STATISTICAL AND COMPUTATIONAL TRADE-OFFS IN ESTIMATION OF SPARSE PRINCIPAL COMPONENTS

BY TENGYAO WANG[*] QUENTIN BERTHET[*,†] RICHARD J. SAMWORTH[*]

University of Cambridge[*]
California Institute of Technology[†]

In recent years, Sparse Principal Component Analysis has emerged as an extremely popular dimension reduction technique for high-dimensional data. The theoretical challenge, in the simplest case, is to estimate the leading eigenvector of a population covariance matrix under the assumption that this eigenvector is sparse. An impressive range of estimators have been proposed; some of these are fast to compute, while others are known to achieve the minimax optimal rate over certain Gaussian or subgaussian classes. In this paper we show that, under a widely-believed assumption from computational complexity theory, there is a fundamental trade-off between statistical and computational performance in this problem. More precisely, working with new, larger classes satisfying a Restricted Covariance Concentration condition, we show that there is an effective sample size regime in which no randomised polynomial time algorithm can achieve the minimax optimal rate. We also study the theoretical performance of a (polynomial time) variant of the well-known semidefinite relaxation estimator, revealing a subtle interplay between statistical and computational efficiency.

1. Introduction. Principal Component Analysis (PCA), which involves projecting a sample of multivariate data onto the space spanned by the leading eigenvectors of the sample covariance matrix, is one of the oldest and most widely-used dimension reduction devices in Statistics. It has proved to be particularly effective when the dimension of the data is relatively small by comparison with the sample size. However, the work of Johnstone and Lu (2009) and Paul (2007) shows that PCA breaks down in the high-dimensional settings that are frequently encountered in many diverse modern application areas. For instance, consider the spiked covariance model where $X_1, \ldots, X_n$ are independent $N_p(0, \Sigma)$ random vectors, with $\Sigma = I_p + \theta v_1 v_1^\top$ for some $\theta > 0$ and an arbitrary unit vector $v_1 \in \mathbb{R}^p$. In this case, v_1 is the leading eigenvector (principal component) of Σ, and the classical PCA estimate would be $\hat{v}_1$, a unit-length leading eigenvector of the sample covariance ma-

1

trix $\hat{\Sigma} := n^{-1} \sum_{i=1}^{n} X_i X_i^{\top}$. In the high-dimensional setting where $p = p_n$ is such that $p/n \to c \in (0, 1)$, Paul (2007) showed that

$$|\hat{v}_1^{\top} v_1| \overset{a.s.}{\to} \begin{cases} \sqrt{\frac{1-c/\theta^2}{1+c/\theta}} & \text{if } \theta > \sqrt{c} \\ 0 & \text{if } \theta \leq \sqrt{c}. \end{cases}$$

In other words, $\hat{v}_1$ is inconsistent as an estimator of v_1 in this asymptotic regime. This phenomenon is related to the so-called 'BBP' transition in random matrix theory (Baik, Ben Arous and Péché, 2005).

Sparse Principal Component Analysis was designed to remedy this inconsistency and to give additional interpretability to the projected data. In the simplest case, it is assumed that the leading eigenvector v_1 of the population covariance matrix Σ belongs to the k-sparse unit Euclidean sphere in $\mathbb{R}^p$, given by

$$(1) \qquad B_0(k) := \left\{ u = (u_1, \ldots, u_p)^{\top} \in \mathbb{R}^p : \sum_{j=1}^{p} \mathbb{1}_{\{u_j \neq 0\}} \leq k, \ \|u\|_2 = 1 \right\}.$$

A remarkable number of recent papers have proposed estimators of v_1 in this setting, including Jolliffe, Trendafilov and Uddin (2003), Zou, Hastie and Tibshirani (2006), d'Aspremont et al. (2007), Johnstone and Lu (2009), Witten, Tibshirani and Hastie (2009), Journée et al. (2010), Birnbaum et al. (2013), Cai, Ma and Wu (2013), Ma (2013), Shen, Shen and Marron (2013) and Vu and Lei (2013).

Sparse PCA methods have gained high popularity in many diverse applied fields where high-dimensional datasets are routinely handled. These include computer vision for online visual tracking (Wang, Lu and Yang, 2013) and pattern recognition (Naikal, Yang and Sastry, 2011), signal processing for image compression (Majumdar, 2009) and Electrocardiography feature extraction (Johnstone and Lu, 2009), and biomedical research for gene expression analysis (Zou, Hastie and Tibshirani, 2006; Chun and Sündüz, 2009; Parkhomenko, Tritchler and Beyene, 2009; Chan and Hall, 2010), RNA-seq classification (Tan, Petersen and Witten, 2014) and metabolomics studies (Genevera and Maletić-Savatić, 2011). In these applications, Sparse PCA is employed to identify a small number of interpretable directions that represent the data succinctly, typically as the first stage of a more involved procedure such as classification, clustering or regression.

The success of the ultimate inferential methods in the types of application described above depends critically on how well the particular Sparse PCA technique involved identifies the relevant meaningful directions in the underlying population. It therefore becomes important to understand the

ways in which our ability to estimate these directions from data depends on the characteristics of the problem, including the sample size, dimensionality, sparsity level and signal-to-noise ratio. Such results form a key component of any theoretical analysis of an inference problem in which Sparse PCA is employed as a first step.

In terms of the theoretical properties of existing methods for Sparse PCA, Ma (2013) was able to show that his estimator attains the minimax rate of convergence over a certain Gaussian class of distributions, provided that k is treated as a fixed constant. Both Cai, Ma and Wu (2013) and Vu and Lei (2013) also study minimax properties, but treat k as a parameter of the problem that may vary with the sample size n. In particular, for a certain class $\mathcal{P}_p(n, k)$ of subgaussian distributions and in a particular asymptotic regime, Vu and Lei (2013) show[1] that

$$\inf_{\hat{v}} \sup_{P \in \mathcal{P}_p(n,k)} \mathbb{E}_P\{1 - (v_1^\top \hat{v})^2\} \asymp \frac{k \log p}{n},$$

where the infimum is taken over all estimators $\hat{v}$; see also Birnbaum et al. (2013). Moreover, they show that the minimax rate is attained by a leading k-sparse eigenvector of $\hat{\Sigma}$, given by

$$(2) \qquad \hat{v}_{\max}^k \in \underset{u \in B_0(k)}{\operatorname{argmax}} \, u^\top \hat{\Sigma} u.$$

The papers cited above would appear to settle the question of sparse principal component estimation (at least in a subgaussian setting) from the perspective of statistical theory. However, there remains an unsettling feature, namely that neither the estimator of Cai, Ma and Wu (2013), nor that of Vu and Lei (2013), is computable in polynomial time[2]. For instance, computing the estimator (2) is an NP-hard problem, and the naive algorithm that searches through all $\binom{p}{k}$ of the $k \times k$ principal submatrices of $\hat{\Sigma}$ quickly becomes infeasible for even moderately large p and k.

Given that Sparse PCA methods are typically applied to massive high-dimensional datasets, it is crucial to understand the rates that can be achieved using only computationally efficient procedures. Specifically, in this paper, we address the question of whether it is possible to find an estimator of v_1 that is computable in (randomised) polynomial time, and that

---

[1]Here and below, $a_n \asymp b_n$ means $0 < \liminf_{n \to \infty} |a_n/b_n| \leq \limsup_{n \to \infty} |a_n/b_n| < \infty$.

[2]Since formal definitions of such notions from computational complexity theory may be unfamiliar to many statisticians, and to keep the paper as self-contained as possible, we provide a brief introduction to this topic in Section 2 of the online supplementary material Wang, Berthet and Samworth (2015).

attains the minimax optimal rate of convergence when the sparsity of v_1 is allowed to vary with the sample size. Some progress in a related direction was made by Berthet and Rigollet (2013a,b), who considered the problem of testing the null hypothesis $H_0 : \Sigma = I_p$ against the alternative $H_1 : v^\top \Sigma v \geq 1 + \theta$ for some $v \in B_0(k)$ and $\theta > 0$. Of interest here is the minimal level $\theta = \theta_{n,p,k}$ that ensures small asymptotic testing error. Under a hypothesis on the computational intractability of a certain well-known problem from theoretical computer science (the 'Planted Clique' detection problem), Berthet and Rigollet showed that for certain classes of distributions, there is a gap between the minimal θ-level permitting successful detection with a randomised polynomial time test, and the corresponding θ-level when arbitrary tests are allowed.

The particular classes of distributions considered in Berthet and Rigollet (2013a,b) were highly tailored to the testing problem, and do not provide sufficient structure to study principal component estimation. The thesis of this paper, however, is that from the point of view of both theory and applications, it is the estimation of sparse principal components, rather than testing for the existence of a distinguished direction, that is the more natural and fundamental (as well as more challenging) problem. Indeed, we observe subtle phase transition phenomena that are absent from the hypothesis testing problem; see Section 4.4 for further details. It is worth noting that different results for statistical and computational trade-offs for estimation and testing were also observed in the context of k-SAT formulas in Feldman, Perkins and Vempala (2015) and Berthet (2015) respectively.

Our first contribution, in Section 2, is to introduce a new Restricted Covariance Concentration (RCC) condition that underpins the classes of distributions $\mathcal{P}_p(n, k, \theta)$ over which we perform the statistical and computational analyses (see (4) for a precise definition). The RCC condition is satisfied by subgaussian distributions, and moreover has the advantage of being more robust to certain mixture contaminations that turn out to be of key importance in the statistical analysis under the computational constraint. We show that subject to mild restrictions on the parameter values,

$$\inf_{\hat{v}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}, v_1) \asymp \sqrt{\frac{k \log p}{n\theta^2}},$$

where $L(u, v) := \{1 - (u^\top v)^2\}^{1/2}$, and where no restrictions are placed on the class of estimators $\hat{v}$. By contrast, in Section 3, we show that a variant $\hat{v}^{\mathrm{SDP}}$ of the semidefinite relaxation estimator of d'Aspremont et al. (2007) and Bach, Ahipaşaoğlu and d'Aspremont (2010), which is computable in

polynomial time, satisfies

$$\sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}^{\mathrm{SDP}}, v_1) \leq (16\sqrt{2} + 2)\sqrt{\frac{k^2 \log p}{n\theta^2}}.$$

Our main result, in Section 4, is that, under a much weaker Planted Clique hypothesis than that in Berthet and Rigollet (2013a,b), for any $\alpha \in (0,1)$, there exists a moderate effective sample size asymptotic regime in which every sequence $(\hat{v}^{(n)})$ of randomised polynomial time estimators satisfies

$$\sqrt{\frac{n\theta^2}{k^{1+\alpha} \log p}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}^{(n)}, v_1) \to \infty.$$

This result shows that there is a fundamental trade-off between statistical and computational efficiency in the estimation of sparse principal components, and that there is in general no consistent sequence of randomised polynomial time estimators in this regime. Interestingly, in a high effective sample size regime, where even randomised polynomial time estimators can be consistent, we are able to show in Theorem 7 that under additional distributional assumptions, a modified (but still polynomial time) version of $\hat{v}^{\mathrm{SDP}}$ attains the minimax optimal rate. Thus, the trade-off disappears for a sufficiently high effective sample size, at least over a subset of the parameter space.

Statistical and computational trade-offs have also recently been studied in the context of convex relaxation algorithms (Chandrasekaran and Jordan, 2013), submatrix signal detection (Ma and Wu, 2015; Chen and Xu, 2014), sparse linear regression (Zhang, Wainwright and Jordan, 2014), community detection (Hajek, Wu and Xu, 2014) and Sparse Canonical Correlation Analysis (Gao, Ma and Zhou, 2014). Given the importance of computationally feasible algorithms with good statistical performance in today's era of Big Data, it seems clear that understanding the extent of this phenomenon in different settings will represent a key challenge for theoreticians in the coming years.

Proofs of our main results are given in the Appendix, while several ancillary results are deferred to the online supplementary material Wang, Berthet and Samworth (2015). We end this section by introducing some notation used throughout the paper. For a vector $u = (u_1, \ldots, u_M)^\top \in \mathbb{R}^M$, a matrix $A = (A_{ij}) \in \mathbb{R}^{M \times N}$ and for $q \in [1, \infty)$, we write $\|u\|_q := \left(\sum_{i=1}^M |u_i|^q\right)^{1/q}$ and $\|A\|_q := \left(\sum_{i=1}^M \sum_{j=1}^N |A_{ij}|^q\right)^{1/q}$ for their (entrywise) ℓ_q-norms. We also write $\|u\|_0 := \sum_{i=1}^M \mathbb{1}_{\{u_i \neq 0\}}$, $\mathrm{supp}(u) := \{i : u_i \neq 0\}$, $\|A\|_0 := \sum_{i=1}^M \sum_{j=1}^N \mathbb{1}_{\{A_{ij} \neq 0\}}$

and $\mathrm{supp}(A) := \{(i,j) : A_{ij} \neq 0\}$. For $S \subseteq \{1, \ldots, M\}$ and $T \subseteq \{1, \ldots, N\}$, we write $u_S := (u_i : i \in S)^\top$ and write $M_{S,T}$ for the $|S| \times |T|$ submatrix of M obtained by extracting the rows and columns with indices in S and T respectively. For positive sequences (a_n) and (b_n), we write $a_n \ll b_n$ to mean $a_n/b_n \to 0$.

2. Restricted Covariance Concentration and minimax rate of estimation. Let $p \geq 2$ and let $\mathcal{P}$ denote the class of probability distributions P on $\mathbb{R}^p$ with $\int_{\mathbb{R}^p} x \, dP(x) = 0$ and such that the entries of $\Sigma(P) := \int_{\mathbb{R}^p} xx^\top \, dP(x)$ are finite. For $P \in \mathcal{P}$, write $\lambda_1(P), \ldots, \lambda_p(P)$ for the eigenvalues of $\Sigma(P)$, arranged in decreasing order. When $\lambda_1(P) - \lambda_2(P) > 0$, the first principal component $v_1(P)$, i.e. a unit-length eigenvector of Σ corresponding to the eigenvalue $\lambda_1(P)$, is well-defined up to sign. In some places below, and where it is clear from the context, we suppress the dependence of these quantities on P, or write the eigenvalues and eigenvectors as $\lambda_1(\Sigma), \ldots, \lambda_p(\Sigma)$ and $v_1(\Sigma), \ldots, v_p(\Sigma)$ respectively. Let $X_1, \ldots, X_n$ be independent and identically distributed random vectors with distribution P, and form the $n \times p$ matrix $\mathbf{X} := (X_1, \ldots, X_n)^\top$. An *estimator* of v_1 is a measurable function from $\mathbb{R}^{n \times p}$ to $\mathbb{R}^p$, and we write $\mathcal{V}_{n,p}$ for the class of all such estimators.

Given unit vectors $u, v \in \mathbb{R}^p$, let $\Theta(u,v) := \cos^{-1}(|u^\top v|)$ denote the acute angle between u and v, and define the loss function

$$L(u,v) := \sin \Theta(u,v) = \{1 - (u^\top v)^2\}^{1/2} = \frac{1}{\sqrt{2}} \|uu^\top - vv^\top\|_2.$$

Note that $L(\cdot, \cdot)$ is invariant to sign changes of either of its arguments. The *directional variance* of P along a unit vector $u \in \mathbb{R}^p$ is defined to be $V(u) := \mathbb{E}\{(u^\top X_1)^2\} = u^\top \Sigma u$. Its empirical counterpart is $\hat{V}(u) := n^{-1} \sum_{i=1}^n (u^\top X_i)^2 = u^\top \hat{\Sigma} u$, where $\hat{\Sigma} := n^{-1} \sum_{i=1}^n X_i X_i^\top$ denotes the sample covariance matrix.

