Full Regularization Path for Sparse Principal Component Analysis

Alexandre d'Aspremont ASPREMON@PRINCETON.EDU
ORFE, Princeton University, Princeton NJ 08544, USA.

Francis R. Bach FRANCIS.BACH@MINES.ORG
CMM, Ecole des Mines de Paris, 35 rue Saint-Honoré, 77305 Fontainebleau, France.

Laurent El Ghaoui ELGHAOUI@EECS.BERKELEY.EDU
EECS, U.C. Berkeley, Berkeley CA 94720, USA.

Abstract

Given a sample covariance matrix, we examine the problem of maximizing the variance explained by a particular linear combination of the input variables while constraining the number of nonzero coefficients in this combination. This is known as sparse principal component analysis and has a wide array of applications in machine learning and engineering. We formulate a new semidefinite relaxation to this problem and derive a greedy algorithm that computes a *full set* of good solutions for all numbers of non zero coefficients, with complexity $O(n^3)$, where n is the number of variables. We then use the same relaxation to derive sufficient conditions for global optimality of a solution, which can be tested in $O(n^3)$. We show on toy examples and biological data that our algorithm does provide globally optimal solutions in many cases.

1. Introduction

Principal component analysis (PCA) is a classic tool for data analysis, visualization or compression and has a wide range of applications throughout science and engineering. Starting from a multivariate data set, PCA finds linear combinations of the variables called *principal components*, corresponding to orthogonal directions maximizing variance in the data. Numerically, a full PCA involves a singular value decomposition of the data matrix.

One of the key shortcomings of PCA is that the factors are linear combinations of *all* original variables; that is, most of factor coefficients (or loadings) are nonzero. This means that while PCA facilitates model interpretation by concentrating the information in a few factors, the factors themselves are still constructed using all variables, hence are often hard to interpret.

In many applications, the coordinate axes involved in the factors have a direct physical interpretation. In financial or biological applications, each axis might correspond to a specific asset or gene. In problems such as these, it is natural to seek a trade-off between the two goals of *statistical fidelity* (explaining most of the variance in the data) and *interpretability* (making sure that the factors involve only a few coordinate axes). Solutions that have only a few nonzero coefficients in the principal components are usually easier to interpret. Moreover, in some applications, nonzero coefficients have a direct cost (*e.g.*, transaction costs in finance) hence there may be a direct trade-off between statistical fidelity and practicality. Thus our aim here is to efficiently derive *sparse principal components*, i.e, a set of sparse vectors that explain a maximum amount of variance. Our belief is that in many applications, the decrease in statistical fidelity required to obtain sparse factors is small and relatively benign. In what follows then, we will focus on the problem of finding sparse factors which explain a maximum amount of variance. This can be written:

$$\max_{\|z\|\leq 1} z^T\Sigma z - \rho\,\mathbf{Card}(z) \qquad (1)$$

in the variable $z \in \mathbf{R}^n$, where $\Sigma \in \mathbf{S}_n$ is the (symmetric) sample covariance matrix, ρ is a parameter controlling sparsity, and $\mathbf{Card}(z)$ denotes the cardinal (ℓ_0 norm) of z, the number of non zero coefficients of z.

While PCA is numerically easy (each factor requires

Appearing in *Proceedings of the 24^{th} International Conference on Machine Learning*, Corvallis, OR, 2007. Copyright 2007 by the author(s)/owner(s).

computing a dominant eigenvector, which can be done in $O(n^2)$), sparse PCA is a NP-hard combinatorial problem—Moghaddam et al. (2006a) show that the subset selection problem for ordinary least squares, which is NP-hard (Natarajan, 1995), can be reduced to sparse PCA. Sometimes ad hoc "rotation" techniques are used to post-process the results from PCA and find interpretable directions underlying a particular subspace (see Jolliffe (1995)). Another simple solution is to *threshold* the loadings (Cadima & Jolliffe, 1995). Kolda and O'Leary (2000) use integer loadings. A more systematic approach to the problem arose in recent years, with various researchers proposing nonconvex algorithms (e.g., SCoTLASS by Jolliffe et al. (2003)). The SPCA algorithm, which is based on the representation of PCA as a regression-type optimization problem (Zou et al., 2004), allows the application of the LASSO (Tibshirani, 1996), a penalization technique based on the ℓ_1 norm. The methods above are all either highly suboptimal (thresholding) or nonconvex (SPCA, ...), hence have unreliable performance.

Recently also, d'Aspremont et al. (2004) derived a semidefinite relaxation for problem (1) which had a complexity of $O(n^{4\sqrt{\log n}})$. Finally, Moghaddam et al. (2006b) used greedy search and branch-and-bound methods to solve small instances of problem (1) exactly and get good solutions for larger ones. Each step of the greedy algorithm has complexity $O(n^3)$, leading to a complexity of $O(n^4)$ for the full path.

Our contribution here is twofold. We first formulate a new semidefinite relaxation to problem (1) and use it to derive a greedy algorithm for computing a *full set* of good solutions (one for each sparsity between 1 and n) at a total numerical cost of $O(n^3)$. We then derive *tractable* sufficient conditions for a vector z to be a *global* optimum of (1). This means in practice that, given a vector z with support I, we can test if z is a globally optimal solution to problem (1) by computing a minimum eigenvalue problem of size $(2m-1)$ where m is the cardinality of z. In particular, we can take any sparsity pattern candidate from any algorithm and test its optimality. Whenever our sufficient condition is fulfilled (which happens somewhat frequently in practice, as shown in Section 7), we have a globally optimal solution to the NP-hard problem in (1).

