A Greedy Anytime Algorithm for Sparse PCA

Guy Holtzman GUYHOL@POST.BGU.AC.IL
Ben-Gurion Unversity of the Negev, Beer-Sheva, Israel

Adam Soffer SOFFER@POST.BGU.AC.IL
Ben-Gurion Unversity of the Negev, Beer-Sheva, Israel

Dan Vilenchik VILENCHI@BGU.AC.IL
Ben-Gurion Unversity of the Negev, Beer-Sheva, Israel

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

The taxing computational effort that is involved in solving some high-dimensional statistical problems, in particular problems involving non-convex optimization, has popularized the development and analysis of algorithms that run efficiently (polynomial-time) but with no general guarantee on statistical consistency. In light of the ever-increasing compute power and decreasing costs, a more useful characterization of algorithms is by their ability to calibrate the invested computational effort with various characteristics of the input at hand and with the available computational resources. We propose a new greedy algorithm for the ℓ_0-sparse PCA problem which supports the calibration principle. We provide both a rigorous analysis of our algorithm in the spiked covariance model, as well as simulation results and comparison with other existing methods. Our findings show that our algorithm recovers the spike in SNR regimes where all polynomial-time algorithms fail while running in a reasonable parallel-time on a cluster.

Keywords: Sparse PCA, non-convex optimization, anytime algorithms, average case analysis

1. Introduction

Principal components analysis (PCA) is the mainstay of modern machine learning and statistical inference, with a wide range of applications involving multivariate data, in both science and engineering (Anderson, 1984; Jolliffe, 2002). The application of PCA to high-dimensional data, where features are plentiful (large p) but samples are relatively scarce (small n) suffers from two major limitations: interpretability and consistency (Bickel and Levina, 2008; Johnstone, 2001; Johnstone and Lu, 2009; Nadler, 2008; Paul, 2007). These limitations encouraged the design of regularized learning schemes, such as the ℓ_0-sparse PCA, or k-sparse PCA as we call it from now on. Given a pair (X, k), where X is an $n \times p$ design matrix and k the desired sparsity level, the goal is to find a unit vector $\mathbf{v}$ that has at most k non-zero entries, a k-sparse vector, such that the variance of X in $\mathbf{v}$'s direction is maximal.

While standard (non-restricted) PCA can be efficiently solved by computing the eigenvectors of a symmetric matrix, its k-sparse variant is NP-hard (Natarajan, 1995). Nevertheless, computationally efficient heuristics were proposed and analyzed under various assumptions on the distribution of X and the parameters n, p and k, e.g. Amini and Wainwright (2009); d'Aspremont et al. (2004); Deshpande and Montanari (2016); Johnstone and Lu (2009); Krauthgamer et al. (2015).

The performance of all the aforecited algorithms features a rather undesirable phase-transition behavior (at least on the benchmark distribution that was studied in each paper). Each algorithm A performs well up to a certain SNR threshold τ_A, and its performance deteriorates as the SNR drops below τ_A. Such a threshold behavior is expected in a worse-case setting, as the problem is NP-hard. However, the results of Berthet and Rigollet (2013a,b); Brennan and Bresler (2019); Ding et al. (2019); Krauthgamer et al. (2015) suggest that the threshold behavior might persist even in the average-case setting, as long as the algorithms belong to the polynomial-time family.

Throughout, we let $\mathbf{v}^* \in \mathbb{R}^p$ denote the solution of the k-sparse PCA problem and $\mathcal{I}^* \subseteq \{1,\ldots,p\}$ the support of $\mathbf{v}^*$. We denote by $\hat{\Sigma} = \frac{1}{n}X^T X$ the sample covariance matrix, assuming X is centered. In what follows, we consider the equivalent problem of finding the support set $\mathcal{I}^*$ rather than $\mathbf{v}^*$.

2. Our contribution

Anytime algorithms provide the ability to achieve results of better quality in return for running time (Zilberstein, 1996). We implement this philosophy in the context of the sparse PCA problem. We propose a new algorithm that consists of a tunable parameter that allows to increase the running time as the SNR weakens. Thus the algorithm maintains a steady success rate and avoids the aforementioned threshold behavior. If necessary, the algorithm invests super polynomial-time.

It will be convenient to reformulate k-sparse PCA as follows. For a fixed symmetric matrix $A \in \mathbb{R}^{p\times p}$, define the mapping $f_{\lambda_1}^{(A)} : 2^{\{1,\ldots,p\}} \to \mathbb{R}$ by $f_{\lambda_1}^{(A)}(\mathcal{S}) = \lambda_1(A_{\mathcal{S}})$, the largest eigenvalue of the principal submatrix $A_{\mathcal{S}}$ of A corresponding to the variables in $\mathcal{S}$. We abbreviate $f_{\lambda_1}^{(A)}$ by f_{λ_1} when A is clear from the context. The k-sparse PCA problem is the solution of

$$\mathcal{I}^* = \underset{\substack{\mathcal{S}\subseteq\{1,\ldots,p\}\\ |\mathcal{S}|=k}}{\operatorname{argmax}} f_{\lambda_1}^{(\hat{\Sigma})}(\mathcal{S}). \tag{1}$$

Our algorithm is composed of two routines. The first, which we call **GreedySprasePCA**, receives a real valued function $f : 2^{\{1,\ldots,p\}} \to \mathbb{R}$ (for example f_{λ_1}), a seed $\mathcal{S}^* \subseteq \{1,\ldots,p\}$ of size $k^* \leq k$, and a solution size k. It greedily completes $\mathcal{S}^*$ to a candidate solution of k-sparse PCA.

GreedySPCA$(f, \mathcal{S}^*, k)$:

1: $k^* \leftarrow |\mathcal{S}^*|$
2: **for all** $i \in \{1,\ldots,p\} \setminus \mathcal{S}^*$ **do**
3: $\quad a_i \leftarrow f(\mathcal{S}^* \cup \{i\})$
4: **end for**
5: sort the a_i's as $a_{i_1} \geq a_{i_2} \geq \ldots \geq a_{i_{p-k^*}}$
6: **return** $\mathcal{S}^* \cup \{i_1,\ldots,i_{k-k^*}\}$

The next routine, **SeedSparsePCA** (SSPCA for short), enumerates over all possible seeds of a given size k^*, completes each one using GreedySPCA, and returns the "best" solution.

SSPCA(f_1, f_2, k, k^*) :

1: **for all** seeds $\mathcal{S}^* \subseteq \{1, \ldots, p\}$ of size k^* **do**
2: $\quad \mathcal{S}^{(\mathcal{S}^*)} \leftarrow$ GreedySPCA$(f_1, \mathcal{S}^*, k)$
3: **end for**
4: **return** $\operatorname{argmax}_{\mathcal{S}^*} f_2(\mathcal{S}^{(\mathcal{S}^*)})$

Note that SSPCA is actually a family of algorithms, depending on the hyper-parameters f_1, f_2. We shortly discuss the choice of these parameters.

The running time of SSPCA is $\binom{p}{k^*} O(pk^* + k \log k)$. By varying k^* one obtains a hierarchy of algorithms, which for our choice of f_1, f_2 ranges from Diagonal Thresholding (Johnstone and Lu, 2009) (for $k^* = 0$) up to the naive exhaustive search (for $k^* = k$). The hierarchy realizes the anytime principle.

An attractive feature of SSPCA is the fact that it is completely white-box with only one simple tunable parameter, k^*. Furthermore, SSPCA can easily be parallelized and run in a multi-core cluster environment. The code we share is written in that way.

The following two conditions are sufficient for SSPCA(f_1, f_2, k, k^*) to recover at least $(\delta - \xi)$-fraction of $\mathcal{I}^*$, for two numbers $\delta, \xi \in [0, 1]$.

C1. There exists a *golden seed* $\mathcal{S}^*$ of size k^* such that GreedySPCA$(f_1, \mathcal{S}^*, k)$ outputs a set $\mathcal{I}$ satisfying $|\mathcal{I} \cap \mathcal{I}^*| \geq \delta k$.

