JMLR: Workshop and Conference Proceedings vol 30:1–21, 2013

Complexity Theoretic Lower Bounds for Sparse Principal Component Detection

Quentin Berthet QBERTHET@PRINCETON.EDU
Philippe Rigollet RIGOLLET@PRINCETON.EDU
Princeton University

Abstract

In the context of sparse principal component detection, we bring evidence towards the existence of a statistical price to pay for computational efficiency. We measure the performance of a test by the smallest signal strength that it can detect and we propose a computationally efficient method based on semidefinite programming. We also prove that the statistical performance of this test cannot be strictly improved by any computationally efficient method. Our results can be viewed as complexity theoretic lower bounds conditionally on the assumptions that some instances of the planted clique problem cannot be solved in randomized polynomial time.

Keywords: Sparse principal component analysis, Polynomial-time reduction, Planted clique.

The modern scientific landscape has been significantly transformed over the past decade by the apparition of massive datasets. From the statistical learning point of view, this transformation has led to a paradigm shift. Indeed, most novel methods consist in *searching* for sparse structure in datasets, whereas *estimating* parameters over this structure is now a fairly well understood problem. It turns out that most interesting structures have a combinatorial nature, often leading to computationally hard problems. This has led researchers to consider various numerical tricks, chiefly convex relaxations, to overcome this issue. While these new questions have led to fascinating interactions between learning and optimization, they do not always come with satisfactory answers from a statistical point of view. The main purpose of this paper is to study one example, namely sparse principal component detection, for which current notions of statistical optimality should also be shifted, along with the paradigm.

Sparse detection problems where one wants to detect the presence of a sparse structure in noisy data falls in this line of work. There has been recent interest in detection problems of the form signal-plus-noise, where the signal is a vector with combinatorial structure (Addario-Berry et al., 2010; Arias-Castro et al., 2011a,b; Arias-Castro and Verzelen, 2013) or even a matrix (Butucea and Ingster, 2013; Sun and Nobel, 2013; Kolar et al., 2011; Balakrishnan et al., 2011). The matrix detection problem was pushed beyond the signal-plus-noise model towards more complicated dependence structures in Arias-Castro et al. (2012, 2013); Berthet and Rigollet (2012). One contribution of this paper is to extend these results to more general distributions.

For matrix problems, and in particular sparse principal component (PC) detection, some computationally efficient methods have been proposed, but they are not proven

© 2013 Q. Berthet & P. Rigollet.

to achieve the optimal detection levels. Johnstone and Lu (2009); Cai et al. (2012); Ma (2013) suggest heuristics for which detection levels are unknown and Berthet and Rigollet (2012) prove suboptimal detection levels for a natural semidefinite relaxation developed by d'Aspremont et al. (2007) and an even simpler, efficient, dual method called Minimum Dual Perturbation (MDP). More recently, d'Aspremont et al. (2012) developed another semidefinite relaxation for sparse PC detection that performs well only outside of the high-dimensional, low sparsity regime that we are interested in. Note that it follows from the results of Amini and Wainwright (2009) that the former semidefinite relaxation is optimal if it has a rank-one solution. Unfortunately, rank-one solutions can only be guaranteed at suboptimal detection levels. This literature hints at a potential cost for computational efficiency in the sparse PC detection problem.

Partial results were obtained by Berthet and Rigollet (2012) who proved that their bound for MDP and SDP are unlikely to be improved, as otherwise they would lead to randomized polynomial time algorithms for instances of the planted clique problem that are believed to be hard. This result only focuses on a given testing method, but suggests the existence of an intrinsic gap between the optimal rates of detection and what is statistically achievable in polynomial time. Such phenomena are hinted at in Chandrasekaran and Jordan (2013) but their these results focus on the behavior of upper bounds. Closer to our goal, is the work of Shalev-Shwartz et al. (2012) that exhibits a statistical price to pay for computational efficiency. In particular, their derive a computational theoretic lower bound using a much weaker conjecture than the hidden clique conjecture that we employ here, namely the existence of one-way permutations. This conjecture is widely accepted and is the basis of many cryptographic protocols. Unfortunately, the lower bound holds only for a synthetic classification problem that is somewhat tailored to this conjecture. It still remains to fully describe a theory, and to develop lower bounds on the statistical accuracy that is achievable in reasonable computational time for natural problems. This article aims to do so for a general sparse PC detection problem.

This paper is organized in the following way. The sparse PC detection problem is formally described in Section 1. Then, we show in Section 2 that our general detection framework is a natural extension of the existing literature, and that all the usual results for classical detection of sparse PC are still valid. Section 3 focuses on testing in polynomial time, where we study detection levels for the semidefinite relaxation developed by d'Aspremont et al. (2007) (It trivially extends to the MDP statistic of Berthet and Rigollet (2012)). These levels are shown to be unimprovable using computationally efficient methods in Section 4. This is achieved by introducing a new notion of optimality that takes into account computational efficiency. Practically, we reduce the planted clique problem, conjectured to be computationally hard already in an *average-case* sense (i.e. over most random instances) to obtaining better rates for sparse PC detection.

Notation. The space of $d \times d$ symmetric real matrices is denoted by $\mathbf{S}_d$. We write $Z \succeq 0$ whenever Z is semidefinite positive. We denote by $\mathbb{N}$ the set of nonnegative integers and define $\mathbb{N}_1 = \mathbb{N} \setminus \{0\}$.

The elements of a vector $v \in \mathbf{R}^d$ are denoted by $v_1, \dots, v_d$ and similarly, a matrix Z has element Z_{ij} on its ith row and jth column. For any $q > 0$, $|v|_q$ denotes the ℓ_q "norm" of a vector v and is defined by $|v|_q = (\sum_j |v_j|^q)^{1/q}$. Moreover, we denote by $|v|_0$ its so-called ℓ_0 "norm", that is its number of nonzero elements. 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. We also define for $q \in [0,2)$ the set $\mathcal{B}_q(R)$ of unit vectors within the ℓ_q-ball of radius $R > 0$

$$\mathcal{B}_q(R) = \{v \in \mathbf{R}^d \,:\, |v|_2 = 1\,, |v|_q \le R\}\,.$$

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 $i \in S$. The vector $\mathbf{1}$ denotes a vector with coordinates all equal to 1. If a vector has an index such as v_i, then we use $v_{i,j}$ to denote its jth element.

The vectors e_i and matrices E_{ij} are the elements of the canonical bases of $\mathbf{R}^d$ and $\mathbf{R}^{d\times d}$. We also define $\mathcal{S}^{d-1}$ as the unit Euclidean sphere of $\mathbf{R}^d$ and $\mathcal{S}_S^{d-1}$ the set of vectors in $\mathcal{S}^{d-1}$ with support $S \subset \{1,\dots,d\}$. The identity matrix in $\mathbf{R}^d$ is denoted by I_d.

A Bernoulli random variable with parameter $p \in [0,1]$ takes values 1 or 0 with probability p and $1-p$ respectively. A Rademacher random variable takes values 1 or -1 with probability $1/2$. A binomial random variable, with distribution $\mathcal{B}(n,p)$ is the sum of n independent Bernoulli random variables with identical parameter p. A hypergeometric random variable, with distribution $\mathcal{H}(N,k,n)$ is the random number of successes in n draws from a population of size N among which are k successes, without replacement. The total variation norm, noted $\|\cdot\|_{\mathsf{TV}}$ has the usual definition.

The trace and rank functionals are denoted by **Tr** and **rank** respectively and have their usual definition. We denote by T^c the complement of a set T. Finally, for two real numbers a and b, we write $a \wedge b = \min(a,b)$, $a \vee b = \max(a,b)$, and $a_+ = a \vee 0$.

1. Problem description

Let $X \in \mathbf{R}^d$ be a centered random vector with unknown distribution $\mathbf{P}$ that has finite second moment along every direction. The first principal component for X is a direction $v \in \mathcal{S}^{d-1}$ such that the variance $\mathsf{V}(v) = \mathbf{E}[(v^\top X)^2]$ along direction v is larger than in any other direction. If no such v exists, the distribution of X is said to be *isotropic*. The goal of sparse principal component detection is to test whether X follows an isotropic distribution $\mathbf{P}_0$ or a distribution $\mathbf{P}_v$ for which there exists a sparse $v \in \mathcal{B}_0(k)$, $k \ll d$, along which the variance is large. Without loss of generality, we assume that under the isotropic distribution $\mathbf{P}_0$, all directions have unit variance and under $\mathbf{P}_v$, the variance along v is equal to $1+\theta$ for some positive θ. Note that since v has unit norm, θ captures the signal strength.