In Sections 2 to 4, we formulate a convex relaxation for the sparse PCA problem and use it in Section 5 to write an efficient algorithm for computing a full set of good solutions to problem (1). In Section 6, we derive tractable, sufficient conditions for global optimality of these solutions. Finally, in Section 7, we test the numerical performance of these results.

Notation For a vector $z \in \mathbf{R}$, we let $\|z\|_1 = \sum_{i=1}^n |z_i|$ and $\|z\| = \left(\sum_{i=1}^n z_i^2\right)^{1/2}$, $\mathbf{Card}(z)$ is the cardinality of z, i.e. the number of nonzero coefficients of z, while the support I of z is the set $\{i : z_i \neq 0\}$ and I^c its complement. For $\beta \in \mathbf{R}$, we write $\beta_+ = \max\{\beta, 0\}$ and for $X \in \mathbf{S}_n$ (the set of symmetric matrix of size $n \times n$) with eigenvalues λ_i, $\mathbf{Tr}(X)_+ = \sum_{i=1}^n \max\{\lambda_i, 0\}$. The vector of all ones in denoted $\mathbf{1}$.

2. Sparse PCA

Let $\Sigma \in \mathbf{S}_n$ be a symmetric matrix. We consider the following sparse PCA problem:

$$\phi(\rho) = \max_{\|z\| \leq 1} z^T \Sigma z - \rho\,\mathbf{Card}(z) \qquad (2)$$

in the variable $z \in \mathbf{R}^n$ where $\rho > 0$ is a parameter controlling sparsity. We assume without loss of generality that $\Sigma \in \mathbf{S}_n$ is positive semidefinite and that the n variables are ordered by decreasing marginal variances, i.e., that $\Sigma_{11} \geq \cdots \geq \Sigma_{nn}$. We also assume that we are given a square root A of the matrix Σ such that $\Sigma = A^T A$, where $A \in \mathbf{R}^{n \times n}$ and we denote by $a_1, \ldots, a_n$ the columns of A. Note that the problem and our algorithms are invariant by permutations of Σ and by the choice of square root A.

Let us first suppose that $\rho \geq \Sigma_{11}$. Since $z^T \Sigma z \leq \Sigma_{11} \|z\|_1^2$ and $\|z\|_1^2 \leq \|z\|_2^2\,\mathbf{Card}(z)$, we always have:

$$\begin{aligned} \phi(\rho) &= \max_{\|z\| \leq 1} z^T \Sigma z - \rho\,\mathbf{Card}(z) \\ &\leq (\Sigma_{11} - \rho)\,\mathbf{Card}(z) \leq 0, \end{aligned}$$

hence the optimal solution to (2) when $\rho \geq \Sigma_{11}$ is $z = 0$.

From now on, we assume $\rho \leq \Sigma_{11}$ in which case the inequality $\|z\| \leq 1$ is tight. We can represent the sparsity pattern of a vector z, by a vector $u \in \{0,1\}^n$ and rewrite (2) in the equivalent form:

$$\begin{aligned} \phi(\rho) &= \max_{u \in \{0,1\}^n} \lambda_{\max}(\mathbf{diag}(u) \Sigma\, \mathbf{diag}(u)) - \rho \mathbf{1}^T u \\ &= \max_{u \in \{0,1\}^n} \lambda_{\max}(\mathbf{diag}(u) A^T A\, \mathbf{diag}(u)) - \rho \mathbf{1}^T u \\ &= \max_{u \in \{0,1\}^n} \lambda_{\max}(A\, \mathbf{diag}(u) A^T) - \rho \mathbf{1}^T u, \end{aligned}$$

using the fact that $\mathbf{diag}(u)^2 = \mathbf{diag}(u)$ for all variables $u \in \{0,1\}^n$. We then have:

$$\begin{aligned} \phi(\rho) &= \max_{u \in \{0,1\}^n} \lambda_{\max}(A\, \mathbf{diag}(u) A^T) - \rho \mathbf{1}^T u \\ &= \max_{\|x\|=1} \max_{u \in \{0,1\}^n} x^T A\, \mathbf{diag}(u) A^T x - \rho \mathbf{1}^T u \\ &= \max_{\|x\|=1} \max_{u \in \{0,1\}^n} \sum_{i=1}^n u_i ((a_i^T x)^2 - \rho), \end{aligned}$$

hence we finally get, after maximizing in u:

$$\phi(\rho) = \max_{\|x\|=1} \sum_{i=1}^{n} ((a_i^T x)^2 - \rho)_+ \tag{3}$$

which is a nonconvex problem in the variable $x \in \mathbf{R}^n$. We observe that if $\Sigma_{ii} = a_i^T a_i < \rho$, we must have $(a_i^T x)^2 \le \|a_i\|^2 \|x\|^2 < \rho$ hence variable i will never be part of the optimal subset and we can remove it.

3. Semidefinite Relaxation

In problem (3), the variable x appears solely through $X = xx^T$, and in this context, it is classical to reformulate the problem using X only. A set of necessary and sufficient conditions for the existence of x of unit norm so that $X = xx^T$ is $\mathbf{Tr}(X) = 1$, $X \succeq 0$ and $\mathbf{Rank}(X) = 1$. We can thus rewrite (3) as:

$$\begin{array}{rll} \phi(\rho) = & \max. & \sum_{i=1}^n (a_i^T X a_i - \rho)_+ \\ & \text{s.t.} & \mathbf{Tr}(X) = 1,\ \mathbf{Rank}(X) = 1 \\ & & X \succeq 0. \end{array}$$

Unfortunately, the function we are *maximizing* $X \mapsto (a_i^T X a_i - \rho)_+$ is still convex in X and not concave. However, we can now show that on the set of positive semidefinite rank one matrices of unit trace, it is equal to a concave function of X.

