# EE 381V Project : A Survey on Sparse PCA

Rashish Tandon
Department of Computer Science
The University of Texas at Austin
rashish@cs.utexas.edu

May 9, 2012

## 1 Introduction

Principal Component Analysis (PCA) is a frequently used tool to analyse, visualize and reduce the dimensionality of data occurring in a variety of fields in science and engineering. Given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ (where $n$ is the number of points and $p$ is the dimensionality), PCA finds a set of $d(\ll p)$ orthonormal vectors $V = \{v_1, v_2, \ldots, v_d\}$ in $\mathbb{R}^p$ such that the $span(V)$ explains the maximum amount of variance in the data (or equivalently, the projection of the data onto $span(V)$ is maximized). This can be cast as a sequence of optimization problems, as follows :

$$\begin{aligned}
v_1 &= \operatorname*{arg\,max}_{\|v\|_2=1} \|Xv\|_2^2 \\
v_2 &= \operatorname*{arg\,max}_{\|v\|_2=1,\, \langle v,v_1\rangle=0} \|Xv\|_2^2 \\
&\vdots \\
v_k &= \operatorname*{arg\,max}_{\|v\|_2=1,\, \langle v,v_i\rangle=0\,\forall\, i<k} \|Xv\|_2^2 \\
&\vdots \\
v_d &= \operatorname*{arg\,max}_{\|v\|_2=1,\, \langle v,v_i\rangle=0\,\forall\, i<d} \|Xv\|_2^2
\end{aligned} \tag{1}$$

If $\mathbf{X}$ has the singular value decomposition (SVD), $X = U\Sigma V^T$, then the solutions to the above optimization problems are simply the right singular vectors $\{v_1, v_2, \ldots, v_d\}$ corresponding to the top-$d$ singular values, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d$.

An alternative view of PCA is that of eigenvector decomposition of a symmetric positive definite matrix. Consider a covariance matrix $\mathbf{A} \in \mathbb{S}_p^+$ which is related to $\mathbf{X}$ as

$$A = X^T X \tag{2}$$

Then, a vector $v \in \mathbb{R}^p$ is a right singular vector of $\mathbf{X}$ with singular value $\sigma$ *iff* $v$ is an eigenvector of $\mathbf{A}$ with eigenvalue $\sigma^2$. Thus, finding the right singular vectors corresponding to the top-$d$ singular values is equivalent to finding the eigenvectors of $\mathbf{A}$ corresponding to the top-$d$ eigenvalues,

$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$. This can then be viewed as repeating the following. Solve the optimization problem :

$$v^* = \arg\max_{\|v\|_2=1} v^T A v \tag{3}$$

followed by a *deflation* of **A**,

$$A = A - (v^{*T} A v^*) v^* v^{*T}. \tag{4}$$

Doing Step (3) followed by Step (4), $d$ times, would yield a set of orthonormal eigenvectors $\{v_1^*, \ldots, v_d^*\}$ corresponding to the top-$d$ eigenvalues of **A**.

A shortcoming of PCA is that in most applications, the *principal components* obtained have a lot of non-zeros. So, even though PCA facilitates representation of the data in a few factors, interpretation of each of the factors becomes difficult. Moreover, in some applications, the cost of analysis subsequent to PCA has a direct dependence on the number of non-zeros in the *principal components*. The problem of Sparse Principal Component Analysis (Sparse PCA) is to obtain a set of *sparse* vectors that explain the maximum amount of variance in the data. However, it is not necessary for the vectors to be orthogonal. Most approaches for Sparse PCA focus on obtaining a sparse solution to the problem in Step (3) followed by a *deflation* of the given matrix.

## 1.1 Surveyed Papers

This section briefly introduces the papers examined for this survey. We studied (Jolliffe *et al.* , 2003), which introduces the Sparse PCA problem as well as provides a simple LASSO based approach to obtain sparse vectors. (Zou *et al.* , 2004) casts Sparse PCA as a *regression-type* problem with an *elastic net* regularization, which is then solved via an alternating minimization scheme. However, both of these methods involve solving a non-convex problem directly, and thus suffer from the problem of running into local minima. (d'Aspremont *et al.* , 2007) proposes a semi-definite programming (SDP) relaxation to the Sparse PCA problem. (Moghaddam *et al.* , 2005) approaches the problem from a combinatorial perspective, and provides a greedy algorithm based on spectral characterizations.