To perform our test, we observe n independent copies $X_1,\dots,X_n$ of X. For any direction $u \in \mathcal{S}^{d-1}$, define the empirical variance along u by

$$\widehat{\mathsf{V}}_n(u) = \frac{1}{n}\sum_{i=1}^n \left(u^\top X_i\right)^2.$$

Clearly the concentration of $\widehat{\mathsf{V}}_n(u)$ around $\mathsf{V}(u)$ will have a significant effect on the performance of our testing procedure. If, for any $u \in \mathcal{S}^{d-1}$, the centered random variable $(u^\top X)^2 - \mathbb{E}[(u^\top X)^2]$ satisfies the conditions for Bernstein's inequality (see, e.g., Massart,

2007, eq. (2.18), p.24) under both $\mathbf{P}_0$ and $\mathbf{P}_v$, then, up to numerical constants, we have

$$\sup_{u\in\mathcal{S}^{d-1}} \mathbf{P}_0^{\otimes n}\Big(\big|\widehat{\mathsf{V}}_n(u)-1\big| > 4\sqrt{\frac{\log(1/\nu)}{n}} + 4\frac{\log(1/\nu)}{n}\Big) \le \nu\,, \qquad \forall \nu>0\,, \quad (1)$$

$$\mathbf{P}_v^{\otimes n}\Big(\widehat{\mathsf{V}}_n(v)-(1+\theta) < -2\sqrt{\frac{2\theta k\log(2/\nu)}{n}} - 4\frac{\log(2/\nu)}{n}\Big) \le \nu\,, \qquad \forall \nu>0,\ v\in\mathcal{B}_0(k)\,. \quad (2)$$

Such inequalities are satisfied if we assume that $\mathbf{P}_0$ and $\mathbf{P}_v$ are sub-Gaussian distributions for example. Rather than specifying such an ad-hoc assumption, we define the following sets of distributions under which the fluctuations of $\widehat{\mathsf{V}}_n$ around V are of the same order as those of sub-Gaussian distributions. As a result, we formulate our testing problem on the unknown distribution $\mathbf{P}$ of X as follows

$$\begin{aligned} H_0 \;&:\; \mathbf{P}\in\mathcal{D}_0 = \big\{\mathbf{P}_0 \,:\, (1) \text{ holds}\big\} \\ H_1 \;&:\; \mathbf{P}\in\mathcal{D}_1^k(\theta) = \bigcup_{v\in\mathcal{B}_0(k)} \big\{\mathbf{P}_v \,:\, (2) \text{ holds}\big\}\,. \end{aligned}$$

Note that distributions in $\mathcal{D}_0$ and $\mathcal{D}_1^k(\theta)$ are implicitly centered at zero.

We argue that interesting testing procedures should be robust and thus perform well uniformly over these distributions. In the rest of the paper, we focus on such procedures. The existing literature on sparse principal component testing, particularly in Berthet and Rigollet (2012) and Arias-Castro et al. (2012) focuses on multivariate normal distributions, yet only relies on the sub-Gaussian properties of the empirical variance along unit directions. Actually, all the distributional assumptions made in Vu and Lei (2012); Arias-Castro et al. (2012) and Berthet and Rigollet (2012) are particular cases of these hypotheses. We will show that concentration of the empirical variance as in (1) and (2) is sufficient to derive the results that were obtained under the sub-Gaussian assumption.

Recall that a test for this problem is a family $\psi=\{\psi_{d,n,k}\}$ of $\{0,1\}$-valued measurable functions of the data $(X_1,\dots,X_n)$. Our goal is to quantify the smallest signal strength $\theta>0$ for which there exists a test ψ with maximum test error bounded by $\delta>0$, i.e.,

$$\sup_{\substack{\mathbf{P}_0\in\mathcal{D}_0\\ \mathbf{P}_1\in\mathcal{D}_1^k(\theta)}} \Big\{\mathbf{P}_0^{\otimes n}(\psi=1)\vee\mathbf{P}_1^{\otimes n}(\psi=0)\Big\} \le \delta\,.$$

To call our problem "sparse", we need to assume somehow that k is rather small. Throughout the paper, we fix a tolerance $0<\delta<1/3$ (e.g., $\delta=5\%$) and focus on the case where the parameters are in the *sparse regime* $R_0\subset\mathbb{N}_1^3$ of positive integers defined by

$$R_0 = \Big\{(d,n,k)\in\mathbb{N}_1^3 \,:\, 15\sqrt{\frac{k\log(6ed/\delta)}{n}} \le 1\,, k\le d^{0.49}\Big\}\,.$$

Note that the constant 0.49 is arbitrary and can be replaced by any constant $C<0.5$.

Definition 1 *Fix a set of parameters $R\subset R_0$ in the sparse regime. Let $\mathcal{T}$ be a set of tests. A function θ^* of $(d,n,k)\in R$ is called* **optimal rate of detection over the class $\mathcal{T}$** *if for any $(d,n,k)\in R$, it holds:*

(i) there exists a test $\psi \in \mathcal{T}$ that discriminates between H_0 and H_1 at level $\bar{c}\theta^*$ for some constant $\bar{c} > 0$, i.e., for any $\theta \geq \bar{c}\theta^*$

$$\sup_{\substack{\mathbf{P}_0 \in \mathcal{D}_0 \\ \mathbf{P}_1 \in \mathcal{D}_1^k(\theta)}} \left\{ \mathbf{P}_0^{\otimes n}(\psi = 1) \vee \mathbf{P}_1^{\otimes n}(\psi = 0) \right\} \leq \delta \,.$$

In this case we say that $\psi \in \mathcal{T}$ discriminates between H_0 and H_1 at rate θ^*.

(ii) for any test $\phi \in \mathcal{T}$, there exists a constant $\underline{c}_\phi > 0$ such that $\theta \leq \underline{c}_\phi \theta^*$ implies

$$\sup_{\substack{\mathbf{P}_0 \in \mathcal{D}_0 \\ \mathbf{P}_1 \in \mathcal{D}_1^k(\theta)}} \left\{ \mathbf{P}_0^{\otimes n}(\phi = 1) \vee \mathbf{P}_1^{\otimes n}(\phi = 0) \right\} \geq \delta \,.$$

Moreover, if both (i) and (ii) hold, we say that ψ is an optimal test over the class $\mathcal{T}$.

This an adaptation of the usual notion of statistical optimality, when one is focusing on the class of measurable functions, for $\psi_{d,n,k} : (X_1, \ldots, X_n) \mapsto \{0, 1\}$, also known as minimax optimality (see, e.g., Tsybakov, 2009, for an introduction). In order to take into account the asymptotic nature of some classes of statistical tests (namely, those that are computationally efficient), we allow the constant $\underline{c}_\phi$ in *(ii)* to depend on the test.

2. Statistically optimal testing

We focus first on the traditional setting where $\mathcal{T}$ contains all sequences $\{\psi_{d,n,k}\}$ of tests.

Denote by $\Sigma = \mathbb{E}[XX^\top]$ the covariance matrix of X and by $\hat{\Sigma}$ its empirical counterpart:

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top \,. \tag{3}$$

Observe that $\mathsf{V}(u) = u^\top \Sigma \, u$ and $\widehat{\mathsf{V}}_n(u) = u^\top \hat{\Sigma} u$, for any $u \in \mathcal{S}^{d-1}$. Maximizing $\widehat{\mathsf{V}}_n(u)$ over $\mathcal{B}_0(k)$ gives the largest empirical variance along any k-sparse direction. It is also known as the k-sparse eigenvalue of $\hat{\Sigma}$ defined by

$$\lambda_{\max}^k(\hat{\Sigma}) = \max_{u \in \mathcal{B}_0(k)} u^\top \hat{\Sigma} \, u \,. \tag{4}$$

The following theorem describes the performance of the test

$$\psi_{d,n,k} = \mathbf{1}\{\lambda_{\max}^k(\hat{\Sigma}) > 1 + \tau\} \,, \quad \tau > 0 \,. \tag{5}$$

Theorem 2 *Assume that $(d, n, k) \in R_0$ and define*

$$\bar{\theta} = 15 \sqrt{\frac{k \log\left(\frac{6ed}{k\delta}\right)}{n}} \,.$$

Then, for $\bar\theta < \theta < 1$, the test ψ defined in (5) with threshold $\tau = 8\sqrt{\frac{k\log\left(\frac{6ed}{k\delta}\right)}{n}}$, satisfies

$$\sup_{\substack{\mathbf{P}_0\in\mathcal{D}_0\\ \mathbf{P}_1\in\mathcal{D}_1^k(\theta)}} \left\{\mathbf{P}_0^{\otimes n}(\psi=1)\vee\mathbf{P}_1^{\otimes n}(\psi=0)\right\}\le\delta\,.$$

Proof Define $\tau_1 = 7\sqrt{k\log(2/\delta)/n}$. For $\mathbf{P}_1\in\mathcal{D}_1^k(\theta)$, by (2), and for $\mathbf{P}_0\in\mathcal{D}_0$, using Lemma 10, we get

$$\mathbf{P}_0^{\otimes n}\Big(\lambda_{\max}^k(\hat\Sigma)\ge 1+\tau\Big)\le\delta\,,\quad \mathbf{P}_1^{\otimes n}\Big(\lambda_{\max}^k(\hat\Sigma)\le 1+\theta-\tau_1\Big)\le\delta\,.$$

To conclude the proof, observe that $\tau\le\bar\theta-\tau_1<\theta-\tau_1$. ■

The following lower bound follows directly from Berthet and Rigollet (2012), Theorem 5.1 and holds already for Gaussian distributions.