Recall the definition of the k-sparse unit ball $B_0(k)$ from (1). Given $\ell \in \{1, \ldots, p\}$ and $C \in (0, \infty)$, we say P satisfies a *Restricted Covariance Concentration* (RCC) condition with parameters p, n, ℓ and C, and write $P \in \mathrm{RCC}_p(n, \ell, C)$, if

$$(3) \quad \mathbb{P}\left\{ \sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq C \max\left(\sqrt{\frac{\ell \log(p/\delta)}{n}}, \frac{\ell \log(p/\delta)}{n} \right) \right\} \leq \delta$$

for all $\delta > 0$. It is also convenient to define

$$\mathrm{RCC}_p(\ell, C) := \bigcap_{n=1}^\infty \mathrm{RCC}_p(n, \ell, C) \quad \text{and} \quad \mathrm{RCC}_p(C) := \bigcap_{\ell=1}^p \mathrm{RCC}_p(\ell, C).$$

The RCC conditions amount to uniform Bernstein-type concentration properties of the directional variance around its expectation along all sparse directions. This condition turns out to be particularly convenient in the study of convergence rates in Sparse PCA, and moreover, as we show in Proposition 1 below, subgaussian distributions satisfy an RCC condition for all sample sizes n and all sparsity levels ℓ. Recall that a mean-zero distribution Q on $\mathbb{R}^p$ is *subgaussian* with parameter[3] $\sigma^2 \in (0, \infty)$, written

$$Q \in \text{subgaussian}_p(\sigma^2),$$

if whenever $Y \sim Q$, we have $\mathbb{E}(e^{u^\top Y}) \leq e^{\sigma^2 \|u\|^2/2}$ for all $u \in \mathbb{R}^p$.

PROPOSITION 1.   *(i) For every $\sigma > 0$, we have*

$$\text{subgaussian}_p(\sigma^2) \subseteq \text{RCC}_p\left(16\sigma^2\left(1 + \frac{9}{\log p}\right)\right).$$

(ii) In the special case where $P = N_p(0, \Sigma)$, we have $P \in \text{RCC}_p\big(8\lambda_1(P)(1 + \frac{9}{\log p})\big)$.

Our convergence rate results for sparse principal component estimation will be proved over the following classes of distributions. For $\theta > 0$, let

$$\mathcal{P}_p(n, k, \theta) := \Big\{ P \in \text{RCC}_p(n, 2, 1) \cap \text{RCC}_p(n, 2k, 1) :$$
$$\text{(4)} \qquad\qquad\qquad v_1(P) \in B_0(k), \lambda_1(P) - \lambda_2(P) \geq \theta \Big\}.$$

Observe that RCC classes have the scaling property that if the distribution of a random vector Y belongs to $\text{RCC}_p(n, \ell, C)$ and if $r > 0$, then the distribution of rY belongs to $\text{RCC}_p(n, \ell, r^2 C)$. It is therefore convenient to fix $C = 1$ in both RCC classes in (4), so that θ becomes a measure of the signal-to-noise level.

For a symmetric $A \in \mathbb{R}^{p \times p}$, define $\hat{v}_{\max}^k(A) := \text{sargmax}_{u \in B_0(k)} u^\top A u$ to be the k-sparse maximum eigenvector of A, where sargmax denotes the smallest element of the argmax in the lexicographic ordering. (This choice ensures that $\hat{v}_{\max}^k(A)$ is a measurable function of A.) Theorem 2 below gives a finite-sample minimax upper bound for estimating $v_1(P)$ over $\mathcal{P}_p(n, k, \theta)$. For similar bounds over Gaussian or subgaussian classes, see Cai, Ma and Wu (2013) and Vu and Lei (2013), who consider the more general problem of principal subspace estimation. As well as working with a larger class of distributions, our different proof techniques facilitate an explicit constant.

---

[3]Note that some authors say that distributions satisfying this condition are subgaussian with parameter σ, rather than σ^2.

THEOREM 2.   *For $2k \log p \leq n$, the k-sparse empirical maximum eigen-vector, $\hat{v}_{\max}^k(\hat{\Sigma})$, satisfies*

$$\sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\big(\hat{v}_{\max}^k(\hat{\Sigma}), v_1(P)\big) \leq 2\sqrt{2}\bigg(1 + \frac{1}{\log p}\bigg)\sqrt{\frac{k \log p}{n\theta^2}} \leq 7\sqrt{\frac{k \log p}{n\theta^2}}.$$

A matching minimax lower bound of the same order in all parameters k, p, n and θ is given below. The proof techniques are adapted from Vu and Lei (2013).

THEOREM 3.   *Suppose that $7 \leq k \leq p^{1/2}$ and $0 < \theta \leq \frac{1}{16(1+\frac{9}{\log p})}$. Then*

$$\inf_{\hat{v} \in \mathcal{V}_{n,p}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\big(\hat{v}, v_1(P)\big) \geq \min\bigg\{\frac{1}{1660}\sqrt{\frac{k \log p}{n\theta^2}}, \frac{5}{18\sqrt{3}}\bigg\}.$$

We remark that the conditions in the statement of Theorem 3 can be strengthened or weakened, with a corresponding weakening or strengthening of the constants in the bound. For instance, a bound of the same order in k, p, n and θ could be obtained assuming only that $k \leq p^{1-\delta}$ for some $\delta > 0$. The upper bound on θ is also not particularly restrictive. For example, if $P = N_p(0, \sigma^2 I_p + \theta e_1 e_1^\top)$, where e_1 is the first standard basis vector in $\mathbb{R}^p$, then it can be shown that the condition $P \in \mathcal{P}_p(n,k,\theta)$ requires that $\theta \leq 1 - \sigma^2$.

3. Computationally efficient estimation.   As was mentioned in the introduction, the trouble with the estimator $\hat{v}_{\max}^k(\hat{\Sigma})$ of Section 2, as well as the estimator of Cai, Ma and Wu (2013), is that there are no known polynomial time algorithms for their computation. In this section, we therefore study the (polynomial time) semidefinite relaxation estimator $\hat{v}^{\mathrm{SDP}}$ defined by the Algorithm 1 below. This estimator is a variant of one proposed by d'Aspremont et al. (2007), whose support recovery properties were studied for a particular class of Gaussian distributions and a known sparsity level by Amini and Wainwright (2009).

To motivate the main step (Step 2) of Algorithm 1, it is convenient to let $\mathcal{M}$ denote the class of $p \times p$ non-negative definite real, symmetric matrices, and let $\mathcal{M}_1 := \{M \in \mathcal{M} : \mathrm{tr}(M) = 1\}$. Let $\mathcal{M}_{1,1}(k^2) := \{M \in \mathcal{M}_1 : \mathrm{rank}(M) = 1, \|M\|_0 = k^2\}$ and observe that

$$\max_{u \in B_0(k)} u^\top \hat{\Sigma} u = \max_{u \in B_0(k)} \mathrm{tr}(\hat{\Sigma} u u^\top) = \max_{M \in \mathcal{M}_{1,1}(k^2)} \mathrm{tr}(\hat{\Sigma} M).$$

In the final expression, the rank and sparsity constraints are non-convex. We therefore adopt the standard semidefinite relaxation approach of dropping

the rank constraint and replacing the sparsity (ℓ_0) constraint with an ℓ_1 penalty to obtain the convex optimisation problem

$$(5) \qquad \max_{M \in \mathcal{M}_1} \big\{ \mathrm{tr}(\hat{\Sigma}M) - \lambda\|M\|_1 \big\}$$

Algorithm 1: Pseudo-code for computing the semidefinite relaxation estimator $\hat{v}^{\mathrm{SDP}}$

Input: $\mathbf{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times p}$, $\lambda > 0, \epsilon > 0$

begin

   Step 1: Set $\hat{\Sigma} \leftarrow n^{-1}\mathbf{X}^\top\mathbf{X}$.

   Step 2: For $f(M) := \mathrm{tr}(\hat{\Sigma}M) - \lambda\|M\|_1$, let $\hat{M}^\epsilon$ be an ϵ-maximiser of f in $\mathcal{M}_1$. In other words, $\hat{M}^\epsilon$ satisfies $f(\hat{M}^\epsilon) \geq \max_{M \in \mathcal{M}_1} f(M) - \epsilon$.

   Step 3: Let $\hat{v}^{\mathrm{SDP}} := \hat{v}^{\mathrm{SDP}}_{\lambda,\epsilon} \in \mathrm{argmax}_{u:\|u\|_2=1} u^\top \hat{M}^\epsilon u$.

end

Output: $\hat{v}^{\mathrm{SDP}}$

We now discuss the complexity of computing $\hat{v}^{\mathrm{SDP}}$ in detail. One possible way of implementing Step 2 is to use a generic interior-point method. However, as shown in Nesterov (2005), Nemirovski (2004) and Bach, Ahipaşaoğlu and d'Aspremont (2010), certain first-order algorithms (i.e. methods requiring $O(1/\epsilon)$ steps to find a feasible point achieving an ϵ-approximation of the optimal objective function value) can significantly outperform such generic interior-point solvers. The key idea in both Nesterov (2005) and Nemirovski (2004) is that the optimisation problem in Step 2 can be rewritten in a saddlepoint formulation:

$$\max_{M \in \mathcal{M}_1} \mathrm{tr}(\hat{\Sigma}M) - \lambda\|M\|_1 = \max_{M \in \mathcal{M}_1} \min_{U \in \mathcal{U}} \mathrm{tr}\big((\hat{\Sigma} + U)M\big),$$

where $\mathcal{U} := \{U \in \mathbb{R}^{p \times p} : U^\top = U, \|U\|_\infty \leq \lambda\}$. The fact that $\mathrm{tr}\big((\hat{\Sigma} + U)M\big)$ is linear in both M and U makes the problem amenable to proximal methods. In Algorithm 2 below, we state a possible implementation of Step 2 of Algorithm 1, derived from the 'basic implementation' in Nemirovski (2004). In the algorithm, the $\|\cdot\|_2$-norm projection $\Pi_{\mathcal{U}}(A)$ of a symmetric matrix $A = (A_{ij}) \in \mathbb{R}^{p \times p}$ onto $\mathcal{U}$ is given by

$$\big(\Pi_{\mathcal{U}}(A)\big)_{ij} := \mathrm{sign}(A_{ij}) \min(|A_{ij}|, \lambda).$$

For the projection $\Pi_{\mathcal{M}_1}(A)$, first decompose $A =: PDP^\top$ for some orthogonal P and diagonal $D = \mathrm{diag}(d)$, where $d = (d_1, \ldots, d_p)^\top \in \mathbb{R}^p$. Now let $\Pi_{\mathcal{W}}(d)$ be the projection image of d on the unit $(p-1)$-simplex

$\mathcal{W} := \{(w_1, \ldots, w_p) : w_j \geq 0, \sum_{j=1}^p w_j = 1\}$. Finally, transform back to obtain $\Pi_{\mathcal{M}_1}(A) := P \operatorname{diag}(\Pi_{\mathcal{W}}(d)) P^\top$. The fact that Algorithm 2 outputs an ϵ-maximiser of the optimisation problem in Step 2 of Algorithm 1 follows from Nemirovski (2004, Theorem 3.2), which implies in our particular case that after N iterations,

$$\max_{M \in \mathcal{M}_1} \min_{U \in \mathcal{U}} \operatorname{tr}\big((\hat{\Sigma} + U)M\big) - \min_{U \in \mathcal{U}} \operatorname{tr}\big((\hat{\Sigma} + U)\hat{M}^\epsilon\big) \leq \frac{\lambda^2 p^2 + 1}{\sqrt{2}N}.$$

Algorithm 2: A possible implementation of Step 2 of Algorithm 1

Input: $\hat{\Sigma} \in \mathcal{M}$, $\lambda > 0$, $\epsilon > 0$.

begin

    Set $M_0 \leftarrow I_p/p$, $U_0 \leftarrow 0 \in \mathbb{R}^{p \times p}$ and $N \leftarrow \left\lceil \frac{\lambda^2 p^2 + 1}{\sqrt{2}\epsilon} \right\rceil$.

    for $t \leftarrow 1$ **to** N **do**

        $U'_t \leftarrow \Pi_{\mathcal{U}}\big(U_{t-1} - \frac{1}{\sqrt{2}}M_{t-1}\big)$, $M'_t \leftarrow \Pi_{\mathcal{M}_1}\big(M_{t-1} + \frac{1}{\sqrt{2}}\hat{\Sigma} + \frac{1}{\sqrt{2}}U_{t-1}\big)$.

        $U_t \leftarrow \Pi_{\mathcal{U}}\big(U_{t-1} - \frac{1}{\sqrt{2}}M'_t\big)$, $M_t \leftarrow \Pi_{\mathcal{M}_1}\big(M_{t-1} + \frac{1}{\sqrt{2}}\hat{\Sigma} + \frac{1}{\sqrt{2}}U'_t\big)$.

    end

    Set $\hat{M}^\epsilon \leftarrow \frac{1}{N}\sum_{t=1}^N M'_t$.

end

Output: $\hat{M}^\epsilon$

In Algorithm 1, Step 1 takes $O(np^2)$ floating point operations; Step 3 takes $O(p^3)$ operations in the worst case, though other methods such as the Lanczos method (Lanczos, 1950; Golub and Van Loan, 1996) require only $O(p^2)$ operations under certain conditions. Our particular implementation (Algorithm 2) for Step 2 requires $O(\frac{\lambda^2 p^2 + 1}{\epsilon})$ iterations in the worst case, though this number may often be considerably reduced by terminating the **for** loop if the primal-dual gap

$$\lambda_1(\hat{U}_t + \hat{\Sigma}) - \{\operatorname{tr}(\hat{M}_t \hat{\Sigma}) - \lambda \|\hat{M}_t\|_1\}$$

falls below ϵ, where $\hat{U}_t := t^{-1}\sum_{s=1}^t U'_s$ and $\hat{M}_t := t^{-1}\sum_{s=1}^t M'_s$. The most costly step within the **for** loop is the eigendecomposition used to compute the projection $\Pi_{\mathcal{M}_1}$, which takes $O(p^3)$ operations. Taking $\lambda := 4\sqrt{\frac{\log p}{n}}$ and $\epsilon := \frac{\log p}{4n}$ as in Theorem 5 below, we find an overall complexity for the algorithm of $O\big(\max(p^5, \frac{np^3}{\log p})\big)$ operations in the worst case.

We now turn to the theoretical properties of the estimator $\hat{v}^{\mathrm{SDP}}$ computed using Algorithm 1. Lemma 4 below is stated in a general, deterministic fashion, but will be used in Theorem 5 below to bound the loss incurred

by the estimator on the event that the sample and population covariance matrices are close in ℓ_∞-norm. See also Vu et al. (2013, Theorem 3.1) for a closely related result in the context of a projection matrix estimation problem. Recall that $\mathcal{M}$ denotes the class of $p \times p$ non-negative definite real, symmetric matrices.

LEMMA 4. *Let $\Sigma \in \mathcal{M}$ be such that $\theta := \lambda_1(\Sigma) - \lambda_2(\Sigma) > 0$. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\hat{\Sigma} := n^{-1}\mathbf{X}^\top \mathbf{X}$. For arbitrary $\lambda > 0$ and $\epsilon > 0$, if $\|\hat{\Sigma} - \Sigma\|_\infty \leq \lambda$, then the semidefinite relaxation estimator $\hat{v}^{\mathrm{SDP}}$ in Algorithm 1 with inputs $\mathbf{X}, \lambda, \epsilon$ satisfies*

$$L\big(\hat{v}^{\mathrm{SDP}}, v_1(\Sigma)\big) \leq \frac{4\sqrt{2}\lambda k}{\theta} + 2\sqrt{\frac{\epsilon}{\theta}}.$$

Theorem 5 below describes the statistical properties of the estimator $\hat{v}^{\mathrm{SDP}}$ over $\mathcal{P}_p(n, k, \theta)$ classes. It reveals in particular that we incur a loss of statistical efficiency of a factor of $\sqrt{k}$ compared with the minimax upper bound in Theorem 2 in Section 2 above. As well as applying Lemma 4 on the event $\{\|\hat{\Sigma} - \Sigma\|_\infty \leq \lambda\}$, the proof relies on Lemma 5 in the online supplementary material Wang, Berthet and Samworth (2015), which relates the event $\{\|\hat{\Sigma} - \Sigma\|_\infty > \lambda\}$ to the $\mathrm{RCC}_p(n, 2, 1)$ condition. Indeed, this explains why we incorporated this condition into the definition of the $\mathcal{P}_p(n, k, \theta)$ classes.