We let $X^{1/2}$ denote the symmetric positive square root (with nonnegative eigenvalues) of a symmetric positive semi-definite matrix X. In particular, if $X = xx^T$ with $\|x\| = 1$, then $X^{1/2} = X = xx^T$, then for all $\beta \in \mathbf{R}$, βxx^T has one eigenvalue equal to β and $n-1$ equal to 0, which implies $\mathbf{Tr}(\beta xx^T)_+ = \beta_+$. We thus get:

$$\begin{array}{rcl} (a_i^T X a_i - \rho)_+ & = & \mathbf{Tr}[(a_i^T xx^T a_i - \rho) xx^T]_+ \\ & = & \mathbf{Tr}(x(x^T a_i a_i^T x - \rho) x^T)_+ \\ & = & \mathbf{Tr}(X^{1/2} a_i a_i^T X^{1/2} - \rho X)_+ \end{array}$$

For any symmetric matrix B, the function $X \mapsto \mathbf{Tr}(X^{1/2} B X^{1/2})_+$ is concave on the set of symmetric positive semidefinite matrices, because:

$$\begin{array}{rcl} \mathbf{Tr}(X^{1/2} B X^{1/2})_+ & = & \max\limits_{\{0 \preceq P \preceq X\}} \mathbf{Tr}(PB) \\ & = & \min\limits_{\{Y \succeq B,\ Y \succeq 0\}} \mathbf{Tr}(YX), \end{array}$$

where this last expression is a concave function of X as a pointwise minimum of affine functions of X. We can now relax problem (3) into a convex optimization problem by simply dropping the rank constraint, to get:

$$\begin{array}{rll} \psi(\rho) = & \max. & \sum_{i=1}^n \mathbf{Tr}(X^{1/2} a_i a_i^T X^{1/2} - \rho X)_+ \\ & \text{s.t.} & \mathbf{Tr}(X) = 1,\ X \succeq 0, \end{array} \tag{4}$$

which is a (convex) program in $X \in \mathbf{S}_n$. In fact, using the above representation of $\mathbf{Tr}(X^{1/2} B X^{1/2})_+$, problem (4) can be written as a semidefinite program:

$$\begin{array}{rll} \psi(\rho) = & \max. & \sum_{i=1}^n \mathbf{Tr}(P_i B_i) \\ & \text{s.t.} & \mathbf{Tr}(X) = 1,\ X \succeq 0,\ X \succeq P_i \succeq 0, \end{array}$$

in the variables $X \in \mathbf{S}_n,\ P_i \in \mathbf{S}_n$. Note that we always have $\psi(\rho) \ge \phi(\rho)$ and when the solution to the above semidefinite program has rank one, $\psi(\rho) = \phi(\rho)$ and the relaxation (4) is tight.

4. Low Rank Optimization

The semidefinite relaxation in (4) can be solved efficiently using SDP solvers such as SDPT3 by Toh et al. (1999) when n is very small: indeed, the complexity of solving (4) without exploiting structure is $O(n^9)$. For larger problems, even storing X and P_i quickly becomes impossible and we need to find a more economical representation. Here, we assume that the matrix X has a low rank representation $X = UU^T$ with $U \in \mathbf{R}^{n \times m}$. Then $X^{1/2} = U(U^T U)^{-1/2} U^T$ and we have for all symmetric matrices B, $X^{1/2} B X^{1/2} = U(U^T U)^{-1/2} U^T B U U (U^T U)^{-1/2}$. The matrix $U(U^T U)^{-1/2}$ has orthogonal columns, thus the positive eigenvalues of $X^{1/2} B X^{1/2}$ are exactly the positive eigenvalues of $U^T B U$. In our situation, the matrix B has at most one positive eigenvalues, and thus we can rewrite problem (4) as:

$$\begin{array}{rll} \psi_m(\rho) = & \max. & \sum_{i=1}^n \lambda_{max}(U^T(a_i a_i^T - \rho \mathbf{I})U)_+ \\ & \text{s.t.} & \mathbf{Tr}(U^T U) = 1, \end{array} \tag{5}$$

in the variable $U \in \mathbf{R}^{n \times m}$. Although we have turned a convex relaxation in X into a nonconvex relaxation in U, Burer and Monteiro (2003) show that if m is strictly larger than the rank of the optimal solution to (4), then problem (5) has a unique, rank-deficient, local minimum $U^\star$ and the corresponding matrix $X = U^\star U^{\star T}$ is a global minimum of the original problem (4). Using this result, in the rest of this paper, we will always optimize problems with very low rank and embed the solution in a higher rank problem (or SDP) to test its optimality and the relaxation's tightness.

5. Greedy Solutions

In this section, we focus on finding a good solution to problem (2) using greedy methods. We first present very simple solutions with complexity $O(n \log n)$ and $O(n^2)$. We then recall a simple greedy algorithm with complexity $O(n^4)$, and show how an approximate greedy algorithm can be used to compute a full set of (approximate) solutions for problem (2), with total complexity $O(n^3)$.

