Optimal detection of sparse principal components in high dimension

Quentin Berthet[*] **and Philippe Rigollet**[†]

Princeton University

Abstract. We perform a finite sample analysis of the detection levels for sparse principal components of a high-dimensional covariance matrix. Our minimax optimal test is based on a sparse eigenvalue statistic. Alas, computing this test is known to be NP-complete in general and we describe a computationally efficient alternative test using convex relaxations. Our relaxation is also proved to detect sparse principal components at near optimal detection levels and performs very well on simulated datasets.

AMS 2000 subject classifications: Primary 62H25; secondary 62F04, 90C22.
Key words and phrases: High-dimensional detection, sparse principal component analysis, spiked covariance model, semidefinite relaxation, minimax lower bound.

1. INTRODUCTION

The sparsity assumption has become preponderant in modern, high-dimensional statistics. In the high dimension, low sample size setting, where consistency seems to be hopeless, sparstiy turns out to be the statistician's salvation. It formalizes the a priori belief that only a few parameters, among a large number of them, are significant for the statistical task at hand. This paper explores a specific high-dimensional problem, namely Principal Component Analysis (PCA). Indeed, classical PCA is known to produce inconsistent estimators of the directions that explain the most variance (Johnstone and Lu, 2009; Nadler, 2008; Paul, 2007) without further assumption. For PCA, the *spiked covariance model* introduced by Johnstone (2001) directly encodes the sparsity assumption. Namely, this model relies on the assumption that there exists a few sparse directions that explain most of the variance. Formally, we assume that the observations are drawn from a multivariate Gaussian distribution with mean zero and covariance matrix given by $I + \theta vv^\top$, where I is the identity matrix and v is a unit norm sparse vector. Akin to other models, the sparsity assumption drives both methods and analysis in a wide variety of applications ranging from signal processing to biology (see, e.g., Alon et al., 1999; Chen, 2011; Jenatton et al., 2009; Wright et al., 2011, for a few examples). Most contributions to this problem have focused on consistent estimation of the sparse principal component v for various performance

[*]Supported in part by a Gordon S. Wu fellowship.
[†]Supported in part by NSF grants DMS-0906424 and CAREER-DMS-1053987.

measures (see, e.g. Amini and Wainwright, 2009; Ma, 2011b; Shen et al., 2009, and the above references).

What if there is no sparse component? In other words, what if $\theta = 0$? From a detection standpoint, one may ask the following question: How much variance should a sparse principal component explain in order to be detectable by a statistical procedure? Answering this question consists in (i) constructing a test that can detect this sparse principal component when the associated variance is above a certain level and (ii) proving that no test can detect such a principal component below a certain level.

Optimal detection levels in a high-dimensional setup have recently received a lot of attention. More precisely, Arias-Castro et al. (2010); Donoho and Jin (2004); Ingster et al. (2010) have studied the detection of a sparse vector corrupted by noise under various sparsity assumptions. More recently, this problem has been extended from vectors to matrices by Butucea and Ingster (2011); Sun and Nobel (2008, 2010) who propose to detect a shifted sub-matrix hidden in a Gaussian or binary random matrix. While the notion of sub-matrix encodes a certain sparsity structure, these two papers focus on the elementwise properties of random matrices, unlike the blooming random matrix theory that focuses on spectral aspects. Arias-Castro et al. (2011) studied a problem somewhere between sparse PCA detection and the shifted sub-matrix problem. Their goal is to detect a shifted off-diagonal sub-matrix hidden in a covariance matrix. Their methods are not spectral either.

We extend the current work on detection in two directions. First, we analyze detection in the framework of sparse PCA, and more precisely, in the spiked covariance model. Second, while all the literature on this topic is presented in an asymptotic framework, we propose a finite sample analysis of our problem with results that hold with high-probability. These results show the delicate interplay between the important parameters of the problem: the ambient dimension, the sample size and the sparsity.

Note that the spiked covariance model is particularly amenable to spectral methods due to its rotational invariance. It turns out that the so-called k-sparse largest eignvalue can be used to construct an optimal test. However, constructing this test raises some technical difficulties and can even be proved to be NP-complete. As a result, a large body of the optimization literature on this topic consists in numerical methods to overcome this issue (see, e.g., d'Aspremont et al., 2008, 2007; Journée et al., 2010; Lu and Zhang, 2011; Ma, 2011a, and references therein). Nevertheless, while these methods do produce a solution, their statistical properties are rarely addressed for the estimation problem and never for the detection problem. One of the approaches introduced by d'Aspremont et al. (2007) uses a convexification technique called *semidefinite programming* (SDP). A major drawback of this technique is that it may not output a vector $\hat{v}$ but a matrix and an ad hoc post-processing step is often required to turn this matrix back into a vector. However, in the context of detection our goal is not estimate the eigenvector v but rather its associated eigenvalue. This notable difference allows us to even bypass SDP optimization, which is known to scale poorly in very high dimension. Inspired by the SDP formulation, we propose a simple test procedure based on the *minimum dual perturbation* (MDP) that is easy to compute and for which we can derive near optimal performance bounds for the detection

problem.

Most of our analysis is performed in the spiked covariance model for Gaussian random vectors. Nevertheless, our results are robust to variations around this model and we spend Section 7 discussing various weaker assumptions under which our results still hold. In particular, we only need that our estimated covariance matrix belongs to a small box around the true covariance matrix with high probability. This setup encompasses biased estimators or adversarial noise.

The rest of the paper is organized as follows. In Section 2, we introduce the detection problem for sparse principal components. In Section 3, we discuss various links with classical and more recent results on random matrix theory and more precisely, the asymptotic effect of low rank perturbations to Wishart matrices. Our main results are contained in Section 4, where in particular, we introduce a new test based on spectral methods and derive the level at which it achieves detection of sparse principal components with high probability. This level is proved to be optimal in a minimax sense in Section 5. Unfortunately, this test cannot be computed efficiently and several relaxations are proposed in Section 6. As mentioned above, Section 7 discusses various weaker assumptions under which our results hold. Finally the performance of our test is illustrated on several numerical examples in Section 8.

NOTATIONS. The space of $d \times d$ symmetric real matrices is denoted by $\mathbf{S}_d$, and the cone of semidefinite positive matrices is denoted by $\mathbf{S}_d^+$. We write equivalently $Z \in \mathbf{S}_d^+$ and $Z \succeq 0$.

For any $q \geq 1$ we denote by $|v|_q$ the ℓ_q norm of a vector v and by extension, we denote by $|v|_0$ its so-called "ℓ_0 norm", that is its number of nonzero elements. The elements of a vector $v \in \mathbf{R}^p$ are denoted by $v_1, \ldots, v_p$ and similarly, a matrix Z has element Z_{ij} on its ith row and jth column. Furthermore, by extension, for $Z \in \mathbf{S}_d$, we denote by $|Z|_q$ the ℓ_q norm of the vector formed by the entries of Z.

The trace and rank functionals are denoted by **Tr** and **rank** respectively and have their usual definition. The identity matrix in $\mathbf{R}^p$ is denoted by I_p. For a finite set S, we denote by $|S|$ its cardinality. We also write A_S for the $|S| \times |S|$ submatrix with elements $(A_{ij})_{i,j \in S}$, and v_S for the vector of $\mathbf{R}^{|S|}$ with elements v_i for $i \in S$. Finally, for two real numbers a and b, we write $a \wedge b = \min(a, b)$, $a \vee b = \max(a, b)$, and $a_+ = a \vee 0$.

2. SPHERICITY TEST WITH SPARSE ALTERNATIVE

Let $X_1, \ldots, X_n$ be n i.i.d. realizations of a random variable X in $\mathbf{R}^p$. Our objective is to test the sphericity hypothesis, i.e., that the distribution of X is invariant by rotation in $\mathbf{R}^p$. For a gaussian distribution, this is equivalent to testing if the covariance matrix of X is of the form $\sigma^2 I_p$ for some known $\sigma^2 > 0$.

Without loss of generality we may assume $\sigma^2 = 1$ so that the covariance matrix is the identity in $\mathbf{R}^p$ under the null hypothesis. Possible alternative hypotheses should encompass the idea that there exists a privileged direction, along which X has more variance. Of course, there are many possible characterizations for this alternative and, in the spirit of sparse PCA, we focus on the case where there the privileged direction is sparse. Therefore, we consider the alternative hypothesis where the covariance matrix is a sparse rank one perturbation of the identity I_p. Formally, let $v \in \mathbf{R}^p$ be such that $|v|_2 = 1, |v|_0 \leq k$, and $\theta > 0$. The hypothesis

testing problem studied throughout this paper is:

$$\begin{aligned} H_0 &: X \sim \mathcal{N}(0, I_p) \\ H_1 &: X \sim \mathcal{N}(0, I_p + \theta\, vv^\top)\,. \end{aligned}$$

Note that the model under H_1 is a generalization of the spiked covariance model since it allows v to be k-sparse on the unit Euclidean sphere. In particular, the statement of H_1 is invariant under rotation on the k relevant variables.

Denote Σ the covariance matrix of X. A useful statistic in these settings is the empirical covariance matrix $\hat{\Sigma}$ defined by

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top\,.$$

It is an unbiased estimator for the covariance matrix of X, the maximum likelihood estimator in the gaussian case, when the mean is known to be 0. Furthermore, it is oftentimes the only data provided to the statistician.