Theorem 3 *For all $\varepsilon>0$, there exists a constant $C_\varepsilon>0$ such that if*

$$\theta<\underline{\theta}_\varepsilon=\sqrt{\frac{k\log\left(C_\varepsilon d/k^2+1\right)}{n}}\,,$$

any test ϕ satisfies

$$\sup_{\substack{\mathbf{P}_0\in\mathcal{D}_0\\ \mathbf{P}_1\in\mathcal{D}_1^k(\theta)}} \left\{\mathbf{P}_0^{\otimes n}(\phi=1)\vee\mathbf{P}_1^{\otimes n}(\phi=0)\right\}\ge\frac{1}{2}-\varepsilon\,.$$

Theorems 2 and 3 imply the following result.

Corollary 4 *The sequence*

$$\theta^*=\sqrt{\frac{k\log d}{n}}\,,\quad (d,n,k)\in R_0\,,$$

is the optimal rate of detection over the class of all tests.

3. Polynomial time testing

It is not hard to prove that approximating $\lambda_{\max}^k(A)$ up to a factor of $m^{1-\varepsilon},\varepsilon>0$, *for any* symmetric matrix A of size $m\times m$ and any $k\in\{1,\dots,m\}$ is NP-hard, by a trivial reduction to CLIQUE (see Håstad, 1996, 1999; Zuckerman, 2006, for hardness of approximation of CLIQUE). Yet, our problem is not worst case and we need not consider *any* matrix A. Rather, here, A is a random matrix and we cannot directly apply the above results.

In this section, we look for a test with good statistical properties and that can be computed in polynomial time. Indeed, finding efficient statistical methods in high-dimension is critical. Specifically, we study a test based on a natural convex (semidefinite) relaxation of $\lambda_{\max}^k(\hat\Sigma)$ developed in d'Aspremont et al. (2007).

For any $A \succeq 0$ let $\mathsf{SDP}_k(A)$ be defined as the optimal value of the following semidefinite program:

$$\begin{aligned} \mathsf{SDP}_k(A) = \max. \quad & \mathbf{Tr}(AZ) \\ \text{subject to} \quad & \mathbf{Tr}(Z) = 1,\ |Z|_1 \le k\,,\ Z \succeq 0 \end{aligned} \tag{6}$$

This optimization problem can be reformulated as a semidefinite program in its canonical form with a polynomial number of constraints and can therefore be solved in polynomial time up to arbitrary precision using interior point methods for example (Boyd and Vandenberghe, 2004), as shown in Appendix A. Consider the following test

$$\psi_{d,n,k} = \mathbf{1}\{\mathsf{SDP}_k^{(n)}(\hat{\Sigma}) > 1 + \tau\}\,, \quad \tau > 0\,, \tag{7}$$

where $\mathsf{SDP}_k^{(n)}$ is a $1/\sqrt{n}$-approximation of SDP_k. Bach et al. (2010) show that $\mathsf{SDP}_k^{(n)}$ can be computed in $\mathcal{O}(kd^3\sqrt{n\log d})$ elementary operations and thus in polynomial time.

Theorem 5 *Assume that (d, n, k) are such that*

$$\tilde{\theta} = 23\sqrt{\frac{k^2\log(4d^2/\delta)}{n}} \le 1\,.$$

Then, for $\theta \in [\tilde{\theta}, 1]$, the test ψ defined in (7) with threshold $\tau = 16\sqrt{\frac{k^2\log(4d^2/\delta)}{n}} + \frac{1}{\sqrt{n}}$, satisfies

$$\sup_{\substack{\mathbf{P}_0 \in \mathcal{D}_0 \\ \mathbf{P}_1 \in \mathcal{D}_1^k(\theta)}} \left\{\mathbf{P}_0^{\otimes n}(\psi = 1) \vee \mathbf{P}_1^{\otimes n}(\psi = 0)\right\} \le \delta\,.$$

Proof Define

$$\tau_0 = 16\sqrt{\frac{k^2\log(4d^2/\delta)}{n}}\,, \qquad \tau_1 = 7\sqrt{\frac{k\log(4/\delta)}{n}}\,.$$

For all $\delta > 0$, $\mathbf{P}_0 \in \mathcal{D}_0, \mathbf{P}_1 \in \mathcal{D}_1^k(\theta)$, by Lemma 11 and Lemma 10, since $\mathsf{SDP}_k(\hat{\Sigma}) \ge \lambda_{\max}^k(\hat{\Sigma})$, it holds

$$\mathbf{P}_0^{\otimes n}\Big(\mathsf{SDP}_k(\hat{\Sigma}) \ge 1 + \tau_0\Big) \le \delta\,, \quad \mathbf{P}_1^{\otimes n}\Big(\mathsf{SDP}_k(\hat{\Sigma}) \le 1 + \theta - \tau_1\Big) \le \delta\,.$$

Recall that $|\mathsf{SDP}_k^{(n)} - \mathsf{SDP}_k| \le 1/\sqrt{n}$ and observe that $\tau_0 + 1/\sqrt{n} = \tau \le \tilde{\theta} - \tau_1 \le \theta - \tau_1$. ■

This size of the detection threshold $\tilde{\theta}$ is consistent with the results of Amini and Wainwright (2009); Berthet and Rigollet (2012) for Gaussian distributions.

Clearly, this theorem, together with Theorem 3, indicate that the test based on SDP may be suboptimal within the class of all tests. However, as we will see in the next section, it can be proved to be optimal in a restricted class of computationally efficient tests.

4. Complexity theoretic lower bounds

It is legitimate to wonder if the upper bound in Theorem 5 is tight. Can faster rates be achieved by this method, or by other, possibly randomized, polynomial time testing methods? Or instead, is this gap intrinsic to the problem? A partial answer to this question is provided in Berthet and Rigollet (2012), where it is proved that the test defined in (7) cannot discriminate at a level significantly lower than $\tilde{\theta}$. Indeed, such a test could otherwise be used to solve instances of the planted clique problem that are believed to be hard. This result is supported by some numerical evidence as well.

In this section, we show that it is true not only of the test based on SDP but of *any* test computable in randomized polynomial time.

4.1. Lower bounds and polynomial time reductions

The upper bound of Theorem 5, if tight, seems to indicate that there is a gap between the detection levels that can be achieved by any test, and those that can be achieved by methods that run in polynomial time. In other words, it indicates a potential statistical cost for computational efficiency. To study this phenomenon, we take the approach favored in theoretical computer science, where our primary goal is to classify problems, rather than algorithms, according to their computational hardness. Indeed, this approach is better aligned with our definition of optimal rate of detection where lower bounds should hold for any tests. Unfortunately, it is difficult to derive a lower bound on the performance of *any* candidate algorithm to solve a given problem. Rather, theoretical computer scientists have developed reductions from problem A to problem B with the following consequence: if problem B can be solved in polynomial time, then so can problem A. Therefore, if problem A is believed to be hard then so is problem B. Note that our reduction requires extra bits of randomness and is therefore a randomized polynomial time reduction.

This question needs to be formulated from a statistical detection point of view. As mentioned above, $\lambda_{\max}^k$ can be proved to be NP-hard to approximate. Nevertheless, such *worst case* results are not sufficient to prove negative results on our *average case* problem. Indeed, the matrix is $\hat{\Sigma}$ is random and we only need to be able to approximate $\lambda_{\max}^k(\hat{\Sigma})$ up to constant factor on most realizations. In some cases, this small nuance can make a huge difference, as problems can be hard in the worst case but easy in average (see, e.g., Boppana (1987) for an illustration on $\mathsf{Graph\ Bisection}$). In order to prove a complexity theoretic lower bound on the sparse principal component detection problem, we will build a reduction from a notoriously hard detection problem: the planted clique problem.