5.1. Sorting and Thresholding

The simplest ranking algorithm is to sort the diagonal of the matrix Σ and rank the variables accordingly. This works intuitively because the diagonal is a rough proxy for the eigenvalues: the Schur-Horn theorem states that the diagonal of a matrix majorizes its eigenvalues. Sorting the diagonal costs $O(n \log n)$. Another quick solution is to compute the dominant eigenvector of Σ and select a sparse vector by thresholding to zero the coefficients whose magnitude is smaller than a certain level. This can be done with cost $O(n^2)$.

5.2. Full greedy solution

Following Moghaddam et al. (2006b), starting from an initial solution of cardinality one at $\rho = \Sigma_{11}$, we can update an increasing sequence of index sets $I_k \subseteq [1, n]$, scanning all the remaining variables to find the index with maximum variance contribution. The algorithm works as follows.

- **Input**: $\Sigma \in \mathbf{R}^{n \times n}$
- **Algorithm**:
  1. Preprocessing. Sort variables by decreasing diagonal elements and permute elements of Σ accordingly. Compute the Cholesky decomposition $\Sigma = A^T A$.
  2. Initialization: $I_1 = \{1\}$, $x_1 = a_1/\|a_1\|$.
  3. Compute
  $$i_k = \underset{i \notin I_k}{\operatorname{argmax}}\, \lambda_{max} \left(\sum_{i \in I_k \cup \{i\}} a_i a_i^T \right)$$
  4. Set $I_{k+1} = I_k \cup \{i_k\}$ and compute x_{k+1} as the dominant eigenvector of $\sum_{i \in I_{k+1}} a_i a_i^T$.
  5. Set $k = k + 1$. If $k < n$ go back to step 3.
- **Output**: sparsity patterns I_k.

By convexity of λ_{max} we have:

$$\lambda_{max} \left(\sum_{i \in I_k \cup \{i\}} a_i a_i^T \right) \geq \lambda_{max} \left(\sum_{i \in I_k} a_i a_i^T \right) + (x_k^T a_i)^2 \tag{6}$$

which means that the variance is increasing with k. At every step, I_k represents the set of nonzero elements (or sparsity pattern) of the current point and we can define z_k as the solution to (2) given I_k, which is:

$$z_k = \underset{\{z_{I_k^c}=0,\ \|z\|=1\}}{\operatorname{argmax}} \quad z^T \Sigma z - \rho k,$$

which means that z_k is formed by padding zeros to the dominant eigenvector of the submatrix Σ_{I_k, I_k}.

5.3. Approximate greedy solution

Computing $n-k$ eigenvalues at each iteration is costly. We can use (6) as a lower bound on those eigenvalues which does not require finding $n - k$ eigenvalues at each iteration, to derive the following algorithm:

- **Input**: $\Sigma \in \mathbf{R}^{n \times n}$
- **Algorithm**:
  1. Preprocessing. Sort variables by decreasing diagonal elements and permute elements of Σ accordingly. Compute the Cholesky decomposition $\Sigma = A^T A$.
  2. Initialization: $I_1 = \{1\}$, $x_1 = a_1/\|a_1\|$.
  3. Compute $i_k = \operatorname{argmax}_{i \notin I_k} (x_k^T a_i)^2$
  4. Set $I_{k+1} = I_k \cup \{i_k\}$ and compute x_{k+1} as the dominant eigenvector of $\sum_{i \in I_{k+1}} a_i a_i^T$.
  5. Set $k = k + 1$. If $k < n$ go back to step 3.
- **Output**: sparsity patterns I_k.

5.4. Computational Complexity

The complexity of computing a greedy regularization path using the classic greedy algorithm in Section 5.2 is $O(n^4)$ (at each step k, it computes $(n-k)$ maximum eigenvalue of matrices with size k). The approximate algorithm in Section 5.3 computes a full path in $O(n^3)$: the first Cholesky decomposition is $O(n^3)$, while the complexity of the k-th iteration is $O(k^2)$ for the maximum eigenvalue problem and $O(n^2)$ for computing all products $(x^T a_j)$.

Also, when the matrix Σ is directly formed as a Gram matrix $A^T A$ with $A \in \mathbf{R}^{p \times n}$ where $p < n$, we can use A as the square root of Σ and the complexity of getting the path up to p non zero elements is then $O(p^3 + p^2 n)$ ($O(p^3)$ for the eigenvalue problems and $O(p^2 n)$ for computing the vector products).

5.5. Interpretation as a Regularization Path

Here, we fix $m = 1$, so that $X = UU^T = xx^T$ has rank one and problem (5) is equivalent to (3):

$$\phi(\rho) = \max_{\|x\|=1} \sum_{i=1}^{n} ((a_i^T x)^2 - \rho)_+ \tag{7}$$

Given $\rho > 0$ and x such that $(a_i^T x)^2 \neq \rho$ for all i, then in a neighborhood of x, the function $\sum_{i=1}^n ((a_i^T x)^2 - \rho)_+$ is equal to $x^T \left(\sum_{i \in I} (a_i a_i^T - \rho \mathbf{I}) \right) x$, where I is the set of indices such that $(a_i^T x)^2 > \rho$. Thus x is a locally optimal solution only if x is the dominant eigenvector of the matrix $\sum_{i \in I} a_i a_i^T$.

Such an eigenvector fulfills the condition $(a_i^T x)^2 > \rho$ for $i \in I$ and $(a_i^T x)^2 < \rho$ for $i \notin I$, if and only if

$$\max_{i \notin I} (a_i^T x)^2 < \rho < \min_{i \in I} (a_i^T x)^2.$$

In Section 5.3, given the sparsity pattern I_k at the k-th iteration, the next index is chosen so that $i_k = \operatorname{argmax}_{i \notin I_k} (x_k^T a_i)^2$, this is equivalent to finding the largest ρ for which the sparsity pattern is still optimal, and hence the algorithm in Section 5.3 implicitly builds a path of locally optimal solutions to problem (7). In Section 6.1, we derive sufficient conditions for the global optimality of such solutions.

