Submitted to the Annals of Statistics

STATISTICAL AND COMPUTATIONAL LIMITS FOR SPARSE MATRIX DETECTION

By T. Tony Cai[¶,**] and Yihong Wu[‖,††]

University of Pennsylvania[**] and Yale University[††]

This paper investigates the fundamental limits for detecting a high-dimensional sparse matrix contaminated by white Gaussian noise from both the statistical and computational perspectives. We consider $p \times p$ matrices whose rows and columns are individually k-sparse. We provide a tight characterization of the statistical and computational limits for sparse matrix detection, which precisely describe when achieving optimal detection is easy, hard, or impossible, respectively. Although the sparse matrices considered in this paper have no apparent submatrix structure and the corresponding estimation problem has no computational issue at all, the detection problem has a surprising computational barrier when the sparsity level k exceeds the cubic root of the matrix size p: attaining the optimal detection boundary is computationally at least as hard as solving the planted clique problem.

The same statistical and computational limits also hold in the sparse covariance matrix model, where each variable is correlated with at most k others. A key step in the construction of the statistically optimal test is a structural property of sparse matrices, which can be of independent interest.

1. Introduction. The problem of detecting sparse signals arises frequently in a wide range of fields and has been particularly well studied in the Gaussian sequence setting (cf. the monograph [40]). For example, detection of unstructured sparse signals under the Gaussian mixture model was studied in [27, 39] for the homoskedastic case and in [15] for the heteroscedastic case, where sharp detection boundaries were obtained and adaptive detection procedures proposed. Optimal detection of structured signals in the Gaussian noise model has also been investigated in [6, 7, 21]. One common feature of these vector detection problems is that the optimal statistical

[¶]The research of T.T. Cai was supported in part by NSF Grant DMS-1712735 and NIH grants R01-GM129781 and R01-GM123056.

[‖]The research of Y. Wu was supported in part by the NSF Grant IIS-1447879, CCF-1527105, an NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship.

AMS 2000 subject classifications: Primary 62H15; secondary 62C20

Keywords and phrases: Minimax rates, computational limits, sparse covariance matrix, sparse detection.

performance can always be achieved by computationally efficient procedures such as thresholding or convex optimization.

Driven by contemporary applications, much recent attention has been devoted to inference for high-dimensional matrices, including covariance matrix estimation, principal component analysis (PCA), image denoising, and multi-task learning, all of which rely on detecting or estimating high-dimensional matrices with low-dimensional structures such as low-rankness or sparsity. For a suite of matrix problems, including sparse PCA [10, 54], biclustering [9, 16, 47], sparse canonical correlation analysis (CCA) [32] and community detection [34], a new phenomenon known as *computational barriers* has been recently discovered, which shows that in certain regimes attaining the statistical optimum is computationally intractable, unless the planted clique problem can be solved efficiently.[1] In a nutshell, the source of computational difficulty in the aforementioned problems is their *submatrix sparsity*, where the signal of interests is concentrated on a submatrix within a large noisy matrix. This combinatorial structure provides a direct connection to, and allows these matrix problems to be reduced in polynomial time from, the planted clique problem, thereby creating computational gaps for not only the detection but also support recovery and estimation.

In contrast, another sparsity structure for matrices postulates the rows and columns are individually sparse, which has been well studied in covariance matrix estimation [13, 22, 28, 42]. The motivation is that in many real-data applications each variable is only correlated with a few others. Consequently, each row and each column of the covariance matrix are individually sparse but, unlike sparse PCA, biclustering, or group-sparse regression, their support sets need not be aligned. Therefore this sparsity model does not postulate any submatrix structure of the signal; indeed, it has been shown for covariance matrix estimation that entrywise thresholding of the sample covariance matrix proposed in [13] attains the minimax estimation rate [22].

The focus of the present paper is to understand the fundamental limits of detecting sparse matrices from both the statistical and computational perspectives. While achieving the optimal estimation rate does not suffer from any computational barrier, it turns out the detection counterpart does when and only when the sparsity level exceeds the *cubic root* of the matrix size. This is perhaps surprising because the sparsity model itself does not

[1]The planted clique problem [2] refers to detecting or locating a clique of size $o(\sqrt{n})$ planted in the Erdös-Rényi random graph $G(n, 1/2)$. Conjectured to be computationally intractable [30, 41], this problem has been frequently used as a basis for quantifying hardness of average-case problems [1, 37].

explicitly enforce any submatrix structure, which has been responsible for problems such as sparse PCA to be reducible from the planted clique. Our main result is a tight characterization of the statistical and computational limits of detecting sparse matrices in both the Gaussian noise model and the covariance matrix model, which precisely describe when achieving optimal detection is easy, hard, and impossible, respectively.

1.1. *Setup.* We start by formally defining the sparse matrix model:

DEFINITION 1. We say a $p \times p$ matrix M is *k-sparse* if all of its rows and columns are k-sparse vectors, i.e., with no more than k non-zeros. Formally, denote the i^{th} row of M by M_{i*} and the i^{th} column by M_{*i}. The following parameter set

$$\mathcal{M}(p,k) = \{M \in \mathbb{R}^{p\times p} : \|M_{i*}\|_0 \le k, \|M_{*i}\|_0 \le k, \forall i \in [p]\}. \tag{1.1}$$

denotes the collection of all k-spares $p\times p$ matrices, where $\|x\|_0 \triangleq \sum_{i\in[p]} \mathbf{1}\{x_i \neq 0\}$ for $x \in \mathbb{R}^p$.

Consider the following "signal + noise" model, where we observe a sparse matrix contaminated with Gaussian noise:

$$X = M + Z \tag{1.2}$$

where M is a $p \times p$ unknown mean matrix, and Z consists of i.i.d. entries normally distributed as $N(0,\sigma^2)$. Without loss of generality, we shall assume that $\sigma = 1$ throughout the paper.

Given the noisy observation X, the goal is to test whether the mean matrix is zero or a k-sparse nonzero matrix, measured in the spectral norm. Formally, we consider the following hypothesis testing problem:

$$H_0 : M = 0 \quad \text{versus} \quad H_1 : M \in \Theta(p,k,\lambda), \tag{1.3}$$

where the mean matrix M belongs to the parameter space

$$\Theta(p,k,\lambda) = \{M \in \mathbb{R}^{p\times p} : M \in \mathcal{M}(p,k), \|M\|_2 \ge \lambda\}. \tag{1.4}$$

Here we use the spectral norm $\|\cdot\|_2$, namely, the largest singular value, to measure the signal strength under the alternative hypothesis. It turns out that if we use the Frobenius norm to define the alternative hypothesis, the sparsity structure does not help detection, in the sense that, the minimal λ required to detect 1-sparse matrices is within a constant factor of that in the

non-sparse case and the matrix problem collapses to the vectorized version (see the supplementary material [20, Section 7] for details).