THEOREM 5. *For an arbitrary $P \in \mathcal{P}_p(n, k, \theta)$ and $X_1, \ldots, X_n \overset{iid}{\sim} P$, we write $\hat{v}^{\mathrm{SDP}}(\mathbf{X})$ for the output of Algorithm 1 with input $\mathbf{X} := (X_1, \ldots, X_n)^\top$, $\lambda := 4\sqrt{\frac{\log p}{n}}$ and $\epsilon := \frac{\log p}{4n}$. If $4 \log p \leq n \leq k^2 p^2 \theta^{-2} \log p$ and $\theta \in (0, k]$, then*

$$(6) \qquad \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\big(\hat{v}^{\mathrm{SDP}}(\mathbf{X}), v_1(P)\big) \leq \min\left\{(16\sqrt{2} + 2)\sqrt{\frac{k^2 \log p}{n\theta^2}}, 1\right\}.$$

We remark that $\hat{v}^{\mathrm{SDP}}$ has the attractive property of being fully adaptive in the sense that it can be computed without knowledge of the sparsity level k. On the other hand, $\hat{v}^{\mathrm{SDP}}$ is not necessarily k-sparse. If a specific sparsity level is desired in a particular application, Algorithm 1 can be modified to obtain a (non-adaptive) k-sparse estimator having similar estimation risk. Specifically, we can find

$$\hat{v}_0^{\mathrm{SDP}} \in \underset{u \in B_0(k)}{\operatorname{argmin}} L(\hat{v}^{\mathrm{SDP}}, u).$$

Since $L(\hat{v}^{\mathrm{SDP}}, u)^2 = 1 - \big(u^\top \hat{v}^{\mathrm{SDP}}\big)^2$, we can compute $\hat{v}_0^{\mathrm{SDP}}$ by setting all but the top k coordinates of $\hat{v}^{\mathrm{SDP}}$ in absolute value to zero and renormalising

the vector. In particular, $\hat{v}_0^{\mathrm{SDP}}$ is computable in polynomial time. We deduce that under the same conditions as in Theorem 5, for any $P \in \mathcal{P}_p(n, k, \theta)$,

$$
\begin{aligned}
&\mathbb{E} L\big(\hat{v}_0^{\mathrm{SDP}}, v_1\big) \\
&\leq \mathbb{E}\big[\big\{L\big(\hat{v}_0^{\mathrm{SDP}}, \hat{v}^{\mathrm{SDP}}\big) + L\big(\hat{v}^{\mathrm{SDP}}, v_1\big)\big\} \mathbb{1}_{\{\|\hat{\Sigma} - \Sigma\|_\infty \leq \lambda\}}\big] + \mathbb{P}\big(\|\hat{\Sigma} - \Sigma\|_\infty > \lambda\big) \\
&\leq 2\mathbb{E}\big\{L\big(\hat{v}_0^{\mathrm{SDP}}, v_1\big) \mathbb{1}_{\{\|\hat{\Sigma} - \Sigma\|_\infty \leq \lambda\}}\big\} + \mathbb{P}\big(\|\hat{\Sigma} - \Sigma\|_\infty > \lambda\big) \\
&\leq (32\sqrt{2} + 3)\sqrt{\frac{k^2 \log p}{n\theta^2}},
\end{aligned}
$$

where the final inequality follows from the proof of Theorem 5.

4. Computational lower bounds in sparse principal component estimation. Theorems 5 and 2 reveal a gap between the provable performance of our semidefinite relaxation estimator $\hat{v}^{\mathrm{SDP}}$ and the minimax optimal rate. It is natural to ask whether there exists a computationally efficient algorithm that achieves the statistically optimal rate of convergence. In fact, as we will see in Theorem 6 below, the effective sample size region over which $\hat{v}^{\mathrm{SDP}}$ is consistent is essentially tight among the class of all *randomised polynomial time algorithms*[4]. Indeed, any randomised polynomial time algorithm with a faster rate of convergence could otherwise be adapted to solve instances of the Planted Clique problem that are believed to be hard; see Section 4.1 below for formal definitions and discussion. In this sense, the extra factor of $\sqrt{k}$ is an intrinsic price in statistical efficiency that we have to pay for computational efficiency, and the estimator $\hat{v}^{\mathrm{SDP}}$ studied in Section 3 has essentially the best possible rate of convergence among computable estimators.

4.1. *The Planted Clique problem.* A *graph* $G := (V(G), E(G))$ is an ordered pair in which $V(G)$ is a countable set, and $E(G)$ is a subset of $\big\{\{x, y\} : x, y \in V(G), x \neq y\big\}$. For $x, y \in V(G)$, we say x and y are *adjacent*, and write $x \sim y$, if $\{x, y\} \in E(G)$. A *clique* C is a subset of $V(G)$ such that $\{x, y\} \in E(G)$ for all distinct $x, y \in C$. The problem of finding a clique of maximum size in a given graph G is known to be NP-*complete* (Karp, 1972). It is therefore natural to consider randomly generated input graphs with a clique 'planted' in, where the signal is much less confounded by the noise. Such problems were first suggested by Jerrum (1992) and Kučera (1995) as a potentially easier variant of the classical Clique problem.

---

[4]In this section, terms from computational complexity theory defined Section 2 of the online supplementary material in Wang, Berthet and Samworth (2015) are written in italics at their first occurrence.

Let $\mathbb{G}_m$ denote the collection of all graphs with m vertices. Define $\mathcal{G}_m$ to be the distribution on $\mathbb{G}_m$ associated with the standard Erdős–Rényi random graph. In other words, under $\mathcal{G}_m$, each pair of vertices is adjacent independently with probability $1/2$. For any $\kappa \in \{1, \ldots, m\}$, let $\mathcal{G}_{m,\kappa}$ be a distribution on $\mathbb{G}_m$ constructed by first picking κ distinct vertices uniformly at random and connecting all edges (the 'planted clique'), then joining each remaining pair of distinct vertices by an edge independently with probability $1/2$. The Planted Clique problem has input graphs randomly sampled from the distribution $\mathcal{G}_{m,\kappa}$. Due to the random nature of the problem, the goal of the Planted Clique problem is to find (possibly randomised) algorithms that can locate a maximum clique K_m with high probability.

It is well known that, for a standard Erdős–Rényi graph, $\frac{|K_m|}{2 \log_2 m} \overset{\text{a.s.}}{\to} 1$ (e.g. Grimmett and McDiarmid, 1975). In fact, if $\kappa = \kappa_m$ is such that $\liminf_{m \to \infty} \frac{\kappa}{2 \log_2 m} > 1$, it can be shown that the planted clique is asymptotically almost surely also the unique maximum clique in the input graph. As observed in Kučera (1995), there exists $C > 0$ such that, if $\kappa > C\sqrt{m \log m}$, then asymptotically almost surely, vertices in the planted clique have larger degrees than all other vertices, in which case they can be located in $O(m^2)$ operations. Alon, Krivelevich and Sudakov (1998) improved the above result by exhibiting a spectral method that, given any $c > 0$, identifies planted cliques of size $\kappa \geq c\sqrt{m}$ asymptotically almost surely.

Although several other polynomial time algorithms have subsequently been discovered for the $\kappa \geq c\sqrt{m}$ case (e.g. Feige and Krauthgamer, 2000; Feige and Ron, 2010; Ames and Vavasis, 2011), there is no known randomised polynomial time algorithm that can detect below this threshold. Jerrum (1992) hinted at the hardness of this problem by showing that a specific Markov chain approach fails to work when $\kappa = O(m^{1/2-\delta})$ for some $\delta > 0$. Feige and Krauthgamer (2003) showed that Lovàcz–Schrijiver semidefinite programming relaxation methods also fail in this regime. Feldman et al. (2013) recently presented further evidence of the hardness of this problem by showing that a broad class of algorithms, which they refer to as 'statistical algorithms', cannot solve the Planted Clique problem with $\kappa = O(m^{1/2-\delta})$ in randomised polynomial time, for any $\delta > 0$. It is now widely accepted in theoretical computer science that the Planted Clique problem is hard, in the sense that the following assumption holds with $\tau = 0$:

(A1)(τ) For any sequence $\kappa = \kappa_m$ such that $\kappa \leq m^\beta$ for some $0 < \beta < 1/2 - \tau$, there is no randomised polynomial time algorithm that can correctly identify the planted clique with probability tending to 1 as $m \to \infty$.

We state the assumption in terms of a general parameter $\tau \in [0, 1/2)$, because it will turn out below that even if only $(A1)(\tau)$ holds for some $\tau \in (0, 1/6)$, there are still regimes of (n, p, k, θ) in which no randomised polynomial time algorithm can attain the minimax optimal rate.

Researchers have used the hardness of the planted clique problem as an assumption to prove various impossibility results in other problems. Examples include cryptographic applications (Juels and Peinado, 2000; Applebaum, Barak and Wigderson, 2010), testing k-wise independence (Alon et al., 2007) and approximating Nash equilibria (Hazan and Krauthgamer, 2011). Recent works by Berthet and Rigollet (2013a,b) and Ma and Wu (2015) used a stronger hypothesis on the hardness of detecting the presence of a planted clique to establish computational lower bounds in sparse principal component detection and sparse submatrix detection problems respectively. Our Assumption $(A1)(0)$ assumes only the computational intractability of identifying the entire planted clique, so in particular, is implied by Hypothesis A_{PC} of Berthet and Rigollet (2013b) and Hypothesis 1 of Ma and Wu (2015).

4.2. *Computational lower bounds.* In this section, we use a reduction argument to show that, under Assumption $(A1)(\tau)$, it is impossible to achieve the statistically optimal rate of sparse principal component estimation using randomised polynomial time algorithms. For $\rho \in \mathbb{N}$, and for $x \in \mathbb{R}$, we let $[x]_\rho$ denote x in its binary representation, rounded to ρ significant figures. Let $[\mathbb{R}]_\rho := \{[x]_\rho : x \in \mathbb{R}\}$. We say $(\hat{v}^{(n)})$ is a *sequence of randomised polynomial time estimators* of $v_1 \in \mathbb{R}^{p_n}$ if $\hat{v}^{(n)}$ is a measurable function from $\mathbb{R}^{n \times p_n}$ to $\mathbb{R}^{p_n}$ and if, for every $\rho \in \mathbb{N}$, there exists a randomised polynomial time algorithm M_{pr} such that for any $\mathbf{x} \in ([\mathbb{R}]_\rho)^{n \times p_n}$ we have $[\hat{v}^{(n)}(\mathbf{x})]_\rho = [M_{\mathrm{pr}}(\mathbf{x})]_\rho$. The sequence of semidefinite programming estimators $(\hat{v}^{\mathrm{SDP}})$ defined in Section 3 is an example of a sequence of randomised polynomial time estimators of $v_1(P)$.

THEOREM 6. *Fix $\tau \in [0, 1/6)$, assume $(A1)(\tau)$, and let $\alpha \in (0, \frac{1-6\tau}{1-2\tau})$. For any $n \in \mathbb{N}$, let $(p, k, \theta) = (p_n, k_n, \theta_n)$ be parameters indexed by n such that $k = O(p^{1/2-\tau-\delta})$ for some $\delta \in (0, 1/2 - \tau)$, $n = o(p \log p)$ and $\theta \leq k^2/(1000p)$. Suppose further that*

$$\frac{k^{1+\alpha} \log p}{n\theta^2} \to 0$$

as $n \to \infty$. Let $\mathbf{X}$ be an $n \times p$ matrix with independent rows, each having distribution P. Then every sequence $(\hat{v}^{(n)})$ of randomised polynomial time

estimators of $v_1(P)$ satisfies

$$\sqrt{\frac{n\theta^2}{k^{1+\alpha}\log p}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\big(\hat{v}^{(n)}(\mathbf{X}), v_1(P)\big) \to \infty$$

as $n \to \infty$.

We note that the choices of parameters in the theorem imply that

$$(7) \qquad \liminf_{n \to \infty} \frac{k^2 \log p}{n\theta^2} \geq \liminf_{n \to \infty} \frac{p}{k^2} = \infty.$$

As remarked in Section 4.1 above, the main interest in this theorem comes from the case $\tau = 0$. Here, our result reveals not only that no randomised polynomial time algorithm can attain the minimax optimal rate, but also that in the effective sample size regime described by (7), and provided the other side conditions of Theorem 6 hold, there is in general no consistent sequence of randomised polynomial time estimators. This is in contrast to Theorem 2, where we saw that consistent estimation with a computationally inefficient procedure is possible in the asymptotic regime (7). A further consequence of Theorem 6 is that, since any sequence $(p, k, \theta) = (p_n, k_n, \theta_n)$ satisfying the conditions of Theorem 6 also satisfies the conditions of Theorem 5 for large n, the conclusion of Theorem 5 cannot be improved in terms of the exponent of k (at least, not uniformly over the parameter range given there). As mentioned in the introduction, for a sufficiently large effective sample size, where even randomised polynomial time estimators can be consistent, the statistical and computational trade-off revealed by Theorems 2 and 6 may disappear. See Section 4.4 below for further details, and Gao, Ma and Zhou (2014) for recent extensions of these results to different classes of distributions.

Even though Assumption (A1)(0) is widely believed, we also present results under the weaker family of conditions (A1)(τ) for $\tau \in (0, 1/6)$ to show that a statistical and computational trade-off still remains for certain parameter regimes even in these settings. The reason for assuming $\tau < 1/6$ is to guarantee that there is a regime of parameters (n, p, k, θ) satisfying the conditions of the theorem. Indeed, if $\tau \in [0, 1/6)$ and $\alpha \in (0, \frac{1-6\tau}{1-2\tau})$, we can set $p = n$, $k = n^{1/2-\tau-\delta}$ for some $\delta \in \big(0, \frac{1}{2} - \tau - \frac{1}{3-\alpha}\big)$, $\theta = k^2/(1000n)$, and in that case,

$$\frac{k^{1+\alpha}\log p}{n\theta^2} = \frac{10^6 n \log n}{k^{3-\alpha}} \to 0,$$

as required.

4.3. *Sketch of the proof of Theorem 6.* The proof of Theorem 6 relies on a randomised polynomial time reduction from the Planted Clique problem to the sparse principal component estimation problem. The reduction is adapted from the 'bottom-left transformation' of Berthet and Rigollet (2013b), and requires a rather different and delicate analysis.

In greater detail, suppose for a contradiction that we were given a randomised polynomial time algorithm $\hat{v}$ for the Sparse PCA problem with a rate $\sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}, v_1) \leq \sqrt{\frac{k^{1+\alpha} \log p}{n\theta^2}}$ for some $\alpha < 1$. Set $m \approx p \log p$ and $\kappa \approx k \log p$, so we are in the regime where $(\mathrm{A1})(\tau)$ holds. Given any graph $G \sim \mathcal{G}_{m,\kappa}$ with planted clique $K \subseteq V(G)$, we draw $n + p$ vertices $u_1, \ldots, u_n, w_1, \ldots, w_p$ uniformly at random without replacement from $V(G)$. On average there are about $\kappa / \log \kappa$ clique vertices in $\{w_1, \ldots, w_p\}$, and our initial aim is to identify a large fraction of these vertices. To do this, we form an $n \times p$ matrix $\mathbf{A} := (\mathbb{1}_{u_i \sim w_j})_{i,j}$, which is an off-diagonal block of the adjacency matrix of G. We then replace each 0 in $\mathbf{A}$ with -1 and flip the signs of each row independently with probability $1/2$ to obtain a new matrix $\mathbf{X}$. Each component of the ith row of $\mathbf{X}$ has a marginal Rademacher distribution, but if u_i is a clique vertex, then the components $\{j : w_j \in K\}$ are perfectly correlated. Writing $\gamma' := (\mathbb{1}_{\{w_j \in K\}})_{j=1,\ldots,p}$, the leading eigenvector of $\mathbb{E}\{\mathbf{X}^\top \mathbf{X}/n | \gamma'\}$ is proportional to γ', which suggests that a spectral method might be able to find $\{w_1, \ldots, w_p\} \cap K$ with high probability. Unfortunately, the joint distribution of the rows of $\mathbf{X}$ is difficult to deal with directly, but since n and p are small relative to m, we can approximate γ' by a random vector γ having independent $\mathrm{Bern}(\kappa/m)$ components. We can then approximate $\mathbf{X}$ by a matrix $\mathbf{Y}$, whose rows are independent conditional on γ and have the same marginal distribution conditional on $\gamma = g$ as the rows of $\mathbf{X}$ conditional on $\gamma' = g$.