C2. $\hat{\Sigma}$ is ξ-*separable* with respect to f_2. Namely, for every two sets $\mathcal{I}, \mathcal{J}$ of size k, if $|\mathcal{I} \cap \mathcal{I}^*| - |\mathcal{J} \cap \mathcal{I}^*| \geq \xi k$ then $f_2(\mathcal{I}) > f_2(\mathcal{J})$.

For C1 and C2 to be meaningful, one should think of golden seeds with δ close to 1, and f_2-separability with ξ close to 0. Proposition 1 formally asserts the sufficiency of these conditions.

The definition of k-sparse PCA in Eq. (1) suggests the choice $f_2 = f_{\lambda_1}$, which is indeed what we chose. For the rigorous analysis, we chose $f_1 = f_{avg}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{i,j \in \mathcal{S}} \hat{\Sigma}_{i,j}$, namely the average row sum in $\hat{\Sigma}_{\mathcal{S}}$. Note that for every $\mathcal{S}$, $f_{\lambda_1}(\mathcal{S}) \geq f_{avg}(\mathcal{S})$ by plugging the characteristic vector of $\mathcal{S}$ in the Rayleigh-quotient definition of λ_1. In the simulation part, we experimented with other functions as well. Details in Section 9.

We analyze SSPCA rigorously in the well-known spiked covariance model, which is formally defined in Section 3. Theorem 2 establishes the scaling of k^* as a function of the parameters (n, p, k), for which condition C1 holds with $\delta = 1$. Theorem 3 and Corollary 4 explicate the gap parameter ξ in condition C2 as a function of (n, p, k), from which the regime where $\xi = o(1)$ is obtained. Together they guarantee the recovery of $(1 - o(1))$-fraction of $\mathcal{I}^*$, up to the information limit. Our results are asymptotic, namely, they hold with probability (w.p.) tending to 1 as the parameters of the problem (n, p, k) go to infinity. The probability is taken only over the choice of the design matrix X.

Figure 1 summarizes simulations that show how our approach implements the anytime paradigm: increasing k^* (and subsequently the running time of SSPCA) translates to the desired increase in the solution quality. We further compared the performance of SSPCA when allowed "polynomial-time" ($k^* = 1, 2$) to three popular polynomial-time algorithms. Figure 1 shows that SSPCA is better than all three. Finally, we show that SSPCA with $k^* = 3$ outperforms the naive exhaustive search when

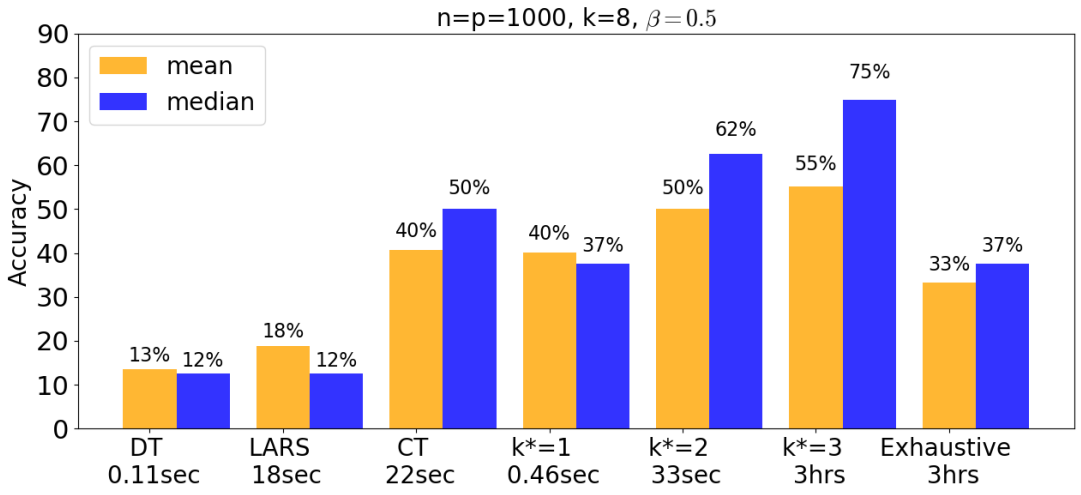

Figure 1: The plot portrays the accuracy averaged over 25 executions in the uniform unbiased spiked covariance model (USPCA), parametrized to suit a weak SNR regime ($n = p = 1000, k = 8, \beta = 0.5$). The compared algorithms are SSPCA with various seed sizes, Diagonal Thresholding (DT) (Johnstone and Lu, 2009), Covariance Thresholding (CT) Bickel and Levina (2008), LARS regression (Zou et al., 2006) and a naive exhaustive search that was allowed the same running time as SSPCA with $k^* = 3$. Full details of the executions are given in Section 9. The average running time is stated below each algorithm name. The reported running time of SSPCA and the preempted exhaustive search is a parallel-time using 90 `Intel Xeon Processor E7-4850 v4 (40M Cache, 2.10 GHz)` cores.

both are running for the same amount of time. The `Python` code, alongside documentation and examples, is available on Github [1].

Finally, let us mention that other greedy algorithms (approximate and exact) have been suggested for the sparse PCA problem, e.g. Asteris et al. (2011); Asteris et al. (2015); Baback et al. (2006); d'Aspremont et al. (2008). However, neither of these algorithms follows the anytime paradigm, and only worst-case guarantees were provided.

Independently of this result, a different anytime algorithm for k-sparse PCA was obtained in Ding et al. (2019). The algorithm also employs a controlled exhaustive search part, but the overall algorithmic approach is different. The algorithm was rigorously analyzed in the spiked covariance model as well and both algorithms have the same asymptotic run-time. In fact, Ding et al. (2019) provide evidence that this run-time is tight.

3. The Spiked Covariance Model

The spiked covariance model was suggested by Johnstone (2001) to model a combined effect of a low-dimensional signal buried in high-dimensional noise. In this paper, we consider the Gaussian case with a single spike, where the population covariance matrix is for the form $\Sigma = \beta \mathbf{v}^* \mathbf{v}^{*T} + I_p$. The parameter $\beta \geq 0$ is the signal strength, $\mathbf{v}^* \in \mathbb{R}^p$ is the planted spike assumed to be a k-sparse

1. https://github.com/sdannyvi/AnytimePCA

unit-length vector. The algorithmic task is to recover $\mathcal{I}^*$, the support of $\mathbf{v}^*$, given n iid samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $\mathcal{N}(0, \Sigma)$. The SNR is governed both by β and k. The larger β and the smaller k the stronger the SNR and the easier the task.

Various efficient algorithms for sparse PCA were rigorously analyzed under different variants of the model just described, e.g. Amini and Wainwright (2009); Cai et al. (2013); Deshpande and Montanari (2016); Johnstone and Lu (2009); Krauthgamer et al. (2015); Shen et al. (2013); Wang et al. (2016). All these algorithms succeed in the regime where the sparsity level satisfies $k = \tilde{O}(\sqrt{\beta^2 n})$ (we use the $\tilde{O}(\cdot)$ notation in the common way, namely logarithmic factors are ignored). The best sparsity asymptotically is achieved for example by Covariance Thresholding (CT) (Deshpande and Montanari, 2016), remaining consistent up to $k_0 \asymp \sqrt{\beta^2 n}$ (the notation $f \asymp g$ stands for $f/g \to c$ for some constant $c > 0$).

It was further shown that under the planted clique hardness assumption there is no polynomial-time algorithm that asymptotically beats k_0 (Berthet and Rigollet, 2013a,b; Brennan and Bresler, 2019). Even without the planted clique assumption, Krauthgamer et al. (2015) show that the SDP relaxation suggested by d'Aspremont et al. (2004) and analyzed by Amini and Wainwright (2009), fails to recover $\mathbf{v}^*$ beyond k_0. The threshold k_0 is commonly referred to as the computational threshold, which we denote from now on by k_{comp}. We informally call the regime $k >> k_{comp}$ the weak SNR regime, and $k \leq k_{comp}$ the strong SNR regime. Finally, an information-limit was proven for $k \geq k_{info} \asymp \beta^2 n / \log p$ (Amini and Wainwright, 2009; Berthet and Rigollet, 2013a; Cai et al., 2015; Wang et al., 2014), and a matching algorithmic result for the naive exhaustive search (Berthet and Rigollet, 2013b; Brennan et al., 2018; Cai et al., 2013; Vu and Lei, 2012).