4.2. The Planted Clique problem

Fix an integer $m \geq 2$ and let $\mathbb{G}_m$ denote the set of undirected graphs on m vertices. Denote by $\mathcal{G}(m, 1/2)$ the distribution over $\mathbb{G}_m$ generated by choosing to connect every pair of vertices by an edge independently with probability $1/2$. For any $\kappa \in \{2, \dots, m\}$, the distribution $\mathcal{G}(m, 1/2, \kappa)$ is constructed by picking κ vertices arbitrarily and placing a clique[1] between them, then connect every other pair of vertices by an edge independently with probability $1/2$. Note that $\mathcal{G}(m, 1/2)$ is simply the distribution of an Erdős-Rényi

1. A clique is a subset of fully connected vertices.

random graph. In the decision version of this problem, called Planted Clique, one is given a graph G on m vertices and the goal is to detect the presence of a planted clique.

Definition 6 *Fix $m \geq \kappa > 2$. Let* Planted Clique *denote the following statistical hypothesis testing problem:*

$$\begin{aligned} H_0^{\mathsf{PC}} &: G \sim \mathcal{G}(m, 1/2) = \mathbf{P}_0^{(G)} \\ H_1^{\mathsf{PC}} &: G \sim \mathcal{G}(m, 1/2, \kappa) = \mathbf{P}_1^{(G)} . \end{aligned}$$

A test for the planted clique problem is a family $\xi = \{\xi_{m,\kappa}\}$, where $\xi_{m,\kappa} : \mathbb{G}_m \to \{0, 1\}$.

The search version of this problem Jerrum (1992); Kučera (1995), consists in finding the clique planted under H_1^{PC}. The decision version that we consider here is traditionally attributed to Saks (see Krivelevich and Vu, 2002, Section 5). It is known (see, e.g., Spencer, 1994) that if $\kappa > 2\log_2(m)$, the planted clique is the only clique of size κ in the graph, asymptotically almost surely (a.a.s.). Therefore, a test based on the largest clique of G allows to distinguish H_0^{PC} and H_1^{PC} for $\kappa > 2\log_2(m)$, a.a.s. This is clearly not a computationally efficient test.

For $\kappa = o(\sqrt{m})$ there is no known polynomial time algorithm that solves this problem. Polynomial time algorithms for the case $\kappa = C\sqrt{m}$ were first proposed in Alon et al. (1998), and subsequently in McSherry (2001); Ames and Vavasis (2011); Dekel et al. (2010); Feige and Ron (2010); Feige and Krauthgamer (2000). It is widely believed that there is no polynomial time algorithm that solves Planted Clique for any κ of order m^c for some fixed positive $c < 1/2$. Recent research has been focused on proving that certain algorithmic techniques, such as the Metropolis process Jerrum (1992) and the Lovàsz-Schrijver hierarchy of relaxations Feige and Krauthgamer (2003) fail at this task. The confidence in the difficulty of this problem is so strong that it has led researchers to prove impossibility results assuming that Planted Clique is indeed hard. Examples include cryptographic applications, in Juels and Peinado (2000), testing for k-wise dependence in Alon et al. (2007), approximating Nash equilibria in Hazan and Krauthgamer (2011) and approximating solutions to the densest κ-subgraph problem by Alon et al. (2011).

We therefore make the following assumption on the planted clique problem. Recall that δ is a confidence level fixed throughout the paper.

Hypothesis A_{PC} For any $a, b \in (0, 1), a < b$ and all randomized polynomial time tests $\xi = \{\xi_{m,\kappa}\}$, there exists a positive constant Γ that may depend on ξ, a, b and such that

$$\mathbf{P}_0^{(G)}(\xi_{m,\kappa}(G) = 1) \vee \mathbf{P}_1^{(G)}(\xi_{m,\kappa}(G) = 0) \geq 1.2\delta\,, \quad \forall\, m^{\frac{a}{2}} < \Gamma\kappa < m^{\frac{b}{2}}\,.$$

Note that $1.2\delta < 1/2$ can be replaced by any constant arbitrary close to $1/2$. Since κ is polynomial in m, here a *randomized polynomial time* test is a test that can be computed in time at most polynomial in m and has access to extra bits of randomness. The fact that Γ may depend on ξ is due to the asymptotic nature of polynomial time algorithms. Below is an equivalent formulation of Hypothesis A_{PC}.

Hypothesis B_{PC} For any $a, b \in (0, 1), a < b$ and all randomized polynomial time tests $\xi = \{\xi_{m,\kappa}\}$, there exists $m_0 \geq 1$ that may depend on ξ, a, b and such that

$$\mathbf{P}_0^{(G)}(\xi_{m,\kappa}(G) = 1) \vee \mathbf{P}_1^{(G)}(\xi_{m,\kappa}(G) = 0) \geq 1.2\delta\,, \quad \forall\, m^{\frac{a}{2}} < \kappa < m^{\frac{b}{2}}\,,\ m \geq m_0\,.$$

Note that we do not specify a computational model intentionally. Indeed, for some restricted computational models, Hypothesis A_{PC} can be proved to be true for all $a < b \in (0,1)$ (Rossman, 2010; Feldman et al., 2013). Moreover, for more powerful computational models such as Turing machines, this hypothesis is conjectured to be true. It was shown in Berthet and Rigollet (2012) that improving the detection level of the test based on SDP would lead to a contradiction of Hypothesis A_{PC} for some $b \in (2/3, 1)$. Herefater, we extend this result to all randomized polynomial time algorithms, not only those based on SDP.

4.3. Randomized polynomial time reduction

Our main result is based on a randomized polynomial time reduction of an instance of the planted clique problem to an instance of the sparse PC detection problem. In this section, we describe this reduction and call it the *bottom-left transformation*. For any $\mu \in (0,1)$, define

$$R_\mu = R_0 \cap \{k \geq n^\mu\} \cap \{n < d\}\,.$$

The condition $k \geq n^\mu$ is necessary since "polynomial time" is an intrinsically asymptotic notion and for fixed k, computing $\lambda_{\max}^k$ takes polynomial time in n. The condition $n < d$ is an artifact of our reduction and could potentially be improved. Nevertheless, it characterizes the high-dimensional setup we are interested in and allows us to shorten the presentation.

Given $(d, n, k) \in R_\mu$, fix integers m, κ such that $n \leq m < d$, $k \leq \kappa \leq m$ and let $G = (V, E) \in \mathbb{G}_{2m}$ be an instance of the planted clique problem with a potential clique of size κ. We begin by extracting a bipartite graph as follows. Choose n right vertices $V_{\texttt{right}}$ at random among the $2m$ possible and choose m left vertices $V_{\texttt{left}}$ among the $2m - n$ vertices that are not in $V_{\texttt{right}}$. The edges of this bipartite graph[2] are $E \cap \{V_{\texttt{left}} \times V_{\texttt{right}}\}$. Next, since $d > m$, add $d - m \geq 1$ new left vertices and place an edge between each new left vertex and every old right vertex independently with probability $1/2$. Label the left (resp. right) vertices using a random permutation of $\{1, \dots, d\}$ (resp. $\{1, \dots, n\}$) and denote by $V' = (\{1, \dots, d\} \times \{1, \dots, n\}, E)$ the resulting $d \times n$ bipartite graph. Note that if G has a planted clique of size κ, then V' has a planted biclique of random size.

Let B denote the $d \times n$ adjacency matrix of V' and let $\eta_1, \dots, \eta_n$ be n i.i.d Rademacher random variables that are independent of all previous random variables. Define

$$X_i^{(G)} = \eta_i(2B_i - 1) \in \{-1, 1\}^d\,,$$

where B_i denotes the i-th column of B. Put together, these steps define the bottom-left transformation $\mathsf{bl} : \mathbb{G}_{2m} \to \mathbf{R}^{d \times n}$ of a graph G by

$$\mathsf{bl}(G) = \left(X_1^{(G)}, \dots, X_n^{(G)}\right) \in \mathbf{R}^{d \times n}\,. \tag{8}$$

Note that $\mathsf{bl}(G)$ can be constructed in randomized polynomial time in d, n, k, κ, m.