6. Tightness

In this section, we derive necessary and sufficient conditions to test the optimality of the solutions to the relaxations detailed in Sections 4 and 5 and the tightness of the relaxation in (4).

6.1. Semidefinite Optimality Conditions

Here, we first derive the Karush-Kuhn-Tucker (KKT) conditions for problem (4) and study the particular case where we are given a rank one solution $X = xx^T$ and need to test if for optimality. Let us start by deriving the dual to problem (4) which is equivalent to:

$$\begin{array}{ll} \text{max.} & \sum_{i=1}^n \mathbf{Tr}(P_i B_i) \\ \text{s.t.} & 0 \preceq P_i \preceq X \\ & \mathbf{Tr}(X) = 1, \; X \succeq 0. \end{array}$$

This is a semidefinite program in the variables $X \in \mathbf{S}_n$, $P_i \in \mathbf{S}_n$, with $B_i = a_i a_i^T - \rho \mathbf{I}$, $i = 1, \ldots, n$. Its dual can be written as:

$$\begin{array}{ll} \text{min.} & \lambda_{\max}\left(\sum_{i=1}^n Y_i\right) \\ \text{s.t.} & Y_i \succeq B_i, \; Y_i \succeq 0, \quad i = 1, \ldots, n. \end{array}$$

in the variables $Y_i \in \mathbf{S}_n$. The KKT conditions for this pair of SDP problems are written:

$$\left\{ \begin{array}{l} \left(\sum_{i=1}^n Y_i\right) X = \lambda_{\max}\left(\sum_{i=1}^n Y_i\right) X \\ X - P_i - Q_i = 0, \; P_i B_i = P_i Y_i, \; Y_i Q_i = 0 \\ Y_i \succeq B_i, \; Y_i, X, P_i, Q_i \succeq 0. \end{array} \right. \tag{8}$$

Suppose now that we are given a sparsity pattern I (obtained using the results detailed in the previous section for example), and that X is a *rank one* matrix with $X = xx^T$. In this case, the constraint $X - P_i = Q_i \succeq 0$ implies that $P_i = \alpha_i xx^T$ with $\alpha_i \in [0, 1]$ and we can further simplify the KKT conditions to get:

$$\left\{ \begin{array}{l} \left(\sum_{i=1}^n Y_i\right) x = \lambda_{\max}\left(\sum_{i=1}^n Y_i\right) x \\ x^T Y_i x = ((a_i^T x)^2 - \rho)_+ \\ Y_i \succeq B_i, \; Y_i \succeq 0. \end{array} \right. \tag{9}$$

In what follows, we show that when (I, ρ) satisfy a basic consistency condition, we can find an explicit expression for the dual variables Y_i corresponding to a rank one solution $X = xx^T$.

6.2. Optimality of Low Rank Solutions

In this section we derive sufficient and necessary optimality conditions for rank one solutions to the convex problem (4). If we can prove that a rank one solution is globally optimal, we also prove that relaxation (4) is *tight*, hence we get a globally optimal solution to the original problem (2).

Again, for a given sparsity pattern I, we let x be the largest eigenvector of $\sum_{i \in I} a_i a_i^T$. We must first identify a set of parameters ρ consistent with the pattern I and the vector x, i.e., such that x is locally optimal for the rank one problem, which means:

$$\max_{i \notin I} (a_i^T x)^2 \leq \rho \leq \min_{i \in I} (a_i^T x)^2. \tag{10}$$

If $\max_{i \notin I} (a_i^T x)^2 > \min_{i \in I} (a_i^T x)^2$, no ρ can be found to match the pattern I. In what follows, we assume that the set of ρ satisfying condition (10) has nonempty interior and derive additional necessary optimality conditions for such pairs (I, ρ) to be globally optimal.

Let us recall that, as in Section 4, the matrix xx^T is optimal for problem (4) if and only if the rank two problem (with $m = 2$) has a local minimum at $U = [x, 0]$. When ρ satisfies condition (10) strictly, in a small neighborhood of $U = [x, 0]$, the function $\psi_m(\rho)$ in (5) can be written as an unconstrained maximization as $\psi_m(\rho) = \max_{U \in \mathbf{R}^{n \times m}} f(U)$ with:

$$f(U) = \sum_{i \in I} \lambda_{max}\left(\frac{U^T B_i U}{\mathbf{Tr}\, U^T U}\right) + \sum_{i \in I^c} \lambda_{max}\left(\frac{U^T B_i U}{\mathbf{Tr}\, U^T U}\right)_+$$

where $B_i = a_i a_i^T - \rho \mathbf{I}$. Since f is twice differentiable at $U = [x, 0]$, a necessary condition for local optimality is that its gradient be zero and its Hessian be negative semidefinite. The gradient of $f(U)$ at $[x, 0]$ is equal to $[2Bx - 2(x^T Bx)x, 0]$ where $B = \sum_{i \in I} B_i$. Since x is an eigenvector of B corresponding to its largest eigenvalue $\sigma = \sum_{i \in I}((a_i^T x)^2 - \rho)$, this gradient is always zero at $[x, 0]$. Furthermore, a second order expansion of f at the point $[x, 0]$ can be written:

$$\begin{array}{l} f(x + dx, dy) - f(x, 0) = 2dx^T (B - \sigma \mathbf{I}) dx \\ + dy^T \left(\sum_{i \in I} \frac{B_i x x^T B_i}{x^T B_i x} - \sigma \mathbf{I} \right) dy + \sum_{i \in I^c} (dy^T B_i dy)_+ \end{array}$$

and a necessary condition for optimality is then:

$$\min_{\{Z_i \succeq B_i, Z_i \succeq 0\}} \lambda_{\max}\left(\sum_{i \in I} \frac{B_i x x^T B_i}{x^T B_i x} + \sum_{i \in I^c} Y_i \right) \leq \sigma \tag{11}$$

