Refining Sparse Principal Components

M. Journée[1], F. Bach[2], P.-A. Absil[3], and R. Sepulchre[1]

[1] Department of Electrical Engineering and Computer Science, University of Liège, Belgium, [M.Journee, R.Sepulchre]@ulg.ac.be
[2] INRIA - Willow project, Dèpartement d'Informatique, Ecole Normale Supérieure, Paris, France, Francis.Bach@mines.org
[3] Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium, absil@inma.ucl.ac.be

Summary. In this paper, we discuss methods to refine locally optimal solutions of sparse PCA. Starting from a local solution obtained by existing algorithms, these methods take advantage of convex relaxations of the sparse PCA problem to propose a refined solution that is still locally optimal but with a higher objective value.

1 Introduction

Principal component analysis (PCA) is a well-established tool for making sense of high dimensional data by reducing it to a smaller dimension. Its extension to *sparse principal component analysis*principal component analysis!sparce, which provides a sparse low-dimensional representation of the data, has attracted alot of interest in recent years (see, e.g., [1, 2, 3, 5, 6, 7, 8, 9]). In many applications, it is in fact worth to sacrifice some of the explained variance to obtain components composed only from a small number of the original variables, and which are therefore more easily interpretable.

Although PCA is, from a computational point of view, equivalent to a singular value decomposition, sparse PCA is a much more difficult problem of NP-hard complexity [8]. Given a data matrix $A \in \mathbf{R}^{m \times n}$ encoding m samples of n variables, most algorithms for sparse PCA compute a unit-norm loading vector $z \in \mathbf{R}^n$ that is only *locally* optimal for an optimization problem aiming at maximizing explained variance penalized for the number of non-zero loadings. This is in particular the case of the SCoTLASS [7], the SPCA [10], the rSVD [9] and the GPower [5] algorithms.

Convex relaxationsconvex relaxation have been proposed in parallel for some of these formulations [2, 1]. To this end, the unit-norm loading vector $z \in \mathbf{R}^n$ is lifted into a symmetric, positive semidefinite, rank-one matrix $Z = zz^T$ with unit trace. The relaxation consists of removing the rank-one constraintrank-one!constraint and accepting any matrix of the *spectahedron*spectahedron

M. Diehl et al. (eds.), *Recent Advances in Optimization and its Applications in Engineering*,
DOI 10.1007/978-3-642-12598-0_14,

$$\mathcal{S} = \{Z \in \mathbf{S}^m \mid Z \succeq 0, \mathrm{Tr}(Z) = 1\},$$

which is a convex set. The solution of these convex problems has usually a rank larger than one. Hence, some post-processing is needed to round this solution to rank-onerank-one!matrices in order to reconstruct a unit-norm vector z.

The aim of this paper is to discuss a way to refine locally optimal solutions of sparse PCA by taking advantage of these convex relaxations. A well-known formulation of sparse PCA is first reviewed and relaxed into a convex program in Section 2. A method that uses both the initial formulation and the relaxation is then discussed in Section 3 in order to improve the quality of the components. Its efficiency is evaluated in Section 4.

2 Formulation and convex relaxation of sparse PCA

Under the assumption that the columns of the data matrix $A \in \mathbf{R}^{m\times n}$ are centered, PCA consists in computing the dominant eigenvectors of the scaled sample covariance matrix $\Sigma = A^T A$. The problem of computing the first principal component can thus be written in the form

$$\max_{\substack{z\in\mathbf{R}^n \\ z^T z=1}} z^T \Sigma z. \tag{1}$$

Several formulations of sparse PCA can be derived from (1) (see, e.g., [5]). A possible one is provided by the optimization problem

$$z^* = \arg\max_{\substack{z\in\mathbf{R}^n \\ z^T z=1}} z^T \Sigma z - \rho\|z\|_0, \tag{2}$$

with $\rho \geq 0$ and where the ℓ_0 "norm" is the number of nonzero coefficients (or *cardinality*) of z. The formulation (2) is essentially the problem of finding an optimal pattern of zeros and nonzeros for the vector z, which is of combinatorial complexity.

Interestingly, as shown in [2, 5], problem (2) can be equivalently rewritten as the maximization of a convex function on the unit Euclidean sphere,

$$x^* = \arg\max_{\substack{x\in\mathbf{R}^m \\ x^T x=1}} \sum_{i=1}^{n} ((a_i^T x)^2 - \rho)_+, \tag{3}$$

where a_i is the ith column of A and $x_+ = \max(0, x)$. The solution z^* of (2) is reconstructed from the solution x^* of (3) as follows,

$$z^* = \frac{[\mathrm{sign}((A^T x^*) \circ (A^T x^*) - \rho)]_+ \circ A^T x^*}{\|[\mathrm{sign}((A^T x^*) \circ (A^T x^*) - \rho)]_+ \circ A^T x^*\|_2},$$

where $\circ$ denotes the matrix element-wise product. The ith component of z^* is thus active (i.e., not constrained to zero) if the condition $(a_i^T x^*)^2 - \rho \geq 0$ holds.