We say that a test *discriminates* between H_0 and H_1 with probability $1-\delta$ if the type I and type II errors both have a probability smaller than δ. Our goal is therefore to find a statistic $\varphi(\hat{\Sigma})$ and levels $\tau_0 < \tau_1$,depending on (p, n, k, δ) such that

$$\begin{aligned} \mathbf{P}_{H_0}(\varphi(\hat{\Sigma}) > \tau_0) &\le \delta \\ \mathbf{P}_{H_1}(\varphi(\hat{\Sigma}) < \tau_1) &\le \delta\,. \end{aligned}$$

Taking $\tau \in [\tau_0, \tau_1]$ gives us control over the type I and type II errors of the test

$$\psi(\hat{\Sigma}) = \mathbf{1}\{\varphi(\hat{\Sigma}) > \tau\}\,,$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. As desired, this test has the property to discriminate between the hypotheses with probability $1-\delta$.

3. LINK WITH RANDOM MATRIX THEORY

Note that under the null hypothesis, the sample covariance matrix $\hat{\Sigma}$ follows a rescaled Wishart distribution. The spectrum of such matrices has been extensively studied and is fairly well understood. We give below a quick overview of the results that are relevant to our problem.

3.1 Spectral methods

It is not hard to see that, under H_1, for any $\theta > 0$, v is an eigenvector associated to the largest eigenvalue of the population covariance matrix Σ, without further assumption. Moreover, if $\hat{\Sigma}$ is close to Σ in spectral norm, then its largest eigenvector should be a good candidate to approximate v. It is therefore natural to consider spectral methods for the spiked covariance model. Understanding the behavior of our test statistic under both the null and the alternative is key in proving that it discriminates between the hypotheses.

Convergence of the empirical covariance matrix to the true covariance matrix in spectral norm has received some attention recently (see, e.g., Bickel and Levina, 2008; Cai et al., 2010; El-Karoui, 2008) under various elementwise sparsity assumption and using thresholding methods. However, since our assumption allows

for relevant variables to produce arbitrary small entries under the alternative hypothesis, we cannot use such results. A natural statistic to discriminate between null and alternative would be, for example, using the largest eigenvalue of the covariance matrix.

Spectral properties of random matrices have received a lot of attention from both a statistical and probabilistic angle. We devote the rest of this section to review some of the classical results from random matrix theory to argue that even in moderate dimensions, the largest eigenvalue cannot discriminate between the null and alternative hypotheses.

It is easy to notice that for any unit vector v

$$\lambda_{\max}(I_p) = 1 \quad \text{and} \quad \lambda_{\max}(I_p + \theta v v^\top) = 1 + \theta \,.$$

If we could allow, for a fixed p, to let n go to infinity, the consistency of the estimator $\hat{\Sigma}$ (for fixed p, entry by entry) and the continuity of the largest eigenvalue as a function of its entries would prove that we have an efficient method to discriminate between the two alternatives, at least asymptotically.

However, in a high dimension setting, where p may grow with n, the behavior of $\lambda_{\max}(\hat{\Sigma})$ under the null hypothesis is different. If $p/n \to \alpha > 0$, Geman (1980) showed that, in accordance with the Marcenko-Pastur distribution, we have

$$\lambda_{\max}(\hat{\Sigma}) \to (1 + \sqrt{\alpha})^2 > 1 \,,$$

were the convergence holds almost surely (see also Bai, 1999; Johnstone, 2001, and references therein). Moreover, Yin et al. (1988) established that $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^4) < \infty$ is a necessary and sufficient condition for this almost sure convergence to hold. Furthermore, as $\hat{\Sigma} \in \mathbf{S}_d^+$, its number of positive eigenvalues is equal to its rank (which is smaller than n), and we have

$$\lambda_{\max}(\hat{\Sigma}) \geq \frac{1}{\mathbf{rank}(\hat{\Sigma})} \sum_{i=1}^{p} \lambda_i(\hat{\Sigma}) \geq \frac{1}{n} \mathbf{Tr}(\hat{\Sigma}) \geq \frac{p}{n} \frac{\mathbf{Tr}\hat{\Sigma}}{np} \,.$$

As the sum of np squared norms of independent standard gaussian vectors, $\mathbf{Tr}(\hat{\Sigma}) \sim \chi^2_{np}$, hence almost surely, for $p/n \to \infty$, we have $\lambda_{\max}(\hat{\Sigma}) \to \infty$ under the null hypothesis.

These two results, indicate that the largest eigenvalue will not be able to discriminate between the two hypotheses unless $\theta > Cp/n$ for some positive constant. In a "large p/small n" scenario, this corresponds to a very strong signal indeed. In the next subsection, we show that such results can be made more formal using perturbation theory.

3.2 Low rank perturbations of Wishart matrices

In a slightly different setting, Baik et al. (2005) established that when adding a finite rank perturbation to a Wishart matrix, a phase transition arises already in the moderate dimensional regime where $p/n \to \alpha \in (0, 1)$. This phenomenon is known as the BBP transition for the name of the authors. A very general class of random matrices exhibits similar behavior, under finite rank perturbations, as shown by Tao (2011). These results are extended to more general distributions in Benaych-Georges et al. (2011).

Assume that

$$W_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top \quad \text{and} \quad \tilde{W}_n = \theta v v^\top + W_n\,.$$

Qualitatively, the BPP phase transition predicates that there exists a critical value θ^* such that if $\theta > \theta^*$, the spectrum of $\tilde{W}_n$ exhibits an isolated eigenvalue significantly larger than the others, and that if $\theta < \theta^*$, the spectrum has a very similar behavior under the two hypotheses.

Even when $p/n \to \alpha \in (0,1)$, Benaych-Georges et al. (2011) show that if $\theta \le \alpha + \sqrt{\alpha}$, the leading eigenvalue will have limit $(1+\sqrt{\alpha})^2$, i.e., the same as under the null hypothesis. Similarly, the random fluctuations around this limit will follow the the Tracy-Widom distribution, for both hypotheses.

The above analysis indicates that detection using the largest eigenvalue is impossible already for moderate dimension, without further assumptions. Nevertheless, resorting to the sparsity assumption allows us to bypass this intrinsic limitation of the largest eigenvalue as a test statistic.

3.3 Sparse eigenvalues

To exploit the sparsity assumption, we use the fact that only a small submatrix of the empirical covariance will be affected by the perturbation. Let A be a $p \times p$ matrix and fix $k < p$. We define the k-sparse largest[1] eigenvalue by

$$\lambda_{\max}^k(A) = \max_{|S|=k} \lambda_{\max}(A_S)\,. \tag{3.1}$$

We have the same equalities as for regular eigenvalues.

$$\lambda_{\max}^k(I_p) = 1 \quad \text{and} \quad \lambda_{\max}^k(I_p + \theta v v^\top) = 1 + \theta\,.$$

However, the k-sparse eigenvalue behaves differently under the two hypotheses as soon as there exists a $k \times k$ matrix with a significantly higher largest eigenvalue. The BBP transition, in a similar setting, indicates that this is true as soon as $\theta > \gamma + \sqrt{\gamma}$, when $k/n \to \gamma > 0$. Therefore, it appears that the sparsity assumption can be exploited us to significantly reduce the dimensionality of the problem.

4. SPARSE PRINCIPAL COMPONENT DETECTION

In sparse principal component detection, we are testing the existence of a sparse direction v with a significantly higher explained variance $v^\top \Sigma v$ than any other direction. Following the motivation of the previous section, we study the properties of the test statistic $\varphi(\hat{\Sigma}) = \lambda_{\max}^k(\hat{\Sigma})$, where we recall that $\lambda_{\max}^k(\hat{\Sigma})$ is the k-sparse eigenvalue of $\hat{\Sigma}$ and can be defined equivalently to (3.1), for any $A \in \mathbf{S}_d^+$, by

$$\lambda_{\max}^k(A) = \max_{\substack{|x|_2=1 \\ |x|_0 \le k}} x^\top A x\,. \tag{4.1}$$

[1] In the rest of the paper, we drop the qualification "largest" since we only refer to this one.

4.1 Deviation bounds for the k-sparse eigenvalue

Finding optimal detection levels amounts to finding the right order of magnitude of the deviations of the test statistic $\lambda^k_{\max}(\hat{\Sigma})$ both under the null and the alternative hypotheses. We begin by the following proposition, which guarantees that our test statistic remains large enough under the alternative hypothesis.

PROPOSITION 4.1. *Under H_1, we have with probability $1-\delta$,*

$$\lambda^k_{\max}(\hat{\Sigma}) \geq 1+\theta-2(1+\theta)\sqrt{\frac{\log(1/\delta)}{n}}\,.$$

PROOF. Under H_1, there exists a unit vector v with sparsity k, such that $X \sim \mathcal{N}(0, I_p + \theta\, vv^\top)$. Therefore, we have

$$\lambda^k_{\max}(\hat{\Sigma}) \geq v^\top \hat{\Sigma} v = \frac{1}{n}\sum_{i=1}^n (X_i^\top v)^2\,,$$

by definition of $\hat{\Sigma}$. Since $X \sim \mathcal{N}(0, I_p + \theta\, vv^\top)$, we have $X^\top v \sim \mathcal{N}(0, 1+\theta)$.

Define the random variable

$$Y = \frac{1}{n}\sum_{i=1}^n \left(\frac{(X_i^\top v)^2}{1+\theta} - 1\right)\,.$$

Using Laurent and Massart (2000, Lemma 1) on concentration of the χ^2 distribution, we get for any $t>0$, that

$$\mathbf{P}\left(Y \leq -2\sqrt{t/n}\right) \leq e^{-t}\,.$$

Hence, taking $t = \log(1/\delta)$, we have $Y \geq -2\sqrt{\log(1/\delta)/n}$ with probability $1-\delta$. This yields the desired inequality. ∎

Note that our proof relies only on the existence of a sparse vector v associated to the eigenvalue $(1+\theta)$ of the population covariance matrix Σ. In particular, the result of Proposition 4.1 extends to more general alternative hypotheses as long as they satisfy this condition. Moreover, observe that the above lower bound is independent of p and k. Proposition 4.1 suggests that the spiked covariance model is well separated from the spherical model where $\theta = 0$. Note that much more than detection can actually be achieved under this model. Indeed, Amini and Wainwright (2009) prove optimal rates of support recovery for v in the case where θ is known and v takes only values in $\{0, \pm 1/\sqrt{k}\}$.