2. The "bottom-left" terminology comes from the fact that the adjacency matrix of this bipartite graph can be obtained as the bottom-left corner of the original adjacency matrix after a random permutation of the row/columns.

4.4. Optimal detection over randomized polynomial time tests

For any $\alpha \in [1,2]$, define the detection level $\theta_\alpha > 0$ by $\theta_\alpha = \sqrt{\frac{k^\alpha}{n}}$.

Up to logarithmic terms, it interpolates polynomially between the statistically optimal detection level θ^* and the detection level $\tilde{\theta}$ that is achievable by the polynomial time test based on SDP. We have $\theta^* = \theta_1\sqrt{\log d}$ and $\tilde{\theta} = C\theta_2\sqrt{\log d}$ for some positive constant C.

Theorem 7 *Fix $\alpha \in [1,2), \mu \in (0, \frac{1}{4-\alpha})$ and define*

$$a = 2\mu\,, \qquad b = 1 - (2-\alpha)\mu\,. \tag{9}$$

For any $\Gamma > 0$, there exists a constant $L > 0$ such that the following holds. For any $(d,n,k) \in R_\mu$, there exists m, κ such that $(2m)^{\frac{a}{2}} \le \Gamma\kappa \le (2m)^{\frac{b}{2}}$, a random transformation $\mathsf{bl} = \{\mathsf{bl}_{d,n,k,m,\kappa}\}$, $\mathsf{bl}_{d,n,k,m,\kappa} : \mathbb{G}_{2m} \to \mathbf{R}^{d\times n}$ that can be computed in polynomial time and distributions $\mathbf{P}_0 \in \mathcal{D}_0, \mathbf{P}_1 \in \mathcal{D}_1^k(L\theta_\alpha)$ such that for any test $\psi = \{\psi_{d,n,k}\}$, we have

$$\mathbf{P}_0^{\otimes n}(\psi_{d,n,k} = 1) \vee \mathbf{P}_1^{\otimes n}(\psi_{d,n,k} = 0) \ge \mathbf{P}_0^{(G)}(\xi_{m,\kappa}(G) = 1) \vee \mathbf{P}_1^{(G)}(\xi_{m,\kappa}(G) = 0) - \frac{\delta}{5}\,,$$

where $\xi_{m,\kappa} = \psi_{d,n,k} \circ \mathsf{bl}_{d,n,k,m,\kappa}$.

Proof Fix $(d,n,k) \in R_\mu, \alpha \in [1,2)$. First, if G is an Erdős-Rényi graph, $\mathsf{bl}(G) = \big(X_1^{(G)}, \dots, X_n^{(G)}\big)$ is an array of n i.i.d. vectors of d independent Rademacher random variables. Therefore $X_1^{(G)} \sim \mathbf{P}_0^{\mathsf{bl}(G)} \in \mathcal{D}_0$.

Second, if G has a planted clique of size κ, let $\mathbf{P}^{\mathsf{bl}(G)}$ denote the joint distribution of $\mathsf{bl}(G)$. The choices of κ and m depend on the relative size of k and n. Our proof relies on the following lemma, proved in appendix C

Lemma 8 *Fix $\beta > 0$ and integers m, κ, n, k such that $1 \le n \le m$, $2 \le k \le \kappa \le m$,*

$$(a)\ \frac{m}{n} \ge \frac{8}{\beta\delta}\,, \qquad (b)\ \frac{n\kappa}{m} \ge 16\log\big(\frac{m}{n}\big)\,, \qquad (c)\ \frac{n\kappa}{m} \ge 8k\,. \tag{10}$$

Moreover, define

$$\bar{\theta} = \frac{(k-1)\kappa}{2m}\,,$$

Let $G \sim \mathcal{G}(2m, 1/2, \kappa)$ and $\mathsf{bl}(G) = \big(X_1^{(G)}, \dots, X_n^{(G)}\big) \in \mathbf{R}^{d\times n}$ be defined in (8). Denote by $\mathbf{P}_1^{\mathsf{bl}(G)}$ the distribution of $\mathsf{bl}(G)$. Then, there exists a distribution $\mathbf{P}_1 \in \mathcal{D}_1^k(\bar{\theta})$ such that

$$\big\|\mathbf{P}_1^{\mathsf{bl}(G)} - \mathbf{P}_1^{\otimes n}\big\|_{\mathsf{TV}} \le \beta\delta\,.$$

Define $N = \lceil 40/\delta \rceil$. Assume first that $k \ge M^{-1} n^{\frac{1}{4-\alpha}}$ where $M > 0$ is a constant to be chosen large enough (see below). Take $\kappa = \max\big(8, M\log(N)\big)Nk\,, m = Nn$. It implies that

$$\bar{\theta} := \frac{(k-1)\kappa}{2m} \ge \frac{Mk^2}{4n} \ge \frac{1}{4M^{1-\frac{\alpha}{2}}}\sqrt{\frac{k^\alpha}{n}}\,.$$

Moreover, under these conditions, it is easy to check that (10) is satisfied with $\beta = 1/5$ since and we are therefore in a position to apply Lemma 8. It implies that there exists $\mathbf{P}_1 \in \mathcal{D}_1^k(\bar\theta)$ such that $\big\|\mathbf{P}_1^{\mathsf{bl}(G)} - \mathbf{P}_1^{\otimes n}\big\|_{\mathsf{TV}} \le \delta/5$.

Assume now that $k < M^{-1} n^{\frac{1}{4-\alpha}}$. Take $m, \kappa \ge 2$ to be the largest integers such that

$$m \le 2N\big(nk^{2-\alpha}\big)^{\frac{1}{2-b}} \quad \Gamma\kappa \le (2m)^{\frac{b}{2}}\,.$$

Note that $\Gamma\kappa \ge (2m)^{\frac{a}{2}}$. Let us now check condition (10). It holds, for M large enough,

$$\begin{aligned}
&(a) \qquad \frac{m}{n} > \frac{N}{n}\big(n^{1+(2-\alpha)\mu}\big)^{\frac{1}{2-b}} = N \ge 40/\delta.\\
&(b) \qquad \frac{n\kappa}{m} \ge \frac{1}{2\Gamma(4N)^{\frac{b}{2}}}\sqrt{\frac{n}{k^{2-\alpha}}} > \frac{M^{1-\frac{\alpha}{2}}}{2\Gamma(4N)^{\frac{b}{2}}} n^{\frac{1}{4-\alpha}} \ge 16\log\Big(\frac{m}{n}\Big)\,.\\
&(c) \qquad \frac{n\kappa}{m} \ge \frac{1}{2\Gamma(4N)^{\frac{b}{2}}}\sqrt{\frac{n}{k^{2-\alpha}}} > \frac{M^{2-\frac{\alpha}{2}}}{2\Gamma(4N)^{\frac{b}{2}}} k \ge 8k\,.
\end{aligned}$$

Under these conditions, (10) is satisfied with $\beta = 1/5$ and we are therefore in a position to apply Lemma 8. It implies that there exists $\mathbf{P}_1 \in \mathcal{D}_1^k(\bar\theta)$ such that $\big\|\mathbf{P}_1^{\mathsf{bl}(G)} - \mathbf{P}_1^{\otimes n}\big\|_{\mathsf{TV}} \le \delta/5$, where $\bar\theta := \frac{(k-1)\kappa}{2m} \ge \frac{1}{8\Gamma(4N)^{\frac{b}{2}}}\sqrt{\frac{k^\alpha}{n}}$, taking $L = \min\Big(\frac{1}{4M^{\alpha-1}}, \frac{1}{8\Gamma(4N)^{\frac{b}{2}}}\Big)$, yields that $\mathbf{P}_1 \in \mathcal{D}_1^k(L\theta_\alpha)$ for any $(d,n,k) \in R_\mu$. Moreover,

$$\mathbf{P}_0^{(G)}(\psi \circ \mathsf{bl}(G) = 1) \vee \mathbf{P}_1^{(G)}(\psi \circ \mathsf{bl}(G) = 0) \le \mathbf{P}_0^{\otimes n}(\psi = 1) \vee \mathbf{P}_1^{\otimes n}(\psi = 0) + \delta/5\,.$$

■

Theorems 5 and 7 imply the following result.