While the boundaries between the different SNR regimes are well understood, at least asymptotically, the following question remains open:

Question: what is the computational complexity required to find the support of $\mathbf{v}^*$ in the weak SNR regime, namely when $k_{comp} \leq k \leq k_{info}$?

The analysis of SSPCA provides an answer to this question (an upper bound).

4. Results

Our results refer to the following choice of hyper-parameters for SSPCA:

$$f_1 = f_{avg}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{i,j \in \mathcal{S}} \hat{\Sigma}_{i,j}, \qquad f_2 = f_{\lambda_1}(\mathcal{S}) = \lambda_1(\hat{\Sigma}_{\mathcal{S}}).$$

Furthermore, we assume the *uniform biased sparse PCA model* (UBSPCA), namely non-zero entries of $\mathbf{v}^*$ are all equal to $1/\sqrt{k}$. In the simulation part we lift the same-sign restriction and use the *uniform unbiased sparse PCA model* (USPCA), where entries equal $\pm 1/\sqrt{k}$.

The next proposition asserts the sufficiency of conditions C1, C2.

Proposition 1 ($\delta, \xi-$Sufficient Conditions) *If $\hat{\Sigma}$ is ξ-separable (Condition C2) and there exists a golden seed $\mathcal{S}_0$ of size k^* such that GreedySPCA$(X, k, \mathcal{S}_0)$ outputs a set $\mathcal{I}_0$ satisfying $|\mathcal{I}_0 \cap \mathcal{I}^*| \geq \delta k$ (Condition C1) then SSPCA outputs a set $\mathcal{I}$ satisfying $|\mathcal{I} \cap \mathcal{I}^*| \geq (\delta - \xi)k$.*

The proof of Proposition 1 is given in Section 5. Our next theorem establishes the scaling of k^* for the existence of a seed from which $\mathcal{I}^*$ is recovered exactly (condition C1 with $\delta = 1$).

Theorem 2 (Golden seed) *Let $\hat{\Sigma}$ be distributed according to the UBSPCA model. Assume that $p/n \to c \geq 0$, and $k \leq \frac{n \cdot \min\{\beta^2, \beta\}}{C \log n}$ for a sufficiently large universal constant C. If*

$$k^* \geq \left\lfloor \frac{Ck^2 \log n}{\beta^2 n} \right\rfloor \tag{2}$$

then w.p. tending to 1 as $(n, p, k) \to \infty$ there exists a seed $\mathcal{S}^* \subseteq \mathcal{I}^*$ of size at most k^* for which the output of GreedySPCA$(f_{avg}, \mathcal{S}^*, k)$ is $\mathcal{I}^*$.

Theorem 2 is proven in Section 6. The next theorem provides a general spectral separability property of spiked covariance matrices. It implies condition C2 as an immediate corollary.

Theorem 3 (Spectral Separation) *Let $\hat{\Sigma}$ be distributed according to the UBSPCA model with $p/n \to c > 0$. Set $\Gamma = C\left(\frac{(1+\beta)k \log n}{n}\right)^{0.5}$ for a suitably chosen constant C. With probability tending to 1 as $(n, p, k) \to \infty$, for every $\delta \in [0, 1]$ and for every set $\mathcal{I} \subseteq \{1, \ldots, p\}$ of size k that satisfies $|\mathcal{I} \cap \mathcal{I}^*| = \delta k$,*

$$\lambda_1(\hat{\Sigma}_{\mathcal{I}}) \in \left[1 + \delta\beta - \Gamma, 1 + \delta\beta + \Gamma + \frac{\beta}{k}\right].$$

Corollary 4 (Condition C2) *Under the conditions of Theorem 3, for every two sets $\mathcal{I}, \mathcal{J}$ of size k, if $|\mathcal{I} \cap \mathcal{I}^*| - |\mathcal{J} \cap \mathcal{I}^*| > \xi k$ then $\lambda_1(\hat{\Sigma}_{\mathcal{I}}) > \lambda_1(\hat{\Sigma}_{\mathcal{J}})$, for ξ that satisfies*

$$\xi = \tfrac{1}{k} + O\left(\sqrt{\frac{(1+\beta)k}{k_{info}}}\right), \tag{3}$$

where $k_{info} \asymp \beta^2 n / \log p$ was defined in Section. 3

Theorem 3 is proven in Section 7 and Corollary 4 is proven in Section 8. Corollary 4 with $\mathcal{I} = \mathcal{I}^*$ was already proven for example in Berthet and Rigollet (2013b); Brennan et al. (2018); Cai et al. (2013); Vu and Lei (2012) and in a more general sparse PCA model. Corollary 4 adds that even if the seed suffices to recover only part of $\mathcal{I}^*$, which might as well be the case in practice (finite problem size), SSPCA will nevertheless pick up this information. This point is demonstrated in Figure 1, where all executions of SSPCA end up with partial recovery.

The next theorem concludes our result and is an immediate corollary of all the statements in this section.

Theorem 5 *Under the conditions of Theorem 2, if k^* satisfies the lower bound in Eq. (2), then w.p. tending to 1 as $(n, p, k) \to \infty$, SSPCA$(f_{avg}, f_{\lambda_1}, k, k^*)$ recovers at least $(1 - O(\sqrt{\alpha}) - 1/k)$-fraction of $\mathcal{I}^*$, where $\alpha = (1+\beta)k/k_{info}$.*

For example, in the regime where DT and SDP fully recover $\mathcal{I}^*$, i.e. $k = O(k_{comp}/\sqrt{\log p})$, SSPCA requires a seed of size 0 to recover $\mathcal{I}^*$ and runs in time $O(p \log p)$. When $k \asymp k_{comp}$ the seed size is $k^* = O(\log n)$ and the running time is quasi-polynomial, $p^{O(\log n)}$. Simulations suggest that up to the computational threshold, even for a fairly large problem size ($n = p = 20,000$), it suffices to choose $k^* = 1$ to recover $\mathcal{I}^*$ exactly (see Figure 2). In the weak SNR regime, i.e. $k = n^{0.5+\varepsilon}$, the seed size scales as $n^{2\varepsilon} \log n$. This means that the computational effort is $\exp\{n^{2\varepsilon} \log n\}$, which

is exponential in $(k/\sqrt{n})^2$ (square the excess above the computational threshold) rather than in k itself (the naive exhaustive search approach). The results in Ding et al. (2019)[Thm 2.14] provide rigorous evidence that the exact scaling that we obtained for k^* in Eq. (2) is asymptotically optimal.

Simulation (Section 9) suggests that SSPCA succeeds also when the UBSPCA assumption is relaxed, namely the same-sign assumption is lifted. In this case, the best performance is achieved when the hyper-parameter f_{avg} is replaced with f_{ℓ_1}, which is the average row ℓ_1-norm rather than the average row sum. The parameter $f_2 = f_{\lambda_1}$ remains the same.

5. Proof of Proposition 1

Suppose by contradiction that conditions C1 and C2 hold but SSPCA outputs a set $\mathcal{J}$ for which $|\mathcal{J} \cap \mathcal{I}^*| < (\delta - \xi)k$. Consider a point in the execution of SSPCA where a golden seed $\mathcal{S}_0$ is explored. By Condition C1, GreedySPCA completes $\mathcal{S}_0$ to a set $\mathcal{I}$ satisfying $|\mathcal{I} \cap \mathcal{I}^*| \geq \delta k$. The latter together with the contradiction assumption give $|\mathcal{I} \cap \mathcal{I}^*| - |\mathcal{J} \cap \mathcal{I}^*| > \delta k - (\delta - \xi)k = \xi k$. In this case C2 guarantees that $\lambda_1(\hat{\Sigma}_{\mathcal{I}}) > \lambda_1(\hat{\Sigma}_{\mathcal{J}})$. Therefore the last line of SSPCA ensures that $\mathcal{J}$ cannot be the output of the algorithm.