We now study the behavior of the k-sparse eigenvalue under the null hypothesis, i.e., for a Wishart matrix with mean I_p. We adapt a technique from Vershynin (2010) to obtain the desired deviation bounds.

PROPOSITION 4.2. *Under H_0, with probability $1-\delta$*

$$\lambda^k_{\max}(\hat{\Sigma}) \leq 1 + 4\sqrt{\frac{k\log(9ep/k)+\log(1/\delta)}{n}} + 4\frac{k\log(9ep/k)+\log(1/\delta)}{n}\,.$$

PROOF. Using a 1/4-net over the unit sphere of $\mathbf{R}^k$, it can be easily shown (see, e.g., Vershynin, 2010) that there exists a subset $\mathcal{N}_k$ of the unit sphere of $\mathbf{R}^k$, with cardinality smaller than 9^k, such that for any $A \in \mathbf{S}_k^+$

$$\lambda_{\max}(A) \le 2 \max_{x \in \mathcal{N}_k} x^\top A x \,. \tag{4.2}$$

Under H_1, we have

$$\lambda_{\max}^k(\hat{\Sigma}) \le 1 + \max_{|S|=k} \left\{ \lambda_{\max}(\hat{\Sigma}_S) - 1 \right\} ,$$

where the maximum in the right-hand side is taken over all subsets of $\{1, \dots, p\}$ that have cardinality k.

Moreover, for all $u \in \mathbf{R}^k, |u|_2 = 1$ and $S \subset \{1, \dots, p\}$ such that $|S| = k$, let $\tilde{u} \in \mathbf{R}^p$ be the vector with support S such that $\tilde{u}_S = u$. We have

$$u^\top \hat{\Sigma}_S u - 1 = \tilde{u}^\top \hat{\Sigma} \tilde{u} - 1 = \frac{1}{n} \sum_{i=1}^n \left[(\tilde{u}^\top X_i)^2 - 1 \right] .$$

Since $|\tilde{u}|_2 = |u|_2 = 1$, Laurent and Massart (2000, Lemma 1) yields for any $t > 0$,

$$\mathbf{P}\left(\frac{1}{n} \sum_{i=1}^n \left[(\tilde{u}^\top X_i)^2 - 1 \right] \ge 2\sqrt{\frac{t}{n}} + 2\,\frac{t}{n} \right) \le e^{-t} . \tag{4.3}$$

For any $S \subset \{1, \dots, p\}$, define $\mathbf{R}^S$ to be the subset of $\mathbf{R}^p$ defined such that $x \in \mathbf{R}^S$ iff $x_j = 0\,, \forall\, j \notin S$. Let $\mathcal{N}_k(S)$ be a subset of the unit sphere of $\mathbf{R}^S$, with cardinality smaller than 9^k such that for any $A \in \mathbf{S}_k^+$, inequality (4.2) holds with $\mathcal{N}_k = \mathcal{N}_k(S)$. Fix $t > 0$ and define the event $\mathcal{A}_S$ by

$$\mathcal{A}_S = \left\{ \lambda_{\max}(\hat{\Sigma}_S) - 1 \ge 4\sqrt{\frac{t}{n}} + 4\,\frac{t}{n} \right\} .$$

Observe that a union bound over the elements of $\mathcal{N}_k(S)$ together with (4.3) yields that for any $t > 0$,

$$\mathbf{P}(\mathcal{A}_S) \le \mathbf{P}\big(\max_{v \in \mathcal{N}_k(S)} \frac{1}{n} \sum_{i=1}^n (v^\top X_i)^2 - 1 \ge 2\sqrt{\frac{t}{n}} + 2\,\frac{t}{n} \big) \le 9^k e^{-t} .$$

Let now $\mathcal{A}$ be the event defined by

$$\mathcal{A} = \bigcup_{|S|=k} \mathcal{A}_S = \left\{ \max_{|S|=k} \left\{ \lambda_{\max}(\hat{\Sigma}_S) - 1 \right\} \ge 4\sqrt{\frac{t}{n}} + 4\,\frac{t}{n} \right\} .$$

Therefore, by a union bound on the $\binom{p}{k}$ subsets S of $\{1, \dots, p\}$ that have cardinality k, we get

$$\mathbf{P}\Big(\lambda_{\max}^k(\hat{\Sigma}) \ge 1 + 4\sqrt{\frac{t}{n}} + 4\,\frac{t}{n} \Big) \le \mathbf{P}(\mathcal{A}) \le \binom{p}{k} 9^k e^{-t} .$$

To conclude our proof, it is sufficient to use the standard inequality $\binom{p}{k} \le \left(\frac{ep}{k}\right)^k$ and to take $t = k \log(9ep/k) + \log(1/\delta)$. ∎

4.2 Hypothesis testing with $\lambda^k_{\max}$

Using these results, we have, with the notations from Section 2,

$$\mathbf{P}_{H_0}(\lambda^k_{\max}(\hat{\Sigma}) > \tau_0) \leq \delta$$
$$\mathbf{P}_{H_1}(\lambda^k_{\max}(\hat{\Sigma}) < \tau_1) \leq \delta\,,$$

where τ_0 and τ_1 are given by

$$\tau_0 = 1 + 4\sqrt{\frac{k\log(9ep/k) + \log(1/\delta)}{n}} + 4\frac{k\log(9ep/k) + \log(1/\delta)}{n}$$
$$\tau_1 = 1 + \theta - 2(1+\theta)\sqrt{\frac{\log(\frac{1}{\delta})}{n}}\,.$$

Whenever $\tau_1 > \tau_0$, we take $\tau \in [\tau_0, \tau_1]$ and define the following test

$$\psi(\hat{\Sigma}) = \mathbf{1}\{\lambda^k_{\max}(\hat{\Sigma}) > \tau\}\,.$$

It follows from the previous subsection that it discriminates between H_1 and H_0 with probability $1-\delta$.

It remains to find for which values of θ the condition $\tau_1 > \tau_0$. It corresponds to our minimum detection level.

THEOREM 4.1. *Assume that k, p, n and δ are such that $\bar{\theta} \leq 1$, where*

$$\bar{\theta} := 4\sqrt{\frac{k\log(\frac{9ep}{k}) + \log(\frac{1}{\delta})}{n}} + 4\frac{k\log(\frac{9ep}{k}) + \log(\frac{1}{\delta})}{n} + 4\sqrt{\frac{\log(\frac{1}{\delta})}{n}}\,. \tag{4.4}$$

Then, for any $\theta > \bar{\theta}$ and for any $\tau \in [\tau_0, \tau_1]$, the test $\psi(\hat{\Sigma}) = \mathbf{1}\{\lambda^k_{\max}(\hat{\Sigma}) > \tau\}$ discriminates between H_0 and H_1 with probability $1-\delta$.

If we consider asymptotic regimes, for large p, n, k, taking $\delta = p^{-\beta}$ with $\beta > 0$, provides a sequence of tests ψ_n that discriminate between H_0 and H_1 with probability converging to 1, for any fixed $\theta > 0$, as soon as

$$\frac{k\log(p)}{n} \to 0\,.$$

5. MINIMAX LOWER BOUNDS FOR DETECTION

The goal of this section is to prove that if $\theta > C\bar{\theta}$ for some $C > 0$, where $\bar{\theta}$ is defined in (4.4), then no test can discriminate between H_0 and H_1 with arbitrarily small probability. We will see that this result can be achieved up to logarithmic terms that vanish for interesting regimes of p, n and k. Throughout this section, assume that $\theta < 1/\sqrt{2}$.

In order to find lower bounds for the probability of error, we study the χ^2 distance between probability measures (see, e.g., Tsybakov, 2009, chapter 2). For any $v \in \mathbf{R}^p$ such that $|v|_2 = 1$, define the matrix $\Sigma_v = I_p + \theta vv^\top$ and let $\mathbf{P}_v$ denote the distribution of a Gaussian random variable $X \sim \mathcal{N}(0, \Sigma_v)$. Moreover,

let $\mathcal{S} = \{S \subset \{1, \dots, p\} : |S| = k\}$ and for any $S \in \mathcal{S}$, define $u(S) \in \mathbf{R}^p$ to be the unit vector with jth coordinate equal to $1/\sqrt{k}$ if $j \in S$ and 0 otherwise. Finally, define the Gaussian mixture $\mathbf{P}_{\mathcal{S}}$ by

$$\mathbf{P}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{S \in \mathcal{S}} \mathbf{P}_{u(S)} \, .$$

The following theorem holds.

THEOREM 5.1. *Fix $\nu > 0$ and define $\underline{\theta} > 0$ by*

$$\underline{\theta} := \sqrt{\frac{k \log\left(\nu p/k^2 + 1\right)}{n}} \, . \tag{5.1}$$

Then there exists a constant $C_\nu > 0$ such that for any $\theta < \underline{\theta} \wedge 1/\sqrt{2}$, it holds

$$\inf_{\Psi} \left\{ \mathbf{P}_0^n(\Psi = 1) \vee \max_{\substack{|v|_2 = 1 \\ |v|_0 \le k}} \mathbf{P}_v^n(\Psi = 0) \right\} \ge C_\nu \, .$$

where the infimum is taken over all possible tests, i.e., measurable functions of the observations that take values in $\{0, 1\}$. Moreover, by taking ν small enough, C_ν can be made arbitrary close to $1/2$.