Corollary 9 *Fix $\alpha \in [1,2), \mu \in (0, \frac{1}{4-\alpha})$. Conditionally on Hypothesis $\mathsf{A_{PC}}$, the optimal rate of detection θ° over the class of randomized polynomial time tests satisfies*

$$\sqrt{\frac{k^\alpha}{n}} \le \theta^\circ \le \sqrt{\frac{k^2 \log d}{n}}\,, \quad (d,n,k) \in R_\mu\,.$$

Proof Let $\mathcal{T}$ denote the class of randomized polynomial time tests. Since bl can be computed in randomized polynomial time, $\psi \in \mathcal{T}$ implies that $\xi = \psi \circ \mathsf{bl} \in \mathcal{T}$. Therefore, for all $(d,n,k) \in R_\mu$,

$$\inf_{\psi \in \mathcal{T}} \mathbf{P}_0^{\otimes n}(\psi = 1) \vee \mathbf{P}_1^{\otimes n}(\psi = 0) \ge \inf_{\xi \in \mathcal{T}} \mathbf{P}_0^{(G)}(\xi(G) = 1) \vee \mathbf{P}_1^{(G)}(\xi(G) = 0) - 0.2\delta = \delta\,.$$

where the last inequality follows from Hypothesis $\mathsf{A_{PC}}$ with a, b as in (9). Therefore $\theta^\circ \ge \theta_\alpha$. The upper bound follows from Theorem 5. ■

The gap between θ° and θ^* in Corollary 4 indicates that the price to pay for using randomized polynomial time tests for the sparse detection problem is essentially of order $\sqrt{k}$.

Acknowledgments

Philippe Rigollet is partially supported by the National Science Foundation grants DMS-0906424 and DMS-1053987. Quentin Berthet is partially supported by a Gordon S. Wu fellowship.

References

Louigi Addario-Berry, Nicolas Broutin, Luc Devroye, and Gábor Lugosi. On combinatorial testing problems. *Annals of Statistics*, 38(5):3063–3092, 08 2010.

Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. In *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, SODA '98, pages 594–598, Philadelphia, PA, USA, 1998. Society for Industrial and Applied Mathematics.

Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *STOC'07—Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, pages 496–505. ACM, New York, 2007.

Noga Alon, Sanjeev Arora, Rajsekar Manokaran, Dana Moshkovitz, and Omri Weinstein. On the inapproximability of the densest κ-subgraph problem. Unpublished, April 2011.

Brendan P.W. Ames and Stephen A. Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Mathematical Programming*, 129:69–89, 2011.

Arash A. Amini and Martin J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *Annals of Statistics*, 37(5B):2877–2921, 03 2009.

E. Arias-Castro, S. Bubeck, and G. Lugosi. Detecting positive correlations in a multivariate sample. *Arxiv Preprint*, 2013. arXiv:1202.5536.

Ery Arias-Castro and Nicolas Verzelen. Community detection in random networks. *Arxiv Preprint*, 02 2013. URL http://arxiv.org/abs/1302.7099.

Ery Arias-Castro, Emmanuel J. Candès, and Arnaud Durand. Detection of an anomalous cluster in a network. *Annals of Statistics*, 39, 01 2011a.

Ery Arias-Castro, Emmanuel J. Candès, and Yaniv Plan. Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.*, 39(5):2533–2556, 2011b. ISSN 0090-5364. doi: 10.1214/11-AOS910. URL http://dx.doi.org/10.1214/11-AOS910.

Ery Arias-Castro, Sébastien Bubeck, and Gábor Lugosi. Detection of correlations. *Ann. Statist.*, 40(1):412–435, 2012.

Francis Bach, Selin Damla Ahipasaoglu, and Alexandre d'Aspremont. Convex relaxations for subset selection. *Arxiv Preprint*, 06 2010. URL http://arxiv.org/abs/1006.3601v1.

Sivaraman Balakrishnan, Mladen Kolar, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Statistical and computational tradeoffs in biclustering. *NIPS 2011 Workshop on Computational Trade-offs in Statistical Learning*, 2011.

Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *ArXiv:1202.5070*, 02 2012. URL http://arxiv.org/abs/1202.5070.

Ravi B. Boppana. Eigenvalues and graph bisection: An average-case analysis. In *Foundations of Computer Science, 1987., 28th Annual Symposium on*, pages 280 –285, oct. 1987.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

Cristina Butucea and Yuri I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli (to appear)*, 2013.

T. Tony Cai, Zongming Ma, and Yihong Wu. Sparse PCA: Optimal rates and adaptive estimation. *Arxiv Preprint*, 11 2012. URL http://arxiv.org/abs/1211.1309.

Venkat Chandrasekaran and Michael I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 2013.

Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49 (3):434–448, July 2007.

Alexandre d'Aspremont, Francis Bach, and Laurent El Ghaoui. Approximation bounds for sparse principal component analysis. *ArXiv:1205.0121*, May 2012. URL http://arxiv.org/abs/1205.0121.

Yael Dekel, Ori Gurel-Gurevich, and Yuval Peres. Finding hidden cliques in linear time with high probability. *Arxiv Preprint*, 2010. URL http://arxiv.org/abs/1010.2997v1.

Persi Diaconis and David Freedman. Finite exchangeable sequences. *Ann. Probab.*, 8(4): 745–764, 1980.

Richard Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

Uriel Feige and Robert Krauthgamer. Finding and certifying a large hidden clique in a semirandom graph. *Random Structures Algorithms*, 16(2):195–208, 2000.

Uriel Feige and Robert Krauthgamer. The probable value of the Lovász-Schrijver relaxations for maximum independent set. *SIAM J. Comput.*, 32(2):345–370 (electronic), 2003.

Uriel Feige and Dorit Ron. Finding hidden cliques in linear time. In *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10)*, Discrete Math. Theor. Comput. Sci. Proc., AM, pages 189–203. Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2010.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for planted clique. In *Proceedings of the Fourty-Fifth Annual ACM Symposium on Theory of Computing, STOC 2013*, 2013.

Johan Håstad. Clique is hard to approximate within $n^{1-\epsilon}$. In *37th Annual Symposium on Foundations of Computer Science (Burlington, VT, 1996)*, pages 627–636. IEEE Comput. Soc. Press, Los Alamitos, CA, 1996.

Johan Håstad. Clique is hard to approximate within $n^{1-\epsilon}$. *Acta Math.*, 182(1):105–142, 1999.

Elad Hazan and Robert Krauthgamer. How hard is it to approximate the best nash equilibrium? *SIAM J. Comput.*, 40(1):79–91, 2011.

Mark Jerrum. Large cliques elude the Metropolis process. *Random Structures Algorithms*, 3(4):347–359, 1992.

Iain M. Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.*, 104(486):682–693, 2009.

Ari Juels and Marcus Peinado. Hiding cliques for cryptographic security. *Des. Codes Cryptogr.*, 20(3):269–280, 2000.

M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh. Minimax localization of structural information in large noisy matrices. *Advances in Neural Information Processing Systems*, 2011.

Michael Krivelevich and Van H. Vu. Approximating the independence number and the chromatic number in expected polynomial time. *J. Comb. Optim.*, 6(2):143–155, 2002.

Luděk Kučera. Expected complexity of graph partitioning problems. *Discrete Appl. Math.*, 57(2-3):193–212, 1995. Combinatorial optimization 1992 (CO92) (Oxford).

Zongming Ma. Sparse principal component analysis and iterative thresholding. *Ann. Statist. (to appear)*, 2013.

Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Frank McSherry. Spectral partitioning of random graphs. In *42nd IEEE Symposium on Foundations of Computer Science (Las Vegas, NV, 2001)*, pages 529–537. IEEE Computer Soc., Los Alamitos, CA, 2001.

Benjamin Rossman. *Average-Case Complexity of Detecting Cliques*. ProQuest LLC, Ann Arbor, MI, 2010. Thesis (Ph.D.)–Massachusetts Institute of Technology.

Shai Shalev-Shwartz, Ohad Shamir, and Eran Tomer. Using more data to speed-up training time. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics April 21-23, 2012 La Palma, Canary Islands.*, volume 22 of *JMLR W&CP*, pages 1019–1027, 2012.

Joel Spencer. *Ten lectures on the probabilistic method*, volume 64 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1994.

Xing Sun and Andrew B. Nobel. On the maximal size of large-average and ANOVA-fit submatrices in a Gaussian random matrix. *Bernoulli*, 19(1):275–294, 2013.

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

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Arxiv Preprint*, 11 2010. URL http://arxiv.org/abs/1011.3027v7.

Vincent Vu and Jing Lei. Minimax rates of estimation for sparse pca in high dimensions. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics April 21-23, 2012 La Palma, Canary Islands.*, volume 22 of *JMLR W&CP*, pages 1278–1286, 2012.

David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, STOC '06, pages 681–690, New York, NY, USA, 2006. ACM.