This necessary condition is also sufficient, because the criterion is equivalent to satisfying the KKT conditions in (9), having set the dual variables:

$$Y_i = \frac{B_i x x^T B_i}{x^T B_i x}, \quad i \in I$$

and primal variable $X = xx^T$. Indeed, we have $Y_i x = B_i x$ and

$$\begin{aligned} a_i^T(Y_i - B_i)a_i &= \rho \tfrac{a_i^T a_i - \rho}{(a_i^T x)^2 - \rho}(a_i^T a_i - (a_i^T x)^2) \\ &\geq 0. \end{aligned}$$

Using a Schur complement, when $Y_i = 0$ for $i \in I^c$, condition (11) can be written:

$$\left(\begin{array}{cc} \mathbf{diag}(x^T B_i x) & (B_i x)_{i \in I}^T \\ (B_i x)_{i \in I} & \sigma \mathbf{I} \end{array} \right) \succeq 0$$

where $\mathbf{diag}(x^T B_i x)$ is the matrix with diagonal elements $x^T B_i x$, when $i \in I$. This is a linear matrix inequality in the single variable ρ:

$$\left(\begin{array}{cc} \mathbf{diag}((x^T a_i)^2 - \rho) & (a_i a_i^T x - \rho x)_{i \in I}^T \\ (a_i a_i^T x - \rho x)_{i \in I} & \sum_i ((x^T a_i)^2 - \rho)\mathbf{I} \end{array} \right) \succeq 0$$

which means that the set of admissible ρ is a convex set and hence an interval.

6.3. Efficient Optimality Conditions

We let $m = \mathbf{Card}(I)$. When $m \ll n$, we can reduce the complexity of finding this interval by projecting the condition on the space spanned by the columns of A indexed by I. We denote by $\tilde{A}$ the submatrix of A obtained by keeping only columns with index in I. We let R be the Cholesky factor of $\tilde{A}^T \tilde{A}$, so $\tilde{A}^T \tilde{A} = RR^T$. With x defined as the dominant eigenvector of $\sum_{i=1}^n (a_i a_i^T - \rho \mathbf{I})$, we have $x = \tilde{A}v$, with $v^T \tilde{A}^T \tilde{A} v = 1$ and the corresponding eigenvalue $\sigma = v^T \tilde{A}^T \tilde{A} \tilde{A}^T \tilde{A} v$. We let D be the diagonal matrix with diagonal elements $v^T \tilde{A}^T \tilde{A} e_i = a_i^T x$. By premultiplying and postmultiplying by $\tilde{A}$, (11) leads to:

$$\tilde{A}^T \tilde{A}(D - \rho v \mathbf{1}^T)(D^2 - \rho \mathbf{I})^{-1}(D - \rho \mathbf{1} v^T)\tilde{A}^T \tilde{A} - \sigma \tilde{A}^T \tilde{A} \preceq 0$$

which, by Schur's complement lemma, becomes:

$$\left(\begin{array}{cc} D^2 & DR^T \\ RD & \sigma \mathbf{I} \end{array} \right) \succeq \rho \left(\begin{array}{cc} \mathbf{I} & \mathbf{1} v^T R^T \\ Rv\mathbf{1}^T & m\mathbf{I} \end{array} \right)$$

By block-diagonalizing the matrix on the right-hand side, using the notation $\Pi = \mathbf{I} - \mathbf{1}\mathbf{1}^T/m$, we get:

$$\left(\begin{array}{cc} \Pi D^2 \Pi & Z^T \\ Z & \sigma \mathbf{I} \end{array} \right) \succeq \rho \left(\begin{array}{cc} \Pi & 0 \\ 0 & m\mathbf{I} \end{array} \right)$$

with $Z = RD - \frac{\sigma}{m} Rv\mathbf{1}^T$ which satisfies $Z\mathbf{1} = 0$. We let $P \in \mathbf{R}^{m \times m-1}$ be an orthonormal basis of vectors orthogonal to the constant vector $\mathbf{1}$, so that $PP^T + \mathbf{1}\mathbf{1}^T/m = \mathbf{I}$. Condition (11) finally implies:

$$\left(\begin{array}{cc} P^T D^2 P & P^T Z^T/\sqrt{m} \\ ZP/\sqrt{m} & \sigma/m\mathbf{I} \end{array} \right) \succeq \rho \mathbf{I}$$

which leads to:

$$\rho \leq \lambda_{min} \left(\begin{array}{cc} P^T D^2 P & P^T Z^T/\sqrt{m} \\ ZP/\sqrt{m} & \sigma/m\mathbf{I} \end{array} \right)$$

which is a minimum eigenvalue problem of dimension $(2m-1)$. Good candidates for Y_i when $i \in I^c$ can then be found by solving:

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}\, Y_i \\ \text{subject to} & Y_i \succeq B_i, \; x^T Y_i x = 0, \; Y_i \succeq 0, \end{array}$$

which has an explicit solution given by:

$$Y_i = \rho \frac{(a_i^T a_i - \rho)}{(\rho - (a_i^T x)^2)} \frac{(\mathbf{I} - xx^T)a_i a_i^T(\mathbf{I} - xx^T)}{\|(\mathbf{I} - xx^T)a_i\|^2}. \tag{12}$$

The total cost of testing the condition is dominated by the Cholesky and QR decompositions and is then $O(m^3)$, but when done sequentially in the greedy path algorithms of Section 5, efficient Cholesky updates may be used to obtain them in $O(m^2)$ for each m, leading to total complexity of $O(n^3)$ for checking optimality for every point on a path. We can summarize the results of this section as follows.