We write for simplicity $\mathbf{P}_S = \mathbf{P}_{u_S}$ when this leads to no confusion. Our proof relies on the following lemma.

LEMMA 5.1. *For any $S, T \in \mathcal{S}$ and any $\theta < 1/2$, it holds*

$$\mathbb{E}_{\mathbf{P}_0} \left(\frac{\mathrm{d}\mathbf{P}_S}{\mathrm{d}\mathbf{P}_0} \frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{P}_0} \right) = \left(1 - \theta^2 (u(S)^\top u(T))^2\right)^{-1/2} \, .$$

PROOF. Fix $S \in \mathcal{S}$ and observe that

$$\frac{\mathrm{d}\mathbf{P}_S}{\mathrm{d}\mathbf{P}_0}(X) = \frac{\det(I_p)^{1/2}}{\det(\Sigma_S)^{1/2}} \frac{\exp(-\frac{1}{2} X^\top \Sigma_{u(S)}^{-1} X)}{\exp(-\frac{1}{2} X^\top I_p^{-1} X)} \, .$$

Furthermore , since $\det(I_p) = 1$ and $|u(S)|_2 = 1$, we get by Sylvester's determinant theorem that

$$\det(\Sigma_{u(S)}) = \det(I_p + \theta u(S) u(S)^\top) = \det(I_1 + \theta \, u(S)^\top u(S)) = 1 + \theta \, .$$

Moreover, the Sherman-Morrison formula yields

$$\Sigma_S^{-1} = (I_p + \theta u(S) u(S)^\top)^{-1} = I_p - \frac{\theta u(S) u(S)^\top}{1 + \theta} \, .$$

By substitution, the above three displays yield

$$\frac{\mathrm{d}\mathbf{P}_S}{\mathrm{d}\mathbf{P}_0}(X) = \frac{1}{\sqrt{1+\theta}} \exp\left(\frac{1}{2} \frac{\theta}{1+\theta} (X^\top u(S))^2 \right)$$

and

$$\frac{\mathrm{d}\mathbf{P}_S}{\mathrm{d}\mathbf{P}_0}\frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{P}_0}(X) = \frac{1}{1+\theta}\exp(X^\top M X)\,, \tag{5.2}$$

where M is defined by

$$M = \frac{1}{2}\frac{\theta}{1+\theta}(u(S)u(S)^\top + u(T)u(T)^\top)\,.$$

Note that M has at most two non-zero eigenvalues given by

$$\lambda_1 = \frac{1}{2}\frac{\theta}{1+\theta}(1+u(S)^\top u(T)) < \frac{1}{2} \quad \text{and} \quad \lambda_2 = \frac{1}{2}\frac{\theta}{1+\theta}(1-u(S)^\top u(T)) < \frac{1}{2}\,,$$

and let Λ denote the diagonal matrix with elements $(\lambda_1, \lambda_2, 0, \dots, 0) \in \mathbf{R}^p$.

Together with (5.2), it yields

$$\begin{aligned}
\mathbb{E}_{\mathbf{P}_0}\left(\frac{\mathrm{d}\mathbf{P}_S}{\mathrm{d}\mathbf{P}_0}\frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{P}_0}\right) &= \frac{1}{1+\theta}\,\mathbb{E}_{\mathbf{P}_0}[\exp(X^\top M X)] \\
&= \frac{1}{1+\theta}\,\mathbb{E}_{\mathbf{P}_0}[\exp(X^\top \Lambda X)] \\
&= \frac{1}{1+\theta}\mathbb{E}_{\mathbf{P}_0}[\exp(\lambda_1 X_1^2)]\,\mathbb{E}_{\mathbf{P}_0}[\exp(\lambda_2 X_2^2)] \\
&= \frac{1}{1+\theta}\,[(1-2\lambda_1)(1-2\lambda_2)]^{-1/2}\,,
\end{aligned}$$

where, in the second equality, the substitution of M by Λ is valid by rotational invariance of the distribution of X under $\mathbf{P}_0$. The last equation yields the desired result. ∎

We now turn to the proof of Theorem 5.1

PROOF. Observe now that

$$\chi^2(\mathbf{P}_\mathcal{S}, \mathbf{P}_0) = \mathbb{E}_{\mathbf{P}_0}\left[\left(\frac{\mathrm{d}\mathbf{P}_\mathcal{S}}{\mathrm{d}\mathbf{P}_0} - 1\right)^2\right] = \frac{1}{|\mathcal{S}|^2}\sum_{S,T\in\mathcal{S}}\mathbb{E}_{\mathbf{P}_0}\left(\frac{\mathrm{d}\mathbf{P}_S}{\mathrm{d}\mathbf{P}_0}\frac{\mathrm{d}\mathbf{P}_T}{\mathrm{d}\mathbf{P}_0}\right) - 1\,.$$

Lemma 5.1 together with the fact $u(S)^\top u(T) = |S\cap T|/k$ yields

$$\chi^2(\mathbf{P}_\mathcal{S}, \mathbf{P}_0) = \sum_{r=0}^{k}\left\{\frac{\mathcal{C}(\mathcal{S}, r)}{|\mathcal{S}|^2}\left(1-\theta^2\frac{r^2}{k^2}\right)^{-1/2}\right\} - 1\,, \tag{5.3}$$

where $\mathcal{C}(\mathcal{S}, r)$ denotes the number of subsets $S, T \in \mathcal{S}$ such that $|S\cap T| = r$.

We now control the term on the right-hand side of (5.3). Let S, T be chosen uniformly at random in $\mathcal{S}$, and observe that $\mathbf{P}(|S\cap T| = r) = \mathbf{P}(R = r)$, where $R = |S \cap \{1, \dots, k\}|$. It yields

$$\begin{aligned}
\chi^2(\mathbf{P}_\mathcal{S}^n, \mathbf{P}_0^n) &= \prod_{i=1}^{n}\left(1+\chi^2(\mathbf{P}_\mathcal{S}, \mathbf{P}_0)\right) - 1 \\
&= \mathbb{E}_{S,T}\left\{\left[1-\theta^2\,\frac{|S\cap T|^2}{k^2}\right]^{-n/2}\right\} - 1 \\
&= \mathbb{E}_R\left\{\left[1-\theta^2\,\frac{R^2}{k^2}\right]^{-n/2}\right\} - 1\,.
\end{aligned}$$

where $\mathbb{E}_S$ denotes the expectation with respect to the random subset S and $\mathbb{E}_R$ the expectation with respect to the hypergeometric random variable R.

Using now the convexity inequality $(1-t)^{-n/2} \le e^{\frac{nt}{2(1-t)}} \le e^{nt}$ valid for $1-t \ge 1/2$, and noticing that $R \le k$, the above display leads to

$$\chi^2(\mathbf{P}_S^n, \mathbf{P}_0^n) \le \mathbb{E}_R\left[\exp\left(\frac{n\theta^2 R}{k}\right)\right] - 1\,. \tag{5.4}$$

Define $\mu^2 = n\theta^2/k$. We have, as in Addario-Berry et al. (2010); Arias-Castro et al. (2011) that

$$\begin{aligned}
\mathbb{E}_R\left[e^{\mu^2 R}\right] &= \mathbb{E}_S\left[\prod_{i=1}^k \exp(\mu^2 \mathbf{1}\{i \in S\})\right] - 1 \\
&\le \prod_{i=1}^k \mathbb{E}_S\left[\exp(\mu^2 \mathbf{1}\{i \in S\})\right] - 1\,, \quad \text{by negative association} \\
&\le \left(\left(e^{\mu^2} - 1\right)\frac{k}{p} + 1\right)^k - 1\,.
\end{aligned}$$

Assume now that $\theta < \underline{\theta}$. It yields

$$\begin{aligned}
\mathbb{E}_R\left[e^{\mu^2 R}\right] &\le \left(\left(e^{\mu^2} - 1\right)\frac{k}{p} + 1\right)^k - 1 \le \left(\left(\frac{\nu p}{k^2}\right)\frac{k}{p} + 1\right)^k - 1 \\
&\le \left(1 + \frac{\nu}{k}\right)^k - 1 \le e^{\nu} - 1\,.
\end{aligned}$$

Together with (5.4) it yields $\chi^2(\mathbf{P}_S^n, \mathbf{P}_0^n) \le e^{\nu} - 1$.

We are now in a position to apply standard results from minimax theory. Note that for all measurable tests Ψ, we have

$$\begin{aligned}
\mathbf{P}_0^n(\Psi = 1) \vee \max_{\substack{|v|_2 = 1 \\ |v|_0 \le k}} \mathbf{P}_v^n(\Psi = 0) &\ge \mathbf{P}_0^n(\Psi = 1) \vee \max_{S \in \mathcal{S}} \mathbf{P}_{u(S)}^n(\Psi = 0) \\
&\ge \mathbf{P}_0^n(\Psi = 1) \vee \mathbf{P}_S^n(\Psi = 0) \\
&\ge \frac{e^{1-e^{\nu}}}{4} \vee \frac{1 - \sqrt{(e^{\nu} - 1)/2}}{2} =: C_{\nu}\,.
\end{aligned}$$

where the last inequality is a direct consequence of Tsybakov (2009, Theorem 2.2, case (iii)). Observe that $C_{\nu} \to 1/2$ if $\nu \to 0$. ∎

We observe a gap between our upper and lower bound, with a term in $\log(p/k)$ in the upper bound, and one in $\log(p/k^2)$ in the lower bound. This gap has been observed in the detection literature before (see, e.g., Baraud, 2002; Verzelen, 2012, for an explicit remark) and, to our knowledge has never been addressed. However, by considering certain regimes for p, n and k, it disappears. Indeed, as soon as $p \ge k^{2+\varepsilon}$, for some $\varepsilon > 0$, upper and lower bounds match up to constants, and the detection rate for the sparse eigenvalue is optimal in a minimax sense. Under this assumption, detection becomes impossible if

$$\theta < C\sqrt{\frac{k \log(p/k)}{n}}\,.$$

for a small enough constant $C > 0$.