For covariance model, the counterpart of the detection problem (1.4) is the following. Consider the Gaussian covariance model, where we observe n independent samples drawn from the p-variate normal distribution $N(0, \Sigma)$ with an unknown covariance matrix Σ. In the sparse covariance matrix model, each coordinate is correlated with at most k others. Therefore each row of the covariance matrix Σ has at most k non-zero off-diagonal entries. This motivates the following detection problem:

$$H_0 : \Sigma = \mathbf{I} \quad \text{versus} \quad H_1 : \|\Sigma - \mathbf{I}\|_2 \geq \lambda,\ \Sigma - \mathbf{I} \text{ is } k\text{-sparse}. \tag{1.5}$$

In the context of covariance matrix model, it is natural to use the spectral norm [13, 22] since it measures the strength of the leading principal component. Under the null hypothesis, the samples are pure noise; under the alternative, there exists at least one significant factor and the entire covariance matrix is k-sparse. The goal is to determine the smallest λ so that the factor can be detected from the samples.

1.2. *Statistical and computational limits.* For ease of exposition, let us focus on the additive Gaussian noise model and consider the following asymptotic regime, wherein the sparsity and the signal level grow polynomially in the dimension as follows:

$$k = p^{\alpha} \quad \text{and} \quad \lambda = p^{\beta}$$

with $\alpha \in [0, 1]$ and $\beta > 0$ held fixed and $p \to \infty$. Theorem 1 in Section 2 implies that the critical exponent of λ behaves according to the following piecewise linear function:

$$\beta^* = \begin{cases} \alpha & \alpha \leq \frac{1}{3} \\ \frac{1+\alpha}{4} & \alpha \geq \frac{1}{3} \end{cases}$$

in the sense that if $\beta > \beta^*$, there exists a test that achieves vanishing probability of error of detection uniformly over all k-sparse matrices; conversely, if $\beta < \beta^*$, no test can outperform random guessing asymptotically.

More precisely, as shown in Figure 1, the phase diagram of α versus β is divided into four regimes:

(I) $\beta > \alpha$: The test based on the largest singular value of the entrywise thresholding estimator succeeds. In particular, we reject if $\|X^{\mathrm{Th}}\|_2 \gtrsim k\sqrt{\log p}$, where $X^{\mathrm{Th}}_{ij} = X_{ij}\mathbf{1}\{|X_{ij}| = \Omega(\sqrt{\log p})\}$.

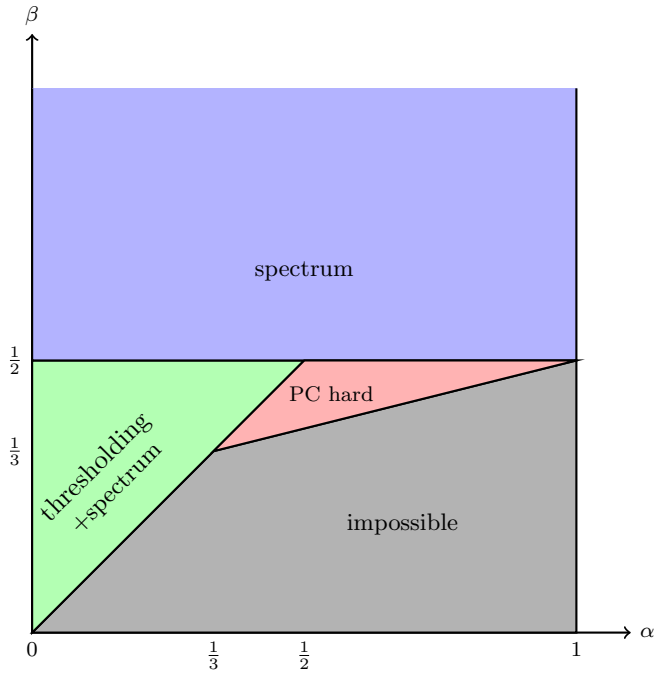


FIG 1. *Statistical and computational limits in detecting sparse matrices.*

(II) $\beta > \frac{1}{2}$: The test based on the large singular value of the direct observation succeeds. In particular, we reject if $\|X\|_2 \gtrsim \sqrt{p}$.
(III) $\frac{1+\alpha}{4} < \beta < \alpha \wedge \frac{1}{2}$: detection is as hard as solving the planted clique problem.
(IV) $\beta < \alpha \wedge \frac{1+\alpha}{4}$: detection is information-theoretically impossible.

As mentioned earlier, the computational intractability in detecting sparse matrices is perhaps surprising because

(a) achieving the optimal estimation rate does not present any computational difficulty;
(b) unlike problems such as sparse PCA, the sparse matrix model in Definition 1 does not explicitly impose any submatrix sparsity pattern as the rows are individually sparse and need not share a common support.

The result in Figure 1 shows that in the moderately sparse regime of $p^{1/3} \ll k \ll p$, outperforming entrywise thresholding is at least as hard as solving planted clique. However, it is possible to improve over entrywise thresholding using computationally inefficient tests. Next we briefly describe the construction of the optimal test which detects the signal when $\lambda \gtrsim$

$(kp\log p)^{1/4}$ and improves over entrywise thresholding which requires $\lambda \gtrsim k\sqrt{\log p}$. This test has two stages: The first stage is a standard χ^2-test, which rejects the null hypothesis if the mean matrix M has a large Frobenius norm, i.e., $\|M\|_{\mathrm{F}} \gtrsim \sqrt{p}$. Under the alternative hypothesis that $\|M\|_2 \geq \lambda$, if the data can survive the χ^2-test, i.e. $\|M\|_{\mathrm{F}} \lesssim \sqrt{p}$, then M has small stable rank, i.e., $\mathsf{sr}(M) \triangleq \|M\|_{\mathrm{F}}^2/\|M\|_2^2 \lesssim r \triangleq \frac{p}{\lambda^2}$. The key observation is that if a matrix M is both k-sparse (in the sense of Definition 1) and has stable rank at most r, then its operator norm is concentrated on a small submatrix, in the sense that there exists subsets I, J of cardinality $m \asymp kr\log p$ such that $\|M_{IJ}\|_2 \geq \|M\|_2$ for some constant c. Thus, in the second stage we apply a scan test that rejects if the maximum spectral norm among all $m \times m$ submatrices exceeds a constant multiple of λ. This succeeds provided that $\lambda \geq \sqrt{m}$, which leads to the condition that $\lambda \gtrsim (kp\log p)^{1/4}$. We emphasize that although scan test is a well-known idea, here both the specification (that is, what class to search over as well as how large the submatrices need to be) and the proof of correctness are new. The crucial difference is the following: scan statistics typically correspond to generalized likelihood ratio test and are defined by maximizing over the corresponding hypothesis class. For instance, in the problem of submatrix detection [14], where the alternative hypothesis is that the mean matrix has a $k \times k$ submatrix with elevated means, the scan test, naturally, searches over all possible $k \times k$ submatrices. In contrast, the scan test described above does not search over all possible support sets of k-sparse matrices (in the sense of both row and column sparsity), but over submatrices of size $m \asymp \frac{kp\log p}{\lambda^2}$, which far exceeds the row/column-wise sparsity k. Such a choice follows from the aforementioned structural property of matrices that are both sparse and of low stable rank (see Theorem 3 for details).