It turns out that the distribution of an appropriately scaled version of an arbitrary row of $\mathbf{Y}$, conditional on $\gamma = g$, belongs to $\mathcal{P}_p(n, k, \theta)$ for g belonging to a set of high probability. We could therefore apply our hypothetical randomised polynomial time Sparse PCA algorithm to the scaled version of the matrix $\mathbf{Y}$ to find a good estimate of γ, and since γ is close to γ', this accomplishes our initial goal. With high probability, the remaining vertices in the planted clique are those having high connectivity to the identified clique vertices in $\{w_1, \ldots, w_p\}$, which contradicts the hypothesis $(\mathrm{A1})(\tau)$.

4.4. *Computationally efficient optimal estimation on subparameter spaces in the high effective sample size regime.* Theorems 2, 3, 5 and 6 enable us to summarise, in Table 1 below, our knowledge of the best possible rate of estimation in different asymptotic regimes, both for arbitrary statistical

TABLE 1
Rate of convergence of best estimator in different asymptotic regimes

	$n \ll \dfrac{k \log p}{\theta^2}$	$\dfrac{k \log p}{\theta^2} \ll n \ll \dfrac{k^2 \log p}{\theta^2}$	$n \gg \dfrac{k^2 \log p}{\theta^2}$
All estimators	$\asymp 1$	$\asymp \sqrt{\dfrac{k \log p}{n \theta^2}}$	$\asymp \sqrt{\dfrac{k \log p}{n \theta^2}}$
Polynomial time estimators	$\asymp 1$	$\asymp 1$	$\lesssim \sqrt{\dfrac{k^2 \log p}{n \theta^2}}$

procedures and for those that are computable in randomised polynomial time. (For ease of exposition, we omit here the additional, relatively mild, side constraints required for the above theorems to hold.) The fact that Theorem 6 is primarily concerned with the setting in which $\frac{k^2 \log p}{n \theta^2} \to \infty$ raises the question of whether computationally efficient procedures could attain a faster rate of convergence in the high effective sample size regime where $n \gg \frac{k^2 \log p}{\theta^2}$.

The purpose of this section is to extend the ideas of Amini and Wainwright (2009) to show that, indeed, a variant of the estimator $\hat{v}^{\mathrm{SDP}}$ introduced in Section 3 attains the minimax optimal rate of convergence in this asymptotic regime, at least over a subclass of the distributions in $\mathcal{P}_p(n, k, \theta)$. Ma (2013) and Yuan and Zhang (2013) show similar results for an iterative thresholding algorithm for other subclasses of $\mathcal{P}_p(n, k, \theta)$ under an extra upper bound condition on $\lambda_2(P)/\lambda_1(P)$; see also Wang, Lu and Liu (2014) and Deshpande and Montanari (2014).

Let $\mathcal{T}$ denote the set of non-negative definite matrices $\Sigma \in \mathbb{R}^{p \times p}$ of the form

$$\Sigma = \theta v_1 v_1^\top + \begin{pmatrix} I_k & 0 \\ 0 & \Gamma_{p-k} \end{pmatrix},$$

where $v_1 \in \mathbb{R}^p$ is a unit vector such that $S := \mathrm{supp}(v_1)$ has cardinality k and where $\Gamma_{p-k} \in \mathbb{R}^{(p-k) \times (p-k)}$ is non-negative definite and satisfies $\lambda_1(\Gamma_{p-k}) \leq 1$. (Here, and in the proof of Theorem 7 below, the block matrix notation refers to the (S, S), (S, S^c), (S^c, S) and (S^c, S^c) blocks.) We now define a subclass of distributions

$$\tilde{\mathcal{P}}_p(n, k, \theta) := \left\{ P \in \mathcal{P}_p(n, k, \theta) : \Sigma(P) \in \mathcal{T}, \min_{j \in S} |v_{1,j}| \geq 16 \sqrt{\frac{k \log p}{n \theta^2}} \right\}.$$

We remark that $\tilde{\mathcal{P}}_p(n, k, \theta)$ is non-empty only if $\sqrt{\frac{k^2 \log p}{n \theta^2}} \leq \frac{1}{16}$, since

$$1 = \|v_{1,S}\|_2 \geq k^{1/2} \min_{j \in S} |v_{1,j}| \geq 16 \sqrt{\frac{k^2 \log p}{n \theta^2}}.$$

This is one reason that the theorem below only holds in the high effective sample size regime. Our variant of $\hat{v}^{\mathrm{SDP}}$ is described in Algorithm 3 below. We remark that $\hat{v}^{\mathrm{MSDP}}$, like $\hat{v}^{\mathrm{SDP}}$, is computable in polynomial time.

Algorithm 3: Pseudo-code for computing the modified semidefinite relaxation estimator $\hat{v}^{\mathrm{MSDP}}$

Input: $\mathbf{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times p}$, $\lambda > 0$, $\epsilon > 0$, $\tau > 0$.
begin

    Step 1: Set $\hat{\Sigma} \leftarrow n^{-1} \mathbf{X}^\top \mathbf{X}$.
    Step 2: For $f(M) := \mathrm{tr}(\hat{\Sigma} M) - \lambda \|M\|_1$, let $\hat{M}^\epsilon$ be an ϵ-maximiser of f in $\mathcal{M}_1$.
    Step 3: Let $\hat{S} \leftarrow \{j \in \{1, \ldots, p\} : \hat{M}_{jj}^\epsilon \geq \tau\}$ and $\hat{v}^{\mathrm{MSDP}} \in \mathbb{R}^p$ by $\hat{v}_{\hat{S}^c}^{\mathrm{MSDP}} \leftarrow 0$
    and $\hat{v}_{\hat{S}}^{\mathrm{MSDP}} \in \mathrm{argmax}_{u \in \mathbb{R}^{|\hat{S}|}} u^\top \hat{\Sigma}_{\hat{S}\hat{S}} u$.

end
Output: $\hat{v}^{\mathrm{MSDP}}$

THEOREM 7. *Assume that $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} P$ for some $P \in \tilde{\mathcal{P}}_p(n, k, \theta)$.*

(a) *Let $\lambda := 4\sqrt{\frac{\log p}{n}}$. The function f in Step 2 of Algorithm 3 has a maximiser $\hat{M} \in \mathcal{M}_{1,1}(k^2)$ satisfying $\mathrm{sgn}(\hat{M}) = \mathrm{sgn}(v_1 v_1^\top)$.*

(b) *Assume that $\log p \leq n$, $\theta^2 \leq Bk^{1/2}$ for some $B \geq 1$ and $p \geq \theta(n/k)^{1/2}$. We write $\hat{v}^{\mathrm{MSDP}}$ for the output of Algorithm 3 with input parameters $\mathbf{X} := (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times p}$, $\lambda := 4\sqrt{\frac{\log p}{n}}$, $\epsilon := (\frac{\log p}{Bn})^{5/2}$ and $\tau := (\frac{\log p}{Bn})^2$. Then*

$$\sup_{P \in \tilde{\mathcal{P}}_p(n,k,\theta)} \mathbb{E}_p\big\{L(\hat{v}^{\mathrm{MSDP}}, v_1)\big\} \leq 6\sqrt{\frac{k \log p}{n\theta^2}}.$$

Theorem 7 generalises Theorem 2 of Amini and Wainwright (2009) in two ways: first, we relax a gaussianity assumption to an RCC condition; second, the leading eigenvector of the population covariance matrix is not required to have non-zero entries equal to $\pm k^{-1/2}$.

5. Numerical experiments. In this section we present the results of numerical experiments to illustrate the results of Theorems 5, 6 and 7. We generate $v_1 \in \mathbb{R}^p$ by setting $v_{1,j} := k^{-1/2}$ for $j = 1, \ldots, k$, and $v_{1,j} := 0$ for $j = k + 1, \ldots, p$. We then draw $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} N_p(0, \Sigma)$, where $\Sigma := I_p + \theta v_1 v_1^\top$ and $\theta = 1$. We apply Algorithm 1 to the data matrix $\mathbf{X} := (X_1, \ldots, X_n)^\top$ and report the average loss of the estimator $\hat{v}^{\mathrm{SDP}}$ over $N_{\mathrm{rep}} := 100$ repetitions. For $p \in \{50, 100, 150, 200\}$ and $k = \lfloor p^{1/2} \rfloor$, we repeat the
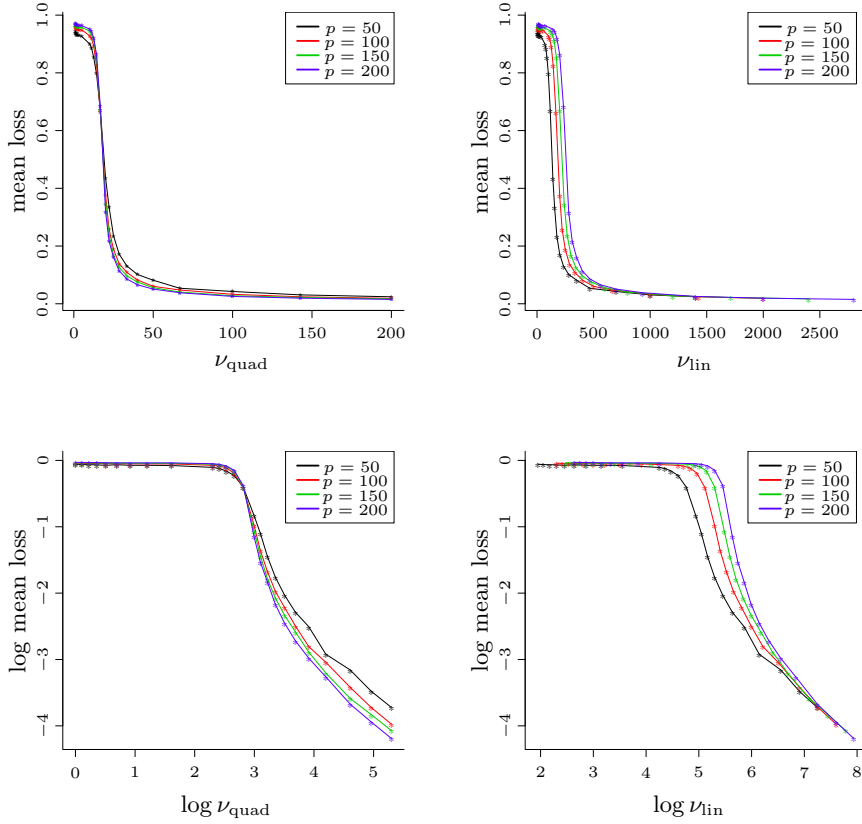
FIG 1. *Average loss of the estimator $\hat{v}^{\mathrm{SDP}}$ over $N_{\mathrm{rep}} = 100$ repetitions against effective sample sizes ν_{quad} (top left) and ν_{lin} (top right). The tail behaviour under both scalings is examined under logarithmic scales in the bottom left and bottom right panels.*

experiment for several choices of n to explore the three parameter regimes described in Table 1. Since the boundaries of these regimes are $n \asymp \frac{k \log p}{\theta^2}$ and $n \asymp \frac{k^2 \log p}{\theta^2}$, we plot the average loss of the experiments against effective samples sizes

$$\nu_{\mathrm{lin}} := \frac{n\theta^2}{k \log p}, \quad \text{and} \quad \nu_{\mathrm{quad}} := \frac{n\theta^2}{k^2 \log p}.$$

The results are shown in Figure 1. The top left panel of Figure 1 shows a sharp phase transition for the average loss, as predicted by Theorems 5 and 6. The right panels of Figure 1 suggest that in the high effective sample size regime, $\hat{v}^{\mathrm{SDP}}$ converges at rate $\sqrt{\frac{k \log p}{n\theta^2}}$ in this setting. This is the same rate as was proved for the modified semidefinite relaxation estimator $\hat{v}^{\mathrm{MSDP}}$ in Theorem 7.

It is worth noting that it is relatively time-consuming to carry out the simulations for the settings in the right-hand tails of the plots in Figure 1. These extreme settings were chosen, however, to illustrate that the linear scaling is the correct one in this tail. For example, when $\nu_{\mathrm{quad}} = 200$ and $p = 200$, we require $n = 207694$, and the pre-processing of the data matrix to obtain the sample covariance matrix is the time-limiting step. In general, in our experience, the semi-definite programming algorithm is certainly not as fast as simpler methods such as diagonal thresholding, but is not prohibitively slow.

APPENDIX A: PROOFS FROM SECTION 2

PROOF OF PROPOSITION 1. (i) Let $P \in \mathrm{subgaussian}_p(\sigma^2)$, and assume that $X_1, \ldots, X_n \overset{iid}{\sim} P$. Then, for any $u \in B_0(\ell)$ and $t \geq 0$, we have

$$\mathbb{P}(u^\top X_1 \geq t) \leq e^{-t^2/\sigma^2} \mathbb{E}(e^{tu^\top X_1/\sigma^2}) \leq e^{-t^2/(2\sigma^2)}.$$

Similarly, $\mathbb{P}(-u^\top X_1 \geq t) \leq e^{-t^2/(2\sigma^2)}$. Write $\mu_u := \mathbb{E}\{(u^\top X_1)^2\}$; since

$$1 + \frac{1}{2}\mu_u t^2 + o(t^2) = \mathbb{E}(e^{tu^\top X_1}) \leq e^{t^2\sigma^2/2} = 1 + \frac{1}{2}\sigma^2 t^2 + o(t^2),$$

as $t \to 0$, we deduce that $\mu_u \leq \sigma^2$. Now, for any integer $m \geq 2$,

$$\mathbb{E}\big(|(u^\top X_1)^2 - \mu_u|^m\big) \leq \int_0^\infty \mathbb{P}\Big\{(u^\top X_1)^2 - \mu_u \geq t^{1/m}\Big\} \, dt + \mu_u^m$$

$$\leq 2 \int_0^\infty e^{-\frac{t^{1/m} + \mu_u}{2\sigma^2}} \, dt + \mu_u^m = m!(2\sigma^2)^m \Big\{ 2e^{-\mu_u/(2\sigma^2)} + \frac{1}{m!}\Big(\frac{\mu_u}{2\sigma^2}\Big)^m \Big\}$$

$$\leq 2m!(2\sigma^2)^m,$$

where the final inequality follows because the function $x \mapsto 2e^{-x} + x^m/m!$ is decreasing on $[0, 1/2]$. This calculation allows us to apply Bernstein's inequality (e.g. van de Geer, 2000, Lemma 5.7, taking $K = 2\sigma^2, R = 4\sigma^2$ in her notation), to deduce that for any $s \geq 0$,

$$\mathbb{P}(|\hat{V}(u) - V(u)| \geq s) \leq 2\exp\left(-\frac{ns^2}{4\sigma^2 s + 32\sigma^4}\right).$$

It follows by Lemma 2 in Section 1 in the supplementary material Wang, Berthet and Samworth (2015), taking $\epsilon = 1/4$ in that result, that if $\eta > 0$ is such that $\ell \log(p/\eta) \leq n$, then for $C := 8\sigma^2$, we have

$$\mathbb{P}\left(\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq 2C\sqrt{\frac{\ell \log(p/\eta)}{n}}\right)$$

$$\leq 2\pi\ell^{1/2}\binom{p}{\ell}\left(\frac{128}{\sqrt{255}}\right)^{\ell-1}\exp\left(-\frac{C^2\ell\log(p/\eta)}{4C\sigma^2\sqrt{\frac{\ell\log(p/\eta)}{n}} + 32\sigma^4}\right)$$

$$\leq 2\pi\ell^{1/2}\left(\frac{e}{\ell}\right)^\ell\left(\frac{128}{\sqrt{255}}\right)^{\ell-1}\eta^\ell \leq e^9\eta,$$