6. Proof of Theorem 2

For convenience, let us assumes w.l.o.g. that the support of $\mathbf{v}^*$ is the first k variables, namely $\mathcal{I}^* = \{1, \ldots, k\}$. Our candidate for a golden seed is any subset $\mathcal{S}^*$ of $\mathcal{I}^*$. For concreteness we fix $\mathcal{S}^* = \{1, \ldots, k^*\}$. We show that when GreedySPCA is called with this subset, then the $k - k^*$ variables that it adds to $\mathcal{S}^*$ in line 6, all belong to $\{1, \ldots, k\}$, thus outputting $\mathcal{I}^*$.

We begin by writing the distribution of the i^{th} sample from $\mathcal{N}(0, \beta \mathbf{v}^* \mathbf{v}^{*T} + I_p)$ explicitly as

$$\mathbf{x}_i = \sqrt{\beta} u_i \, \mathbf{v}^* + \boldsymbol{\xi}_i, \tag{4}$$

where $\boldsymbol{\xi}_i \in \mathbb{R}^p$ is a noise vector whose entries are all i.i.d. $\mathcal{N}(0, 1)$, and $u_i \sim \mathcal{N}(0, 1)$. Furthermore, all the u_i's and $\boldsymbol{\xi}_i$'s are independent of each other.

By the greedy rule in line 3 of GreedySPCA, the $k - k^*$ variables in $\{1, \ldots, p\} \setminus \mathcal{S}^*$ that will be chosen are those with largest value of $f_{avg}(\mathcal{S}^* \cup \{i\})$. We rewrite $f_{avg}(\mathcal{S}^* \cup \{i\})$ as

$$f_{avg}(\mathcal{S}^* \cup \{i\}) = \tfrac{1}{k^*+1} \sum_{s,t \in \mathcal{S}^* \cup \{i\}} \hat{\Sigma}_{s,t} = \tfrac{k^*}{k^*+1} f_{avg}(\mathcal{S}^*) + \tfrac{2}{k^*+1} \underbrace{\left(\sum_{s \in \mathcal{S}^*} \hat{\Sigma}_{is} \right)}_{c_i(\mathcal{S}^*)} + \tfrac{1}{k^*+1} \hat{\Sigma}_{ii} := \tag{5}$$

$$:= \tfrac{k^*}{k^*+1} f_{avg}(\mathcal{S}^*) + \tfrac{1}{k^*+1} \left(2c_i(\mathcal{S}^*) + \hat{\Sigma}_{ii} \right).$$

The only part in Eq. (5) that depends on i is its total covariance with $\mathcal{S}^*$ (which we denote by c_i), and its variance $\hat{\Sigma}_{ii}$. In high level, the algorithm succeeds since c_i is much bigger than c_j for all pairs $i \in \mathcal{I}^*, j \notin \mathcal{I}^*$. We now turn to make this argument formal.

Lemma 6 *Under the conditions of Theorem 2, for a fixed $\mathcal{S}^* \subseteq \mathcal{I}^*$ of size k^*, w.p. at least $1 - 1/n$, every $i \in \mathcal{I}^*$ satisfies $c_i \geq 0.4\beta k^*/k$.*

Lemma 7 *Under the conditions of Theorem 2, for a fixed $\mathcal{S}^* \subseteq \mathcal{I}^*$ of size k^*, w.p. at least $1-1/n$, every $j \notin \mathcal{I}^*$ satisfies $c_j \leq 0.3\beta k^*/k$.*

Lemma 8 *Under the conditions of Theorem 2, with probability at least $1-1/n$, for every $i \in \mathcal{I}^*, j \notin \mathcal{I}^*$, $\hat{\Sigma}_{jj} - \hat{\Sigma}_{ii} \leq 0.1\beta k^*/k$.*

We use Lemmas 6, 7 and 8 to complete the proof of the theorem. Let $\Delta_{ij} = f_{avg}(\mathcal{S}^* \cup \{i\}) - f_{avg}(\mathcal{S}^* \cup \{j\})$. To prove that GreedySPCA outputs $\mathcal{I}^*$ we need to show that $\Delta_{ij} > 0$ for every $i \in \mathcal{I}^*, j \notin \mathcal{I}^*$. Using Lemmas 6–8, we have $k \cdot \Delta_{ij} \geq 2(0.4\beta k^*/k - 0.3\beta k^*/k) - 0.1\beta k^*/k \geq 0.1\beta k^*/k > 0$. In the last inequality we assumed $k^* \geq 1$.

The case $k^* = 0$ corresponds to the regime $k = O(k_{comp}/\sqrt{\log p})$. In this regime the k largest diagonal entries belong to $\mathcal{I}^*$ (Johnstone and Lu, 2009). Indeed, when $k^* = 0$ then $f_{avg}(\{i\})$ is simply $\hat{\Sigma}_{ii}$, and GreedySPCA is no other than Diagonal Thresholding.

We turn to prove Lemmas 6–8. In the proof we use the following two auxiliary facts. The first is a large deviation result for a Chi-square random variable.

Lemma 9 *(Laurent and Massart (2000)) Let $X \sim \chi^2_n$. For all $x \geq 0$,*

$$Pr[X \geq n + 2\sqrt{nx} + x] \leq e^{-x}, \qquad Pr[X \leq n - 2\sqrt{nx}] \leq e^{-x}.$$

The second fact records a well-known argument about the inner-product of two multivariate Gaussians. Its short proof can be found in Appendix C.

Lemma 10 *Let $\{x_i, y_i\}_{i=1}^n$ be standard i.i.d. Gaussian random variables. Then $\sum_{i=1}^n x_i y_i$ is distributed like the product of two independent random variables $\|\mathbf{x}\| \cdot \tilde{y}$, where $\mathbf{x} = (x_1, \ldots, x_n)$, $\|\mathbf{x}\|^2 \sim \chi^2_n$ and $\tilde{y} \sim \mathcal{N}(0,1)$.*

6.1. Proof of Lemma 6

We start by explicitly writing $c_i(\mathcal{S}^*)$ from Eq. (5) for a fixed set $\mathcal{S}^*$ of size k^*. Let $\mathbf{r}^{(i)}$ denote the i^{th} row of the $p \times n$ design matrix X. For every candidate $i \notin \mathcal{S}^*$,

$$c_i(\mathcal{S}^*) = \sum_{j \in \mathcal{S}^*} \hat{\Sigma}_{ij} = \tfrac{1}{n} \sum_{j \in \mathcal{S}^*} \mathbf{r}^{(i)} \cdot (\mathbf{r}^{(j)})^T = \tfrac{1}{n} \mathbf{r}^{(i)} \left(\sum_{j \in \mathcal{S}^*} \mathbf{r}^{(j)} \right)^T .$$

Following the distribution rule of X given in Eq. (4), all entries of the vector $\mathbf{s} = \sum_{j \in \mathcal{S}^*} \mathbf{r}^{(j)}$ are i.i.d. with

$$\mathbf{s}_\ell \sim \sqrt{\beta} u_\ell \sum_{j \in \mathcal{S}^*} \mathbf{v}^*_j + \sqrt{k^*} w_\ell,$$

where $u_\ell \sim \mathcal{N}(0,1)$ is defined in Eq. (4) and $w_\ell \sim \mathcal{N}(0,1)$ independently of u_ℓ ($\sqrt{k^*} w_\ell$ is derived from $\sum_{j \in \mathcal{S}^*} (\boldsymbol{\xi}_\ell)_j$). The product $\mathbf{r}^{(i)} \mathbf{s}^T$ is distributed as

$$\mathbf{r}^{(i)} \mathbf{s}^T \sim \frac{1}{n} \sum_{\ell=1}^{n} \left(\sqrt{\beta} u_\ell \mathbf{v}^*_i + y_\ell \right) \left(\sqrt{\beta} u_\ell \left(\sum_{j \in \mathcal{S}^*} \mathbf{v}^*_j \right) + \sqrt{k^*} w_\ell \right) \tag{6}$$