Sparse PCA is a hard non-convex optimization problem, and so, its analysis becomes difficult. All of the above methods lack any guarantees in general for the recovery of sparse principal components. However, the SDP relaxation for Sparse PCA has been analysed for a rather special covariance matrix - the *spiked covariance model* - which is a covariance matrix in which a base matrix has been perturbed by the addition of a $k$-sparse maximal eigenvector. (Amini & Wainwright, 2009) analyses the SDP relaxation for recovery of the signed support set of this maximal eigenvector.

Sparse PCA has been a subject of extensive study over the past few years. There are a few other approaches that have been proposed but were not studied as part of this survey. These include a general power method for Sparse PCA (Journée *et al.* , 2010) and a probabilistic formulation for Sparse PCA (Guan & Dy, 2009). Also (Mackey, 2009) discusses a number of deflation techniques that are better suited for Sparse PCA in comparison to the simple deflation of Step (4) (also called *Hotelling's deflation*) which may not necessarily preserve the eigenvectors and positive semi-definiteness of the deflated matrix when used with a non-eigenvector. On a different note, recently, (Vu & Lei, 2012) have provided an information-theoretic lower bound (under some additional assumptions) on the estimation error $\|\hat{\theta}\hat{\theta}^T - \theta\theta^T\|_F$, for a covariance matrix with a sparse leading eigenvector $\theta$ and any estimator $\hat{\theta}$.

### 1.2 Extension

As the extension, we ran experiments on the *spiked covariance model* to compare the probability of signed support recovery of the regression based Sparse PCA (called SPCA), greedy approach to Sparse PCA (called GSPCA) and the SDP based approach to Sparse PCA (called DSPCA).

The rest of this survey is organized as follows. Section 2 discusses SCoTLASS- the LASSO based approach of Jolliffe *et al.* (2003). Section 3 discusses SPCA - the regression based approach of Zou *et al.* (2004). Section 4 discusses GSPCA - the greedy approach of Moghaddam *et al.* (2005). Section 5 discusses DSPCA - the SDP based approach of d'Aspremont *et al.* (2007). Section 6 discusses the results of the analysis in Amini & Wainwright (2009). Finally, Section 7 presents the results of our experiments on the *spike covariance model.*

## 2 Sparse PCA via LASSO (SCoTLASS)

Jolliffe *et al.* (2003) propose solving the sequence of problems in (1) with an additional $\ell_1$-constraint to promote sparsity. Given a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and some constant $t$, for each $i = 1, \ldots, d$ solve :

$$\begin{aligned} &\arg\max_{v_i} \|Xv_i\|_2^2 \\ &\qquad \text{s.t. } v_i^T v_i = 1,\ v_i^T v_j = 0 \, \forall \, j < i,\ \|v_i\|_1 \leq t \end{aligned} \tag{5}$$

Equivalently, given a covariance matrix $\mathbf{A} \in \mathbb{S}_p^+$ $(= X^T X)$, for each $i = 1, \ldots, d$ solve :

$$\begin{aligned} &\arg\max_{v_i} v_i^T A v_i \\ &\qquad \text{s.t. } v_i^T v_i = 1,\ v_i^T v_j = 0 \, \forall \, j < i,\ \|v_i\|_1 \leq t \end{aligned} \tag{6}$$

The following simple observations can be made :

1. If $t \geq \sqrt{p}$, then (6) is equivalent to PCA (since for any vector $v$ with $\|v\|_2 = 1$, $\|v\|_1 \leq \sqrt{p}\|v\|_2 = \sqrt{p}$).
2. If $t < 1$, then no solution exists (since for any $v$ with $\|v\|_2 = 1$, $\|v\|_1 \geq \|v\|_2 = 1$).
3. If $t = 1$, then $v_i$ has only one non-zero entry (since this would force $\|v\|_1 = \|v\|_2 = 1$).

The problem (6) is non-convex. Jolliffe *et al.* (2003) suggest a simple projected gradient descent approach to solve it, by moving the $\ell_1$-constraint into the objective as a penalty function. Even though SCoTLASS succeeds empirically in finding sparse vectors, its performance has been found to be inferior to other methods in terms of the amount of variance explained. Moreover, it can run into local optima, which forces restarts to obtain adequate solutions.