6. SEMIDEFINITE METHODS FOR SPARSE PRINCIPAL COMPONENT TESTING

Computing the largest k-sparse eigenvalue $\lambda^k_{\max}$ of a symmetric matrix A is, in general, a hard computational problem. To see this, consider the particular case where A is a $p \times p$ symmetric matrix with values in $\{0,1\}$ and $A_{ii} = 1$ for all diagonal entries, so that A corresponds to the adjacency matrix of an undirected graph. It is not hard to see that $\lambda^k_{\max}(A) \leq k$, with equality if and only if the graph of A contains a clique of size k. Yet, it is a well known fact of computational complexity (see, e.g., Sipser, 1996) that the decision problem associated to finding whether a graph contains a clique of size k is NP-complete. Note that if k were fixed, this problem would actually be polynomial in the size p of the graph since there are "only" $\binom{p}{k} \leq p^k$ subgraphs to enumerate. However, the exponential dependence in k is clearly concerning even for moderate values of k.

6.1 Semidefinite relaxation for $\lambda^k_{\max}$

Semidefinite programming (SDP) is the matrix equivalent of linear programming. Define the Euclidean scalar product in $\mathbf{S}^+_d$ by $\langle A, B\rangle = \mathbf{Tr}\,(AB)$. A semidefinite program can be written in the canonical form.

$$\begin{aligned} \mathsf{SDP} = \max. \quad & \mathbf{Tr}(CX) \\ \text{subject to} \quad & \mathbf{Tr}(A_i X) \leq b_i,\ \forall i \in \{1,\dots,m\} \\ & X \succeq 0 \end{aligned} \tag{6.1}$$

As convex problems, they are computationally efficient and can be solved using interior point or first order methods (see, e.g., Boyd and Vandenberghe, 2004; Nesterov and Nemirovskii, 1987). Using SDP relaxations of problems with non-convex constraints is a common method to find an approximate solution. Tightness bounds are sometimes proven (see, e.g., Goemans (1995) for the MAXCUT problem). A major breakthrough for sparse PCA was achieved by d'Aspremont et al. (2007), who introduced a SDP relaxation for $\lambda^k_{\max}$, but tightness of this relaxation is, to this day, unknown. Our task is not as difficult though. Indeed, we only need to prove that the SDP objective criterion has significantly different behavior under H_0 and H_1.

Making the change of variables $Z = xx^\top$, in (4.1) yields

$$\begin{aligned} \lambda^k_{\max}(A) = \max. \quad & \mathbf{Tr}(AZ) \\ \text{subject to} \quad & \mathbf{Tr}(Z) = 1,\ |Z|_0 \leq k^2 \\ & Z \succeq 0,\ \mathbf{rank}(Z) = 1\,. \end{aligned}$$

Note that this problem contains two sources of non-convexity: the ℓ_0 norm constraint and the rank constraint. We make two relaxations in order to have a convex feasible set.

First, for a semidefinite matrix Z, with trace 1, and sparsity k^2, the Cauchy-Schwarz inequality yields $|Z|_1 \leq k$, which is substituted to the cardinality constraint in this relaxation. Simply dropping the rank constraint leads to the fol-

lowing relaxation of our original problem:

$$\begin{aligned} \mathsf{SDP}_k(A) = \max. \quad & \mathbf{Tr}(AZ) \\ \text{subject to} \quad & \mathbf{Tr}(Z) = 1,\ |Z|_1 \le k \\ & Z \succeq 0\,. \end{aligned} \tag{6.2}$$

Note that this optimization problem is convex since it consists in minimizing a linear objective over a convex set. Moreover, it is a standard exercise to show that it can be expressed in the canonical form (6.1). As such, it can be solved efficiently using any of the aforementioned algorithms.

As a relaxation of the original problem, for any $A \succeq 0$, it holds

$$\lambda^k_{\max}(A) \le \mathsf{SDP}_k(A)\,. \tag{6.3}$$

Since we have proved in Section 4 that $\lambda^k_{\max}(\hat{\Sigma})$ takes large values under H_1, this inequality tells us that using $\mathsf{SDP}_k(\hat{\Sigma})$ as a test statistic will be to our advantage under H_1. It remains to show that it stays small under H_0. This can be achieved by using the dual formulation of the SDP.

LEMMA 6.1.(Bach et al., 2010). *For a given $A \succeq 0$, we have by duality*

$$\mathsf{SDP}_k(A) = \min_{U \in \mathbf{S}_p} \{\lambda_{\max}(A+U) + k|U|_\infty\}\,.$$

Together with (6.3), Lemma 6.1 implies that for any $z \ge 0$ and any matrix U such that $|U|_\infty \le z$, it holds

$$\lambda^k_{\max}(A) \le \mathsf{SDP}_k(A) \le \lambda_{\max}(A+U) + kz\,. \tag{6.4}$$

A direct consequence of (6.4) is that the functional $\lambda^k_{\max}$ is robust to perturbations by matrices that have small $|\cdot|_\infty$-norm. Formally, let $A \succeq 0$ be such that its largest eigenvector has ℓ_0 norm bounded by k. Then, for any matrix N, (6.4) yields

$$\lambda^k_{\max}(A+N) \le \lambda_{\max}((A+N)-N) + k|N|_\infty = \lambda^k_{\max}(A) + k|N|_\infty\,.$$

6.2 High probability bounds for convex relaxation

We now study the properties of $\mathsf{SDP}_k(\hat{\Sigma})$ and other computationally efficient variants as test statistics for our detection problem. Recall that $\mathsf{SDP}_k(\hat{\Sigma}) \ge \lambda^k_{\max}(\hat{\Sigma})$. In view of (6.3), the following proposition follows directly from Proposition 4.1.

PROPOSITION 6.1. *Under H_1, we have, with probability $1-\delta$*

$$\mathsf{SDP}_k(\hat{\Sigma}) \ge 1 + \theta - 2(1+\theta)\sqrt{\frac{\log(1/\delta)}{n}}\,.$$

Akin to Proposition 4.1, Proposition 6.1 shows that H_1 is the easy case. Indeed, under H_1, the lower deviations of $\mathsf{SDP}_k(\hat{\Sigma})$ remain small and do not depend on k or p. We now turn to the upper deviations under H_0.

PROPOSITION 6.2. *Under H_0, we have, with probability $1-\delta$,*

$$\mathsf{SDP}_k(\hat{\Sigma}) \le 1 + 2\sqrt{\frac{k^2\log(4p^2/\delta)}{n}} + 2\frac{k\log(4p^2/\delta)}{n} + 2\sqrt{\frac{\log(2p/\delta)}{n}} + 2\frac{\log(2p/\delta)}{n}\,.$$

PROOF. Let $st_z(A)$ be the soft-threshold of A, with threshold z defined by $(st_z(A))_{ij} = \mathrm{sign}\,(A_{ij})(|A_{ij}| - z)_+$. It follows from (6.4) that

$$\lambda_{\max}^k(A) \le \mathsf{SDP}_k(A) \le \lambda_{\max}(st_z(A)) + kz\,. \tag{6.5}$$

Let $\hat{\Delta} = \mathrm{diag}(\hat{\Sigma})$ be the diagonal matrix with the same diagonal entries as $\hat{\Sigma}$, and $\hat{\Psi} = \hat{\Sigma} - \hat{\Delta}$ the matrix of its off-diagonal entries, so that $\hat{\Sigma} = \hat{\Delta} + \hat{\Psi}$. Since $\hat{\Psi}$ and $\hat{\Delta}$ have disjoint supports, it follows that

$$st_z(\hat{\Sigma}) = st_z(\hat{\Delta}) + st_z(\hat{\Psi})\,. \tag{6.6}$$

We first control the largest off-diagonal element of $\hat{\Sigma}$ by bounding $|\hat{\Psi}|_\infty$ with high probability. For every i, j, we have

$$\begin{aligned}\hat{\Psi}_{ij} &= \frac{1}{n}\sum_{k=1}^n X_{ki}X_{kj} \\ &= \frac{1}{2}\left[\sum_{k=1}^n[\frac{1}{2}(X_{ki}+X_{kj})^2 - 1] - \sum_{k=1}^n[\frac{1}{2}(X_{ki}-X_{kj})^2 - 1]\right]\,.\end{aligned}$$

Under H_0, we have $X \sim \mathcal{N}(0, I_p)$, so by Laurent and Massart (2000, Lemma 1), it holds for $t > 0$ that

$$\mathbf{P}\Big(|\hat{\Psi}_{ij}| \ge 2\sqrt{\frac{t}{n}} + 2\frac{t}{n}\Big) \le 4e^{-t}\,.$$

Hence, by union bound on the off-diagonal terms, we get

$$\mathbf{P}\Big(\max_{i<j}|\hat{\Psi}_{ij}| \ge 2\sqrt{\frac{t}{n}} + 2\frac{t}{n}\Big) \le 2p^2e^{-t}\,.$$

Taking $t = \log(4p^2/\delta)$ yields with probability $1-\delta/2$, that $|\hat{\Psi}|_\infty \le z$, where

$$z = 2\sqrt{\frac{\log(4p^2/\delta)}{n}} + 2\frac{\log(4p^2/\delta)}{n}\,. \tag{6.7}$$

Next, we control the largest diagonal element of $\hat{\Sigma}$ as follows. We have by definition of $\hat{\Delta}$, for every i

$$\hat{\Delta}_{ii} = \frac{1}{n}\sum_{j=1}^n X_{ji}^2\,.$$

Applying Laurent and Massart (2000, Lemma 1) and a union bound over the p diagonal terms, we get

$$\mathbf{P}\Big(\max_{1\le i\le p}\hat{\Delta}_{ii} \ge 1 + 2\sqrt{\frac{t}{n}} + 2\frac{t}{n}\Big) \le p\,e^{-t}\,.$$

Taking $t = \log(2p/\delta)$ yields yields with probability $1 - \delta/2$

$$\max_{1\le i\le p} \hat{\Delta}_{ii} \le 1 + 2\sqrt{\frac{\log(2p/\delta)}{n}} + 2\frac{\log(2p/\delta)}{n}\,. \tag{6.8}$$

To conclude the proof of Proposition 6.2, observe that (6.5) and (6.6) implies that for all $z \ge 0$, we have

$$\begin{aligned}\lambda^k_{\max}(\hat{\Sigma}) \le \mathsf{SDP}_k(\hat{\Sigma}) &\le \lambda_{\max}(st_z(\hat{\Sigma})) + kz \\ &\le \lambda_{\max}(st_z(\hat{\Delta})) + \lambda_{\max}(st_z(\hat{\Psi})) + kz\,.\end{aligned}$$

where the last inequality follows from (6.6) and the triangle inequality for the operator norm.