The variable $y_\ell = (\boldsymbol{\xi}_\ell)_i \sim \mathcal{N}(0,1)$. We rearrange the sum as four components, corresponding to the pure signal part, cross noise-signal and pure noise:

$$\sum_{\ell=1}^{n} \sqrt{\beta} u_\ell \mathbf{v}_i^* \cdot \sqrt{\beta} u_\ell \sum_{j \in \mathcal{S}^*} \mathbf{v}_j^* = \frac{\beta k^* \mathbf{v}_i^*}{\sqrt{k}} \sum_{\ell=1}^{n} u_\ell^2, \tag{7}$$

$$\sum_{\ell=1}^{n} \sqrt{\beta} u_\ell \mathbf{v}_i^* \cdot \sqrt{k^*} w_\ell = \sqrt{\beta k^*} \mathbf{v}_i^* \sum_{\ell=1}^{n} u_\ell w_\ell, \tag{8}$$

$$\sum_{\ell=1}^{n} y_\ell \sqrt{\beta} u_\ell \sum_{j \in \mathcal{S}^*} \mathbf{v}_j^* = \frac{\sqrt{\beta} k^*}{\sqrt{k}} \sum_{\ell=1}^{n} y_\ell u_\ell, \tag{9}$$

$$\sum_{\ell=1}^{n} y_\ell \sqrt{k^*} w_\ell = \sqrt{k^*} \sum_{\ell=1}^{n} y_\ell w_\ell. \tag{10}$$

We now bound each term separately. To lower bound Eq. (7), we use the fact that $\sum_{\ell=1}^{n} u_\ell^2$ in Eq. (7) is distributed χ_n^2. The second inequality in Lemma 9 with $x = 0.05n$ gives $Pr[\chi_n^2 \leq 0.8n] \leq e^{-n/100}$. Therefore, w.p. at least $1 - e^{-n/100}$,

$$\tfrac{1}{n}(7) \geq 0.8\beta k^*/k \tag{11}$$

Moving to Eq. (8), according to Lemma 10, the product term in Eq. (8) is distributed as $\sqrt{\chi_n^2}\mathcal{N}(0,1)$. For $\frac{1}{n}(8) > 0.1\beta k^*/k$ to hold, the following has to happen,

$$\sqrt{\chi_n^2}|\mathcal{N}(0,1)| > \frac{\beta n \sqrt{k^*}}{10k} = \frac{\beta n \sqrt{k^*}}{30k\sqrt{\log n}} \cdot \sqrt{9 \log n}.$$

Using standard tail-bounds for Gaussians, $Pr[|\mathcal{N}(0,1)| > \sqrt{9 \log n}] \leq n^{-4}$. Next we bound $Pr[\chi_n^2 \geq \beta^2 n^2 k^*/(900k^2 \log n)]$. Substituting the value of k^* from Eq. (2) we have that

$$\frac{\beta^2 n^2 k^*}{900k^2 \log n} \geq \frac{\beta^2 n^2}{900k^2 \log n} \cdot \frac{Ck^2 \log n}{\beta^2 n} = Cn/900.$$

Choosing $C \geq 1800$ for example and using Lemma 10 gives $Pr[\chi_n^2 \geq 2n] \leq e^{-n/4}$. To conclude, w.p. at least $1 - n^{-4} - e^{-n/4}$ we get

$$\tfrac{1}{n}|(8)| \leq \frac{0.1\beta k^*}{k} \tag{12}$$

Moving to Eq. (9), according to Lemma 10, the sum-product term in Eq. (9) is distributed as $\sqrt{\chi_n^2}\mathcal{N}(0,1)$. Using standard tail-bounds for Gaussians, $Pr[|\mathcal{N}(0,1)| \geq \sqrt{6 \log n}] \leq 2n^{-3}$, and according to Lemma 9, $Pr[\chi_n^2 \geq 2n] \leq e^{-n/4}$. Therefore w.p. at least $1 - 2n^{-3} - e^{-n/4}$ we get

$$\tfrac{1}{n}(9) \leq \frac{1}{n} \frac{\sqrt{\beta} k^*}{\sqrt{k}} \sqrt{6 \log n} \sqrt{2n} \leq \frac{0.1\beta k^*}{k} \tag{13}$$

The last inequality is true when $k \leq \beta n/(1200 \log n)$, which holds by our choice of k.

Moving to Eq. (10), we similarly have that w.p. at least $1 - 2n^{-3} - e^{-n/4}$

$$\tfrac{1}{n}(10) \leq \frac{1}{n} \sqrt{k^*} \sqrt{6 \log n} \sqrt{2n} \leq \frac{0.2k^*\beta}{k}. \tag{14}$$

The last inequality in Eq. (14) holds whenever $k^* \geq 200k^2 \log n/(\beta^2 n)$, which is what Eq. (2) says.

Finally, the lower bound on k^* in Eq. (2) makes sense as long as $k^* \leq k$, which translates to requiring $k \leq \beta^2 n/(C \log n)$.

To conclude, w.p. at least $1 - 3n^{-3}$, for a fixed $i \in \mathcal{I}^*$,

$$c_i \geq (11) - (12) - (13) - (14) \geq \frac{0.4k^*\beta}{k}.$$

The lemma now follows from taking the union bound over the $k - k^* \leq p$ indices in $\mathcal{I}^* \setminus \mathcal{S}^*$, together with the fact that $p = O(n)$.

7. Proof outline of Theorem 3

We provide the general outline of the proof. The complete proof is given in Appendix D. Fix a set $\mathcal{I} \subseteq \{1, \ldots, p\}$ s.t. $|\mathcal{I} \cap \mathcal{I}^*| = \delta k$. The matrix $\hat{\Sigma}_{\mathcal{I}}$ can be written as $\hat{\Sigma}_{\mathcal{I}} = N + S$ where N is composed of the noise part, Eq. (10), and S is composed of the signal and noise-signal cross terms, Eq. (7)–(9). N is easily seen to be symmetric, and in fact it follows a Wishart distribution. The matrix $S = \hat{\Sigma}_{\mathcal{I}} - N$ is the difference of two symmetric matrices, hence symmetric as well. The proof follows Weyl's inequality for Hermitian matrices,

$$\lambda_k(N) + \lambda_1(S) \leq \lambda_1(\hat{\Sigma}_{\mathcal{I}}) \leq \lambda_1(N) + \lambda_1(S).$$

The bound on the Wishart part, $\lambda_1(N)$ and $\lambda_k(N)$, is taken from (Davidson and Szarek, 2001, Theorem II.13). To upper bound $\lambda_1(S)$ we use Gershgorin's circle theorem, which says that every eigenvalue λ of an $n \times n$ matrix A satisfies at least one of the n inequalities for $i = 1, \ldots, n$,

$$|\lambda - A_{ii}| \leq \sum_{j \neq i} |A_{ij}|.$$

To ower bound $\lambda_1(S)$ we use its Rayleigh quotient definition: $\lambda_1(S)$ is the argmax of $\mathbf{x}^T S \mathbf{x}$ over all unit vectors $\mathbf{x} \in \mathbb{R}^k$. In particular, for $\mathbf{x}_0 = (\delta k)^{-0.5} \mathbf{1}_{\mathcal{I} \cap \mathcal{I}^*}$ ($\mathbf{1}_Q$ is the characteristic vector of a set Q), the value of $\mathbf{x}_0^T S \mathbf{x}_0$ is a lower bound on $\lambda_1(S)$. The latter is simply the average row sum in the $\delta k \times \delta k$ submatrix $S_{\mathcal{I} \cap \mathcal{I}^*}$.