The crucial structural property of sparse matrices used above is established using a celebrated result of Rudelson and Vershynin [51] in randomized numerical linear algebra which shows that the Gram matrix of any matrix M of low stable rank can be approximated by that of a small submatrix of M. The existence of such small submatrix is shown by means of probabilistic method but does not provide a constructive method to find it, which, as Figure 1 suggests, is likely to be computationally intractable. We mention that the computational hardness in the hard regime follows straightforwardly from the corresponding results from submatrix detection and sparse PCA, although the sparsity is in fact much bigger than the row-wise sparsity k. The underlying reason is that in the moderately sparse case ($k = \Omega(p^{1/3})$), the least-favorable (up to constant) prior has a submatrix structure of size approximately $\sqrt{kp}$.

To conclude this part, we note that, the same statistical and computational limits in Figure 1 also apply to detecting sparse covariance matrices when λ is replaced by $\lambda\sqrt{n}$, under appropriate assumptions on the sample size; see Section 6 for details.

1.3. *Related work.* As opposed to the vector case, there exist various notions of sparsity for matrices as motivated by specific applications, including

- Vector sparsity: the total number of nonzeros in the matrix is constrained [23], e.g., in robust PCA.
- Row sparsity: each row of the matrix is sparse, e.g. matrix denoising [43].
- Group sparsity: each row of the matrix is sparse and shares a common support, e.g., group-sparse regression [46].
- Submatrix sparsity: the matrix is zero except for a small submatrix, e.g., sparse PCA [11, 18], biclustering [9, 14, 47, 54], sparse SVD [55], sparse CCA [32], and community detection [35].

The sparse matrix model (Definition 1) studied in this paper is stronger than the vector or row sparsity and weaker than submatrix sparsity.

The statistical and computational aspects of detecting matrices with submatrix sparsity has been investigated in the literature for the Gaussian mean, covariance and the Bernoulli models. In particular, for the spiked covariance model where the leading singular vector is assumed to be sparse, the optimal detection rate has been obtained in [11, 19]. Detecting submatrices in additive Gaussian noise was studied by Butucea and Ingster [14] who not only found the optimal rate but also determined the sharp constants. In the random graph (Bernoulli) setting, the problem of detecting the presence of a small denser community planted in an Erdös-Rényi graph was studied in [8]; here the entry of the mean adjacency matrix is p on a small submatrix and $q < p$ everywhere else. The computational lower bounds in all three models were established in [10, 35, 47] by means of reduction to the planted clique problem.

Another work that is closely related to the present paper is [3, 4], where the goal is to detect covariance matrices with sparse correlation. Specifically, in the n-sample Gaussian covariance model, the null hypothesis is the identity covariance matrix and the alternative hypothesis consists of covariances matrices whose off-diagonals are equal to a positive constant on a submatrix and zero otherwise. Assuming various combinatorial structure of the support set, the optimal tradeoff between the sample size, dimension, sparsity and the correlation level has been studied. One can apply the results from [4] in the special case of k-subsets to yield a lower bound for testing sparse covari-

ance matrices, which turns out to be highly suboptimal; see Section 2 for a detailed comparison. Other work on testing high-dimensional covariance matrices that do not assume sparse alternatives include testing independence and sphericity, with specific focus on asymptotic power analysis and the limiting distribution of test statistics [17, 24, 49, 50]. Finally, we mention yet another two-dimensional detection problem in Gaussian noise [5], where the sparse alternative corresponds to paths in a large graph.

1.4. *Notation and organization.* We introduce the main notation used in this paper: For any sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \gtrsim b_n$ if $a_n \geq cb_n$ holds for all n and some absolute constant $c > 0$, $a_n \lesssim b_n$ if $a_n \gtrsim b_n$, and $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. In addition, we use $\asymp_k$ to indicate that the constant depends only on k.

For any $q \in [1, \infty]$, the $\ell_q \to \ell_q$ induced operator norm of an matrix M is defined as $\|M\|_q \triangleq \max_{\|x\|_{\ell_q} \leq 1} \|Mx\|_{\ell_q}$. In particular, $\|M\|_2$ is the spectral norm, i.e., the largest singular value of M, and $\|M\|_1$ (resp. $\|M\|_\infty$) is the largest ℓ_1-norm of the columns (resp. rows) of M. For any $p \times p$ matrix M and $I, J \subset [p]$, let M_{IJ} denote the submatrix $(M_{ij})_{i \in I, j \in J}$. Let $\mathbf{I}$ and $\mathbf{J}$ denote the identity and the all-one matrix. Let $\mathbf{1}$ denote the all-one vector. Let S_p denotes the set of $p \times p$ positive-semidefinite matrices.

The rest of the paper is organized as follows: Section 2 presents the main results of the paper in terms of the minimax detection rates for both the Gaussian noise model and the covariance matrix model. Minimax upper bounds together with the testing procedures for the mean model are presented in Section 3, shown optimal by the lower bounds in Section 4; in particular, Section 3.1 introduces a structural property of sparse matrices which underpins the optimal tests in the moderately sparse regime. Results for the covariance model are given in Section 5 together with additional proofs. Section 6 discusses the computational aspects and explains how to deduce the computational limit in Figure 1 from that of submatrix detection and sparse PCA.

2. Main results. We begin with the Gaussian noise model. To quantify the fundamental limit of the hypothesis testing problem (1.3), we define $\epsilon^*(p, k, \lambda)$ as the optimal sum of Type-I and Type-II probability of error:

$$\epsilon^*(p, k, \lambda) = \inf_{\phi} \left\{ \mathbb{P}_0(\phi = 1) + \sup_{M \in \Theta(p,k,\lambda)} \mathbb{P}_M(\phi = 0) \right\} \tag{2.1}$$

where $\mathbb{P}_M$ denotes the distribution of the observation $X = M + Z$ conditioned on the mean matrix M, and the infimum is taken over all decision

rules $\phi : \mathbb{R}^{p\times p} \to \{0,1\}$.