Note now that if we take z as in (6.7), then $st_z(\hat{\Psi}) = 0$ with probability $1-\delta/2$. Furthermore, since $\hat{\Delta}$ is a non negative diagonal matrix, then

$$\lambda_{\max}(st_z(\hat{\Delta})) \le \lambda_{\max}(\hat{\Delta}) = \max_{1\le i\le p} \hat{\Delta}_{ii}\,. \tag{6.9}$$

Using that (6.8) holds with probability $1 - \delta/2$, and a simple union bound allows us to ensure that the desired inequality is valid with probability $1 - \delta$. ∎

6.3 Hypothesis testing with convex methods

Using the notation from Section 2, the results of the previous subsection can be written as

$$\begin{aligned}\mathbf{P}_{H_0}(\mathsf{SDP}_k(\hat{\Sigma}) > \tilde{\tau}_0) &\le \delta \\ \mathbf{P}_{H_1}(\mathsf{SDP}_k(\hat{\Sigma}) < \tilde{\tau}_1) &\le \delta\,,\end{aligned}$$

where $\tilde{\tau}_0$ and $\tilde{\tau}_1$ are given by

$$\begin{aligned}\tilde{\tau}_0 &= 1 + 2\sqrt{\frac{k^2\log(4p^2/\delta)}{n}} + 2\frac{k\log(4p^2/\delta)}{n} + 2\sqrt{\frac{\log(2p/\delta)}{n}} + 2\frac{\log(2p/\delta)}{n} \\ \tilde{\tau}_1 &= 1 + \theta - 2(1+\theta)\sqrt{\frac{\log(1/\delta)}{n}}\,.\end{aligned}$$

Whenever $\tilde{\tau}_1 > \tilde{\tau}_0$, we take $\tau \in [\tilde{\tau}_0, \tilde{\tau}_1]$ and define the following computationally efficient test

$$\tilde{\psi}(\hat{\Sigma}) = \mathbf{1}\{\mathsf{SDP}_k(\hat{\Sigma}) > \tau\}\,.$$

It discriminates between H_1 and H_0 with probability $1 - \delta$.

It remains to find for which values of θ the condition $\tilde{\tau}_1 > \tilde{\tau}_0$ holds. It corresponds to our minimum detection level.

THEOREM 6.1. *Assume that p, n, k and δ are such that $\tilde{\theta} \le 1$, where*

$$\tilde{\theta} := 2\sqrt{\frac{k^2\log(4p^2/\delta)}{n}} + 2\frac{k\log(4p^2/\delta)}{n} + 2\sqrt{\frac{\log(2p/\delta)}{n}} + 2\frac{\log(2p/\delta)}{n} + 4\sqrt{\frac{\log(\frac{1}{\delta})}{n}}\,. \tag{6.10}$$

Then, for any $\theta > \tilde{\theta}$, any $\tau \in [\tilde{\tau}_0, \tilde{\tau}_1]$, the test $\tilde{\psi}(\hat{\Sigma}) = \mathbf{1}\{\mathsf{SDP}_k(\hat{\Sigma}) > \tau\}$ discriminates between H_0 and H_1 with probability $1 - \delta$.

If we consider asymptotic regimes, for large p, n, k, taking $\delta = p^{-\beta}$ with $\beta > 0$, provides a sequence of tests $\tilde{\psi}_n$ that discriminate between H_0 and H_1 with probability converging to 1, for any fixed $\theta > 0$, as soon as

$$\frac{k^2 \log(p)}{n} \to 0\,.$$

Note that, compared to Theorem 4.1, the price to pay for using this convex relaxation is to multiply the minimum detection level by a factor $\sqrt{k}$. Of course, in most examples, k remains small so that this is not a very high price.

6.4 Simple methods

While the SDP relaxation proposed in the previous subsection is provably computationally efficient, it is also known to scale poorly on very large problems. Simple heuristics such as the diagonal method of Johnstone (2001) become more attractive for larger problems. A careful inspection of the proofs in the previous subsection is very informative. It indicates that our results not only hold for the test $\tilde{\psi}(\hat{\Sigma})$ but for a simpler statistic arising from the dual formulation (6.5). Indeed, to control the behavior of $\mathsf{SDP}_k(\hat{\Sigma})$ under H_0, we showed that it was no larger than the *minimum dual perturbation* $\mathsf{MDP}_k(\hat{\Sigma})$ defined by

$$\mathsf{MDP}_k(\hat{\Sigma}) = \min_{z \geq 0} \left\{ \lambda_{\max}(st_z(\hat{\Sigma})) + kz \right\}\,. \tag{6.11}$$

Clearly $\mathsf{MDP}_k(\hat{\Sigma}) \geq \mathsf{SDP}_k(\hat{\Sigma}) \geq \lambda^k_{\max}(\hat{\Sigma})$ so that both Proposition 6.1 and Proposition 6.2 still hold for $\mathsf{SDP}(\hat{\Sigma})$ replaced by $\mathsf{MDP}(\hat{\Sigma})$. As a result, for any $\theta > \tilde{\theta}$ the test $\hat{\psi}(\hat{\Sigma}) = \mathbf{1}\{\mathsf{MDP}_k(\hat{\Sigma}) > \tau\}$ discriminates between H_0 and H_1 with probability $1 - \delta$.

Actually, a detection level of the same order as $\tilde{\theta}$ holds already for an even simpler test statistic: the largest diagonal element of $\hat{\Sigma}$. This method called *Johnstone's diagonal method* was first proposed by Johnstone (2001) and later studied by Amini and Wainwright (2009). For the problem of detection considered here, it dictates to employ the test statistic

$$\mathsf{D}(\hat{\Sigma}) = \max_{1 \leq i \leq p} \hat{\Sigma}_{ii}\,.$$

Using even simpler techniques than in Propositions 6.1 and 6.2, it is not hard to show that

$$\begin{aligned} \mathbf{P}_{H_0}(\mathsf{D}(\hat{\Sigma}) > \tau_0^d) &\leq \delta \\ \mathbf{P}_{H_1}(\mathsf{D}(\hat{\Sigma}) < \tau_1^d) &\leq \delta\,, \end{aligned}$$

for levels τ_0^d and τ_1^d given by

$$\begin{aligned} \tau_0^d &= 1 + \frac{1}{k}\theta - 2\left(1 + \frac{1}{k}\theta\right)\sqrt{\frac{\log(1/\delta)}{n}} \\ \tau_1^d &= 1 + 2\sqrt{\frac{\log(p/\delta)}{n}} + 2\frac{\log(p/\delta)}{n}\,. \end{aligned}$$

However, as we shall see in Section 8, MDP_k behaves much better than D in practice. It was proved by Amini and Wainwright (2009) that if the SDP (6.2) has a

solution of rank one then it is strictly better than Johnstone's diagonal method. While they study a support recovery problem different from the detection problem considered here, it seems to indicate that the two methods are qualitatively different. However, the assumption that the SDP (6.2) has a solution of rank one is very strong and unnecessary in our problem. Indeed, while rank one solutions are amenable to extracting sparse eigenvectors, we only need the value of the objective at its maximum and not the solution itself. In this sense, we avoid the main limitation of SDP relaxations to vector problems.

7. GENERALIZATION WITH WEAKENED ASSUMPTIONS

In this section we investigate two extensions of our original problem. For simplicity, we denote by $*\mathsf{DP}_k$ any of the two functionals MDP_k or SDP_k.

7.1 Adversarial noise

While the results for $\lambda_{\max}^k$ rely heavily on the fact that the X_i are Gaussian random vectors, it is not the case for those on the convex relaxation. We can find that under much weaker assumptions, the results for detection using the SDP statistic are still valid. We also describe an adversarial noise setting where we prove that these detection levels are optimal.

In this setting, for an original covariance matrix Σ, assume that

$$\hat{\Sigma} = \Sigma + N\,. \tag{7.1}$$

Where the only assumption on N is that $|N|_\infty \le \sqrt{\log(p/\delta)/n}$ with probability $1-\delta$. Up to a constant, this is a generalization of our initial setting, and can describe a situation where the data is censured, akin to the setting of Loh and Wainwright (2011). Here, however, the situation is more general, as the censured entries are not necessarily chosen randomly.

We show below that the high probability bounds for $\lambda_{\max}^k(\hat{\Sigma}), \mathsf{SDP}_k(\hat{\Sigma})$ and $\mathsf{MDP}_k(\hat{\Sigma})$ under H_0 and H_1 that were constructed before depend only on this very mild assumption.