## 3 Sparse PCA via Regularized Linear Regression (SPCA)

Zou *et al.* (2004) express PCA as a regression type problem, on which sparsity can then be enforced through appropriate $\ell_1$ penalties. We already know that for a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, the following optimization problem gives us the top-$d$ principal components :

$$\arg\min_{V \in \mathbb{R}^{p \times d}} \|X - XVV^T\|_F^2 \quad \text{s.t. } V^T V = I \tag{7}$$

This is seen as follows. Writing $X$ as a collection of $p$-dimensional points, $X^T = [x_1\, x_2\, \ldots\, x_n]$, we see that (7) is equivalent to

$$\begin{aligned} &\underset{V \in \mathbb{R}^{p\times d}}{\arg\min} \sum_{i=1}^{n} \|x_i^T - x_i^T V V^T\|_2^2 \quad \text{s.t. } V^T V = I \\ \Rightarrow\ &\underset{V \in \mathbb{R}^{p\times d}}{\arg\max} \sum_{i=1}^{n} \|x_i - V V^T x_i\|_2^2 \quad \text{s.t. } V^T V = I \end{aligned} \tag{8}$$

Now, for any orthogonal matrix $V_{p\times d} = [v_1\, v_2\, \ldots\, v_d]$,

$$V V^T x_i = \sum_{k=1}^{d} v_k (v_k^T x_k) = proj_V(x_i) \tag{9}$$

where $proj_V(x_i)$ denotes the projection of $x_i$ onto the $span(V)$. So, our optimization problem becomes

$$\underset{V \in \mathbb{R}^{p\times d}}{\arg\min} \sum_{i=1}^{n} \|x_i - proj_V(x_i)\|_2^2 \quad \text{s.t. } V^T V = I \tag{10}$$

which we know yields the top-$d$ principal components.

Zou *et al.* (2004) show that if we split (7) as an optimization over two matrices $(\mathbf{P}, \mathbf{Q})$ and add an additional $\lambda\|Q\|_F^2$ penalty, the solution remains the same. This is described in Theorem 1.

**Theorem 1.** *Let $\{v_1\, v_2\, \ldots\, v_d\}$ be the top-$d$ principal components of $X_{n\times p}$. Let $P_{p\times d} = [p_1\, p_2\, \ldots\, p_d]$ and $Q_{p\times d} = [q_1\, q_2\, \ldots\, q_d]$. For any $\lambda \geq 0$, let*

$$(\hat{P}, \hat{Q}) = \underset{P,Q}{\arg\min}\ \|X - XQP^T\|_F^2 + \lambda\|Q\|_F^2 \quad s.t.\ P^T P = I \tag{11}$$

*Then, $\dfrac{\hat{q}_i}{\|\hat{q}_i\|_2} = v_i$, $\forall\, i = 1, \ldots, d$.*

Now, to enforce sparse solutions, we can simply add an $\ell_1$ penalization on the columns of $\mathbf{Q}$ to get :

$$\underset{P,Q}{\arg\min}\ \|X - XQP^T\|_F^2 + \lambda\|Q\|_F^2 + \sum_{i=1}^{d} \mu_i \|q_i\|_1 \quad \text{s.t. } P^T P = I \tag{12}$$

Even though the optimization problem (12) is still non-convex, it is useful since we have decoupled the orthogonality and sparsity conditions. So, (12) can be solved via an alternating minimization approach *i.e.* starting from an arbitrary $(\mathbf{P}_0, \mathbf{Q}_0)$, at the $t^{th}$ iteration solve for $\mathbf{Q}_t$ assuming a fixed $\mathbf{P}_{t-1}$, then solve for $\mathbf{P}_t$ assuming a fixed $\mathbf{Q}_t$.

**For a fixed orthogonal P**. To solve for $\mathbf{Q}$, we need to solve

$$\underset{Q}{\arg\min}\ \|X - XQP^T\|_F^2 + \lambda\|Q\|_F^2 + \sum_{i=1}^{d} \mu_i \|q_i\|_1 \tag{13}$$

Let $\mathbf{P}_\perp$ be an orthogonal matrix such that $[P; P_\perp]$ is $p \times p$ orthogonal. Then, by projecting the rows of $(X - XQP^T)$ onto $\mathbf{P}$ and $\mathbf{P}_\perp$, we get

$$\begin{aligned} \|X - XQP^T\|_F^2 &= \|(X - XQP^T)P_\perp\|_F^2 + \|(X - XQP^T)P\|_F^2 \\ &= \|XP_\perp\|_F^2 + \|XP - XQ\|_F^2 \end{aligned} \tag{14}$$

Since $\|XP_{\perp}\|_F^2$ is independent of $\mathbf{Q}$, (13) becomes

$$\arg\min_Q \|XP - XQ\|_F^2 + \lambda\|Q\|_F^2 + \sum_{i=1}^{d} \mu_i \|q_i\|_1 \tag{15}$$

Now, (15) has the form of a regression problem with an elastic-net penalization (both $\ell_1$ and $\ell_2$ penalties). This problem can solved efficiently by the LARS-EN algorithm (Zou & Hastie, 2005).