For the purpose of relaxing (2) into a convex program, the unit-norm vector x is lifted into a matrix $X = xx^T$. The formulation (3) is so rewritten in terms of a matrix variable X as follows,

$$\begin{array}{ll} \max\limits_{X \in \mathbf{S}^m} & \sum_{i=1}^n (a_i^T X a_i - \rho)_+ \\ \text{s.t.} & \mathrm{Tr}(X) = 1, \\ & X \succeq 0, \\ & \mathrm{rank}(X) = 1, \end{array} \tag{4}$$

where $\mathbf{S}^m$ denotes the set of symmetric matrices in $\mathbf{R}^{m \times m}$.The problem (4) is relaxed into a convex program in two steps. First, the nonconvex rank constraint is removed. Then, the convex objective function

$$f_{cvx}(X) = \sum_{i=1}^n (a_i^T X a_i - \rho)_+$$

is replaced by the concave function

$$f_{ccv}(X) = \sum_{i=1}^n \mathrm{Tr}(X^{\frac{1}{2}}(a_i a_i^T - \rho I)X^{\frac{1}{2}})_+,$$

where $\mathrm{Tr}(X)_+$ denotes the sum of the positive eigenvalues of X. Observe that maximizing a concave function over a convex set is indeed a convex program. Since the values $f_{cvx}(X)$ and $f_{ccv}(X)$ are equal for matrices X that are rank one, the convex relaxation of the sparse PCA formulation (2),

$$\begin{array}{ll} \max\limits_{X \in \mathbf{S}^m} & \sum_{i=1}^n \mathrm{Tr}(X^{\frac{1}{2}}(a_i a_i^T - \rho I)X^{\frac{1}{2}})_+ \\ \text{s.t.} & \mathrm{Tr}(X) = 1, \\ & X \succeq 0, \end{array} \tag{5}$$

is tight for solutions of rank one. We refer to [1] for more details on the derivation of (5).

3 A procedure to refine the components

Several methods have been proposed to compute locally optimal solutions of the NP-hard formulation (2) of sparse PCA. For instance, the greedy algorithm of [2] sequentially increments the cardinality of the solution with the component of z that maximizes the objective function in (2). The GPower algorithm of [5] exploits the convexity of the objective function to generalize the well-known power method in the present context.

In parallel, a method for solving the convex relaxation (5) in an efficient manner is discussed in the recent paper [4]. This method parameterizes the positive semidefinite matrix variable X as the product $X = YY^T$ where the

number of independent columns of $Y \in \mathbf{R}^{m \times p}$ fixes the rank of X. The parameter p enables to interpolate between the initial combinatorial problem (i.e., $p = 1$) and the convex relaxation (i.e., $p = n$). In practice, the dimension p is incremented until a sufficient condition is satisfied for Y to provide a solution YY^T of (5). Since this often holds for $p \ll n$, the reduction of per-iteration numerical complexity for solving (5) can be significant: from $\mathcal{O}(n^2)$ for traditional convex optimization tools to $\mathcal{O}(np)$ for the algorithm of [4].

Starting from a locally optimal solution of the sparse PCA formulation (2), the proposed method for improving the quality of this solution works in two steps. First, solve the convex relaxation (5) with the algorithm of [4] that increases the rank of the variable X from one until a sufficiently accurate solution is found. Then, in order to recover a rank-one matrix from this solution of rank $p \geq 1$, solve the optimization problem,

$$\begin{array}{ll} \max\limits_{X \in \mathbf{S}^m} & \mu f_{cvx}(X) + (1-\mu) f_{ccv}(X) \\ \text{s.t.} & \mathrm{Tr}(X) = 1, \\ & X \succeq 0, \end{array} \tag{6}$$

for the parameter μ that is gradually increased from 0 to 1. In the case $\mu = 0$, (6) is the convex relaxation (5). In the other limit case $\mu = 1$, problem (6) amounts to maximize a convex function on a convex set, which has local solutions at all the extreme points of this set. Solving a sequence of problems of the form of (6) for an increasing value of μ from zero to one converges to the extreme points of the spectahedron that are all rank-one matrices. Hence, this process reduces the rank of the solution of the convex relaxation (5) from $p \geq 1$ to one. This rank-one solution is hoped to have a larger objective value than the rank-one matrix chosen to initialize the resolution of (5). The algorithm of [4] can be used to solve (6) for any value of μ.