Our main result is a tight characterization of the optimal detection threshold for λ. Define the following upper bound

$$\lambda_1(k,p) \triangleq \begin{cases} k\sqrt{\log p} & k \le (\frac{p}{\log p})^{\frac{1}{3}} \\ \left(kp\log\frac{ep}{k}\right)^{\frac{1}{4}} & k \ge (\frac{p}{\log p})^{\frac{1}{3}} \end{cases} \tag{2.2}$$

and the lower bound

$$\lambda_0(k,p) \triangleq \begin{cases} k\sqrt{\log\left(\frac{p\log p}{k^3}\right)} & k \le (p\log p)^{\frac{1}{3}} \\ \left(kp\log\frac{ep}{k}\right)^{\frac{1}{4}} & k \ge (p\log p)^{\frac{1}{3}} \end{cases}. \tag{2.3}$$

It can be verified that (2.2) and (2.3) differ by at most a factor of $O\left(\sqrt{\frac{\log p}{\log\log p}}\right)$.

THEOREM 1 (Gaussian noise model). *There exists absolute constant k_0, c_0, c_1, such that the following holds for all $k_0 \le k \le p$:*

1. *For any $c > c_1$, if*

$$\lambda \ge c\lambda_1(k,p), \tag{2.4}$$

then $\epsilon^*(k,p,\lambda) \le \epsilon_1(c)$, where $\epsilon_1(c) \to 0$ as $c \to \infty$.

2. *Conversely, for any $c > c_0$, if*

$$\lambda \le c\lambda_0(p,k), \tag{2.5}$$

then $\epsilon^*(k,p,\lambda) \ge \epsilon_0(c) - o_{p\to\infty}(1)$, where $\epsilon_0(c) \to 1$ as $c \to 0$.

To parse the result of Theorem 1, let us denote by $\lambda^*(p,k)$ the optimal detection threshold, i.e., the minimal value of λ so that the optimal probability of error $\epsilon^*(p,k,\lambda)$ is at most a constant, say, 0.1. Then we have the following characterization:

- High sparsity: $k \le p^{1/3-\delta}$:

$$\lambda^* \asymp_\delta k\sqrt{\log p}$$

- Moderate sparsity: $k \gtrsim (p\log p)^{1/3}$:

$$\lambda^* \asymp \left(kp\log\frac{ep}{k}\right)^{\frac{1}{4}}$$

- Boundary case: $(\frac{p}{\log p})^{1/3} \lesssim k \lesssim (p \log p)^{1/3}$:

$$k\sqrt{\log \frac{ep \log p}{k^3}} \lesssim \lambda^* \lesssim \left(kp \log \frac{ep}{k}\right)^{\frac{1}{4}},$$

  where the upper and lower bounds are within a factor of $O\left(\sqrt{\frac{\log p}{\log \log p}}\right)$.

Furthermore, two generalizations of Theorem 1 will be evident from the proof: (a) the upper bound in Theorem 1 as well as the corresponding optimal tests apply as long as the noise matrix consists of independent entries with subgaussian distribution with constant proxy variance; (b) the lower bound in Theorem 1 continues to hold up even if the mean matrix is constrained to be symmetric. Thus, symmetry does not improve the minimax detection rate.

Next we turn to the sparse covariance model: Given n independent samples drawn from $N(0, \Sigma)$, the goal is to test the following hypothesis

$$H_0 : \Sigma = \mathbf{I} \quad \text{versus} \quad H_1 : \Sigma \in \Xi(p, k, \lambda, \tau), \tag{2.6}$$

where the parameter space for sparse covariances matrices is

$$\Xi(p, k, \lambda, \tau) = \{\Sigma \in \mathsf{S}_p : \Sigma \in \mathcal{M}(p, k), \|\Sigma - \mathbf{I}\|_2 \geq \lambda, \|\Sigma\| \leq \tau\}. \tag{2.7}$$

In other words, under the alternative, the covariance is equal to identity plus a sparse perturbation. Throughout the paper, the parameter τ is assumed to be a constant.

Define the minimax probability of error as:

$$\epsilon_n^*(p, k, \lambda) = \inf_{\phi} \left\{ \mathbb{P}_{\mathbf{I}}(\phi = 1) + \sup_{\Sigma \in \Xi(p,k,\lambda,\tau)} \mathbb{P}_{\Sigma}(\phi = 0) \right\} \tag{2.8}$$

where $\phi \in \{0, 1\}$ is a function of the samples $(X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$.

Analogous to Theorem 1, the next result characterizes the optimal detection threshold for sparse covariance matrices.

THEOREM 2 (Covariance model). *There exists absolute constants k_0, C, c_0, c_1, such that the following holds for all $k_0 \leq k \leq p$.*

1. *Assume that $n \geq C \log p$. For any $c > c_1$, if*

$$\lambda \geq \frac{c}{\sqrt{n}} \lambda_1(k, p), \tag{2.9}$$

   then $\epsilon_n^*(k, p, \lambda) \leq \epsilon_1(c)$, where $\epsilon_1(c) \to 0$ as $c \to \infty$.

2. *Assume that*

$$n \geq C\lambda_0(p,k)^2 \log p \tag{2.10}$$

and

$$n \geq C \cdot \begin{cases} \frac{k^6}{p}\left(\frac{p}{k^3}\right)^{2\delta} \log^2 p & k \leq p^{1/3} \\ p & k \geq p^{1/3} \end{cases}, \tag{2.11}$$

where δ *is any constant in* $(0, \frac{2}{3}]$. *If*

$$\lambda \leq \frac{c}{\sqrt{n}}\lambda_0(k,p), \tag{2.12}$$

then $\epsilon^*(k,p,\lambda) \geq \epsilon_0(c) - o_{p\to\infty}(1)$, *where* $\epsilon_0(c) \to 1$ *as* $c \to 0$

In comparison with Theorem 1, we note that the rate-optimal lower bound in Theorem 2 holds under the assumption that the sample size is sufficiently large. In particular, the condition (2.10) is very mild because, by the assumption that $\|\Sigma\|_2$ is at most a constant, in order for the right hand side of (2.12) to be bounded, it is necessary to have $n \geq \lambda_0(p,k)^2$. The extra assumption (2.11), when $k \geq p^{1/4}$, does impose a non-trivial constraint on the sample size. This assumption is due to the current lower bound technique based on the χ^2-divergence. In fact, the lower bound in [17] for testing covariance matrix without sparsity uses the same method and also requires $n \gtrsim p$.