**For a fixed Q**. To solve for $\mathbf{P}$, we need to solve

$$\arg\min_P \|X - XQP^T\|_F^2 \quad \text{s.t. } P^TP = I \tag{16}$$

This can be solved by computing the SVD of $X^TXQ$ as $X^TXQ = U\Sigma V^T$ and setting $P = UV^T$. This is justified in Lemma 1 (from Zou *et al.* (2004)).

**Lemma 1.** *Given* $\mathbf{M}_{n\times p}$ *and* $\mathbf{N}_{p\times k}$*, consider*

$$\hat{P} = \arg\min_P \|M - NP^T\|_F^2 \quad \textit{s.t. } P^TP = I \tag{17}$$

*If* $M^TN$ *has the SVD,* $M^TN = U\Sigma V^T$*, then* $\hat{P} = UV^T$*.*

Finally, empirically, SPCA is seen to outperform SCoTLASS. For example, on the Pitprops dataset, which is a classic example for testing out Sparse PCA algorithms, SPCA yields solutions which are sparser than SCoTLASS (a total of 18 non-zeros vs. 47 non-zeros, respectively) and which account for greater variation in the data than what is accounted for by SCoTLASS (75.8% variation vs. 69.3%, respectively).

# 4 A Spectral Approach to Sparse PCA (GSPCA)

Moghaddam *et al.* (2005) adopt a combinatorial approach to tackling Sparse PCA. Given $\mathbf{A} \in \mathbb{S}_p^+$, the focus is on solving the following problem :

$$\begin{aligned} &\arg\max_{v\in\mathbb{R}^p} v^TAv \\ &\qquad \text{s.t. } v^Tv = 1,\ \|v\|_0 \le k \end{aligned} \tag{18}$$

If we can obtain a solution to (18), then a set of sparse vectors can be obtained by repeatedly solving (18) coupled with an appropriate *deflation* step.

Let $v^*$ be the optimal for (18) and $S$ be its support *i.e.* $S = supp(v^*) = \{j \mid v_j^* \neq 0\}$. Clearly, we would have $|S| \le k$. Then, it is easy to see that $v_S^*$ (*i.e.* $v^*$ restricted to its support set) would be the optimum of

$$\begin{aligned} &\arg\max_{v\in\mathbb{R}^{|S|}} v^TA_{SS}v \\ &\qquad \text{s.t. } v^Tv = 1 \end{aligned} \tag{19}$$

where $\mathbf{A}_{SS}$ is the principal submatrix of $\mathbf{A}$ corresponding to the set $S$. However, (19) is simply the maximum eigenvalue problem for the matrix $\mathbf{A}_{SS}$. Its optimum value is $\lambda_{\max}(A_{SS})$, and is achieved for the principal eigenvector of $\mathbf{A}_{SS}$. In other words, $v^*_S$ would be the principal eigenvector of $\mathbf{A}_{SS}$. Note that $v^*_{S^c} = 0$.

Thus, the following brute force scheme would find the optimum for (18) : Search over all possible support sets $S$ s.t. $|S| \leq k$ and pick the $S$ with the maximum value for $\lambda_{\max}(A_{SS})$. For this $S$, set $v^*_S$ to be the principal eigenvector of $A_{SS}$ and $v^*_{S^c} = 0$. Ofcourse, such an approach would be computationally infeasible owing to the large number of possible support sets ( $= O(n^k)$).

Additionally, the above discussion also suggests a simple *renormalization* step to improve any candidate solution to (18) that may have been obtained by some other method. Given a candidate solution to (18), $\tilde{x}$, let $\tilde{S} = supp(\tilde{x})$. Then, compute $\lambda_{\max}(A_{\tilde{S}\tilde{S}})$. If $\tilde{x}^T A \tilde{x} \neq \lambda_{\max}(A_{\tilde{S}\tilde{S}})$, then we know that $\tilde{x}$ is not optimal for its given support set $\tilde{S}$, and we may replace $\tilde{x}_{\tilde{S}}$ by the principal eigenvector of $\mathbf{A}_{\tilde{S}\tilde{S}}$.

Using the Courant-Fischer variational form for the eigenvectors of a symmetric matrix, it is possible to relate the eigenvalues of a symmetric matrix to the eigenvalues of its principal submatrix. This is described in Lemma 2.

**Lemma 2.** *Let* $\mathbf{A}$ *be an* $n \times n$ *symmetric matrix and* $\mathbf{A}_k$ *be any* $k \times k$ *principal submatrix of* $\mathbf{A}$*, for* $1 \leq k \leq n$*. Also, let* $\lambda_i(X)$ *correspond to the* $i^{th}$ *largest eigenvalue of any matrix* $\mathbf{X}$*. Then, for each* $1 \leq i \leq k$*,*

$$\lambda_i(A) \leq \lambda_i(A_k) \leq \lambda_{i+n-k}(A) \tag{20}$$

For a proof of this, see Horn & Johnson (1985, Sec. 4.3.15). The proof uses the simple idea of enforcing additional constraints on the variational forms to obtain the eigenvalues of the corresponding principal submatrix as a solution.

