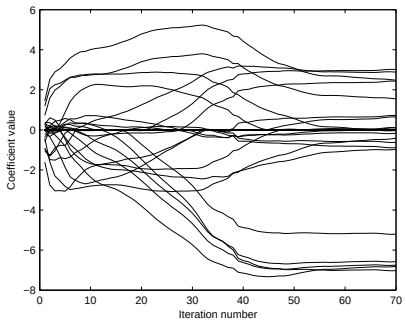


Figure 7. Coefficient values as functions of iteration number for the face data set. Typically, coefficients vary considerably before convergence.