The results of Theorems 1 and 2 also demonstrate the phenomenon of the *separation of detection and estimation*, which is well-known in the Gaussian sequence model. The minimax estimation of sparse matrices has been systematically studied by Cai and Zhou [22] in the covariance model, where it is shown that entrywise thresholding achieves the minimax rate in the spectral norm loss of $k\sqrt{\frac{\log p}{n}}$ provided that $n \gtrsim k^2 \log^3 p$ and $\log n \lesssim \log p$; similar rate of $k\sqrt{\log p}$ also holds for the Gaussian noise model. In view of this result, an interesting question is whether a "plug-in" approach for testing, namely, using the spectral norm of the minimax estimator as the test statistic, achieves the optimal detection rate. This method is indeed optimal in the very sparse regime of $k \ll p^{1/3}$, but fails to achieve the optimal detection rate in the moderately sparse regime of $k \gg p^{1/3}$, which, in turn, can be attained by a computationally intensive test procedure. This observation should be also contrasted with the behavior in the vector case. To detect the presence of a k-sparse p-dimensional vector in Gaussian noise, entrywise thresholding, which is the optimal estimator for all sparsity levels, achieves

the minimax detection rate in ℓ_2-norm when $k \ll \sqrt{p}$, while the χ^2-test, which disregards sparsity, is optimal when $k \gg \sqrt{p}$.

It is instructive to compare Theorem 2 with the results in [3, 4], who considered the following type of hypotheses testing problem with observation $(X_1, \ldots, X_n) = \Sigma$:

$$H_0 : \Sigma = \mathbf{I}, \quad \text{versus} \quad H_1 : \Sigma = (1-\rho)\mathbf{I} + \rho \mathbf{1}_S \mathbf{1}_S^\top, \text{ for some } S \in \mathcal{C} \tag{2.13}$$

where $\rho > 0$, $\mathcal{C}$ is a collection of subsets of $[p]$, and $\mathbf{1}_S$ is the indicator vector of S. In other words, for $i \neq j$, $\Sigma_{ij} = \rho$ if both i and j belong to some $S \in \mathcal{C}$ and zero otherwise. The instantiation that is relevant to the present paper is the collection of k-sets, i.e., $\mathcal{C} = \binom{p}{[k]}$, which is a smaller subset of the class of sparse covariance matrix (2.7) considered in this paper (with $\lambda = \rho k$). Therefore none of the upper bounds in [3, 4] applies. On the other hand, the lower bound from [4] yield a valid lower bound here, which gives $\lambda = \Omega(\frac{p}{k\sqrt{n}})$ if $\sqrt{p} \ll k \ll p$ and $\lambda = \Omega\big(\sqrt{\frac{k}{n}\log\frac{p}{k^2}}\big)$ if $k \ll \sqrt{p}$ and $k \ll n \log \frac{p}{k^2}$ (cf. [4, Eqn. (4.3) and (4.6)] respectively). This is highly suboptimal compared to Theorem 2 in both the moderately sparse or highly sparse regimes. In fact, since (2.13) in the case of k-sets is a special instance of sparse PCA, one can consider the best lower bound that detecting sparse principal components gives, in which case the eigenvector need not be binary-valued. The sharp detection rate was found in [19, Proposition 2] (see also [12] for an earlier suboptimal result) to be $\sqrt{\frac{k}{n}\log\frac{ep}{k}}$. Again, this yields a suboptimal lower bound and, in turn, shows the fundamental difference between the sparse PCA structure (submatrix sparsity) and that of sparse covariance matrices in this paper. In terms of the support set of the matrix, the analogy is that the former corresponds to k-cliques and the latter corresponds to k-regular graphs.

Finally, we compare the three models for sparse matrices of increasingly stronger structural assumptions, namely, (a) k-sparse matrices (Definition 1); (b) $k \times k$ submatrices; (c) k-sparse principal component (rank-one). In the normal mean model, the minimax rate of testing (against the zero null as in (1.3)) and estimation are summarized in Table 1, both with respect to the spectral norm.[2] Except for estimation in the k-sparse model where entrywise thresholding is optimal, attaining the optimal rate in all other problems demonstrates computational hardness.

3. Test procedures and upper bounds. In this section we consider the two sparsity regimes separately and design the corresponding rate-

[2]For the estimation rate of $k \times k$ submatrices, see [48, Example 1].

	testing	estimation
k-sparse matrices	$k \wedge (kp)^{1/4}$	k
$k \times k$ submatrix	$\sqrt{k}$	
k-sparse principal component		

TABLE 1

Rates of testing and estimation in various sparsity models (modulo logarithmic factors).

optimal testing procedures. In the highly sparse regime of $k \lesssim (\frac{p}{\log p})^{\frac{1}{3}}$, tests based on componentwise thresholding turns out to achieve the optimal rate of detection. In the moderately sparse regime of $k \gtrsim (\frac{p}{\log p})^{\frac{1}{3}}$, chi-squared test combined with the structural property in Section 3.1 is optimal.

3.1. *A structural property of sparse matrices.* Before we proceed to the construction of the rate-optimal tests, we first present a structural property of sparse matrices, which may be of independent interest. Recall that a matrix M is k-sparse in the sense of Definition 1 if its rows and columns are sparse but need not to have a common support. If in addition M has low rank, then the support sets of its rows must be highly aligned, and hence M has a sparse eigenvector and M is in fact supported on a submatrix. The main result of this section is an extension of this result to approximately low-rank matrices, in the sense of *stable rank* (also known as numerical rank):

$$\mathsf{sr}(M) \triangleq \frac{\|M\|_{\mathrm{F}}^2}{\|M\|_2^2}, \tag{3.1}$$

which is always a lower bound of $\mathsf{rank}(M)$.

The following lemma shows that for any sparse matrix of low stable rank, a constant fraction of its operation norm is concentrated on a small submatrix. The key ingredient of the proof is a celebrated result of Rudelson-Vershynin [51] in randomized numerical linear algebra which shows that the Gram matrix of any matrix M of stable rank at most r can be approximated by that of a submatrix of M formed by $O(r \log r)$ rows. The following is a restatement of [51, Theorem 3.1] without the normalization:

LEMMA 1. *There exists an absolute constant C_0 such that the following holds. Let $y \in \mathbb{R}^n$ be a random vector with covariance matrix $\Sigma = \mathbb{E}[yy^\top]$. Assume that $\|y\|_2 \le K$ holds almost surely. Let $y_1, \ldots, y_d$ be iid copies of y. Then*

$$\mathbb{E}\left\|\frac{1}{d}\sum_{i=1}^d y_i y_i^\top - \Sigma\right\|_2 \le C_0 K \sqrt{\|\Sigma\|_2 \frac{\log d}{d}},$$

provided that the right-hand side is less than $\|\Sigma\|_2$.