A simple consequence of Lemma 2 is the following. For any symmetric matrix $\mathbf{A}$, let $\mathbf{A}_{\setminus j}$ correspond to the principal submatrix of $\mathbf{A}$ with the $j^{th}$ row and column removed. Then,

$$\lambda_{\max}(A_{\setminus j}) \leq \lambda_{\max}(A) \tag{21}$$

In other words, adding a row and corresponding column to a matrix can only increase the maximum eigenvalue and likewise, deleting a row and corresponding column can only decrease it. Thus, for the problem (18), the cardinality constraint $\|v\|_0 \leq k$ is a tight equality at the optimum. Moreover, this motivates the following *greedy* strategy to find a solution to (18) :

1. Start with an empty set, $S = \{\}$.
2. Choose j which maximizes $\lambda_{\max}(A_{S\cup\{j\}})$.
3. Set $S \leftarrow S \cup \{j\}$. If $|S| = k$, Stop. Else, Goto Step 2.
4. Set $v^*_S$ to be the principal eigenvector of $A_{SS}$ and $v^*_{S^c} = 0$.

The solution obtained by the above greedy strategy may not be globally optimal, but will be locally optimal *i*.e. optimal for its support set. We can also adopt the above strategy in reverse *i*.e. start with $S = \{1, 2, \ldots, p\}$ and then delete elements based on our greedy strategy. We may also combine these two *i*.e. run a forward pass and a backward pass, and then choose the better of the two solutions. This method is called Greedy SPCA (or GSPCA). Empirically (in Moghaddam *et al.* (2005)), GSPCA is seen perform better or comparably to SPCA.

Lemma 2 also yields a lower bound on the optimum value of our problem (18) as :

$$\lambda_k(A) \leq \lambda_{\max}(A_k) \tag{22}$$

This may be useful in governing the choice of $k$ in practical applications.

# 5 SDP Relaxations for Sparse PCA (DSPCA)

The primary focus in d'Aspremont *et al.* (2007) is also problem (18). We restate it here for completeness.

$$\begin{aligned}\arg\max_{v\in\mathbb{R}^p} \, & v^T A v \\ \text{s.t. } & v^T v = 1, \ \|v\|_0 \leq k\end{aligned}$$

d'Aspremont *et al.* (2007) propose a convex relaxation to this problem (by means of an SDP) as follows. First, an equivalent problem is considered where the vector $v$ has been transformed into a matrix $\mathbf{X}$.

$$\begin{aligned}\arg\max_{X\in\mathbb{S}_p} \ & Tr(AX) \\ \text{s.t. } & Tr(X) = 1, \ Card(X) \leq k^2 \\ & X \succeq 0, \ rank(X) = 1\end{aligned} \tag{23}$$

where $Card(X)$ is the number of non-zero elements in the matrix $\mathbf{X}$.

Problems (18) and (23) are equivalent as $X \succeq 0$ and $rank(X) = 1$ enforce that $X = vv^T$, while if $X = vv^T$, then $Tr(X) = 1$ enforces $v^T v = 1$. Also, if $X = vv^T$, then $Card(X) \leq k^2$ is equivalent to $\|v\|_0 \leq k$. The advantage of the form of (23) is that the convex maximization objective from (18) is now linear. However, the constraints $Card(X) \leq k^2$ and $rank(X) = 1$ make the problem non-convex, and so must be relaxed. $Card(X) \leq k^2$ is relaxed by an analogue of the $\ell_1$ constraint for matrices (*i.e.* a sum of the absolute values of the matrix), a weaker but convex constraint : $\mathbf{1}^T|X|\mathbf{1} \leq k$. Note that $k^2$ becomes $k$ due to the fact that for any vector $u$, $\|u\|_1 \leq \sqrt{\|u\|_0}\|u\|_2$. So, for the matrix $\mathbf{X}$, we would have $\mathbf{1}^T X \mathbf{1} \leq k\|X\|_F = k$, as for $X = xx^T$, $\|X\|_F = \sqrt{x^T x} = 1$. The rank constraint is simply dropped. This gives us a convex relaxation (an SDP) :

$$\begin{aligned}\arg\max_{X\in\mathbb{S}_p} \ & Tr(AX) \\ \text{s.t. } & Tr(X) = 1, \ 1^T|X|1 \leq k \\ & X \succeq 0\end{aligned} \tag{24}$$