PROPOSITION 7.1. *Under H_1, we have with probability $1-\delta$*

$$*\mathsf{DP}_k(\hat{\Sigma}) \ge \lambda_{\max}^k(\hat{\Sigma}) \ge 1 + \theta - k\sqrt{\frac{\log(p/\delta)}{n}}\,.$$

PROOF. Recall that for any v such that $|v|_0 \le k$, we have

$$\begin{aligned} *\mathsf{DP}_k(\hat{\Sigma}) \ge \lambda_{\max}^k(\hat{\Sigma}) &\ge v^\top \hat{\Sigma} v \ge v^\top (I_p + \theta v v^\top) v + v^\top N v \\ &\ge 1 + \theta - |N|_\infty |v|_1^2 \\ &\ge 1 + \theta - k|N|_\infty\,, \end{aligned}$$

which yields the desired result. ∎

PROPOSITION 7.2. *Under H_0, we have with probability $1-\delta$*

$$\lambda_{\max}^k(\hat{\Sigma}) \le *\mathsf{DP}_k(\hat{\Sigma}) \le 1 + k\sqrt{\frac{\log(p/\delta)}{n}}\,.$$

PROOF. It follows from (6.4) that

$$\lambda^k_{\max}(\hat{\Sigma}) \le *\mathsf{DP}_k(\hat{\Sigma}) \le \lambda_{\max}(I_p) + k|N|_\infty \,,$$

which yields the desired result. ∎

The following theorem follows from Proposition 7.1 and Proposition 7.2. We omit its proof.

THEOREM 7.1. *Let ψ^{adv} be the test defined by*

$$\psi^{adv}(\hat{\Sigma}) = \mathbf{1}\left\{ *\mathsf{DP}_k(\hat{\Sigma}) > 1 + \frac{k}{2}\sqrt{\frac{\log(p/\delta)}{n}} \right\}.$$

Then the test ψ^{adv} discriminates between H_0 and H_1 with probability $1-\delta$ as soon as

$$\theta > 2k\sqrt{\frac{\log(p/\delta)}{n}}\,.$$

The lower bound proved in Section 5 can be extended to encompass the adversarial setup of this section. The next theorem gives a lower bound on the detection level that holds for an adversarial noise that is bounded in $|\cdot|_\infty$ norm. Note that the lower bound in Theorem 7.2 below is not minimax since there exists one model under which all tests cannot discriminate between H_0 and H_1 with probability less than $1/2$. The model is the following. Let $v = (v_1, \dots, v_p)^\top \in \mathbf{R}^p$ be such that $v_j = 1/\sqrt{k}$ if $j \le k$ and $v_j = 0$ otherwise. Define the random matrix N that takes values $\pm\frac{\theta}{2}vv^\top$, each with probability $1/2$.

THEOREM 7.2. *There exists an adversarial model of the form (7.1) where $|N|_\infty \le \sqrt{\log(p)/n}$ almost surely, such that for any test $\psi(\hat{\Sigma}) \in \{0,1\}$ it holds*

$$\mathbf{P}_{H_1}(\psi(\hat{\Sigma}) = 0) \vee \mathbf{P}_{H_0}(\psi(\hat{\Sigma}) = 1) = 1/2\,.$$

as soon as $\theta \le 2k\sqrt{\log(p)/n}$.

PROOF. Note first that the matrix N defined above satisfies the assumptions on the noise since almost surely, we have

$$|N|_\infty = \frac{\theta}{2k} \le \sqrt{\frac{\log(p)}{n}}\,.$$

Therefore, it holds that

$$\mathbf{P}_{H_0}(\hat{\Sigma} = I_p + \frac{\theta}{2}vv^\top) = \frac{1}{2}$$
$$\mathbf{P}_{H_1}(\hat{\Sigma} = I_p + \frac{\theta}{2}vv^\top) = \frac{1}{2}\,.$$

Therefore, if $\psi(I_p + \frac{\theta}{2}vv^\top) = 1$, then $\mathbf{P}_{H_0}(\psi(\hat{\Sigma}) = 1) = 1/2$ and if $\psi(I_p + \frac{\theta}{2}vv^\top) = 0$, then $\mathbf{P}_{H_1}(\psi(\hat{\Sigma}) = 0) = 1/2$. ∎

Note that unlike Theorem 5.1, Theorem 7.2 gives a lower bound only for tests that depend on $\hat{\Sigma}$. Nonetheless, in the adversarial model (7.1), these are the only tests that make sense since $\hat{\Sigma}$ is the only observation available.

7.2 Sparsity in terms of ℓ_1 norm

Sparsity in terms of ℓ_0 norm is actually very stringent and hardly occurs in real datasets. Rather, it may be more realistic to perform the test

$$\begin{aligned} H_0 &: X \sim \mathcal{N}(0, I_p) \\ \tilde{H}_1 &: X \sim \mathcal{N}(0, I_p + \theta\, vv^\top)\,. \end{aligned}$$

where $|v|_1^2 = \omega$ for some small $\omega > 0$. This allows for vectors $v \in \mathbf{R}^p$ that have ordered coordinates that decay fast enough but never take value zero. It is not hard to see that our analysis extends to this case and the following theorem holds by following the same steps as the proof of Theorem 6.1. We provide it without proof. Recall that $\tilde{\theta}$ is defined in (6.10).

THEOREM 7.3. *Assume that p, n, δ and $k = \omega$ are such that $\tilde{\theta} \le 1$. Then, for any $\theta > \tilde{\theta}$ and for any $\tau \in [\tilde{\tau}_0, \tilde{\tau}_1]$, the test $\mathbf{1}\{*\mathsf{DP}_\omega(\hat{\Sigma}) > \tau\}$ discriminates between H_0 and $\tilde{H}_1$ with probability $1 - \delta$.*

8. NUMERICAL EXPERIMENTS

Computation costs are a crucial element in this study. In Bach et al. (2010), the SDP relaxation with accuracy ε is shown to have a total complexity of $\mathcal{O}(kp^3\sqrt{\log(p)}/\varepsilon)$. This is achieved by minimizing a smooth approximation of the dual function, using first order methods from Nesterov (2003). However, this polynomial cost is already prohibitive in a very high-dimensional setting, and we study only tests based on the MDP_k statistic. The purpose of this section is to illustrate the empirical behavior of tests based on MDP_k and to compare it with the diagonal method.

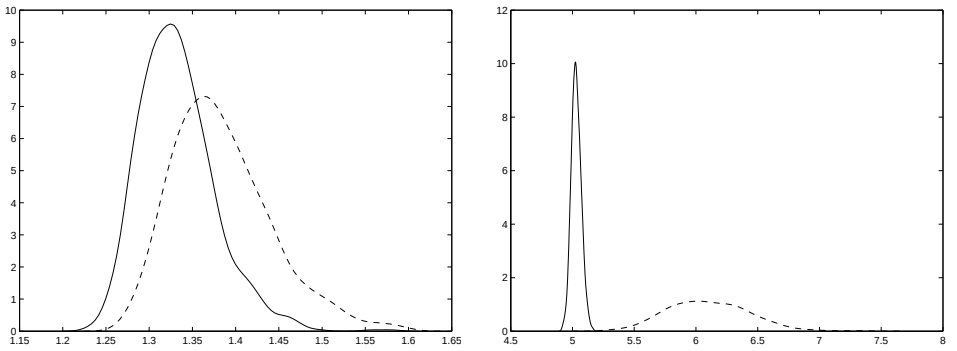
8.1 Comparison of different methods

We simulate $N = 1,000$ samples of n independent random vectors $X_1^0, \dots, X_n^0 \sim \mathcal{N}(0, I_p)$ and $X_1^1, \dots, X_n^1 \sim \mathcal{N}(0, I_p + \theta vv^\top)$, for random unit vectors v supported on $S = \{1, \dots, k\}$. The vector v_S is distributed uniformly on the unit sphere of dimension k.

It yields N empirical covariance matrices $\hat{\Sigma}_1^0, \dots, \hat{\Sigma}_N^0$ under H_0 and N of them, $\hat{\Sigma}_1^1, \dots, \hat{\Sigma}_N^1$ under H_1. We compare the D and MDP_k statistics for these samples and compare their densities. We take $\theta = 4$ and observe that the D statistic yields two distributions under H_0 and H_1 that are hard to distinguish (Figure 1, left). In particular, it is clear that the statistic D cannot discriminate between H_0 and H_1 for $\theta = 4$, with this set of parameters. However, the distributions of $\mathsf{MDP}_k(\hat{\Sigma})$ under H_0 and H_1 have almost disjoint support so that it can discriminate between the two hypothesis with probability close to one.

8.2 Tightness of error bounds

In Section 6, we prove that both the D and MDP_k statistics discriminate between H_0 and H_1 with high probability as long as $\theta \ge Ck\sqrt{\log(p/k)/n}$. The previous subsection indicates that MDP_k actually performs better than D. It is pertinent to wonder if it performs as well as $\lambda_{\max}^k$. Answering this question would actually require implementing $\lambda_{\max}^k$, which is impossible even for a moderate problem size.



Diagonal statistic D MDP_k statistic

FIGURE 1. *For $p = 500$, $n = 200$, $k = 30$, $N = 1,000$, estimated densities for the two statistics, under H_0 (whole line) and under H_1 (dashed line)*

For MDP_k to be considered a performant approximation of $\lambda^k_{\max}$, it would need to discriminate between H_0 and H_1 with high probability as soon as θ is of the order $\sqrt{k\log(p/k)/n}$, which is the minimax optimal detection level that is also achieved by $\lambda^k_{\max}$. This behavior can be illustrated by showing a phase transition for the probability of error in the testing problem, as a function of θ, for different choices of (p, n, k). More precisely, there should exist a critical value θ_{crit} and a constant C_{crit}, such that $\theta > \theta_{\text{crit}} = C_{\text{crit}}\sqrt{k\log(p/k)/n}$, the probability of type II error is very close to 0. Moreover, as a constant, C_{crit} should not depend on (p, n, k).