The computations that lead to the lower and upper bounds on $\lambda_1(S)$ are similar to those in the proof of Theorem 2.

8. Proof outline of Corollary 4

Take $\mathcal{I}, \mathcal{J} \subseteq \{1, \ldots, p\}$ that satisfy $|\mathcal{I} \cap \mathcal{I}^*| = \delta_1 > \delta_2 = |\mathcal{J} \cap \mathcal{I}^*|$. According to Theorem 3, for $\lambda_1(\hat{\Sigma}_{\mathcal{I}}) > \lambda_1(\hat{\Sigma}_{\mathcal{J}})$ to hold, it suffices to require

$$1 + \delta_2 \beta + \frac{\beta}{k} + \Gamma < 1 + \delta_1 \beta - \Gamma.$$

Rearranging we get,

$$\frac{2\Gamma}{\beta} + \frac{1}{k} < \delta_1 - \delta_2 := \xi.$$

The corollary follows immediately from the definition of Γ and k_{info}.

9. Simulations

We turn to evaluate the performance of SSPCA both in the strong and weak SNR regimes. The following points summarize the way we ran simulations:

- The spike follows the UNBSPCA distribution: $\mathbf{v}^* = \left(\pm\frac{1}{\sqrt{k}}, \ldots, \pm\frac{1}{\sqrt{k}}, 0, \ldots, 0\right)$, where the signs of non-zero entries are randomly chosen. Accordingly, we change the choice of f_1 in GreedySPCA from f_{avg} to f_{ℓ_1}, which measures the ℓ_1 norm of $\hat{\Sigma}_{S^*\cup\{i\}}$, rather than the sum. Furthermore, in the weak SNR regime, it makes sense to ignore the diagonal of $\hat{\Sigma}$. Therefore we only use c_i from Eq. (5) to choose i.

- We keep $p = n$ and fix $\beta = 0.5$. The choice of 0.5 is somewhat arbitrary and any value below $\sqrt{p/n} = 1$ is suitable. When β exceeds $\sqrt{p/n}$ the problem is computationally easy for all k up to the information limit (Krauthgamer et al., 2015)[Thm 1.1].

- The *success rate* of an algorithm on a given input is defined to be $\frac{1}{k}|\mathcal{I} \cap \mathcal{I}^*|$, where $\mathcal{I} \subseteq \{1, \ldots, p\}$ is the algorithm's guess of $\mathbf{v}^*$'s support.

- The algorithm DT has no tunable parameters – it simply returns the indices of the k largest diagonal entries. The performance of CT, on the other hand, depends crucially on the chosen threshold. When running CT we loop over 50 thresholds, the empirical percentiles of the off-diagonal entries of the input covariance matrix. We choose the best result as CT's output. Also, the output of CT is a vector (a guess for $\mathbf{v}^*$). We convert the vector to a set $\mathcal{I}$ by taking the indices of the k largest entries in absolute value.

- We compared SSPCA against the well-known sparse PCA algorithm described in Zou et al. (2006). This algorithm casts PCA as a regression problem and uses both ridge and lasso penalties. LARS (Efron et al., 2004) is then used to obtain the optimal solution. We ran Python's implementation of this algorithm using a grid search for the ridge and lasso penalties in the rectangle $[0, 2] \times [0, 2]$, discretized to 100 equidistant points. The algorithm is in module `sklearn.decomposition.SparsePCA` (Pedregosa et al., 2011).

Figure 1 compares the performance of all aforementioned algorithms in a certain weak SNR configuration, $n = p = 1000, k = 8, \beta = 0.5$. Among the polynomial-time algorithms (we include in this category SSPCA with $k^* = 1, 2$), SSPCA with $k^* = 2$ performs best. When running for super polynomial-time, SSPCA with $k^* = 3$ is superior to the naive exhaustive search, when both are given the same time budget T (3 parallel-hours on 90 cores).

Our next experiment demonstrates the existence of golden seeds in the weak SNR regime. The empirical boundary of the strong/weak SNR regime is charted by the success rate curve of CT. Figure 2 shows the performance (y-axis) of CT as k increases. Three configurations are plotted $n = p = 10,000, 15,000, 20,000$. The x-axis is scaled by $\sqrt{n}$ to defuse the dependence on n. Indeed all three CT lines overlap as expected (due to scaling), and the phase transition to the hard regime occurs when k is in the window $[0.2\sqrt{n}, 0.3\sqrt{n}]$. The plot also includes the performance of DT, lagging behind, and GreedySPCA initialized once with a seed of size $k^* = 1$ and second with $k^* = k/3$. In both cases the seed is a random subset of $\mathcal{I}^*$.

As evident from Figure 2, the performance of GreedySPCA with seeds of size $k^* = 1$ is similar to CT. This is somewhat surprising when comparing to the asymptotic lower bound given by Eq. (2),

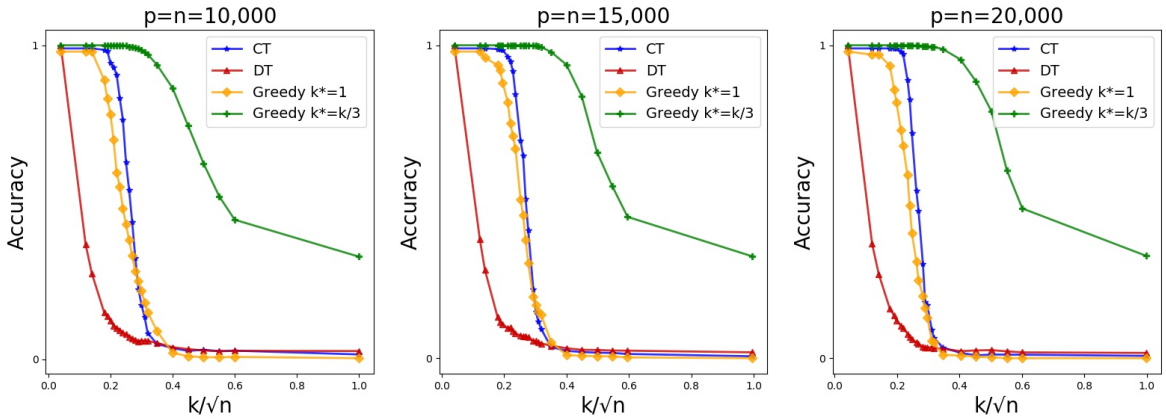

Figure 2: The success rate of DT, CT and GreedySPCA as a function of k. Every point is an average of 25 executions, with $n = p$ samples. GreedySPCA was initialized with $k^* = 1$ or $k^* = k/3$ random entries from $\mathcal{I}^*$.

which would be of order $\log n$ at the computational threshold k_{comp}. The right-most (green) line in Figure 2 extends with an accuracy of roughly 100% into the weak SNR regime, thus showing the existence of golden seeds in that regime.

10. Discussion

In this paper, we presented a family of anytime algorithms for the k-sparse PCA problem, which follow the same simple white-box greedy template that we called GreedySPCA and SSPCA.

GreedySPCA performs a bulk greedy choice, and instead we could have grown the solution iteratively, adding in iteration $r = 1, \ldots, k-k^*$ the variable i_r which maximizes $f_1(\mathcal{S}^* \cup \{i_1, \ldots, i_{r-1}\})$. The iterative variant is exactly the well-known greedy algorithm of Nemhauser, Wolsey and Fisher, which was proposed for sub-modular function optimization (Nemhauser et al., 1978). The only difference is that Nemhauser et al. start with an empty seed.

Nemhauser et al. proved that if f_1 is sub-modular and monotone, then the iterative greedy algorithm finds a solution which is a $(1 - \frac{1}{e})$-approximation of the optimum. However, the $(1 - \frac{1}{e})$-approximation ratio is useless in many cases, as the guaranteed value is lower than a random solution (see Theorem 3). Our proof shows that GreedySPCA recovers $\mathcal{I}^*$ exactly when called with the right seed without the sub-modularity assumption on f_1. Simulations that we ran in the spiked covariance model with the iterative version (with seed) resulted in very similar performance compared to the bulk version.