Similarly, if $\ell \log(p/\eta) > n$, then

$$\mathbb{P}\left(\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq 2C\frac{\ell\log(p/\eta)}{n}\right)$$

$$\leq 2\pi\ell^{1/2}\binom{p}{\ell}\left(\frac{128}{\sqrt{255}}\right)^{\ell-1}\exp\left(-\frac{C^2\ell^2\log^2(p/\eta)}{4C\sigma^2\ell\log(p/\eta) + 32\sigma^4 n}\right) \leq e^9\eta.$$

Setting $\delta := e^9\eta$, we find (noting that we only need to consider the case $\delta \in (0, 1]$) that

$$\mathbb{P}\left\{\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq 16\sigma^2\left(1 + \frac{9}{\log p}\right)\max\left(\sqrt{\frac{\ell\log(p/\delta)}{n}}, \frac{\ell\log(p/\delta)}{n}\right)\right\}$$

$$\leq \mathbb{P}\left\{\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq 16\sigma^2\max\left(\sqrt{\frac{\ell\log(e^9 p/\delta)}{n}}, \frac{\ell\log(e^9 p/\delta)}{n}\right)\right\} \leq \delta.$$

(ii) By Lemma 1 of Laurent and Massart (2000), if $Y_1, \ldots, Y_n$ are independent χ_1^2 random variables, then for all $a > 0$,

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n Y_i - n\right| \geq a\right) \leq 2e^{-\frac{n}{2}(1+a-\sqrt{1+2a})} \leq 2e^{-n\min(\frac{a}{4}, \frac{a^2}{16})}.$$

Setting $\eta := e^{-n \min(\frac{a}{4}, \frac{a^2}{16})}$, we deduce that

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^n Y_i - n\right| \geq 4\max\left(\sqrt{\frac{\log(1/\eta)}{n}}, \frac{\log(1/\eta)}{n}\right)\right\} \leq 2\eta.$$

Hence, using Lemma 2 again, and by a similar calculation to Part (i),

$$\mathbb{P}\left\{\sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq 8\lambda_1(P)\max\left(\sqrt{\frac{\log(1/\eta)}{n}}, \frac{\log(1/\eta)}{n}\right)\right\} \leq e^9 p^\ell \eta.$$

The result follows on setting $\delta := e^9 p^\ell \eta$.   □

PROOF OF THEOREM 2. Fix an arbitrary $P \in \mathcal{P}_p(n, k, \theta)$. For notational simplicity, we write $v := v_1(P)$ and $\hat{v} := \hat{v}_{\max}^k(\hat{\Sigma})$ in this proof. We now exploit the Curvature Lemma of Vu et al. (2013, Lemma 3.1), which is closely related to the Davis–Kahan $\sin\theta$ theorem (Davis and Kahan, 1970; Yu, Wang and Samworth, 2015). This lemma gives that

$$\|\hat{v}\hat{v}^\top - vv^\top\|_2^2 \leq \frac{2}{\theta}\text{tr}\big(\Sigma(vv^\top - \hat{v}\hat{v}^\top)\big) \leq \frac{2}{\theta}\text{tr}\big((\Sigma - \hat{\Sigma})(vv^\top - \hat{v}\hat{v}^\top)\big).$$

When $\hat{v}\hat{v}^\top \neq vv^\top$, we have that $\frac{vv^\top - \hat{v}\hat{v}^\top}{\|vv^\top - \hat{v}\hat{v}^\top\|_2}$ has rank 2, trace 0 and has non-zero entries in at most $2k$ rows and $2k$ columns. It follows that its non-zero eigenvalues are $\pm 1/\sqrt{2}$, so it can be written as $(xx^\top - yy^\top)/\sqrt{2}$ for some $x, y \in B_0(2k)$. Thus

$$\mathbb{E}L(\hat{v}, v) = \mathbb{E}\frac{1}{\sqrt{2}}\|\hat{v}\hat{v}^\top - vv^\top\|_2 \leq \frac{1}{\theta}\mathbb{E}\text{tr}\big((\Sigma - \hat{\Sigma})(xx^\top - yy^\top)\big)$$

$$\leq \frac{2}{\theta}\mathbb{E}\sup_{u \in B_0(2k)}|\hat{V}(u) - V(u)| \leq 2\sqrt{2}\left(1 + \frac{1}{\log p}\right)\sqrt{\frac{k\log p}{n\theta^2}},$$

where we have used Proposition 1 in Section 1 in the online supplementary material Wang, Berthet and Samworth (2015) to obtain the final inequality.   □

PROOF OF THEOREM 3. Set $\sigma^2 := \frac{1}{8(1+\frac{9}{\log p})} - \theta$. We have by Proposition 1(ii) that $N_p(0, \sigma^2 I_p + \theta v_1 v_1^\top) \in \mathcal{P}_p(n, k, \theta)$ for any unit vector $v_1 \in B_0(k)$. Define $k_0 := k - 1$ and $p_0 := p - 1$. Applying the variant of the Gilbert–Varshamov lemma given as Lemma 3 in Section 1 in the online supplementary material Wang, Berthet and Samworth (2015) with $\alpha := 1/2$ and $\beta := 1/4$, we can construct a set $\mathcal{N}_0$ of k_0-sparse vectors in $\{0, 1\}^{p_0}$ with

cardinality at least $(p_0/k_0)^{k_0/8}$, such that the Hamming distance between every pair of distinct points in $\mathcal{N}_0$ is at least k_0. For $\epsilon \in (0,1]$ to be chosen later, define a set of k-sparse vectors in $\mathbb{R}^p$ by

$$\mathcal{N} := \left\{ \begin{pmatrix} \sqrt{1-\epsilon^2} \\ k_0^{-1/2}\epsilon u_0 \end{pmatrix} : u_0 \in \mathcal{N}_0 \right\}.$$

Observe that if u, v are distinct elements of $\mathcal{N}$, then

$$L(u,v) = \{1 - (u^\top v)^2\}^{1/2} \geq \{1 - (1 - \epsilon^2/2)^2\}^{1/2} \geq \frac{\sqrt{3}\epsilon}{2},$$

and similarly $L(u,v) \leq \epsilon$. For $u \in \mathcal{N}$, let P_u denote the multivariate normal distribution $N_p(0, \sigma^2 I_p + \theta uu^\top)$. For any estimator $\hat{v} \in \mathcal{V}_{n,p}$, we define $\hat{\psi}_{\hat{v}} := \text{sargmin}_{u \in \mathcal{N}} L(\hat{v}, u)$, where sargmin denotes the smallest element of the argmin in the lexicographic ordering. Note that $\{\hat{\psi}_{\hat{v}} \neq u\} \subseteq \{L(\hat{v}, u) \geq \sqrt{3}\epsilon/4\}$. We now apply the generalised version of Fano's lemma given as Lemma 4 in Section 1 in the online supplementary material Wang, Berthet and Samworth (2015). Writing $D(P\|Q)$ for the Kullback–Leibler divergence between two probability measures defined on the same space (a formal definition is given just prior to Lemma 4), we have

$$\inf_{\hat{v} \in \mathcal{V}_{n,p}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}, v_1(P)) \geq \inf_{\hat{v} \in \mathcal{V}_{n,p}} \max_{u \in \mathcal{N}} \mathbb{E}_{P_u} L(\hat{v}, u)$$

$$\geq \frac{\sqrt{3}\epsilon}{4} \inf_{\hat{v} \in \mathcal{V}_{n,p}} \max_{u \in \mathcal{N}} P_u^{\otimes n}(\hat{\psi}_{\hat{v}} \neq u)$$

$$(8) \qquad \geq \frac{\sqrt{3}\epsilon}{4} \left(1 - \frac{\max_{u,v \in \mathcal{N}, u \neq v} D(P_v^{\otimes n}\|P_u^{\otimes n}) + \log 2}{(k_0/8)\log(p_0/k_0)} \right).$$

We can compute, for distinct points $u, v \in \mathcal{N}$,

$$D(P_v^{\otimes n}\|P_u^{\otimes n}) = nD(P_v\|P_u) = \frac{n}{2}\text{tr}\big((\sigma^2 I_p + \theta uu^\top)^{-1}(\sigma^2 I_p + \theta vv^\top) - I_p\big)$$

$$= \frac{n}{2}\text{tr}\big((\sigma^2 I_p + \theta uu^\top)^{-1}\theta(vv^\top - uu^\top)\big)$$

$$= \frac{n\theta}{2}\text{tr}\left(\left(\frac{1}{\sigma^2}I_p - \frac{\theta}{\sigma^2(\sigma^2 + \theta)}uu^\top\right)(vv^\top - uu^\top)\right)$$

$$(9) \qquad = \frac{n\theta^2}{2\sigma^2(\sigma^2 + \theta)}L^2(u,v) \leq \frac{n\theta^2\epsilon^2}{2\sigma^2(\sigma^2 + \theta)}.$$

Let $\epsilon := \min\{\sqrt{a/(3b)}, 1\}$, where

$$a := 1 - \frac{8\log 2}{k_0 \log(p_0/k_0)} \quad \text{and} \quad b := \frac{4n\theta^2}{\sigma^2(\sigma^2 + \theta)k_0 \log(p_0/k_0)}.$$

Then from (8) and (9), we find that

$$\inf_{\hat{v} \in \mathcal{V}_{n,p}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\big(\hat{v}, v_1(P)\big) \geq \min\bigg\{ \frac{1}{1660} \sqrt{\frac{k \log p}{n\theta^2}}, \frac{5}{18\sqrt{3}} \bigg\},$$

as required. $\square$

APPENDIX B: PROOFS FROM SECTION 3

PROOF OF LEMMA 4. For convenience, we write $v := v_1(\Sigma)$, $\hat{v}$ for $\hat{v}^{\mathrm{SDP}}$ and $\hat{M}$ for $\hat{M}^\epsilon$ in this proof. We first study $vv^\top - \hat{M}$, where $\hat{M} \in \mathcal{M}_1$ is computed in Step 2 of Algorithm 1. By the Curvature Lemma of Vu et al. (2013, Lemma 3.1),

$$\|vv^\top - \hat{M}\|_2^2 \leq \frac{2}{\theta} \mathrm{tr}\big(\Sigma(vv^\top - \hat{M})\big).$$

Moreover, since $vv^\top \in \mathcal{M}_1$, we have the basic inequality

$$\mathrm{tr}(\hat{\Sigma}\hat{M}) - \lambda\|\hat{M}\|_1 \geq \mathrm{tr}(\hat{\Sigma}vv^\top) - \lambda\|vv^\top\|_1 - \epsilon.$$

Let S denote the set of indices corresponding to the non-zero components of v, and recall that $|S| \leq k$. Since by hypothesis $\|\hat{\Sigma} - \Sigma\|_\infty \leq \lambda$, we have

$$\|vv^\top - \hat{M}\|_2^2 \leq \frac{2}{\theta}\big\{ \mathrm{tr}\big(\hat{\Sigma}(vv^\top - \hat{M})\big) + \mathrm{tr}\big((\Sigma - \hat{\Sigma})(vv^\top - \hat{M})\big) \big\}$$

$$\leq \frac{2}{\theta}\big(\lambda\|vv^\top\|_1 - \lambda\|\hat{M}\|_1 + \epsilon + \|\hat{\Sigma} - \Sigma\|_\infty\|vv^\top - \hat{M}\|_1 \big)$$

$$\leq \frac{2\lambda}{\theta}\big(\|v_S v_S^\top\|_1 - \|\hat{M}_{S,S}\|_1 + \|v_S v_S^\top - \hat{M}_{S,S}\|_1 \big) + \frac{2\epsilon}{\theta}$$

$$\leq \frac{4\lambda}{\theta}\|v_S v_S^\top - \hat{M}_{S,S}\|_1 + \frac{2\epsilon}{\theta} \leq \frac{4\lambda k}{\theta}\|vv^\top - \hat{M}\|_2 + \frac{2\epsilon}{\theta}.$$

We deduce that

$$\|vv^\top - \hat{M}\|_2 \leq \frac{4\lambda k}{\theta} + \sqrt{\frac{2\epsilon}{\theta}}.$$

On the other hand,

$$\|vv^\top - \hat{M}\|_2^2 = \mathrm{tr}\big((vv^\top - \hat{M})^2\big) = 1 - 2v^\top \hat{M} v + \mathrm{tr}(\hat{M}^2)$$

$$\geq 1 - 2\hat{v}^\top \hat{M}\hat{v} + \mathrm{tr}(\hat{M}^2) = \|\hat{v}\hat{v}^\top - \hat{M}\|_2^2.$$

We conclude that

$$L(\hat{v}, v) = \frac{1}{\sqrt{2}}\|\hat{v}\hat{v}^\top - vv^\top\|_2 \leq \frac{1}{\sqrt{2}}(\|\hat{v}\hat{v}^\top - \hat{M}\|_2 + \|vv^\top - \hat{M}\|_2)$$

$$\leq \sqrt{2}\|vv^\top - \hat{M}\|_2 \leq \frac{4\sqrt{2}\lambda k}{\theta} + 2\sqrt{\frac{\epsilon}{\theta}},$$

as required. $\square$

PROOF OF THEOREM 5. Fix $P \in \mathcal{P}_p(n, k, \theta)$. By Lemma 4, and by Lemma 5 in Section 1 of the online supplementary material Wang, Berthet and Samworth (2015),

$$\mathbb{E}L\big(\hat{v}^{\mathrm{SDP}}, v_1(P)\big) = \mathbb{E}\big\{L\big(\hat{v}^{\mathrm{SDP}}, v_1(P)\big)\mathbb{1}_{\{\|\hat{\Sigma}-\Sigma\|_\infty \leq \lambda\}}\big\}$$
$$+ \mathbb{E}\big\{L\big(\hat{v}^{\mathrm{SDP}}, v_1(P)\big)\mathbb{1}_{\{\|\hat{\Sigma}-\Sigma\|_\infty > \lambda\}}\big\}$$

$$(10) \qquad \leq \frac{4\sqrt{2}\lambda k}{\theta} + 2\sqrt{\frac{\epsilon}{\theta}} + \mathbb{P}\bigg(\sup_{u \in B_0(2)}\big|\hat{V}(u) - V(u)\big| > 2\sqrt{\frac{\log p}{n}}\bigg)$$

Since $P \in \mathrm{RCC}_p(n, 2, 1)$, we have for each $\delta > 0$ that

$$\mathbb{P}\bigg\{\sup_{u \in B_0(2)}\big|\hat{V}(u) - V(u)\big| > \max\bigg(\sqrt{\frac{2\log(p/\delta)}{n}}, \frac{2\log(p/\delta)}{n}\bigg)\bigg\} \leq \delta.$$

Set $\delta := \sqrt{\frac{k^2 \log p}{n\theta^2}}$. Since $4\log p \leq n$, which in particular implies $n \geq 3$, we have

$$\frac{2\log(p/\delta)}{n} \leq \frac{1}{2} + \frac{1}{n}\log\Big(\frac{n\theta^2}{k^2 \log p}\Big) \leq \frac{1}{2} + \frac{\log n}{n} - \frac{1}{n}\log\log 2 \leq 1.$$

Moreover, since $n \leq k^2 p^2 \theta^{-2} \log p$,

$$2\log(p/\delta) = 2\log p + \log\Big(\frac{n\theta^2}{k^2 \log p}\Big) \leq 4\log p.$$

We deduce that

$$(11) \qquad \mathbb{P}\bigg(\sup_{u \in B_0(2)}\big|\hat{V}(u) - V(u)\big| > 2\sqrt{\frac{\log p}{n}}\bigg) \leq \sqrt{\frac{k^2 \log p}{n\theta^2}}.$$

The desired risk bound follows from (10), the fact that $\theta \leq k$, and (11). $\square$

APPENDIX C: PROOFS FROM SECTION 4

PROOF OF THEOREM 6. Suppose, for a contradiction, that there exist an infinite subset $\mathcal{N}$ of $\mathbb{N}$, $K_0 \in [0, \infty)$ and a sequence $(\hat{v}^{(n)})$ of randomised polynomial time estimators of $v_1(P)$ satisfying

$$\sup_{P \in \mathcal{P}_p(n, k, \theta)} \mathbb{E}_P L\big(\hat{v}^{(n)}(\mathbf{X}), v_1(P)\big) \leq K_0\sqrt{\frac{k^{1+\alpha} \log p}{n\theta^2}}$$

for all $n \in \mathcal{N}$. Let $L := \lceil \log p_n \rceil$, let $m = m_n := \lceil 10Lp_n/9 \rceil$ and let $\kappa = \kappa_n := Lk_n$. We claim that Algorithm 4 below is a randomised polynomial time

---

Algorithm 4: Pseudo-code for a planted clique algorithm based on a hypothetical randomised polynomial time sparse principal component estimation algorithm.