Similarly, a penalized form of problem (18),

$$\begin{aligned}\arg\max_{v\in\mathbb{R}^p} \, & v^T A v - \rho\|v\|_0^2 \\ \text{s.t. } & v^T v = 1\end{aligned} \tag{25}$$

can be relaxed to

$$\begin{aligned}\arg\max_{X\in\mathbb{S}_p} \ & Tr(AX) - \rho \, 1^T|X|1 \\ \text{s.t. } & Tr(X) = 1, \ X \succeq 0\end{aligned} \tag{26}$$

Note that the optimal values of problems (24) and (26) are upper bounds on the optimal values of problems (18) and (25) respectively. If the optimal solution $\mathbf{X}$(in both cases) is of rank one *i*.e. $X = xx^T$, then we can take $x$ to be the solution of the corresponding original problem. In this case, we are guaranteed that $Card(X) = \|x\|_0^2$. If the optimal $\mathbf{X}$ is not of rank one, then we retain the dominant eigenvector $x$ as an approximate solution to the corresponding original problem. However, here, the dominant eigenvector cannot be guaranteed to be as sparse as the matrix $\mathbf{X}$.

Empirically, DSPCA is seen to perform better than SPCA and comparable to GSPCA. On the Pitprops dataset, it produces solutions which are sparser than SPCA (a total of 12 non-zeros vs. 18 non-zeros, respectively) and which account for slightly more variation in the data than what is accounted for by SPCA.

## 6 Known Results on the Spiked Covariance Model

Amini & Wainwright (2009) provide a theoretical analysis of the SDP relaxation (26) on the *spiked covariance model*. The setting considered here is as follows.

We have a covariance matrix $\Sigma_p$, which has been constructed by perturbing a base matrix by adding a maximal sparse eigenvector $z^* \in \mathbb{R}^p$. Formally, for the model considered in their paper,

$$\Sigma_p = \beta z^* z^{*T} + \begin{bmatrix} I & 0 \\ 0 & \Gamma_{p-k} \end{bmatrix} = \beta z^* z^{*T} + \Gamma \tag{27}$$

where $z^*$ is $k$-sparse with non-zero values in its first $k$ entries. Additionally, for each $i = 1, \ldots, k$, $z_i^* \in \frac{1}{\sqrt{k}}\{-1, 1\}$ and thus, $\|z\|_2 = 1$. The matrix $\Gamma_{p-k} \in \mathbb{S}_{p-k}^{+}$. $\Gamma_{p-k}$ is also required to satisfy the following two conditions :

(C1) $|||\sqrt{\Gamma_{p-k}}|||_{\infty,\infty} = O(1)$

(C2) $\lambda_{\max}(\Gamma_{p-k}) \leq \min\{1, \lambda_{\min}(\Gamma_{p-k}) + \frac{\beta}{8}\}$

Note that $\sqrt{\Gamma_{p-k}}$ is the symmetric square root of $\Gamma_{p-k}$ and the $\ell_{\infty,\infty}$-induced operator is defined as :

$$|||X|||_{\infty,\infty} = \max_i \sum_j |X_{ij}| \tag{28}$$

Now, suppose $n$-samples, $\{x_1, x_2, \ldots, x_n\}$, are drawn from $\mathcal{N}(0, \Sigma_p)$ and the empirical covariance is computed as

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \tag{29}$$

Then, the SDP relaxation (26) is solved with the matrix $\mathbf{A}$ being $\hat{\boldsymbol{\Sigma}}$. This model is analysed for the recovery of the signed support of $z^*$, the maximal eigenvector of $\Sigma_p$. The signed support set is denoted by $S_{\pm}(z^*)$.

Condition (C2) is required to guarantee that $z^*$ is the maximal eigenvector of $\Sigma_p$. This will be true as for $\beta > 0$, $z^*$ has eigenvalue $(1 + \beta)$, whereas, by the bound, all other eigenvalues are small (less than 1). Condition (C1) is required to provide a bound on $\|\sqrt{\Gamma_{p-k}}u\|_\infty$, which in turn is required to provide a bound on $\|\hat{z} - z^*\|_\infty$ in the analysis, for the estimate $\hat{z}$ obtained through the SDP (and thus guarantee signed support recovery, as for signed support recovery, we must have $\|\hat{z} - z^*\|_\infty \leq 1/\sqrt{k}$).

The main result of Amini & Wainwright (2009) is stated in Theorem 2.