Figure 1 illustrates the proposed procedure in the case of a random Gaussian matrix $A \in \mathbb{R}^{150 \times 50}$. Because any matrix of the spectahedron has non-negative eigenvalues with the sum being one, the maximum eigenvalue can be used to monitor the rank: a matrix of the spectahedron is rank one if and only if its maximum eigenvalue is one. The homotopy methodHomotopy method (i.e., solving (6) for an increasing value of μ) is compared against the best rank-one least squares approximation of the solution of (5), i.e., the matrix $\tilde{X} = xx^T$ where x is the unit-norm dominant eigenvector of X. Let $f_{EVD}(X)$ denote the function

$$f_{EVD}(X) = f_{ccv}(\tilde{X}) = f_{cvx}(\tilde{X}).$$

The continuous plots of Figure 1 display the evolution of both functions $f_{ccv}(X)$ and $f_{EVD}(X)$ during the resolution of the convex program (5), i.e., $\mu = 0$ in (6). Point A represents a rank-one solution that is locally optimal for the sparse PCA formulation (2) and obtained, for instance, with the GPower algorithm [4]. When solving the convex relaxation (5), the rank of the matrix X is gradually incremented until a solution is identified (point B/B'). The

dashed plots illustrate the resolution of (6) for a parameter μ that is gradually increased from 0 to 1 (by steps of 0.05). For a sufficiently large value of μ, problem (6) has a rank-one solution (point C). The objective value in C is clearly larger than that of the initialization A as well as than that of the best rank-one least-squares approximation B'. This improvement results most probably from the fact that the homotopy method takes the objective function into account whereas the least-squares approximation does not.

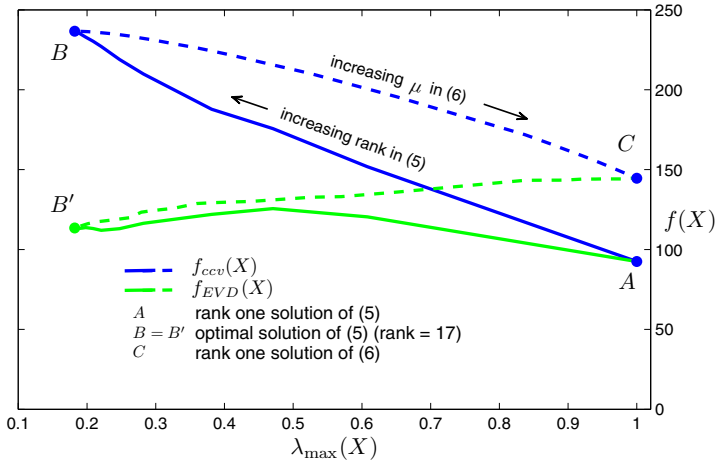

Fig. 1. Evolution of the functions $f_{ccv}(X)$ and $f_{EVD}(X)$ in two situations. Continuous plots: resolution of the convex program (5) ($\mu = 0$ in (6)). Dashed plots: projection of the solution of (5) on a rank-one matrix by gradual increase of μ in (6).

4 Numerical experiments

In Table 1, we compare the objective value obtained by the GPower algorithm which computes a locally optimal solution of the sparse PCA problem (3), the objective value of the best rank-one approximation of the solution of the convex relaxation (5) and finally the objective value of the proposed homotopy method, i.e., we compare the objective values at the points A, B' and C in Figure 1. Each value in Table 1 is an average on 20 instances for each problem dimension. The data is systematically generated according to a Gaussian distribution of zero mean and unit variance. The proposed homotopy method is shown to improve the objective value by several percents. Such an improvement might be significant for applications for which it is crucial to identify

the best solution of sparse PCA. Compressed sensing is such an application [1].

Dimension	f_A	$f_{B'}$	$(f_{B'} - f_A)/f_A$	f_C	$(f_C - f_A)/f_A$
50×25	3.9757	4.0806	+ 2.64 %	4.1216	+ 3.67 %
100×50	3.6065	3.7038	+ 2.70 %	3.8276	+ 6.13 %
200×100	2.9963	2.8711	- 4.18 %	3.1904	+6.48 %
400×200	3.9549	4.1089	+3.89 %	4.2451	+ 7.34 %
800×200	5.6032	5.6131	+0.18 %	5.8754	+ 4.86 %
800×400	3.0541	3.0688	+ 0.48 %	3.4014	+11.37 %

Table 1. Average objective values at the points A, B' and C of Figure (1) for Gaussian data matrices of various size. The GPower algorithm of [5] is used to compute the rank-one solution A.

5 Acknowledgements

Michel Journée is a research fellow of the Belgian National Fund for Scientific Research (FNRS). This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors.

References

1. A. d' Aspremont, F. R. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9:1269–1294, 2008.
2. A. d' Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49:434–448, 2007.
3. J. Cadima and I. T. Jolliffe. Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, 22:203–214, 1995.
4. M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre. Low-rank optimization for semidefinite convex problems. *Submitted to SIAM Journal on Optimization (preprint available on ArXiv)*, 2008.
5. M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *Accepted to Journal of Machine Learning Research (preprint available on ArXiv)*, 2008.
6. I. T. Jolliffe. Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, 22:29–35, 1995.
7. 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):531–547, 2003.

8. B. Moghaddam, Y. Weiss, and S. Avidan. Spectral bounds for sparse PCA: Exact and greedy algorithms. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 915–922. MIT Press, Cambridge, MA, 2006.
9. H. Shen and J. Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034, 2008.
10. H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.