Input: $m \in \mathbb{N}$, $\kappa \in \{1, \ldots, m\}$, $G \in \mathbb{G}_m$, $L \in \mathbb{N}$
begin

> **Step 1:** Let $n \leftarrow \lfloor 9m/(10L) \rfloor$, $p \leftarrow p_n$, $k \leftarrow \lfloor \kappa/L \rfloor$. Draw $u_1, \ldots, u_n, w_1, \ldots, w_p$ uniformly at random without replacement from $V(G)$. Form
> $\mathbf{A} = (A_{ij}) \leftarrow (\mathbb{1}_{\{u_i \sim w_j\}}) \in \mathbb{R}^{n \times p}$ and $\mathbf{X} \leftarrow \mathrm{diag}(\xi_1, \ldots, \xi_n)(2\mathbf{A} - \mathbf{1}_{n \times p})$, where $\xi_1, \ldots, \xi_n$ are independent Rademacher random variables (independent of $u_1, \ldots, u_n, w_1, \ldots, w_p$), and where every entry of $\mathbf{1}_{n \times p} \in \mathbb{R}^{n \times p}$ is 1.
>
> **Step 2:** Use the randomised estimator $\hat{v}^{(n)}$ to compute $\hat{v} = \hat{v}^{(n)}(\mathbf{X}/\sqrt{750})$.
>
> **Step 3:** Let $\hat{S} = \hat{S}(\hat{v})$ be the lexicographically smallest k-subset of $\{1, \ldots, p\}$ such that $(\hat{v}_j : j \in \hat{S})$ contains the k largest coordinates of $\hat{v}$ in absolute value.
>
> **Step 4:** For $u \in V(G)$ and $W \subseteq V(G)$, let
> $\mathrm{nb}(u, W) := \mathbb{1}_{\{u \in W\}} + \sum_{w \in W} \mathbb{1}_{\{u \sim w\}}$. Set
> $\hat{K} := \{u \in V(G) : \mathrm{nb}(u, \{w_j : j \in \hat{S}\}) \geq 3k/4\}$.

end
Output: $\hat{K}$

---

algorithm that correctly identifies the Planted Clique problem on m_n vertices and a planted clique of size κ_n with probability tending to 1 as $n \to \infty$. Since $\kappa_n = O(m_n^{1/2 - \tau - \delta} \log m_n)$, this contradicts Assumption (A1)(τ). We prove the claim below.

Let $G \sim \mathbb{G}_{m,\kappa}$, and let $K \subseteq V(G)$ denote the planted clique. Note that the matrix $\mathbf{A}$ defined in Step 1 of Algorithm 4 is the off-diagonal block of the adjacency matrix of G associated with the bipartite graph induced by the two parts $\{u_i : i = 1, \ldots, n\}$ and $\{w_j : j = 1, \ldots, p\}$. Let $\boldsymbol{\epsilon}' = (\epsilon_1', \ldots, \epsilon_n')^\top$ and $\boldsymbol{\gamma}' = (\gamma_1', \ldots, \gamma_p')^\top$, where $\epsilon_i' := \mathbb{1}_{\{u_i \in K\}}$, $\gamma_j' := \mathbb{1}_{\{w_j \in K\}}$, and set $S' := \{j : \gamma_j' = 1\}$.

It is convenient at this point to introduce the notion of a *Graph Vector distribution*. We say Y has a p-variate Graph Vector distribution with parameters $g = (g_1, \ldots, g_p)^\top \in \{0, 1\}^p$ and $\pi_0 \in [0, 1]$, and write $Y \sim \mathrm{GV}_p^g(\pi_0)$, if we can write

$$Y = \xi\{(1 - \epsilon)R + \epsilon(g + \tilde{R})\},$$

where ξ, ϵ and R are independent, where ξ is a Rademacher random variable, where $\epsilon \sim \mathrm{Bern}(\pi_0)$, where $R = (R_1, \ldots, R_p)^\top \in \mathbb{R}^p$ has independent Rademacher components, and where $\tilde{R} = (\tilde{R}_1, \ldots, \tilde{R}_p)^\top$ with $\tilde{R}_j := (1 - g_j)R_j$.

Let $(\boldsymbol{\epsilon}, \boldsymbol{\gamma})^\top = (\epsilon_1, \ldots, \epsilon_n, \gamma_1, \ldots, \gamma_p)^\top$ be $n + p$ independent $\mathrm{Bern}(\kappa/m)$ random variables. For $i = 1, \ldots, n$, let $Y_i := \xi_i\{(1 - \epsilon_i)R_i + \epsilon_i(\boldsymbol{\gamma} + \tilde{R}_i)\}$

so that, conditional on $\boldsymbol{\gamma}$, the random vectors $Y_1, \ldots, Y_n$ are independent, each distributed as $\mathrm{GV}_p^{\gamma}(\kappa/m)$. As shorthand, we denote this conditional distribution as $Q_{\boldsymbol{\gamma}}$, and write $S := \{j : \gamma_j = 1\}$. Note that by Lemma 6 in Section 1 of the online supplementary material Wang, Berthet and Samworth (2015), $Q_{\boldsymbol{\gamma}} \in \cap_{\ell=1}^{\lfloor 20p/(9k) \rfloor} \mathrm{RCC}_p(\ell, 750)$.

Let $\boldsymbol{Y} := (Y_1, \ldots, Y_n)^{\top}$. Recall that if P and Q are probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$, the *total variation distance* between P and Q is defined by

$$d_{\mathrm{TV}}(P, Q) := \sup_{B \in \mathcal{B}} |P(B) - Q(B)|.$$

Writing $\mathcal{L}(Z)$ for the distribution (or law) of a generic random element Z, and using elementary properties of the total variation distance given in Lemma 9 in Section 1 in the online supplementary material Wang, Berthet and Samworth (2015), we have

$$d_{\mathrm{TV}}\big(\mathcal{L}(\mathbf{X}), \mathcal{L}(\boldsymbol{Y})\big) \leq d_{\mathrm{TV}}\Big(\mathcal{L}\big(\boldsymbol{\epsilon}', \boldsymbol{\gamma}', (R_{ij}), (\xi_i)\big), \mathcal{L}\big(\boldsymbol{\epsilon}, \boldsymbol{\gamma}, (R_{ij}), (\xi_i)\big)\Big)$$

$$(12) \qquad = d_{\mathrm{TV}}\big(\mathcal{L}(\boldsymbol{\epsilon}', \boldsymbol{\gamma}'), \mathcal{L}(\boldsymbol{\epsilon}, \boldsymbol{\gamma})\big) \leq \frac{2(n+p)}{m} \leq \frac{9(n+p)}{5p \log p}.$$

Here, the penultimate inequality follows from Diaconis and Freedman (1980, Theorem 4). In view of (12), we initially analyse Steps 2, 3 and 4 in Algorithm 4 with $\mathbf{X}$ replaced by $\boldsymbol{Y}$. Observe that $\mathbb{E}(Y_i | \boldsymbol{\gamma}) = 0$ and, writing $\Delta := \mathrm{diag}(\boldsymbol{\gamma}) \in \mathbb{R}^{p \times p}$, we have

$$\Sigma_{\boldsymbol{\gamma}} := \mathrm{Cov}(Y_i | \boldsymbol{\gamma}) = \mathbb{E}\big\{(1 - \epsilon_i) R_i R_i^{\top} + \epsilon_i (\boldsymbol{\gamma} + \tilde{R}_i)(\boldsymbol{\gamma} + \tilde{R}_i)^{\top} \big| \boldsymbol{\gamma}\big\}$$

$$= I_p + \frac{\kappa}{m}(\boldsymbol{\gamma}\boldsymbol{\gamma}^{\top} - \Delta).$$

Writing $N_{\boldsymbol{\gamma}} := \sum_{j=1}^{p} \gamma_j$, it follows that the largest eigenvalue of $\Sigma_{\boldsymbol{\gamma}}$ is $1 + \frac{\kappa}{m}(N_{\boldsymbol{\gamma}} - 1)$, with corresponding eigenvector $\boldsymbol{\gamma}/N_{\boldsymbol{\gamma}}^{1/2} \in B_0(N_{\boldsymbol{\gamma}})$. The other eigenvalues are 1, with multiplicity $p - N_{\boldsymbol{\gamma}}$, and $1 - \frac{\kappa}{m}$, with multiplicity $N_{\boldsymbol{\gamma}} - 1$. Hence $\lambda_1(\Sigma_{\boldsymbol{\gamma}}) - \lambda_2(\Sigma_{\boldsymbol{\gamma}}) = \frac{\kappa}{m}(N_{\boldsymbol{\gamma}} - 1)$. Define

$$\Gamma_0 := \left\{ g \in \{0, 1\}^p : \left| N_g - \frac{p\kappa}{m} \right| \leq \frac{k}{20} \right\},$$

where $N_g := \sum_{j=1}^{p} g_j$. We note that by Bernstein's inequality (e.g. Shorack and Wellner, 1986, p. 855) that

$$(13) \qquad \mathbb{P}(\boldsymbol{\gamma} \in \Gamma_0) \geq 1 - 2e^{-k/800}.$$

If $g \in \Gamma_0$, the conditional distribution of $Y_1/\sqrt{750}$ given $\gamma = g$ belongs to $\mathcal{P}_p(n, k, \theta)$ for $\theta \leq \frac{\kappa}{750m}(N_g - 1)$ and all large $n \in \mathcal{N}$. By hypothesis, it follows that for $g \in \Gamma_0$,

$$\mathbb{E}\Big\{ L\big(\hat{v}^{(n)}(\boldsymbol{Y}/\sqrt{750}), v_1(Q_{\boldsymbol{\gamma}})\big) \;\Big|\; \boldsymbol{\gamma} = g \Big\} \leq K_0 \sqrt{\frac{k^{1+\alpha} \log p}{n\theta^2}}$$

for all large $n \in \mathcal{N}$. Then by Lemma 7 in Section 1 in the online supplementary material Wang, Berthet and Samworth (2015), for $\hat{S}(\cdot)$ defined in Step 3 of Algorithm 4, for $g \in \Gamma_0$, and large $n \in \mathcal{N}$,

$$\mathbb{E}\Big\{ \big|S \setminus \hat{S}\big(\hat{v}^{(n)}(\boldsymbol{Y}/\sqrt{750})\big)\big| \;\Big|\; \boldsymbol{\gamma} = g \Big\}$$
$$\leq 2N_g \mathbb{E}\Big\{ L\big(\hat{v}^{(n)}(\boldsymbol{Y}/\sqrt{750}), v_1(Q_{\boldsymbol{\gamma}})\big)^2 \;\Big|\; \boldsymbol{\gamma} = g \Big\}$$
$$\leq 2N_g K_0 \sqrt{\frac{k^{1+\alpha} \log p}{n\theta^2}}$$

We deduce by Markov's inequality that for $g \in \Gamma_0$, and large $n \in \mathcal{N}$,

$$(14) \quad \mathbb{P}\Big\{ \big|S \cap \hat{S}\big(\hat{v}^{(n)}(\boldsymbol{Y}/\sqrt{750})\big)\big| \leq 16N_{\boldsymbol{\gamma}}/17 \;\Big|\; \boldsymbol{\gamma} = g \Big\} \leq 34K_0 \sqrt{\frac{k^{1+\alpha} \log p}{n\theta^2}}$$

Let

$$\Omega_{0,n} := \{\boldsymbol{\gamma} \in \Gamma_0\} \cap \big\{ \big|S \cap \hat{S}\big(\hat{v}^{(n)}(\boldsymbol{Y}/\sqrt{750})\big)\big| > 16N_{\boldsymbol{\gamma}}/17 \big\}$$
$$\Omega'_{0,n} := \{\boldsymbol{\gamma}' \in \Gamma_0\} \cap \big\{ \big|S \cap \hat{S}\big(\hat{v}^{(n)}(\boldsymbol{X}/\sqrt{750})\big)\big| > 16N_{\boldsymbol{\gamma}'}/17 \big\} =: \Omega'_{1,n} \cap \Omega'_{2,n},$$

say, where $N_{\boldsymbol{\gamma}'} := \sum_{j=1}^p \gamma'_j$. When $n \in \mathcal{N}$ is sufficiently large, we have on the event $\Omega'_{0,n}$ that

$$(15) \qquad\qquad \big|\{j \in \hat{S}\big(\hat{v}^{(n)}(\boldsymbol{X}/\sqrt{750})\big) : w_j \in K\}\big| > 3k/4.$$

Now set

$$\Omega'_{3,n} := \Big\{ \mathrm{nb}(u, \{w_j : j \in S'\}) \leq \frac{k}{2} \text{ for all } u \in V(G) \setminus K \Big\}.$$

Recall the definition of $\hat{K}$ from Step 4 of Algorithm 4. We claim that for sufficiently large $n \in \mathcal{N}$,

$$\Omega'_{0,n} \cap \Omega'_{3,n} \subseteq \{\hat{K} = K\}.$$

To see this, note that for $n \in \mathcal{N}$ sufficiently large, on $\Omega'_{0,n}$ we have $K \subseteq \hat{K}$ by (15). For the reverse inclusion, note that if $u \in V(G) \setminus K$, then on $\Omega'_{0,n} \cap \Omega'_{3,n}$, we have for sufficiently large $n \in \mathcal{N}$ that

$$
\text{nb}\Big(u, \big\{ w_j : j \in \hat{S}\big(\hat{v}^{(n)}(\boldsymbol{X}/\sqrt{750}) \big) \big\} \Big)
$$
$$
\leq \big| \{ w_j : j \in \hat{S} \} \setminus K \big| + \text{nb}\big(u, \{ w_j : j \in \hat{S} \} \cap K \big)
$$
$$
\leq \big| \{ w_j : j \in \hat{S} \} \setminus K \big| + \text{nb}\big(u, \{ w_j : j \in S' \} \big) < \frac{k}{4} + \frac{k}{2} = \frac{3k}{4}.
$$

This establishes our claim. We conclude that for sufficiently large $n \in \mathcal{N}$,

$$
(16) \quad \mathbb{P}(\hat{K} \neq K) \leq \mathbb{P}\big((\Omega'_{0,n} \cap \Omega'_{3,n})^c \big) \leq \mathbb{P}\big((\Omega'_{0,n})^c \big) + \mathbb{P}\big(\Omega'_{1,n} \cap (\Omega'_{3,n})^c \big).
$$

Now, by Lemma 9, we have

$$
(17) \qquad |\mathbb{P}(\Omega'_{0,n}) - \mathbb{P}(\Omega_{0,n})| \leq d_{\text{TV}}\big(\mathcal{L}(\mathbf{X}, \boldsymbol{\gamma}'), \mathcal{L}(\mathbf{Y}, \boldsymbol{\gamma}) \big) \leq \frac{9(n+p)}{5p \log p}.
$$