Acknowledgment

This project was supported by ISF grant number 1388/16. We thank Jonathan Rosenblatt and ISF grant 924/16 for the computing equipment required for this research.

References

A. Amini and M. Wainwright. High dimensional analysis of semidefinite relaxations for sparse principal component analysis. *Annals of Statistics*, 37(5B):2877–2921, 2009. doi: 10.1214/08-AOS664.

T.W. Anderson. *An introduction to multivariate statistical analysis*. Wiley series in probability and mathematical statistics. Wiley, 2nd edition, 1984. ISBN 0471889873.

M. Asteris, D. S. Papailiopoulos, and G. N. Karystinos. Sparse principal component of a rank-deficient matrix. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 673–677, 2011.

M. Asteris, D. Papailiopoulos, A. Kyrillidis, and Alexandros Dimakis. Sparse pca via bipartite matchings. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 766–774. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5901-sparse-pca-via-bipartite-matchings.pdf.

M. Baback, Y. Weiss, and S. Avidan. Spectral bounds for sparse pca: Exact and greedy algorithms. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 915–922. MIT Press, 2006. URL http://papers.nips.cc/paper/2780-spectral-bounds-for-sparse-pca-exact-and-greedy-algorithms.pdf.

Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *Annals of Statistics*, 41(4):1780–1815, 08 2013a. doi: 10.1214/13-AOS1127.

Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT*, pages 1046–1066, 2013b.

J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36:199–227, 2008. doi: 10.1214/009053607000000758.

M. Brennan and G. Bresler. Optimal average-case reductions to sparse pca: From weak assumptions to strong hardness. In *COLT*, 02 2019.

M. Brennan, G. Bresler, and W. Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 48–166. PMLR, 06–09 Jul 2018.

T. Cai, Z. Ma, and Y. Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013. doi: 10.1214/13-AOS1178.

T. Cai, Z. Ma, and Y. Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, 161(3-4):781–815, 4 2015. ISSN 0178-8051. doi: 10.1007/s00440-014-0562-z.

A. d'Aspremont, L. El-Ghaoui, M. Jordan, and G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2004. doi: 10.1137/050645506.

A. d'Aspremont, F. Bach, and L. El-Ghaoui. Optimal solutions for sparse principal component analysis. *J. Mach. Learn. Res.*, 9:12691294, June 2008. ISSN 1532-4435.

K. Davidson and S. Szarek. Local operator theory, random matrices and Banach spaces. In Lindenstrauss, editor, *Handbook on the Geometry of Banach spaces*, volume 1, pages 317–366. Elsevier Science, 2001.

Y. Deshpande and A. Montanari. Sparse pca via covariance thresholding. *J. Mach. Learn. Res.*, 17 (1):4913–4953, January 2016. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2946645.3007094.

Y. Ding, D. Kunisky, A. Wein, and A. Bandeira. Subexponential-time algorithms for sparse pca, 2019. URL https://arxiv.org/abs/1907.11635.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 4 2004. doi: 10.1214/009053604000000067. URL https://doi.org/10.1214/009053604000000067.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29:295–327, 2001. doi: 10.1214/aos/1009210544.

I. M. Johnstone and A. Lu. On Consistency and Sparsity for Principal Components Analysis in High Dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009. doi: 10.1198/jasa.2009.0121.

I.T. Jolliffe. *Principal Component Analysis*. Springer series in statistics. Springer, 2nd edition, 2002.

R. Krauthgamer, B. Nadler, and D. Vilenchik. Do semidefinite relaxations solve sparse pca up to the information limit? *Ann. Statist.*, 43(3):1300–1322, 06 2015. doi: 10.1214/15-AOS1310. URL https://doi.org/10.1214/15-AOS1310.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

B. Nadler. Finite sample approximation results for principal component analysis: a matrix perturbation approach. *Annals of Statistics*, 36:2791–2817, 2008. doi: 10.1214/08-AOS618.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions–i. *Math. Program.*, 14(1):265–294, December 1978. ISSN 0025-5610. doi: 10.1007/BF01588971. URL https://doi.org/10.1007/BF01588971.

D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17:1617–1642, 2007.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

D. Shen, H. Shen, and J.S. Marron. Consistency of sparse pca in high dimension, low sample size contexts. *Journal of Multivariate Analysis*, 115:317 – 333, 2013. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2012.10.007. URL http://www.sciencedirect.com/science/article/pii/S0047259X12002308.

V. Vu and J. Lei. Minimax rates of estimation for sparse pca in high dimensions. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1278–1286, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

T. Wang, Q. Berthet, and R. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. http://arxiv.org/abs/1408.5369, August 2014.

T. Wang, Q. Berthet, and R. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.*, 44(5):1896–1930, 10 2016. doi: 10.1214/15-AOS1369. URL https://doi.org/10.1214/15-AOS1369.

S. Zilberstein. Using anytime algorithms in intelligent systems. *AI Magazine*, 17(3):73, Mar. 1996. doi: 10.1609/aimag.v17i3.1232. URL https://www.aaai.org/ojs/index.php/aimagazine/article/view/1232.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006. doi: 10.1198/106186006X113430. URL https://doi.org/10.1198/106186006X113430.

Appendix A. Proof of Lemma 7

For $i \notin \mathcal{I}^*$ the terms in Eq. (7) and (8) are 0 (because $\mathbf{v}_i^* = 0$). The proof of Lemma 7 is identical to the proof leading to the bound on Eq. (9) given in Eq. (13) and to the bound on Eq. (10) given in Eq. (14). A union bound is then taken over the at most p variables in $\{1, \ldots, p\} \setminus \mathcal{I}^*$.

Appendix B. Proof of Lemma 8

For $j \notin \mathcal{I}^*$, the distribution of $\hat{\Sigma}_{jj} \sim \frac{\chi_n^2}{n}$ (From Eq. (6)). Lemma 9 entails that for a fixed j, w.p. at least $1 - n^{-3}$, $\hat{\Sigma}_{jj} \leq 1 + \sqrt{\frac{9 \log n}{n}}$.
For $i \in \mathcal{I}^*$,

$$\hat{\Sigma}_{ii} \sim \frac{\beta}{k} \frac{\chi_n^2}{n} + 2 \frac{\sqrt{\beta}}{\sqrt{k}} \frac{\mathcal{N}(0,1)\sqrt{\chi_n^2}}{n} + \frac{\chi_n^2}{n}$$

Using Lemma 9 and standard tails on the Gaussian, we obtain that w.p. at least $1 - O(n^{-3})$, $\hat{\Sigma}_{ii} \geq 1 + \frac{\beta}{k} - \sqrt{\frac{36 \log n}{n}}$. Using the union bound we get that w.p. at least $1 - O(n^{-1})$, for every

pair $i \in \mathcal{I}^*, j \notin \mathcal{I}^*$,

$$\hat{\Sigma}_{jj} - \hat{\Sigma}_{ii} \leq \sqrt{\frac{100 \log n}{n}} - \frac{\beta}{k} \leq \sqrt{\frac{100 \log n}{n}} \leq \frac{0.1\beta k^*}{k}.$$

The last inequality holds if $k^* \geq \sqrt{L}$, where L is the lower bound on k^* in Eq. (2). However, $k^* \geq L$ implies $k^* \geq \sqrt{L}$ since k^* is an integer.

Appendix C. Proof of Lemma 10

For every fixed realization of $\mathbf{x}$, we have $x_i y_i \sim \mathcal{N}(0, x_i^2)$ and by the independence of the y_i's,

$$\sum_{i=1}^{n} x_i y_i \sim \mathcal{N}(0, \|\mathbf{x}\|^2) = \|\mathbf{x}\| \cdot \mathcal{N}(0,1) := \|\mathbf{x}\| \cdot \tilde{y}.$$

The lemma follows by observing that $\|\mathbf{x}\|^2 \sim \chi^2_n$.