Proposition 1 *Given a sparsity pattern I, setting x to be the largest eigenvector of $\sum_{i \in I} a_i a_i^T$. If there is a parameter ρ_I such that:*

$$\max_{i \notin I}(a_i^T x)^2 \leq \rho_I \leq \min_{i \in I}(a_i^T x)^2. \tag{13}$$

and

$$\left(\begin{array}{cc} \mathbf{diag}((x^T a_i)^2) - \rho_I \mathbf{I} & (a_i a_i^T x)_{i \in I}^T - \rho_I \mathbf{1} x^T \\ (a_i a_i^T x)_{i \in I} - \rho_I x \mathbf{1}^T & \sum_i (x^T a_i)^2 \mathbf{I} - \rho_I \mathbf{I} \end{array} \right) \succeq 0 \tag{14}$$

with

$$\lambda_{\max}\left(\sum_{i \in I} \frac{B_i x x^T B_i}{x^T B_i x} + \sum_{i \in I^c} Y_i \right) \leq \sigma$$

with Y_i given in (12), then the vector z such that:

$$z = \underset{\{z_{I^c} = 0, \; \|z\| = 1\}}{\operatorname{argmax}} \; z^T \Sigma z,$$

which is formed by padding zeros to the dominant eigenvector of the submatrix $\Sigma_{I,I}$ is a global solution to problem (2) for $\rho = \rho_I$. When $\mathbf{Card}(I) << n$, we can replace condition (14) by a minimum eigenvalue problem of size $(2m-1)$.

Note that (13) corresponds to local optimality conditions for the rank-one problem while (14) corresponds to global optimality of such locally optimal solutions.

6.4. Solution Improvements & Randomization

When these conditions are not satisfied, we can use the low-rank algorithms of Section 4. Also, El Ghaoui (2006) shows that we can get bounds on the suboptimality of the solution by randomization.

7. Numerical Results

In this section, we first compare the various methods detailed here on artificial examples, then test their performance on a biological data set.

7.1. Artificial Data

We generate a matrix U of size 150 with uniformly distributed coefficients in $[0, 1]$. We let $v \in \mathbf{R}^{150}$ be a sparse vector with:

$$v_i = \begin{cases} 1 & \text{if } i \leq 50 \\ 1/(i-50) & \text{if } 50 < i \leq 100 \\ 0 & \text{otherwise} \end{cases}$$

We form a test matrix $\Sigma = U^T U + \sigma v v^T$, where σ is the signal-to-noise ratio. We first compare the relative performance of the algorithms in Section 5 at identifying the correct sparsity pattern in v given the matrix Σ. The resulting ROC curves are plotted in figure 1 for $\sigma = 2$. On this example, the computing time for the approximate greedy algorithm in Section 5.3 was 3 seconds versus 37 seconds for the full greedy solution in Section 5.2. Both algorithms produce the same answer. We can also see that both sorting and thresholding ROC curves are dominated by the greedy algorithms. We then plot the variance versus cardinality tradeoff curves for various values of the signal-to-noise ratio. In figure 2, dashed lines correspond to *potentially* suboptimal points while bold lines correspond to *provably* optimal ones. We notice that the proportion of optimal points increases with the signal-to-noise ratio. Next, we use the DSPCA algorithm of d'Aspremont et al. (2004) to find optimal solutions where the greedy codes have failed to obtain a provably globally optimal solutions. In figure 3 we plot the variance versus cardinality tradeoff curve for $\sigma = 10$. The dashed line corresponds to suboptimal points while bold points correspond to optimal ones, both computed using the approximate greedy algorithm in Section 5.3. The circles correspond to additional globally optimal points obtained using DSPCA.

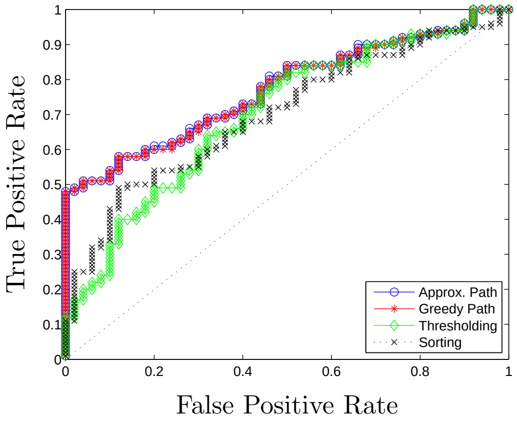

Figure 1. ROC curves for sorting, thresholding, fully greedy solutions (Section 5.2) and approximate greedy solutions (Section 5.3) for $\sigma = 2$.