In order to substantiate such effects, it is actually more pertinent to use a reciprocal setting. For fixed $\theta_0 = 1$ and several choices of parameters p, k, we exhibit a phase transition for the probability of error in the testing problem, as a function of

$$\eta = \eta(n) = \frac{k}{n}\log\left(\frac{p}{k}\right) .$$

In this setting, there should exist a critical value η_{crit}, such that when $\eta < \eta_{\text{crit}} = \theta_0^2/C^2_{\text{crit}}$, the probability of error is very close to 0.

To achieve this goal, we simulate $N = 1,400$ samples of n independent random variables $X^0_1, \dots, X^0_n \sim \mathcal{N}(0, I_p)$. It yields $\hat{\Sigma}^0_1, \dots, \hat{\Sigma}^0_N$ that are drawn under H_0, and used to estimate the quantiles $q_{.01}, q_{.05}$ at 1% and 5% for the MDP_k statistic. The same process is repeated under H_1 to estimate the probability of type II error $\mathbf{P}_{H_1}(\mathsf{MDP}_k(\hat{\Sigma}) > q_\alpha)$. To that end, we simulate $X^1_1, \dots, X^1_n \sim \mathcal{N}(0, I_p + \theta vv^\top)$, for random unit vectors v supported on $S = \{1, \dots, k\}$. The restriction of v to S is distributed uniformly on the unit sphere of dimension k. To display a one-dimensional dependence, k is chosen equal to the integer part of $\sqrt{p}$.

Figure 2 illustrates a phase transition for the probability of testing error, at a critical level $\eta_{\text{crit}} \simeq 0.1$ independent of (p, n, k). The concomitance of these curves for different choices of (p, n, k) indicates that η is the correct scaling factor for the MDP_k statistic. This suggests that the upper bound for convex detection that we prove in (6.10) is pessimistic and that $\mathsf{MDP}_k(\hat{\Sigma})$ is an even better proxy for $\lambda^k_{\max}(\hat{\Sigma})$ than predicted by our theory.

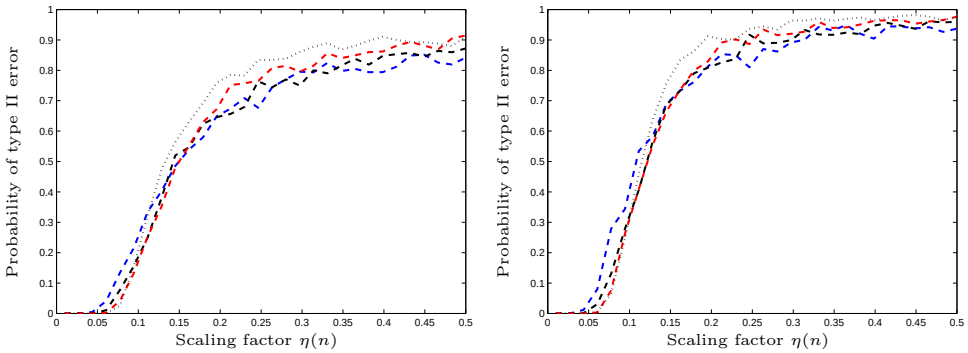


FIGURE 2. *Type II errors as a function of* $\eta(n)$ *for* $p = \{50, 100, 200, 500\}$, $k = \lfloor\sqrt{p}\rfloor$, $N = 1,400$. *Left:* $\alpha = 5\%$, *right:* $\alpha = 1\%$

REFERENCES

ADDARIO-BERRY, L., BROUTIN, N., DEVROYE, L. and LUGOSI, G. (2010). On combinatorial testing problems. *Annals of Statistics*, **38** 3063–3092. Available from: arXiv:0908.3437v2.

ALON, U., BARKAI, N., NOTTERMAN, D. A., GISH, K., YBARRA, S., MACK, D. and LEVINE, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, **96** 6745–6750.

AMINI, A. A. and WAINWRIGHT, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Annals of Statistics*, **37** 2877–2921. Available from: arXiv:0803.4026.

ARIAS-CASTRO, E., BUBECK, S. and LUGOSI, G. (2011). Detection of correlations. *Annals of Statistics*. Available from: arXiv:1106.1193.

ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2010). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *Annals of Statistics*, **39** 2533–2556. Available from: arXiv:1007.1434v1.

BACH, F., AHIPASAOGLU, S. D. and D'ASPREMONT, A. (2010). Convex relaxations for subset selection. *Arxiv Preprint*. Available from: arXiv:1006.3601v1.

BAI, Z. D. (1999). Methodologies in spectral analysis of large-dimensional random matrices, a review. *Statist. Sinica*, **9** 611–677. With comments by G. J. Rodgers and Jack W. Silverstein; and a rejoinder by the author.

BAIK, J., AROUS, G. B. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, **33** 1643–1697.

BARAUD, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, **8** pp. 577–606.

BENAYCH-GEORGES, F., GUIONNET, A. and MAÏDA, M. (2011). Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electron. J. Prob.*, **16** 1621–1662. Available from: arXiv:1009.0145v4.

BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, **36** 2577–2604. Available from: arXiv:0901.3079v1.

BOYD, S. and VANDENBERGHE, L. (2004). *Convex optimization*. Cambridge University Press, Cambridge.

BUTUCEA, C. and INGSTER, Y. I. (2011). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Arxiv Preprint*. Available from: arXiv:1109.0898v1.

CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics*, **38** 2118–2144. Available from: arXiv:1010.3866v1.

CHEN, X. (2011). Adaptive elastic-net sparse principal component analysis for pathway association. *Statistical Applications in Genetics and Molecular Biology*, **10** 48.

D'ASPREMONT, A., BACH, F. and GHAOUI, L. E. (2008). Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, **9** 1269–1294. Available from: arXiv:0707.0705.

D'ASPREMONT, A., GHAOUI, L. E., JORDAN, M. I. and LANCKRIET, G. R. G. (2007). A

direct formulation for sparse pca using semidefinite programming. *SIAM Review*, **49** 434–448. Available from: arXiv:cs/0406021v3.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, **32** 962–994. Available from: arXiv:math/0410072v1.

El-Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Annals of Statistics*, **36** 2757–2790. Available from: arXiv:math/0609418.

Geman, S. (1980). A limit theorem for the norm of random matrices. *Ann. Probab.*, **8** 252–261.

Goemans, M. (1995). Improved approximation algorithms for maximum cut and satisfability problems using semidenite programming. *J. ACM*.

Ingster, Y. I., Tsybakov, A. B. and Verzelen, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.*, **4** 1476–1526. Available from: arXiv:1009.1706v1.

Jenatton, R., Obozinski, G. and Bach, F. (2009). Structured sparse principal component analysis. *Arxiv Preprint*. Available from: arXiv:0909.1440v1.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, **29** 295–327.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.*, **104** 682–693.

Journée, M., Nesterov, Y., Richtárik, P. and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *J. Mach. Learn. Res.*, **11** 517–553.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, **28** 1302–1338.

Loh, P.-L. and Wainwright, M. J. (2011). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *NIPS*. Available from: arXiv:1109.3714v1.

Lu, Z. and Zhang, Y. (2011). An augmented lagrangian approach for sparse principal component analysis. *Mathematical Programming* 1–45. 10.1007/s10107-011-0452-4.

Ma, S. (2011a). Alternating direction method of multipliers for sparse principal component analysis. *SIAM Review*. Available from: arXiv:1111.6703v1.

Ma, Z. (2011b). Sparse principal component analysis and iterative thresholding. *Arxiv Preprint*. Available from: arXiv:1112.2432v1.

Nadler, B. (2008). Finite sample approximation results for principal component analysis: a matrix perturbation approach. *Ann. Statist.*, **36** 2791–2817.

Nesterov, Y. (2003). *Introductory Lectures on Convex Optimization*. Springer.

Nesterov, Y. and Nemirovskii, A. (1987). *Interior-point polynomial algorithms in convex programming*, vol. 13. Society for Industrial Mathematics.

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

Shen, D., Shen, H. and Marron, J. (2009). Consistency of sparse pca in high dimension, low sample size contexts. *Annals of Statistics*, **37** 4104–4130. Available from: arXiv:1104.4289.

Sipser, M. (1996). *Introduction to the Theory of Computation*. 1st ed. International Thomson Publishing.

Sun, X. and Nobel, A. B. (2008). On the size and recovery of submatrices of ones in a random binary matrix. *J. Mach. Learn. Res.*, **9** 2431–2453.

Sun, X. and Nobel, A. B. (2010). On the maximal size of large-average and anova-fit submatrices in a gaussian random matrix. *Bernoulli*. Available from: arXiv:1009.0562v1.

Tao, T. (2011). Outliers in the spectrum of i.i.d. matrices with bounded rank perturbations. *Probab. Theory Related Fields*. Available from: arXiv:1012.4818v4.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics, Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *Arxiv Preprint*. Available from: arXiv:1011.3027v7.

Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. *Electron. J. Stat.*, **6** 38–90. Available from: arXiv:1008.0526v3.

Wright, J., Ganesh, A., Yang, A., Zhou, Z. and Ma, Y. (2011). Sparsity and robustness in face recognition. *Arxiv Preprint*. Available from: arXiv:1111.1014.

Yin, Y. Q., Bai, Z. D. and Krishnaiah, P. R. (1988). On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probab. Theory Related Fields*, **78** 509–521.