**Theorem 2.** *Assuming conditions (C1) and (C2) on* $\Sigma_p$ *and that* $\rho = \dfrac{\beta}{2k}$ *in (26) and* $k = O(\log p)$*, there exists a constant* $\theta > 0$ *such that if*

$$n > \theta k \log(p - k) \tag{30}$$

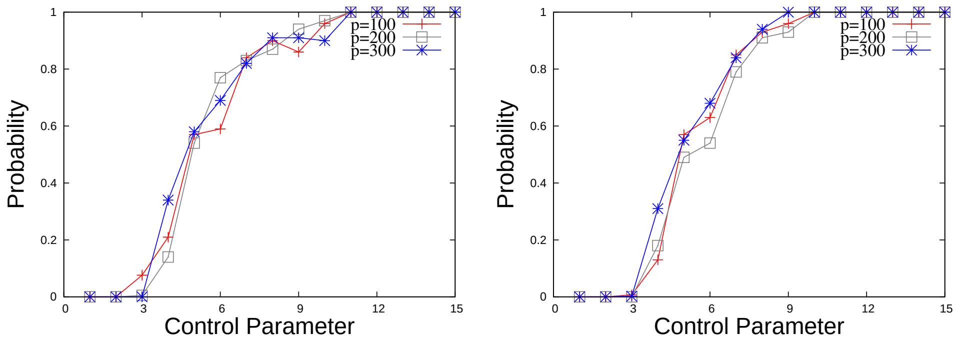
*and the SDP (26) has a rank-one solution* $\hat{z}$*, then* $P(S_{\pm}(\hat{z}) = S_{\pm}(z^*))$ *converges to 1.*

The high level proof idea for proving this theorem is the standard primal-dual witness approach for a candidate solution obtained through an oracle which satisfies the exact signed recovery criterion and the rank 1 criterion.

## 7 Extension : Empirical Comparison on the Spiked Covariance Model

As an extension, we performed empirical evaluations on the *spiked covariance model*, to compare the probability of signed recovery for DSPCA, GSPCA and SPCA. In our experiments, we took the value $\beta = 3$. $\Gamma_{p-k}$ was taken to be a diagonal matrix with each diagonal entry chosen uniformly randomly from $[0.5, 0.975]$. The non-zero entries of $z^*$ were chosen uniformly randomly from $\dfrac{1}{\sqrt{k}}\{-1, 1\}$. The number of non-zeros in $z^*$, $k$ was taken as $log(p)$ and $\sqrt{p}$ in the experiments. The dimension of the problem, $p$ was taken as 100, 200 and 300. It is easy to verify that our setup satisfies the conditions of Amini & Wainwright (2009), except for $k$ being taken as $\sqrt{p}$.

The results of our experiments are presented as plots between the probability of signed recovery, $P(S_{\pm}(\hat{z}) = S_{\pm}(z^*))$ and the control parameter $\alpha = \frac{n}{k \log(p-k)}$. The probability was estimated by 100 independent trials for each tuple of $(p, k, \alpha)$ values. Similar experiments have been done in Amini & Wainwright (2009), but only for DSPCA with $\Gamma_{p-k} = I$.