Moreover, by a union bound and Hoeffding's inequality, for large $n \in \mathcal{N}$,

$$
(18) \qquad \mathbb{P}\big(\Omega'_{1,n} \cap (\Omega'_{3,n})^c \big) \leq \sum_{g \in \Gamma_0} \mathbb{P}\big((\Omega'_{3,n})^c | \boldsymbol{\gamma} = g \big) \mathbb{P}(\boldsymbol{\gamma} = g) \leq m e^{-k/800}.
$$

We conclude by (16), (17), (13), (14) and (18) that for large $n \in \mathcal{N}$,

$$
\mathbb{P}(\hat{K} \neq K) \leq \frac{9(n+p)}{5p \log p} + 2e^{-k/800} + 34 K_0 \sqrt{\frac{k^{1+\alpha} \log p}{n\theta^2}} + m e^{-k/800} \to 0
$$

as $n \to \infty$. This contradicts Assumption (A1)(τ), and therefore completes the proof. $\square$

PROOF OF THEOREM 7. Setting $\delta := p^{-1}$ in (3), there exist events Ω_1 and Ω_2, each with probability at least $1 - p^{-1}$, such that on Ω_1 and Ω_2, we respectively have
(19)

$$
\sup_{u \in B_0(2k)} |\hat{V}(u) - V(u)| \leq 2\sqrt{\frac{k \log p}{n}} \quad \text{and} \quad \sup_{u \in B_0(2)} |\hat{V}(u) - V(u)| \leq 2\sqrt{\frac{\log p}{n}}.
$$

Let $\Omega_0 := \Omega_1 \cap \Omega_2$. We work on Ω_0 henceforth. The main ingredient for proving both parts of the theorem is the following weak-duality inequality:

$$
\max_{M \in \mathcal{M}_1} \text{tr}(\hat{\Sigma} M) - \lambda \|M\|_1 = \max_{M \in \mathcal{M}_1} \min_{U \in \mathcal{U}} \text{tr}\big((\hat{\Sigma} - U) M \big)
$$
$$
(20) \qquad\qquad\qquad \leq \min_{U \in \mathcal{U}} \max_{M \in \mathcal{M}_1} \text{tr}\big((\hat{\Sigma} - U) M \big) = \min_{U \in \mathcal{U}} \lambda_1(\hat{\Sigma} - U).
$$

It is convenient to denote $\gamma := \sqrt{\frac{k^2 \log p}{n\theta^2}}$, and note that

$$\gamma \leq \frac{\sqrt{k}}{16} \min_{j \in S} |v_{1,j}| \leq \frac{1}{16}\|v_{1,S}\|_2 = \frac{1}{16}.$$

Proof of (a). From (20), it suffices to exhibit a primal-dual pair $(\hat{M}, \hat{U}) \in \mathcal{M}_1 \times \mathcal{U}$, such that

(C1) $\hat{M} = \hat{v}\hat{v}^\top$ with $\mathrm{sgn}(\hat{v}) = \mathrm{sgn}(v_1)$.
(C2) $\mathrm{tr}(\hat{\Sigma}\hat{M}) - \lambda\|\hat{M}\|_1 = \lambda_1(\hat{\Sigma} - \hat{U})$.

We construct the primal-dual pair as follows. Define

$$\hat{U} := \begin{pmatrix} \lambda\,\mathrm{sgn}(v_{1,S})\mathrm{sgn}(v_{1,S})^\top & \hat{\Sigma}_{SS^c} - \Sigma_{SS^c} \\ \hat{\Sigma}_{S^cS} - \Sigma_{S^cS} & \hat{\Sigma}_{S^cS^c} - \Sigma_{S^cS^c} \end{pmatrix}.$$

By (19) and Lemma 5, we have that $\|\hat{\Sigma} - \Sigma\|_\infty \leq 4\sqrt{\frac{\log p}{n}} \leq \lambda$, so $U \in \mathcal{U}$. Let $w = (w_1, \ldots, w_k)$ be a unit-length leading eigenvector of $\Sigma_{SS} - \hat{U}_{SS}$ such that $w^\top v_{1,S} \geq 0$. Then, define $\hat{v}$ componentwise by

$$\hat{v}_S \in \operatorname*{argmax}_{\substack{u \in \mathbb{R}^k, \|u\|_2 = 1 \\ u^\top w \geq 0}} u^\top\left(\hat{\Sigma}_{SS} - \hat{U}_{SS}\right)u, \quad \hat{v}_{S^c} = 0,$$

and set $\hat{M} := \hat{v}\hat{v}^\top$. Note that our choices above ensure that $\hat{M} \in \mathcal{M}_1$. To verify (C1), we now show that $\mathrm{sgn}(\hat{v}_S) = \mathrm{sgn}(w) = \mathrm{sgn}(v_{1,S})$. By a variant of the Davis–Kahan theorem (Yu, Wang and Samworth, 2015, Theorem 2),

$$\|w - \hat{v}_S\|_\infty \leq \|w - \hat{v}_S\|_2 \leq \sqrt{2}L(\hat{v}_S, w) \leq \frac{2\sqrt{2}\|\hat{\Sigma}_{SS} - \Sigma_{SS}\|_{\mathrm{op}}}{\theta}$$

$$(21) \qquad\qquad \leq \frac{2\sqrt{2}}{\theta} \sup_{u \in B_0(2k)} |\hat{V}(u) - V(u)| \leq 4\sqrt{2}\gamma k^{-1/2},$$

where the final inequality uses (19). But w is also a leading eigenvector of

$$\frac{1}{\theta}(\Sigma_{SS} - \hat{U}_{SS} - I_k) = v_{1,S}v_{1,S}^\top - 4\gamma ss^\top,$$

where $s := \frac{\mathrm{sgn}(v_{1,S})}{\|\mathrm{sgn}(v_{1,S})\|}$. Write $s = \alpha v_{1,S} + \beta v_\perp$ for some $\alpha, \beta \in \mathbb{R}$ with $\alpha^2 + \beta^2 = 1$, and a unit vector $v_\perp \in \mathbb{R}^k$ orthogonal to $v_{1,S}$. Then

$$v_{1,S}v_{1,S}^\top - \gamma ss^\top = \begin{pmatrix} v_{1,S} & v_\perp \end{pmatrix} \begin{pmatrix} 1 - 4\gamma\alpha^2 & -4\gamma\alpha\beta \\ -4\gamma\alpha\beta & -4\gamma\beta^2 \end{pmatrix} \begin{pmatrix} v_{1,S}^\top \\ v_\perp^\top \end{pmatrix}$$

$$= \begin{pmatrix} v_{1,S} & v_\perp \end{pmatrix} \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix} \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix} \begin{pmatrix} v_{1,S}^\top \\ v_\perp^\top \end{pmatrix},$$

where $d_1 \geq d_2$ and $\begin{pmatrix} a_1 & a_2 \end{pmatrix}^\top$, $\begin{pmatrix} b_1 & b_2 \end{pmatrix}^\top$ are eigenvalues and corresponding unit-length eigenvectors of the middle matrix on the right-hand side of the first line. Direct computation yields that $d_1 \geq 1/2 > 0 \geq d_2$ and

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \propto \begin{pmatrix} 1 - 4\gamma\alpha^2 + 4\gamma\beta^2 + \sqrt{16\gamma\beta^2 + (1 - 4\gamma)^2} \\ -8\gamma\alpha\beta \end{pmatrix}.$$

Consequently, w is a scalar multiple of

$$(22) \qquad a_1 v_{1,S} + a_2 v_\perp = \left\{ 1 + 4\gamma + \sqrt{16\gamma\beta^2 + (1 - 4\gamma)^2} \right\} v_{1,S} - 8\gamma\alpha s.$$

Since

$$\left\{ 1 + 4\gamma + \sqrt{16\gamma\beta^2 + (1 - 4\gamma)^2} \right\} \min_{j \in S} |v_{1,j}| \geq 2 \min_{j \in S} |v_{1,j}| \geq 32\gamma k^{-1/2}$$
$$> 8\gamma\alpha\|s\|_\infty,$$

we have $\mathrm{sgn}(w) = \mathrm{sgn}(v_{1,S})$. Hence by (22),

$$\min_{j=1,\dots,k} |w_j| \geq \frac{\left\{ 1 + 4\gamma + \sqrt{16\gamma\beta^2 + (1 - 4\gamma)^2} \right\} \min_{j \in S} |v_{1,j}| - 8\gamma\alpha\|s\|_\infty}{\|a_1 v_{1,S} + a_2 v_\perp\|_2}$$

$$(23) \qquad \geq \frac{(32 - 8\alpha)\gamma k^{-1/2}}{1 + 4\gamma + \sqrt{16\gamma\beta^2 + (1 - 4\gamma)^2}} \geq \frac{12\gamma k^{-1/2}}{1 + 4\gamma} \geq \frac{48}{5}\gamma k^{-1/2},$$

By (21) and (23), we have $\min_j |w_j| > \|w - \hat{v}_S\|_\infty$. So $\mathrm{sgn}(\hat{v}_S) = \mathrm{sgn}(w) = \mathrm{sgn}(v_{1,S})$ as desired.

It remains to check condition (C2). Since $\mathrm{sgn}(\hat{v}_S) = \mathrm{sgn}(v_{1,S})$, we have

$$\mathrm{tr}(\hat{\Sigma}\hat{M}) - \lambda\|\hat{M}\|_1 = \mathrm{tr}(\hat{\Sigma}_{SS}\hat{v}_S\hat{v}_S^\top) - \mathrm{tr}(\hat{U}_{SS}\hat{v}_S\hat{v}_S^\top)$$
$$= \hat{v}_S^\top(\hat{\Sigma}_{SS} - \hat{U}_{SS})\hat{v}_S = \lambda_1(\hat{\Sigma}_{SS} - \hat{U}_{SS}).$$

Moreover,

$$\hat{\Sigma} - \hat{U} = \begin{pmatrix} \hat{\Sigma}_{SS} - \hat{U}_{SS} & 0 \\ 0 & \Gamma_{p-k} \end{pmatrix}.$$

As $\lambda_1(\Gamma_{p-k}) \leq 1$ by assumption, it suffices to show that $\lambda_1(\hat{\Sigma}_{SS} - \hat{U}_{SS}) \geq 1$. By Weyl's inequality (see e.g. Horn and Johnson (2012, Theorem 4.3.1))

$$\lambda_1(\hat{\Sigma}_{SS} - \hat{U}_{SS}) \geq \lambda_1(\Sigma_{SS} - \hat{U}_{SS}) - \|\hat{\Sigma}_{SS} - \Sigma_{SS}\|_{\mathrm{op}}$$

$$(24) \qquad \geq 1 + \theta\lambda_1(v_{1,S}v_{1,S}^\top - 4\gamma s s^\top) - 2\sqrt{\frac{k \log p}{n}} \geq 1 + \frac{3\theta}{8} > 1,$$

as required.

Proof of (b). We claim first that $\hat{S} = S$. Let $\phi^* := f(\hat{M})$ be the optimal value of the semidefinite programme (5). From (24), we have $\phi^* \geq 1 + 3\theta/8$. The proof strategy here is to use dual matrices $\hat{U}$ defined in part (a) and $\hat{U}'$ to be defined below to respectively bound $\operatorname{tr}(\hat{M}^\epsilon_{S^c S^c})$ from above and bound $\operatorname{tr}(\hat{M}^\epsilon_{rr})$ from below for each $r \in S$. We then check that for the choice of ϵ we have in the theorem, the diagonal entries of $\hat{M}^\epsilon$ are above the threshold $\log p/(6n)$ precisely when they belong to the (S, S)-block of the matrix.

From (20), and using the fact that $\operatorname{tr}(AB) \leq \operatorname{tr}(A)\lambda_1(B)$ for all symmetric matrices A and B, we have

$$
\begin{aligned}
\operatorname{tr}(\hat{\Sigma}\hat{M}^\epsilon) - \lambda\|\hat{M}^\epsilon\|_1 &\leq \operatorname{tr}\big((\hat{\Sigma} - \hat{U})\hat{M}^\epsilon\big) \\
&= \operatorname{tr}\big((\hat{\Sigma}_{SS} - \hat{U}_{SS})\hat{M}^\epsilon_{SS}\big) + \operatorname{tr}\big(\Sigma_{S^c S^c}\hat{M}^\epsilon_{S^c S^c}\big) \\
&\leq \operatorname{tr}(\hat{M}^\epsilon_{SS})\phi^* + \operatorname{tr}(\hat{M}^\epsilon_{S^c S^c})\lambda_1(\Gamma_{p-k}) \\
&= \phi^* - \operatorname{tr}(\hat{M}^\epsilon_{S^c S^c})(\phi^* - 1) \leq \phi^* - 3\theta\operatorname{tr}(\hat{M}^\epsilon_{S^c S^c})/8.
\end{aligned}
$$

On the other hand, $\operatorname{tr}(\hat{\Sigma}\hat{M}^\epsilon) - \lambda\|\hat{M}^\epsilon\|_1 \geq \phi^* - \epsilon$. It follows that

$$
(25) \qquad \operatorname{tr}(\hat{M}^\epsilon_{S^c S^c}) \leq \frac{8\epsilon}{3\theta} \leq \frac{1}{6}\left(\frac{\log p}{Bn}\right)^2 < \tau.
$$

Next, fix an arbitrary $r \in S$ and define $S_0 := S \setminus \{r\}$. Define $\hat{U}'$ by

$$
\hat{U}'_{ij} := \begin{cases} \lambda\operatorname{sgn}(\hat{M}_{ij}) & \text{if } i, j \in S_0 \\ \hat{\Sigma}_{ij} - \Sigma_{ij} & \text{otherwise.} \end{cases}
$$

We note that on Ω_0, we have $\hat{U}' \in \mathcal{U}$. Again by (20),

$$
\begin{aligned}
\operatorname{tr}(\hat{\Sigma}\hat{M}^\epsilon) - \lambda\|\hat{M}^\epsilon\|_1 &\leq \operatorname{tr}\big((\hat{\Sigma} - \hat{U}')\hat{M}^\epsilon\big) \\
&= \operatorname{tr}\big((\hat{\Sigma}_{S_0 S_0} - \hat{U}_{S_0 S_0})\hat{M}^\epsilon_{S_0 S_0}\big) + \sum_{\substack{(i,j) \in S \times S \\ i=r \text{ or } j=r}} \Sigma_{ij}\hat{M}^\epsilon_{ji} + \operatorname{tr}\big(\Sigma_{S^c S^c}\hat{M}^\epsilon_{S^c S^c}\big)
\end{aligned}
$$

$$
\begin{aligned}
(26) \\
&\leq \operatorname{tr}(\hat{M}^\epsilon_{S_0 S_0})\lambda_1\big(\hat{\Sigma}_{S_0 S_0} - \hat{U}_{S_0 S_0}\big) + \sum_{\substack{(i,j) \in S \times S \\ i=r \text{ or } j=r}} \Sigma_{ij}\hat{M}^\epsilon_{ji} + \operatorname{tr}(\hat{M}^\epsilon_{S^c S^c})\lambda_1(\Gamma_{p-k}).
\end{aligned}
$$

We bound the three terms of (26) separately. By Lemma 8 in Section 1 in the online supplementary material Wang, Berthet and Samworth (2015),

$$
\lambda_1(\hat{\Sigma}_{S_0 S_0} - \hat{U}_{S_0 S_0}) \leq \lambda_1(\hat{\Sigma}_{SS} - \hat{U}_{SS}) - \big\{\lambda_1(\hat{\Sigma}_{SS} - \hat{U}_{SS}) - \lambda_2(\hat{\Sigma}_{SS} - \hat{U}_{SS})\big\}\min_{j \in S}\hat{v}_j^2.
$$

From (21) and (23),

$$\min_j |\hat{v}_j| \geq \min_j |w_j| - \|w - \hat{v}_S\|_\infty \geq 3.9\gamma k^{-1/2}.$$