THEOREM 3 (Concentration of operator norm on small submatrices). *Let* $k \in [p]$. *Let* M *be a* $p \times p$ k*-sparse matrix (not necessarily symmetric) in the sense that all rows and columns are* k*-sparse. Let* $r = \mathsf{sr}(M)$. *Then there exist* $I, J \subset [p]$, *such that*

$$\|M_{IJ}\|_2 \geq \frac{1}{8}\|M\|_2, \quad |I| \leq Ckr, \quad |J| \leq Ckr \log r$$

where C *is an absolute constant.*

REMARK 1. The intuition behind the above result is the following: consider the ideal case where X is low-rank, say, $\mathsf{rank}(X) \leq r$. Then its right singular vector belongs to the span of at most r rows and is hence kr-sparse; so is the left singular vector. Theorem 3 extends this simple observation to stable rank with an extra log factor. Furthermore, the result in Theorem 3 cannot be improved beyond this log factor. To see this, consider a matrix M consisting of an $m \times m$ submatrix with independent Bern(q) entries and zero elsewhere, where $q = k/(2m) \ll 1$. Then with high probability, M is k-sparse, $\|M\|_2 \approx qm$, and $\|M\|_{\mathrm{F}}^2 \approx qm^2$. Although the rank of M is approximately m, its stable rank is much lower $\mathsf{sr}(M) \approx \frac{1}{q}$, and the leading singular vector of M is m-sparse, with $m = \Theta(k\,\mathsf{sr}(M))$. In fact, this example plays a key role in constructing the least favorable prior for proving the minimax lower bound in Section 4.

PROOF. Denote the i^{th} row of M by M_{i*}. Denote the j^{th} row of M by M_{*j}. Let

$$I_0 \triangleq \{i \in [p] : \|M_{i*}\|_2 \geq \tau\}$$
$$J_0 \triangleq \{j \in [p] : \|M_{*j}\|_2 \geq \tau\},$$

where $\tau > 0$ is to be chosen later. Then

$$|I_0| \vee |J_0| \leq \frac{\|M\|_{\mathrm{F}}^2}{\tau^2}. \tag{3.2}$$

Since the operator norm and Frobenius norm are invariant under permutation of rows and columns, we may and will assume that I_0, J_0 corresponds to the first few rows or columns of M. Write $M = \left(\begin{smallmatrix} A & C \\ D & B \end{smallmatrix}\right)$ where $B = M_{I_0^{\mathrm{c}} J_0^{\mathrm{c}}}$. Since each row of B is k-sparse, by the Cauchy-Schwarz inequality its ℓ_1-norm is at most $\sqrt{k}\tau$. Consequently its $\ell_\infty \to \ell_\infty$ operator norm satisfies

$\|B\|_\infty = \max_i \|B_{i*}\|_1 \le \sqrt{k}\tau$. Likewise, $\|B\|_1 = \max_j \|B_{*j}\|_1 \le \sqrt{k}\tau$. By duality (see, e.g., [33, Corollary 2.3.2]),

$$\|B\|_2 \le \sqrt{\|B\|_1 \|B\|_\infty} \le \sqrt{k}\tau. \tag{3.3}$$

Let $X = (A\ C)$ and $Y = \binom{A}{D}$. By triangle inequality, we have $\|M\|_2 \le \|X\|_2 + \|Y\|_2 + \|B\|_2$. Setting $\tau = \frac{\|M\|_2}{2\sqrt{k}}$, we have $\|B\|_2 \le \|M\|_2/2$ and hence $\|X\|_2 \vee \|Y\|_2 \ge \frac{\|M\|_2}{4}$. Without loss of generality, assume henceforth $\|X\|_2 \ge \frac{\|M\|_2}{4}$. Set $I = I_0$.

Note that $X \in \mathbb{R}^{\ell \times p}$, where $\ell = |I| \le \frac{\|M\|_{\mathrm{F}}^2}{\tau^2} = \frac{4k\|M\|_{\mathrm{F}}^2}{\|M\|_2^2} = 4L$. Furthermore, $\mathsf{sr}(X) = \frac{\|X\|_{\mathrm{F}}^2}{\|X\|_2^2} \le \frac{\|M\|_{\mathrm{F}}^2}{\|M\|_2^2/16} = 16r$. Next we show that X has a submatrix formed by a few columns whose operator norm is large. We proceed as in the proof of [51, Theorem 1.1]. Write

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_\ell^\top \end{bmatrix}, \quad \tilde{X} = \frac{1}{\sqrt{d}} \begin{bmatrix} y_1^\top \\ \vdots \\ y_d^\top \end{bmatrix}.$$

Define the random vector y by $\mathbb{P}\left\{y = \frac{\|X\|_{\mathrm{F}}}{\|x_i\|_2} x_i\right\} = \frac{\|x_i\|_2^2}{\|X\|_{\mathrm{F}}^2}$ and let $y_1, \dots, y_d$ which are iid copies of y. Then $X^\top X = \mathbb{E}[yy^\top]$ and $\tilde{X}^\top \tilde{X} = \frac{1}{d}\sum_{i=1}^d y_i y_i^\top$. Furthermore, $\|y\|_2 \le \|X\|_F$ almost surely and $\left\|\mathbb{E}[yy^\top]\right\|_2 = \|X\|_2^2$. By Lemma 1,

$$\mathbb{E}\left\|\tilde{X}^\top \tilde{X} - X^\top X\right\|_2 \le C_0 \sqrt{\frac{\log d}{d}} \|X\|_{\mathrm{F}} \|X\|_2 \le \frac{1}{4}\|X\|_2^2,$$

where the last inequality follows by choosing $d = \lceil Cr \log r \rceil$ with C being a sufficiently large universal constant. Therefore there exists a realization of $\tilde{X}$ so that the above inequality holds. Let J be the column support of $\tilde{X}$. Since the rows of $\tilde{X}$ are scaled version of those of X which are k-sparse, we have $|J| \le dk$. Let v denote a leading right singular vector of $\tilde{X}$, i.e., $\tilde{X}^\top \tilde{X} v = \|\tilde{X}\|_2^2 v$ and $\|v\|_2 = 1$. Then $\mathrm{supp}(v) \subset J$. Note that

$$\begin{aligned} \|Xv\|_2^2 &= v^\top X^\top X v = v^\top \tilde{X}^\top \tilde{X} v + v^\top (X^\top X - \tilde{X}^\top \tilde{X}) v \\ &\ge \|\tilde{X}\|_2^2 - \|X^\top X - \tilde{X}^\top \tilde{X}\|_2 \\ &\ge \|X\|_2^2 - 2\|X^\top X - \tilde{X}^\top \tilde{X}\|_2 \\ &\ge \frac{1}{2}\|X\|_2^2. \end{aligned}$$