Appendix D. Proof of Theorem 3

We prove that w.p. tending to 1 as $(n, p, k) \to \infty$, for every set $\mathcal{I} \subseteq \{1, \ldots, p\}$ of size k that satisfies $|\mathcal{I} \cap \mathcal{I}^*| = \delta k$, for every $\delta \in [0, 1]$:

$$\lambda_1(\hat{\Sigma}_S) \in [1 + \delta\beta - (1 + 2\sqrt{\beta})\Phi, 1 + \delta\beta + (1 + 2\sqrt{\beta})\Phi + \frac{\beta}{k}], \tag{15}$$

Where $\Phi = \sqrt{\frac{8k \log n}{n}}$.

Fix a set $\mathcal{I} \subseteq \{1, \ldots, p\}$ s.t. $|\mathcal{I} \cap \mathcal{I}^*| = \delta k$. The matrix $\hat{\Sigma}_{\mathcal{I}}$ can be written as $\hat{\Sigma}_{\mathcal{I}} = N + S$ where N is composed of the noise part, Eq. (10), and S is composed of the signal and noise-signal cross terms, Eq. (7)–(9). N is easily seen to be symmetric (in fact it follows a Wishart distribution), and therefore the matrix $S = \hat{\Sigma}_{\mathcal{I}} - N$, the difference of two symmetric matrices, is symmetric as well. Weyl's inequality, applicable for Hermitian matrices, implies that

$$\lambda_k(N) + \lambda_1(S) \leq \lambda_1(\hat{\Sigma}_{\mathcal{I}}) \leq \lambda_1(N) + \lambda_1(S) \tag{16}$$

D.1. Bounding $\lambda_1(N)$ and $\lambda_k(N)$

The matrix $N \in \mathbb{R}^{k \times k}$ follows a Wishart distribution, and by (Davidson and Szarek, 2001, Theorem II.13),

$$\Pr[\lambda_1(N) \geq (1 + \sqrt{k/n} + t)^2 \vee \lambda_k(N) \leq (1 - \sqrt{k/n} - t)^2] \leq e^{-nt^2/2}.$$

Plugging in $t = \sqrt{6k \log n / n}$ we obtain that w.p. at least $1 - n^{-3k}$,

$$\lambda_1(N) \leq \left(1 + \sqrt{\frac{k}{n}} + \sqrt{\frac{6k \log n}{n}}\right)^2 \leq 1 + \sqrt{\frac{8k \log n}{n}} = 1 + \Phi. \tag{17}$$

and similarly,

$$\lambda_k(N) \geq 1 - \Phi. \tag{18}$$

Taking the union bound over all $\binom{p}{k} \leq p^k$ possible sub-matrices N, the bounds hold w.p. at least $1 - n^{-3k} p^k \geq 1 - n^{-1}$.

D.2. Upper Bounding $\lambda_1(S)$

Recall the parameters $\mathbf{u} = (u_1, \ldots, u_n)$ and $\boldsymbol{\xi}_i$ from definition of the single-spike distribution given in Eq. (4). We start with a certain property of $\hat{\Sigma}$ that we require during the proof of the upper and lower bound on $\lambda_1(S)$. We say that $\hat{\Sigma}$ is *typical* if $\|\mathbf{u}\|^2 \leq n + 6\sqrt{n \log n}$ and if $(\boldsymbol{\xi}_1)_i \leq 2\sqrt{\log n}$ for every $i = 1, \ldots, p$. Lemmas 9 and 10 guarantee that $\hat{\Sigma}$ is typical w.p. at least $1 - n^{-1}$. In what follows we condition on this fact.

To upper bound the largest eigenvalue of S we use Gershgorin's circle theorem, which says that every eigenvalue λ of an $n \times n$ matrix A satisfies at least one of the n inequalities for $i = 1, \ldots, n$,

$$|\lambda - A_{ii}| \leq \sum_{j \neq i} |A_{ij}|. \tag{19}$$

Each inequality defines a Gershgorin's disc, and every λ belongs to at least one disc. We next show that all discs are almost identical, and evaluate their center and radius.

Decompose each entry S_{ij} according to the three sums Eq. (7)-(9) (plugging $k^* = 1$). To bound the sums in Eq.(8) and (9) we note that both involve the term u_ℓ, which does not depend on i or j. Therefore we may rotate the distribution to point in the direction of $\mathbf{u}$. According to Lemma 10, the sum-product (8) is then distributed $\|\mathbf{u}\| \cdot (\boldsymbol{\xi}_1)_i$ and Eq.(9) is distributed $\|\mathbf{u}\| \cdot (\boldsymbol{\xi}_1)_j$. Using the fact that $\hat{\Sigma}$ is typical we obtain the following bounds:

$$\tfrac{1}{n}|(8)| \sim \tfrac{1}{n}\sqrt{\beta}\mathbf{v}_i^* \|\mathbf{u}\| |(\boldsymbol{\xi}_1)_j| \leq \sqrt{\frac{8\beta \log n}{nk}}.$$

Similarly

$$\tfrac{1}{n}|(9)| \leq \sqrt{\frac{8\beta \log n}{nk}}, \qquad \tfrac{1}{n}|(7)| = \left(1 \pm \sqrt{\frac{36 \log n}{n}}\right)\frac{\beta}{k}.$$

Putting everything together, and letting $\delta_i = 1$ if $i \in \mathcal{I}^*$ and 0 otherwise, if $\hat{\Sigma}$ is typical then for every $i, j \in \mathcal{I}$,

$$S_{ij} = \delta_i \delta_j \frac{\beta}{k} + \Delta_{ij}, \qquad |\Delta_{ij}| \leq \Delta := \sqrt{\frac{36\beta \log n}{nk}}. \tag{20}$$

To bound the radius of the i^{th} disc, $\sum_j |S_{ij}|$, we need to account for $|\mathcal{I} \cap \mathcal{I}^*| = \delta k$ indices $j \in \mathcal{I}^*$ and $(1 - \delta)k$ indices $j \notin \mathcal{I}^*$. Plugging (20) in (19), we obtain that

$$|\lambda - S_{ii}| \leq \delta k \left(\frac{\beta}{k} + \Delta\right) + (1 - \delta)k \cdot \Delta = \delta\beta + \Delta k.$$

Rearranging we get

$$\lambda_1(S) \leq S_{ii} + \delta\beta + \Delta k \leq \frac{\beta}{k} + \Delta + \delta\beta + \Delta k \leq \delta\beta + \left(\frac{\beta}{k} + 2\Delta k\right). \tag{21}$$

D.3. Lower Bounding $\lambda_1(S)$

To lower bound the largest eigenvalue of S we use the Rayleigh quotient definition, namely $\lambda_1(S)$ is the argmax of $\mathbf{x}^T S \mathbf{x}$ over all unit vectors $\mathbf{x} \in \mathbb{R}^k$. In particular, for $\mathbf{x}_0 = (\delta k)^{-0.5} \mathbf{1}_{\mathcal{I} \cap \mathcal{I}^*}$ ($\mathbf{1}_Q$ is the characteristic vector of a set Q), the value of $\mathbf{x}_0^T S \mathbf{x}_0$ is a lower bound on $\lambda_1(S)$. The latter

is simply the average row sum in the $\delta k \times \delta k$ submatrix $S_{\mathcal{I}\cap\mathcal{I}^*}$. If $\hat{\Sigma}$ is typical, then according to Eq. (20),

$$\lambda_1(S) \geq \delta k \cdot \left(\frac{\beta}{k} - \Delta\right) \geq \delta\beta - \Delta k. \tag{22}$$

To conclude the proof of the theorem, note that $\Delta k \leq 3\sqrt{\beta}\Phi$. Putting Equations (17),(18),(21),(22) together, we get that w.p. at least $1 - 2n^{-1}$,

$$\lambda_1(\hat{\Sigma}_S) \in [1 + \delta\beta - (1 + 3\sqrt{\beta})\Phi, 1 + \delta\beta + (1 + 3\sqrt{\beta})\Phi + \frac{\beta}{k}].$$