Appendix A. Canonical representation of $\mathsf{SDP}_k(A)$

In this subsection, we give a reformulation of the optimization problem (6) as a semidefinite program in its canonical form with a polynomial number of constraints. As a result, it can be solved in polynomial time up to arbitrary precision using interior point methods for example (see, e.g., Boyd and Vandenberghe, 2004):

$$\begin{aligned}
\mathsf{SDP}_k(A) = \max. \quad & \sum_{i,j} A_{ij}(z^+_{ij} - z^-_{ij}) \\
\text{subject to} \quad & z^+_{ij} = z^+_{ji} \geq 0,\ z^-_{ij} = z^-_{ji} \geq 0 \\
& \sum_i (z^+_{ii} - z^-_{ii}) = 1,\ \sum_{i,j} (z^+_{ij} + z^-_{ij}) \leq k \\
& \sum_{i>j} (z^+_{ij} - z^-_{ij})(E_{ij} + E_{ji}) + \sum_\ell (z^+_{\ell\ell} - z^-_{\ell\ell}) E_{\ell\ell} \succeq 0\,.
\end{aligned}$$

Appendix B. Technical lemmas used to compute the detection thresholds

Lemma 10 *For all $\mathbf{P}_0 \in \mathcal{D}_0$, and $t > 0$, it holds*

$$\mathbf{P}_0\Big(\lambda^k_{\max}(\hat{\Sigma}) > 1 + 4\sqrt{\frac{t}{n}} + t\frac{t}{n}\Big) \leq \Big(\frac{ed}{k}\Big)^k 9^k e^{-t}\,.$$

Proof We define the following events, for all $S \subset \{1, \ldots, d\}$, $u \in \mathbf{R}^p$, and $t > 0$

$$\begin{aligned}
\mathcal{A} &= \Big\{\lambda^k_{\max}(\hat{\Sigma}) \geq 1 + 4\sqrt{\frac{t}{n}} + 4\frac{t}{n}\Big\} \\
\mathcal{A}_S &= \Big\{\lambda_{\max}(\hat{\Sigma}_S) \geq 1 + 4\sqrt{\frac{t}{n}} + 4\frac{t}{n}\Big\} \\
\mathcal{A}_u &= \Big\{u^\top \hat{\Sigma} u \geq 1 + 2\sqrt{\frac{t}{n}} + 2\frac{t}{n}\Big\}\,.
\end{aligned}$$

By union on all sets of cardinal k, it holds

$$\mathcal{A} \subset \bigcup_{|S|=k} \mathcal{A}_S\,.$$

Furthermore, let $\mathcal{N}_S$, be a minimal covering 1/4-net of $\mathcal{S}^S$, the set of unit vectors with support included in S. It is a classical result that $|\mathcal{N}_S| \leq 9^k$ as shown in Vershynin (2010) and that it holds

$$\lambda_{\max}(\hat{\Sigma}_S - I_S) \leq 2 \max_{u \in \mathcal{N}_S} u^\top (\hat{\Sigma} - I_p) u\,.$$

Therefore it holds

$$\mathcal{A}_S \subset \bigcup_{u \in \mathcal{N}_S} \mathcal{A}_u\,.$$

Hence, by union bound

$$\mathbf{P}_0(\mathcal{A}) \leq \sum_{|S|=k} \sum_{u \in \mathcal{N}_S} \mathbf{P}_0(\mathcal{A}_u)\,.$$

By definition of $\mathcal{D}_0$, $\mathbf{P}_0(A_u) \le e^{-t}$ for $|u|_2 = 1$. The classical inequality $\binom{d}{k} \le \left(\frac{ed}{k}\right)^k$ yields the desired result. ■

Lemma 11 *For all $\mathbf{P}_0 \in \mathcal{D}_0$, and $\delta > 0$, it holds*

$$\mathbf{P}_0\Big(\mathsf{SDP}_k(\hat{\Sigma}) \le 1+2\sqrt{\frac{k^2\log(4d^2/\delta)}{n}}+2\frac{k\log(4d^2/\delta)}{n}+2\sqrt{\frac{\log(2d/\delta)}{n}}+2\frac{\log(2d/\delta)}{n}\Big) \ge 1-\delta\,.$$

Proof We decompose $\hat{\Sigma}$ as the sum of its diagonal and off-diagonal matrices, respectively $\hat{\Delta}$ and $\hat{\Psi}$. Taking $U = -\hat{\Psi}$ in the dual formulation of the semidefinite program (see Bach et al., 2010; Berthet and Rigollet, 2012) yields

$$\mathsf{SDP}_k(\hat{\Sigma}) = \min_{U\in\mathbf{S}^d}\left\{\lambda_{\max}(\hat{\Sigma}+U) + k|U|_\infty\right\} \le |\hat{\Delta}|_\infty + k|\hat{\Psi}|_\infty\,. \tag{11}$$

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

$$\begin{aligned}\hat{\Psi}_{ij} &= \frac{1}{2}\Big[\frac{1}{n}\sum_{\ell=1}^n[\frac{1}{2}(X_{\ell,i}+X_{\ell,j})^2-1] - \frac{1}{n}\sum_{\ell=1}^n[\frac{1}{2}(X_{\ell,i}-X_{\ell,j})^2-1]\Big]\\ &= \frac{1}{2}\Big[\frac{1}{n}\sum_{\ell=1}^n\Big[\Big(\frac{e_i^\top+e_j^\top}{\sqrt{2}}X_\ell\Big)^2-1\Big] - \frac{1}{n}\sum_{k=1}^n\Big[\Big(\frac{e_i^\top-e_j^\top}{\sqrt{2}}X_\ell\Big)^2-1\Big]\Big]\,.\end{aligned}$$

By definition of $\mathcal{D}_0$, it holds for $t > 0$ that

$$\mathbf{P}_0\Big(|\hat{\Psi}_{ij}| \ge 2\sqrt{\frac{t}{n}}+2\frac{t}{n}\Big) \le 4e^{-t}\,.$$

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

$$\mathbf{P}_0\Big(\max_{i<j}|\hat{\Psi}_{ij}| \ge 2\sqrt{\frac{t}{n}}+2\frac{t}{n}\Big) \le 2d^2e^{-t}\,.$$

Taking $t = \log(4p^2/\delta)$ yields that under $\mathbf{P}_0$ with probability $1-\delta/2$,

$$|\hat{\Psi}|_\infty \le 2\sqrt{\frac{\log(4d^2/\delta)}{n}}+2\frac{\log(4d^2/\delta)}{n}\,. \tag{12}$$

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

$$\hat{\Delta}_{ii} = \frac{1}{n}\sum_{\ell=1}^n(e_i^\top X_\ell)^2\,.$$

Similarly, by union bound over the p diagonal terms, it holds

$$\mathbf{P}_0\Big(|\hat{\Delta}|_\infty \ge 1+2\sqrt{\frac{t}{n}}+2\frac{t}{n}\Big) \le d\,e^{-t}\,.$$

Taking $t = \log(2p/\delta)$ yields, under $\mathbf{P}_0$ with probability $1 - \delta/2$,

$$|\hat{\Delta}|_\infty \leq 1 + 2\sqrt{\frac{\log(2d/\delta)}{n}} + 2\frac{\log(2d/\delta)}{n}\,. \tag{13}$$

The desired result is obtained by plugging (12) and (13) into (11). ■

Appendix C. Proof of Lemma 8

Proof Let $S \subset \{1, \ldots, n\}$ (resp. $T \subset \{1, \ldots, d\}$) denote the (random) right (resp. left) vertices of V' that are in the planted biclique.

Define the random variables

$$\begin{aligned} \varepsilon'_i &= \mathbf{1}\{i \in S\}, \quad i = 1, \ldots, n \\ \gamma'_j &= \mathbf{1}\{j \in T\}, \quad j = 1, \ldots, d\,. \end{aligned}$$

On the one hand, if $i \notin S$, i.e., if $\varepsilon'_i = 0$, then $X_i^{(G)}$ is a vector of independent Rademacher random variables. On the other hand, if $i \in S$, i.e., if $\varepsilon'_i = 1$ then, for any $j = 1, \ldots, d$,

$$X_{i,j}^{(G)} = Y'_{i,j} = \begin{cases} \eta_i & \text{if } \gamma'_j = 1\,, \\ r_{ij} & \text{otherwise,} \end{cases}$$

where $r = \{r_{ij}\}_{ij}$ is a $n \times d$ matrix of i.i.d Rademacher random variables.