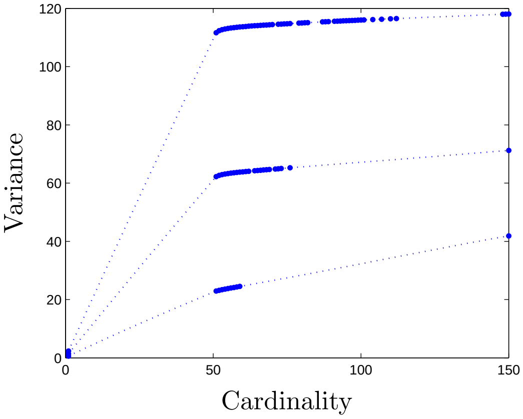

Figure 2. Variance versus cardinality tradeoff curves for $\sigma = 10$ (bottom), $\sigma = 50$ and $\sigma = 100$ (top). Optimal points are in bold.

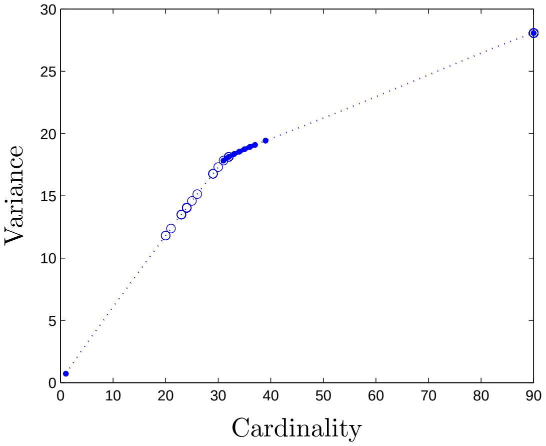

Figure 3. Variance versus cardinality tradeoff curve for $\sigma = 10$. Greedy optimal points are in bold, while additional DSPCA optimal points are plotted as circles.

7.2. Biological Data

We run the algorithm of Section 5.3 on two gene expression data sets, one on Colon cancer from Alon et al. (1999), the other on Lymphoma from Alizadeh et al. (2000). We plot the variance versus cardinality tradeoff curve in figure 4. Again, the dashed line corresponds to suboptimal points while bold lines correspond to optimal ones. In both cases, we consider the 500 genes with largest variance. In Table 1, we also compare the 20 most important genes selected by the second sparse PCA factor on the colon cancer data set, with the top 10 genes selected by the RankGene software by (Su et al., 2003). We observe that 6 genes (out of an original 4027 genes) were both in the top 20 sparse PCA genes and in the top 10 Rankgene genes.

Rank	Rankgene	GAN	Description
3	8.6	J02854	Myosin regul.
6	18.9	T92451	Tropomyosin
7	31.5	T60155	Actin
8	25.1	H43887	Complement fact. D prec.
10	2.1	M63391	Human desmin
12	32.3	T47377	S-100P Prot.

Table 1. 6 genes (out of 4027) that were both in the top 20 sparse PCA genes and in the top 10 Rankgene genes.

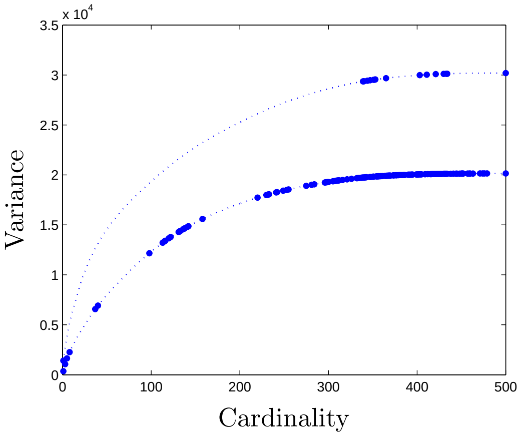

Figure 4. Variance versus cardinality tradeoff curve for two gene expression data sets, lymphoma (top) and colon cancer (bottom). Optimal points are in bold.

References

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., & Rosenwald, A. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, *403*, 503–511.

Alon, A., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Cell Biology*, *96*, 6745–6750.

Burer, S., & Monteiro, R. (2003). A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, *95*, 329–357.

Cadima, J., & Jolliffe, I. T. (1995). Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, *22*, 203–214.

d'Aspremont, A., El Ghaoui, L., Jordan, M., & Lanckriet, G. R. G. (2004). A direct formulation for sparse PCA using semidefinite programming. *To appear in SIAM Review.*

El Ghaoui, L. (2006). On the quality of a semidefinite programming bound for sparse principal component analysis. *ArXiV math.OC/0601448.*

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

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

Kolda, T. G., & O'Leary, D. P. (2000). Algorithm 805: computation and uses of the semidiscrete matrix decomposition. *ACM Transactions on Mathematical Software*, *26*, 415–435.

Moghaddam, B., Weiss, Y., & Avidan, S. (2006a). Generalized spectral bounds for sparse LDA. *Proc. ICML.*

Moghaddam, B., Weiss, Y., & Avidan, S. (2006b). Spectral bounds for sparse PCA: Exact and greedy algorithms. *Advances in Neural Information Processing Systems*, *18*.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM J. Comput.*, *24*, 227–234.

Su, Y., Murali, T. M., Pavlovic, V., Schaffer, M., & Kasif, S. (2003). Rankgene: identification of diagnostic genes based on expression data. *Bioinformatics*, *19*, 1578–1579.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal statistical society, series B*, *58*, 267–288.

Toh, K. C., Todd, M. J., & Tutuncu, R. H. (1999). SDPT3 – a MATLAB software package for semidefinite programming. *Optimization Methods and Software*, *11*, 545–581.

Zou, H., Hastie, T., & Tibshirani, R. (2004). Sparse principal component analysis. *Journal of Computational and Graphical Statistics.*