Therefore $\|X_{*J}\|_2 \ge \|Xv\|_2 \ge \frac{1}{\sqrt{2}}\|X\|_2 \ge \frac{1}{4\sqrt{2}}\|M\|_2$. The proof is completed by noting that $X_{*J} = M_{IJ}$. □

3.2. *Highly sparse regime.* It is has been shown that, in the covariance model, entrywise thresholding is rate-optimal for estimating the matrix itself with respect to the spectral norm [22]. It turns out that in the very sparse regime entrywise thresholding is optimal for testing as well. Define

$$\hat{M} = (X_{ij}\mathbf{1}\{|X_{ij}| \geq \tau\}).$$

and the following test

$$\psi(X) = \mathbf{1}\Big\{\|\hat{M}\|_2 \geq \lambda\Big\}. \tag{3.4}$$

THEOREM 4. *For any $\epsilon \in (0,1)$, if*

$$\lambda > 2k\sqrt{2\log\frac{4p^2}{\epsilon}} \tag{3.5}$$

then the test (3.4) with $\tau = \sqrt{2\log\frac{4p^2}{\epsilon}}$ satisfies

$$\mathbb{P}_0(\psi = 1) + \sup_{M\in\Theta(p,k,\lambda)} \mathbb{P}_M(\psi = 0) \leq \epsilon$$

for all $1 \leq k \leq p$.

PROOF. Denote the event $E = \{\|Z\|_{\ell_\infty} \leq \tau\}$. Conditioning on E, for any k-sparse matrix $M \in \mathcal{M}(p,k)$, we have $\hat{M} \in \mathcal{M}(p,k)$ and

$$\|\hat{M} - M\|_2 \leq k\tau. \tag{3.6}$$

To see this, note that for any i,j, $\hat{M}_{ij} = 0$ whenever $M_{ij} = 0$. Therefore $\|\hat{M}_{i*} - M_{i*}\|_{\ell_1} \leq k\|Z\|_{\ell_\infty} \leq k\tau$ and, consequently, $\|\hat{M} - M\|_1 = \max_i \|\hat{M}_{i*} - M_{i*}\|_{\ell_1} \leq k\tau$. Similarly, $\|\hat{M} - M\|_\infty = \max_j \|\hat{M}_{*j} - M_{*j}\|_{\ell_1} \leq k\tau$. Therefore (3.6) follows from the fact that $\|\cdot\|_2^2 \leq \|\cdot\|_1\|\cdot\|_\infty$ for matrix induced norms. Therefore if $\lambda > 2k\tau$, then

$$\mathbb{P}_0(\psi = 1) + \sup_{M\in\Theta(p,k,\lambda)} \mathbb{P}_M(\psi = 0) \leq 2\mathbb{P}\left\{\|Z\|_{\ell_\infty} > \tau\right\} \leq 4p^2 e^{-\tau^2/2}.$$

This completes the proof. □

3.3. *Moderately sparse regime.* Our test in the moderately sparse regime relies on the existence of sparse approximate eigenvectors established in

Theorem 3. More precisely, the test procedure is a combination of the matrix-wise χ^2-test and the scan test based on the largest spectral norm of $m \times m$ submatrices, which is detailed as follows: Let

$$m = C\sqrt{\frac{kp}{\log \frac{ep}{k}}}.$$

where C is the universal constant from Theorem 3. Define the following test statistic

$$T_m(X) = \max\{\|X_{IJ}\|_2 : I, J \subset [p], |I| = |J| = m\} \tag{3.7}$$

and the test

$$\psi(X) = \mathbf{1}\{\|X\|_{\mathrm{F}}^2 \geq p^2 + s\} \vee \mathbf{1}\{T_m(X) \geq t\} \tag{3.8}$$

where

$$s \triangleq 2\log\frac{1}{\epsilon} + 2p\sqrt{\log\frac{1}{\epsilon}}, \quad t \triangleq 2\sqrt{m} + 4\sqrt{m\log\frac{ep}{m}}. \tag{3.9}$$

THEOREM 5. *There exists a universal constant C_0 such that the following holds. For any $\epsilon \in (0, 1/2)$, if*

$$\lambda \geq C_0\left\{kp\log\frac{1}{\epsilon}\log\left(\frac{p}{k}\log\frac{1}{\epsilon}\right)\right\}^{\frac{1}{4}}, \tag{3.10}$$

then the test (3.8) satisfies

$$\mathbb{P}_0(\psi = 1) + \sup_{M \in \Theta(p,k,\lambda)} \mathbb{P}_M(\psi = 0) \leq \epsilon$$

holds for all $1 \leq k \leq p$.

PROOF. First consider the null hypothesis, where $M = 0$ and $X = Z$ has iid standard normal entries so that $\|Z\|_{\mathrm{F}}^2 - p^2 = O_P(p)$. By standard concentration equality for χ^2 distribution, we have

$$\mathbb{P}\left\{|\|Z\|_{\mathrm{F}}^2 - p^2| > s\right\} \leq \epsilon,$$

where

$$s \triangleq 2\log\frac{1}{\epsilon} + 2p\sqrt{\log\frac{1}{\epsilon}}.$$

Consequently the false alarm probability satisfies

$$\mathbb{P}_0(\psi = 1) \le \underbrace{\mathbb{P}\left\{\|Z\|_{\mathrm{F}}^2 - p^2 > C_0 p\right\}}_{\le \epsilon} + \binom{p}{m}^2 \mathbb{P}\left\{\|W\|_2 \ge t\right\}.$$

where $t = 2\sqrt{m} + 4\sqrt{m \log \frac{ep}{m}}$ and $W \triangleq Z_{[m],[m]}$. By the Davidson-Szarek inequality [26, Theorem II.7], $\|W\|_2 \overset{\text{s.t.}}{\le} N(2\sqrt{m}, 1)$. Then $\mathbb{P}\{\|W\|_2 \ge t\} \le (\frac{em}{p})^m$. Hence the false alarm probability vanishes.