We can therefore write

$$X_i^{(G)} = (1 - \varepsilon'_i) r_i + \varepsilon'_i Y'_i\,, \quad i = 1, \ldots, n\,,$$

where $Y'_i = (Y'_{i,1}, \ldots, Y'_{i,d})^\top$ and $r_i^\top$ is the ith row of r.

Note that the ε'_is are not independent. Indeed, they correspond to n draws *without* replacement from an urn that contains $2m$ balls (vertices) among which κ are of type 1 (in the planted clique) and the rest are of type 0 (outside of the planted clique). Denote by $\mathbf{p}_{\varepsilon'}$ the joint distribution of $\varepsilon' = (\varepsilon'_1, \ldots, \varepsilon'_n)$ and define their "with replacement" counterparts as follows. Let $\varepsilon_1, \ldots, \varepsilon_n$ be n i.i.d. Bernoulli random variables with parameter $p = \frac{\kappa}{2m} \leq \frac{1}{2}$. Denote by $\mathbf{p}_\varepsilon$ the joint distribution of $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$.

We also replace the distribution of the γ'_js as follows. Let $\gamma = (\gamma_1, \ldots, \gamma_n)$ have conditional distribution given ε be given by

$$\mathbf{p}_{\gamma|\varepsilon}(A) = \mathbf{P}\Big(\gamma' \in A \Big| \sum_{i=1}^{d} \gamma' \geq k, \varepsilon' = \varepsilon\Big)\,.$$

Define $\big(X_1, \ldots, X_n\big)$ by

$$X_i = (1 - \varepsilon_i) r_i + \varepsilon_i Y_i\,, \quad i = 1, \ldots, n\,,$$

where $Y_i \in \mathbf{R}^d$ has coordinates given by

$$Y_{i,j} = \begin{cases} \eta_i & \text{if } \gamma_j = 1 \\ r_{ij} & \text{otherwise} \end{cases}$$

With this construction, the X_is are iid. Moreover, as we will see, the joint distribution $\mathbf{P}_1^{\mathsf{bl}(G)}$ of $\mathsf{bl}(G) = \big(X_1^{(G)}, \dots, X_n^{(G)}\big)$ is close in total variation to the joint distribution $\mathbf{P}_1^{\otimes n}$ of $\big(X_1, \dots, X_n\big)$.

Note first that Markov's inequality yields

$$\mathbf{P}\Big(\sum_{i=1}^n \varepsilon_i > \frac{\kappa}{2}\Big) \le \frac{2np}{\kappa} = \frac{n}{m}\,. \tag{14}$$

Moreover, given $\sum_{i=1}^n \varepsilon_i = s$, we have $\sum_{i=1}^d \gamma_i \ge U \sim \mathcal{H}(2m-n, \kappa - s, n)$. It follows from Diaconis and Freedman (1980, Theorem (4)), that

$$\Big\|\mathcal{H}(2m-n,\kappa-s,n) - \mathcal{B}\Big(n, \frac{\kappa-s}{2m-n}\Big)\Big\|_{\mathsf{TV}} \le \frac{4n}{2m-n} \le \frac{4n}{m}\,.$$

Together with the Chernoff-Okamoto inequality (Dudley, 1999, Equation (1.3.10)), it yields

$$\mathbf{P}\Big(U < \frac{n(\kappa-s)}{2m-n} - \sqrt{\frac{n(\kappa-s)}{2m-n}\log\big(\frac{m}{n}\big)}\Big|\sum_{i=1}^n \varepsilon_i = s\Big) \le \frac{n}{m} + \frac{4n}{m} = \frac{5n}{m}\,.$$

Combined with (14) and view of (10)(b,c), it implies that with probability $1 - 6n/m$, it holds

$$\sum_{j=1}^d \gamma_j \ge U \ge \frac{n\kappa}{4m} - \sqrt{\frac{n\kappa}{4m}\log\big(\frac{m}{n}\big)} \ge \frac{n\kappa}{8m} \ge k\,. \tag{15}$$

Denote by $\mathbf{p}$ the joint distribution of $(\varepsilon_1, \dots, \varepsilon_n, \gamma_1, \dots, \gamma_d)$ and by $\mathbf{p}'$ that of $(\varepsilon_1', \dots, \varepsilon_n', \gamma_1', \dots, \gamma_d')$. Using again Diaconis and Freedman (1980, Theorem (4)) and (10)(a), we get

$$\|\mathbf{p}' - \mathbf{p}\|_{\mathsf{TV}} \le \frac{6n}{m} + \|\mathbf{p}_{\varepsilon'} - \mathbf{p}_{\varepsilon}\|_{\mathsf{TV}} \le \frac{6n}{m} + \frac{4n}{2m} = \frac{8n}{m} \le \beta\delta\,.$$

Since the conditional distribution of $\big(X_1, \dots, X_n\big)$ given (ε, γ) is the same as that of $\mathsf{bl}(G)$ given (ε', γ'), we have

$$\|\mathbf{P}_1^{\mathsf{bl}(G)} - \mathbf{P}_1^{\otimes n}\|_{\mathsf{TV}} = \|\mathbf{p}' - \mathbf{p}\|_{\mathsf{TV}} \le \beta\delta\,.$$

It remains to prove that $\mathbf{P}_1 \in \mathcal{D}_1^k(\bar\theta)$. Fix $\nu > 0$ and define $Z \in \mathcal{B}_0(k)$ by

$$Z_j = \begin{cases} \gamma_j/\sqrt{k}\,, & \text{if } \sum_{i=1}^j \gamma_i \le k \\ 0 & \text{otherwise.} \end{cases}$$

Denote by $S_Z \subset \{1, \dots, d\}$, the support of Z. Next, observe that for any $x, \theta > 0$, it holds

$$\inf_{v \in \mathcal{B}_0(k)} \mathbf{P}_1^{\otimes n}\Big(\widehat{\mathsf{V}}_n(v) - (1+\theta) < -x\Big) \le \mathbf{P}_1^{\otimes n}\Big(\widehat{\mathsf{V}}_n(Z) - (1+\theta) < -x\Big)\,. \tag{16}$$

Moreover, for any $i = 1, \dots, n$

$$(Z^\top X_i)^2 = \frac{1}{k}\Big(k\varepsilon_i\eta_i + (1-\varepsilon_i)\sum_{j\in S_Z} r_{ij}\Big)^2 = \varepsilon_i k + (1-\varepsilon_i)\frac{1}{k}\Big(\sum_{j\in S_Z} r_{ij}\Big)^2\,.$$

Therefore, since Z is independent of the r_{ij}s, the following equality holds in distribution:

$$(Z^\top X_i)^2 \overset{\text{dist.}}{=} 1 + \varepsilon_i(k-1) + \frac{2(1-\varepsilon_i)}{k} \sum_{\ell=1}^{\binom{k}{2}} \omega_{i,\ell}\,,$$

where $\omega_{i,\ell}, i, \ell \geq 1$ is a sequence of i.i.d Rademacher random variables that are independent of the ε_is. Note that by Hoeffding's inequality, it holds with probability at least $1-\nu/2$,

$$\frac{2}{nk} \sum_{i=1}^{n} \sum_{\ell=1}^{\binom{k}{2}} \omega_{i,\ell} \geq -\frac{4}{nk} \sqrt{2n \binom{k}{2} \log(2/\nu)} \geq -4\sqrt{\frac{\log(2/\nu)}{n}}\,.$$

Moreover, it follows from the Chernoff-Okamoto inequality (Dudley, 1999, Equation (1.3.10)) that with probability at least $1-\nu/2$, it holds

$$\frac{k-1}{n} \sum_{i=1}^{n} \varepsilon_i \geq \frac{(k-1)}{n} np - \frac{k-1}{n} \sqrt{2np \log(2/\nu)}\,.$$

Put together, the above two displays imply that with probability $1-\nu$, it holds

$$\begin{aligned}
\widehat{\mathsf{V}}_n(Z) &> 1 + \frac{(k-1)\kappa}{2m} - \frac{k-1}{n}\sqrt{\frac{n\kappa}{m}\log(2/\nu)} - 4\sqrt{\frac{\log(2/\nu)}{n}} \\
&\geq 1 + \frac{(k-1)\kappa}{2m} - \sqrt{2k\frac{(k-1)\kappa}{2m}\frac{\log(2/\nu)}{n}} - 4\sqrt{\frac{\log(2/\nu)}{n}} \\
&= 1 + \bar{\theta} - \sqrt{2k\bar{\theta}\frac{\log(2/\nu)}{n}} - 4\sqrt{\frac{\log(2/\nu)}{n}}\,.
\end{aligned}$$

Together with (16), this completes the proof. ■