Also, by Weyl's inequality,

$$\lambda_1(\hat{\Sigma}_{SS} - \hat{U}_{SS}) - \lambda_2(\hat{\Sigma}_{SS} - \hat{U}_{SS})$$
$$\geq \lambda_1(\Sigma_{SS} - \hat{U}_{SS}) - \lambda_2(\Sigma_{SS} - \hat{U}_{SS}) - 2\|\hat{\Sigma}_{SS} - \Sigma_{SS}\|_{\mathrm{op}}$$
$$\geq \theta\Big\{\lambda_1(v_{1,S}v_{1,S}^\top - \gamma ss^\top) - \lambda_2(v_{1,S}v_{1,S}^\top - \gamma ss^\top)\Big\} - 4\sqrt{\frac{k\log p}{n}}$$
$$\geq \theta(1/2 - 4\gamma k^{-1/2}) \geq \theta/4.$$

It follows that

$$(27) \qquad \lambda_1(\hat{\Sigma}_{S_0 S_0} - \hat{U}_{S_0 S_0}) \leq \phi^* - 3.8\gamma^2 k^{-1}\theta.$$

For the second term in (26), observe that

$$\sum_{\substack{(i,j)\in S\times S \\ i=r \text{ or } j=r}} \Sigma_{ij}\hat{M}_{ij}^\epsilon \leq (1 + \theta v_{1,r}^2)\hat{M}_{rr}^\epsilon + 2\sum_{i\in S, i\neq r}\theta v_{1,i} v_{1,r}\hat{M}_{i,r}^\epsilon$$

$$(28) \qquad \leq \hat{M}_{rr}^\epsilon + 2\theta|v_{1,r}|\cdot\|v_1\|_1\sqrt{\hat{M}_{rr}^\epsilon} \leq \hat{M}_{rr}^\epsilon + 2\theta\sqrt{k}\sqrt{\hat{M}_{rr}^\epsilon},$$

where the penultimate inequality uses the fact that $\hat{M}_{ir}^\epsilon \leq \sqrt{\hat{M}_{ii}^\epsilon \hat{M}_{rr}^\epsilon} \leq \sqrt{\hat{M}_{rr}^\epsilon}$ for a non-negative definite matrix $\hat{M}^\epsilon$. Substituting (27) and (28) into (26),

$$\mathrm{tr}(\hat{\Sigma}\hat{M}^\epsilon) - \lambda\|\hat{M}^\epsilon\|_1 \leq \mathrm{tr}(\hat{\Sigma}_{S_0 S_0}^\epsilon)\big(\phi^* - \tfrac{3.8\gamma^2\theta}{k}\big) + \hat{M}_{rr}^\epsilon + 2\theta\sqrt{k\hat{M}_{rr}^\epsilon} + \mathrm{tr}(\hat{M}_{S^c S^c}^\epsilon)$$
$$\leq \phi^* - 3.8\gamma^2 k^{-1}\theta\,\mathrm{tr}(\hat{M}_{S_0 S_0}^\epsilon) + 2\theta\sqrt{k\hat{M}_{rr}^\epsilon}$$
$$\leq \phi^* - 3.8\gamma^2 k^{-1}\theta\big\{1 - \mathrm{tr}(\hat{M}_{S^c S^c}^\epsilon)\big\} + 2\theta(\sqrt{k} + 1.9\gamma^2)\sqrt{\hat{M}_{rr}^\epsilon}.$$

By definition, $\mathrm{tr}(\hat{\Sigma}\hat{M}^\epsilon) - \lambda\|\hat{M}^\epsilon\|_1 \geq \phi^* - \epsilon$, so together with (25), we have

$$\sqrt{\hat{M}_{rr}^\epsilon} \geq \frac{3.8\gamma^2 k^{-1}\theta(1 - \tfrac{8\epsilon}{3\theta}) - \epsilon}{2\theta(\sqrt{k} + 1.9\gamma^2)} \geq \frac{1.9\gamma^2 k^{-1}(1 - \tfrac{8\epsilon}{3\theta})}{(\sqrt{k} + \tfrac{1.9}{256})} - \frac{\epsilon}{2\theta}$$

$$\geq 1.8\gamma^2 k^{-3/2}\Big(1 - \frac{8\epsilon}{3\theta}\Big) - \frac{\epsilon}{2\theta}$$

$$(29) \quad \geq \frac{1.8 k^{1/2}\log p}{n\theta^2}\bigg\{1 - \frac{1}{6}\Big(\frac{\log p}{Bn}\Big)^2\bigg\} - \frac{1}{32}\Big(\frac{\log p}{Bn}\Big)^2 \geq \frac{1.4\log p}{Bn} > \tau^{1/2}.$$

From (25) and (29) we conclude that $\hat{S} = S$, as claimed.

To conclude, by Yu, Wang and Samworth (2015, Theorem 2), on Ω_0,

$$L(\hat{v}^{\mathrm{MSDP}}, v_1) = L(\hat{v}_S^{\mathrm{MSDP}}, v_{1,S}) \leq \frac{2\|\hat{\Sigma}_{SS} - \Sigma_{SS}\|_{\mathrm{op}}}{\lambda_1(\Sigma_{SS}) - \lambda_2(\Sigma_{SS})} \leq 4\sqrt{\frac{k \log p}{n\theta^2}},$$

where we used (19) and Lemma 5 in the online supplementary material Wang, Berthet and Samworth (2015) in the final bound.

For the final part of the theorem, when $p \geq \theta\sqrt{n/k}$,

$$\sup_{P \in \tilde{\mathcal{P}}_p(n,k,\theta)} \mathbb{E}_p\big\{L(\hat{v}^{\mathrm{MSDP}}, v_1)\big\} \leq 4\sqrt{\frac{k \log p}{n\theta^2}} + \mathbb{P}(\Omega_0^c)$$

$$\leq 4\sqrt{\frac{k \log p}{n\theta^2}} + 2p^{-1} \leq 6\sqrt{\frac{k \log p}{n\theta^2}},$$

as desired.   □

ACKNOWLEDGEMENTS

SUPPLEMENTARY MATERIAL

Supplement A: Supplementary material to 'Statistical and computational trade-offs in estimation of sparse principal components' (). Ancillary results and a brief introduction to computational complexity theory.

REFERENCES

Alon, N., Andoni, A., Kaufman, T., Matulef, K., Rubinfeld, R., and Xie, N. (2007) Testing k-wise and almost k-wise independence. *Proceedings of the thirty-ninth ACM Symposium on Theory of Computing*, 496–505.

Alon, N., Krivelevich, M. and Sudakov, B. (1998) Finding a large hidden clique in a random graph. *Proceedings of the ninth annual ACM-SIAM Symposium on Discrete Algorithms*, 594–598.

Ames, B. P. W. and Vavasis, S. A. (2011) Nuclear norm minimization for the planted clique and biclique problems. *Math. Program.*, **129**, 69–89.

Amini, A. A. and Wainwright, M. J. (2009) High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.*, **37**, 2877–2921.

Applebaum, B., Barak, B. and Wigderson, A. (2010) Public-key cryptography from different assumptions. *Proceedings of the forty-second ACM Symposium on Theory of Computing*, 171–180.

Bach, F., Ahipaşaoğlu, S. D., d'Aspremont, A. (2010) Convex relaxations for subset selection. Available at `http://arxiv.org/abs/1006.3601`.

Baik, J., Ben Arous, G. and Péché, S. (2005) Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, **33**, 1643–1697.

Berthet, Q. (2015) Optimal testing for planted satisfiability problems. *Electron. J. Statist.*, **9**, 298–317.

Berthet, Q. and Rigollet P. (2013) Optimal detection of sparse principal components in high dimension. *Ann. Statist.*, **41**, 1780–1815.

Berthet, Q. and Rigollet P. (2013) Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res. W&CP*, **30**, 1046–1066.

Birnbaum, A., Johnstone , I. M., Nadler , B. and Paul, D. (2013) Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.*, **41**, 1055–1084.

Cai, T. T., Ma, Z. and Wu, Y. (2013) Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.*, **41**, 3074–3110.

Chan, Y.-b. and Hall, P. (2010) Using evidence of mixed populations to select variables for clustering very high-dimensional data. *J. Amer. Statist. Assoc.*, **105**, 798–809.

Chandrasekaran, V. and Jordan, M. I. (2013) Computational and statistical tradeoffs via convex relaxation. *Proc. Nat. Acad. Sci.*, **110**, E1181–E1190.

Chen, Y. and Xu, J. (2014) Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. Available at `http://arxiv.org/abs/1402.1267`.

Chun, H. and Sündüz, K. (2009) Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*, **182**, 79–90.

d'Aspremont, A. El Ghaoui, L., Jordan, M. I. and Lanckriet, G. R. G. (2007) A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.*, **49**, 434–448.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a pertubation. III. *SIAM J. Numer. Anal.* **7**, 1–46.

Deshpande, Y. and Montanari, A. (2014) Sparse PCA via covariance thresholding. Available at `http://arxiv.org/abs/1311.5179`.

Diaconis, P. and Freedman D. (1980) Finite exchangeable sequences. *Ann. Probab.*, **8**, 745–764.

Feldman, V., Grigorescu, E., Reyzin, L., Vempala, S. S. and Xiao, Y. (2013) Statistical algorithms and a lower bound for detecting planted cliques. *Proceedings of the forty-fifth annual ACM Symposium on Theory of Computing*, 655–664.

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

Feige, U. and Krauthgamer, R. (2003) The probable value of the Lovàsz–Schrijver relaxations for a maximum independent set. *SIAM J. Comput.*, **32**, 345–370.

Feige, U. and Ron, D. (2010) Finding hidden cliques in linear time. *Discrete Math. Theor. Comput. Sci. Proc.*, 189–204.

Feldman, V., Perkins, W. and Vempala, S. (2015) On the Complexity of Random Satisfiability Problems with Planted Solutions. *Proceedings of the forty-seventh annual ACM Symposium on Theory of Computing*, to appear.

Gao, C., Ma, Z. and Zhou, H. H. (2014) Sparse CCA: Adaptive estimation and computational barriers. Available at `http://arxiv.org/abs/1409.8565`.

Allen, G. I., and Maletić-Savatić, M. (2011) Sparse non-negative generalized PCA with
    applications to metabolomics. *Bioinformatics* **27**, 3029–3035.
Golub, G. H. and Van Loan, C. F. (1996) *Matrix Computations*. The Johns Hopkins
    University Press, Baltimore, Maryland.
Grimmett, G. R. and McDiarmid C. J. H. (1975) On colouring random graphs. *Math.
    Proc. Cambridge Philos. Soc.*, **77**, 313–324.
Hajek, B., Wu, Y. and Xu, J. (2014) Computational lower bounds for community detection
    on random graphs. Available at `http://arxiv.org/abs/1406.6625`.
Hazan, E. and Krauthgamer, R. (2011) How hard is it to approximate the best Nash
    equilibrium? *SIAM J. Comput.*, **40**, 79–91.
Horn, R. A. and Johnson, C. R. (2012) *Matrix Analysis*. Cambridge University Press.
Jerrum, M. (1992) Large cliques elude the Metropolis process. *Random Structures Algo-
    rithms*, **3**, 347–359.
Johnstone, I. M. and Lu, A. Y. (2009) On consistency and sparsity for principal compo-
    nents analysis in high dimensions. *J. Amer. Statist. Assoc.*, **104**, 682–693.
Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003) A modified principal component
    technique based on the LASSO, *J. Comput. Graph. Statist.*, **12**, 531–547.
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.
Juels, A. and Peinado, M. (2000) Hiding cliques for cryptographic security. *Des. Codes
    Cryptogr.*, **20**, 269–280.
Karp, R. M. (1972) Reducibility among combinatorial problems. In R. E. Miller et al.
    (Eds.), *Complexity of Computer Computations*, 85–103. Springer, New York.
Kučera, L. (1995) Expected complexity of graph partitioning problems. *Discrete Appl.
    Math.*, **57**, 193–212.
Lanczos, C. (1950) An iteration method for the solution of the eigenvalue problem of linear
    differential and integral operators. *J. Res. Nat. Bur. Stand.*, **45**, 255–282.
Laurent, B. and Massart, P. (2000) Adaptive estimation of a quadratic functional by model
    selection. *Ann. Statist.*, **28**, 1302–1338.
Ma, Z. (2013) Sparse principal component analysis and iterative thresholding. *Ann.
    Statist.*, **41**, 772–801.
Ma, Z. and Wu, Y. (2015) Computational barriers in minimax submatrix detection. *Ann.
    Statist.*, to appear.
Majumdar, A. (2009) Image compression by sparse PCA coding in curvelet domain. *Signal,
    image and video processing*, **3**, 27–34.
Naikal, N., Yang A. Y. and Sastry S. S. (2011) Informative feature selection for object
    recognition via sparse PCA. *Computer Vision (ICCV), 2011 IEEE International Con-
    ference*, 818–825.
Nemirovski, A. (2004) Prox-method with rate of convergence $O(1/t)$ for variational in-
    equalities with Lipschitz continuous monotone operators and smooth convex-concave
    saddle point problems. *SIAM J. Optim.* **15**, 229–251.
Nesterov, Yu. (2005) Smooth minimization of nonsmooth functions. *Math. Program., Ser.
    A*, **103**, 127–152.
Parkhomenko, E., Tritchler, D., and Beyene, J. (2009) Sparse canonical correlation analysis
    with application to genomic data integration. *Stat. Appl. Genet. Mol. Biol.*, **8**, 1–34.
Paul, D. (2007) Asymptotics of sample eigenstructure for a large dimensional spiked co-
    variance model. *Statist. Sinica*, **17**, 1617–1642.
Shen, D., Shen, H. and Marron, J. S. (2013) Consistency of sparse PCA in high dimension,
    low sample size contexts. *J. Multivariate Anal.*, **115**, 317–333.
Shorack, G. R. and Wellner, J. A. (1986) *Empirical Processes with Applications to Statis-*

tics. Wiley, New York.

Tan, K. M., Petersen, A. and Witten, D. (2014) Classification of RNA-seq Data. In S. Datta and D. Nettleton (Eds.) *Statistical Analysis of Next Generation Sequencing Data*, 219–246. Springer, New York.

van de Geer, S. (2000) *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes*. Springer, New York.

Vu, V. Q., Cho, J., Lei, J. and Rohe, K. (2013) Fantope Projection and Selection: A near-optimal convex relaxation of sparse PCA. *Advances in Neural Information Processing Systems (NIPS)*, **26**, 2670–2678.

Vu, V. Q. and Lei, J. (2013) Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.*, **41**, 2905–2947.

Wang, D., Lu, H.-C. and Yang, M.-H. (2013) Online object tracking with sparse prototypes. *IEEE transactions on image processing*, **22**, 314–325.

Wang, T., Berthet, Q. and Samworth, R. J. (2015) Supplementary material to 'Statistical and computational trade-offs in estimation of sparse principal components'. *Online supplement*.

Wang, Z., Lu, H. and Liu, H. (2014) Tighten after Relax: Minimax-Optimal Sparse PCA in Polynomial Time. *Advances in Neural Information Processing Systems (NIPS)*, **27**, 3383–3391.

Witten, D. M., Tibshirani, R. and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.

Yu, Y., Wang, T. and Samworth, R. J. (2015) A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, **102**, 315–323.

Yuan, X.-T. and Zhang, T. (2013) Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.*, **14**, 899–925.

Zhang, Y., Wainwright, M. J. and Jordan, M. I. (2014) Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *J. Mach. Learn. Res. W&CP*, **35**, 921–948.

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal components analysis. *J. Comput. Graph. Statist.*, **15**, 265–86.

STATISTICAL LABORATORY
WILBERFORCE ROAD
CAMBRIDGE, CB3 0WB
UNITED KINGDOM
E-MAIL: r.samworth@statslab.cam.ac.uk
E-MAIL: t.wang@statslab.cam.ac.uk
E-MAIL: q.berthet@statslab.cam.ac.uk
URL: http://www.statslab.cam.ac.uk/~rjs57
URL: http://www.statslab.cam.ac.uk/~tw389
URL: http://www.statslab.cam.ac.uk/~qb204