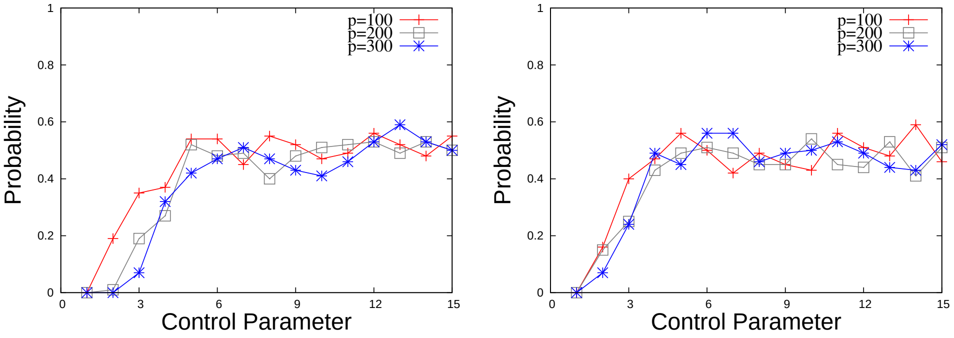


(a) $k = \log(p)$

(b) $k = \sqrt{p}$

Figure 1: Performance of DSPCA on the *spiked covariance model*

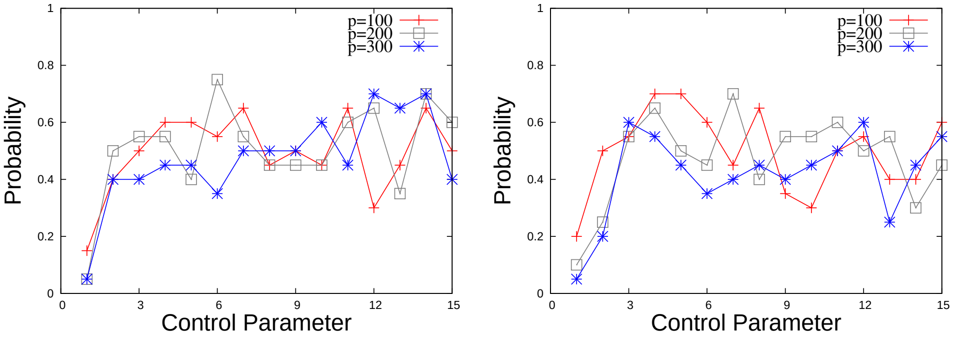
In case of DSPCA (Fig. 1), the behaviour is as expected, and the probability of success (signed recovery) goes to 1 as the control parameter increases. Even though $k = \sqrt{p}$ violates the $k = O(\log p)$ condition from Theorem 2, empirically, the probability still goes to 1. A similar observation was also made in Amini & Wainwright (2009) and an open question here is whether the analysis of Amini & Wainwright (2009) can be extended for a general $k$.



(a) $k = \log(p)$

(b) $k = \sqrt{p}$

Figure 2: Performance of SPCA on the *spiked covariance model*



(a) $k = \log(p)$

(b) $k = \sqrt{p}$

Figure 3: Performance of GSPCA on the *spiked covariance model*

SPCA (Fig. 2) and GSPCA (Fig. 3) perform inferior to DSPCA on this model. SPCA is seen to err with probability almost 0.5, even for larger values of the control parameter. GSPCA also does not perform much better, and shows a stronger fluctuation in the probability of success, whereas SPCA tends to gradually converge around the 0.5 value.

For SPCA, we used the LARS-based SPCA implementation in Matlab of Sjöstrand (2005). This implements a constrained version of SPCA (constraint on the $\ell_1$ norm). The upper bound for the $\ell_1$ norm here was taken as $1.2 * \sqrt{k}$ ($\|z^*\|_1 = \sqrt{k}$, so this was relaxed slightly). For GSPCA, we implemented it ourselves, utilizing the ARPACK toolbox in Matlab for eigenvalue computations. For DSPCA, we used the implementation of d'Aspremont *et al.* (2007), with $\rho = \frac{\beta}{2k}$.

## References

Amini, Arash A., & Wainwright, M. J. 2009. High-dimensional analysis of semidefinite programming relaxations for sparse principal component analysis. *Annals of Statistics*, **37**, 2877–2921.

d'Aspremont, Alexandre, Ghaoui, Laurent El, Jordan, Michael I., & Lanckriet, Gert R. G. 2007.

A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, **49**(3), 434–448.

Guan, Yue, & Dy, Jennifer G. 2009. Sparse Probabilistic Principal Component Analysis. *Journal of Machine Learning Research - Proceedings Track*, **5**, 185–192.

Horn, Roger A., & Johnson, Charles R. 1985. *Matrix analysis*. Repr. with corr. edn. Cambridge University Press, Cambridge [Cambridgeshire] ; New York :.

Jolliffe, Ian T, Trendafilov, Nickolay T, & Uddin, Mudassir. 2003. A Modified Principal Component Technique Based on the LASSO. *Journal of Computational and Graphical Statistics*, **12**(3), 531–547.

Journée, Michel, Nesterov, Yurii, Richtárik, Peter, & Sepulchre, Rodolphe. 2010. Generalized Power Method for Sparse Principal Component Analysis. *Journal of Machine Learning Research*, **11**, 517–553.

Mackey, Lester. 2009. Deflation Methods for Sparse PCA. *Pages 1017–1024 of: Advances in Neural Information Processing Systems 21.*

Moghaddam, Baback, Weiss, Yair, & Avidan, Shai. 2005. Spectral Bounds for Sparse PCA: Exact and Greedy Algorithms. *In: Advances in Neural Information Processing Systems 18.* MIT Press.

Sjöstrand, K. 2005. *Matlab implementation of LASSO, LARS, the elastic net and SPCA*. Version 2.0.

Vu, Vincent Q., & Lei, Jing. 2012. Minimax Rates of Estimation for Sparse PCA in High Dimensions. *CoRR*, **abs/1202.0786**.

Zou, Hui, & Hastie, Trevor. 2005. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, **67**, 301–320.

Zou, Hui, Hastie, Trevor, & Tibshirani, Robert. 2004. Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15**, 2006.