Next consider the alternative hypothesis, where, by assumption, M is row/column k-sparse and $\|M\|_2 \ge \lambda$. To begin, suppose that $\|M\|_{\mathrm{F}} \ge 2\sqrt{s}$. Then since $\|X\|_{\mathrm{F}}^2 - p^2 = \|M\|_{\mathrm{F}}^2 + 2\langle M, Z\rangle + \|Z\|_{\mathrm{F}}^2 - p^2$, we have

$$\begin{aligned}\mathbb{P}\{\|M + Z\|_{\mathrm{F}}^2 - p^2 < s\} &\le \mathbb{P}\left\{\|M\|_{\mathrm{F}}^2 + 2\langle M, Z\rangle < 2s\right\} + \mathbb{P}\left\{\|Z\|_{\mathrm{F}}^2 - p^2 < -s\right\} \\ &\le \exp(-s^2/8) + \epsilon.\end{aligned}$$

Therefore, as usual, if $\|M\|_{\mathrm{F}}$ is large, the χ^2-test will succeeds with high probability. Next assume that $\|M\|_{\mathrm{F}} < 2\sqrt{s}$. Therefore M is approximately low-rank, in the sense that

$$\mathsf{sr}(M) \le r \triangleq \frac{4s}{\lambda^2}.$$

By Theorem 3, there exists an absolute constant C and $I, J \subset [p]$ of cardinality at most

$$m = Ckr \log r = Ck \frac{4s}{\lambda^2} \log \frac{4s}{\lambda^2},$$

such that $\|M_{IJ}\|_2 \ge \frac{1}{8}\lambda$. Therefore the statistic defined in (3.7) satisfies $T_m(X) \ge \|X_{IJ}\|_2 \ge \frac{\lambda}{8} - \|Z_{IJ}\|_2$. Therefore $T_m(X) \ge \frac{\lambda}{8} - 3\sqrt{m}$ with probability at least $1 - \exp(-\Omega(m))$. Choose λ so that $\frac{\lambda}{8} - 3\sqrt{m} \ge t$. Since $t + 3\sqrt{m} = 5\sqrt{m} + 4\sqrt{m \log \frac{ep}{m}} \le 9\sqrt{m \log \frac{ep}{m}}$, it suffices to ensure that $\lambda \ge c_0\sqrt{m \log \frac{ep}{m}}$ for some absolute constant c_0. Plugging the expression of m, we found a sufficient condition is $\lambda \ge C_0(ks \log \frac{es}{k})^{\frac{1}{4}}$ for some absolute constant C_0. The proof is completed by noting that $s \le 2p(\log \frac{1}{\epsilon} + \log \frac{1}{\epsilon})$ and $s \mapsto s \log \frac{es}{k}$ is increasing. □

4. Minimax lower bound. In this section we prove the lower bound part of Theorem 1. The key step is to specify a prior π_1 under which the matrix is k-sparse with high probability and bound the χ^2-divergence between the null distribution and the mixture of the alternatives. Strictly speaking,

π_1 does not directly qualify as a prior for the alternative hypothesis since it is not exactly supported on the alternative parameter set; nevertheless, by conditioning it can be modified to be a valid prior (c.f. [52, Theorem 2.15 (i)] or Lemma 5 in the supplementary material [20, Section 8]).

4.1. *Least favorable prior.* Let I be chosen uniformly at random from all subsets of $[p]$ of cardinality m. Let $u = (u_1, \dots, u_p)$ be independent Rademacher random variables. Let B be a $p \times p$ matrix with i.i.d. Bern$(\frac{k}{m})$ entries and let (u, I, B) be independent. Let U_I denote the diagonal matrix defined by $(U_I)_{ii} = u_i \mathbf{1}\{i \in I\}$. Let $t > 0$ be specified later. Let the prior π_1 be the distribution of the following random sparse matrix:

$$M = tU_I B U_I, \tag{4.1}$$

i.e., $M_{ij} = t\mathbf{1}\{i \in I\}\mathbf{1}\{j \in I\}u_i u_j b_{ij}$. Therefore the non-zero pattern of M has the desired marginal distribution Bern$(\frac{k}{p})$, but the entries of M are *dependent*. Alternatively, M can be generated as follows: First choose an $m \times m$ principal submatrix with a uniformly chosen support I, fill it with i.i.d. Bern$(\frac{k}{m})$ entries, then pre- and post-multiply by a diagonal matrix consisting of independent Rademacher variables, which used to randomize the sign of the leading eigenvector. By construction, with high probability, the matrix M is $O(k)$-sparse and, furthermore, its operator norm satisfies $\|M\|_2 \geq kt$. Furthermore, the corresponding eigenvector is approximately $\mathbf{1}_J$, which is m-sparse.

The construction of this prior is based on the following intuition. The operator norm of a matrix highly depends on the correlation of the rows. Given the ℓ_2-norm of the rows, the largest spectral norm is achieved when all rows are aligned (rank-one), while the smallest spectral norm is achieved when all rows are orthogonal. In the sparse case, aligned support results in large spectral norm while disjoint support in small spectral norm. However, if all rows are aligned, then the signal is prominent enough to be distinguished from noise. Note that a submatrix structure strikes a precise balance between the extremal cases of completely aligned and disjoint support, which enforces that the row support sets are *contained* in a set of cardinality m, which is much larger than the row sparsity k but much smaller than the matrix size p. In fact, the optimal choice of the submatrix size given by $m \asymp k^2 \wedge \sqrt{kp}$, which matches the structural property given in Theorem 3. The structure of the least favorable prior, in a way, shows that the optimality of tests based on concentration on small submatrices is not a coincidence.

Another perspective is that the sparsity constraint on the matrix forces the marginal distribution of each entry in the nonzero pattern $(\mathbf{1}\{M_{ij} \neq 0\})$

to be $\mathrm{Bern}(\frac{k}{p})$. However, if all the entries were independent, then it would be very easy to test from noise. Indeed, perhaps the most straightforward choice of prior is $M_{ij} \overset{\text{i.i.d.}}{\sim} t \cdot \mathrm{Bern}(\frac{k}{p})$, where $t \asymp \frac{k}{p}$. However, the linear test statistic based on $\sum_{ij} M_{ij}$ succeeds unless $\lambda \lesssim 1$. We can improve the prior by randomize the eigenvector, i.e., $M_{ij} \overset{\text{i.i.d.}}{\sim} t u_i u_j \mathrm{Bern}(\frac{k}{p})$, but the χ^2-test in Theorem 5 succeeds unless $\lambda \lesssim \sqrt{k}$, which still falls short of the desired $\lambda \asymp (kp)^{1/4}$. Thus, we see that the coupling between the entries is useful to make the mixture distribution closer to the null hypothesis.

4.2. *Key lemmas.* The main tool for our lower bound is the χ^2-divergence, defined by $\chi^2(P \,\|\, Q) \triangleq \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q$ if $P \ll Q$ and $+\infty$ otherwise. The χ^2-divergence is related to the total variation via the following inequality [29, p. 1496]:

$$\chi^2 \geq \mathrm{TV} \log \frac{1 + \mathrm{TV}}{1 - \mathrm{TV}}. \tag{4.2}$$

Therefore the total variation distance cannot goes to one unless the χ^2-divergence diverges. Furthermore, if χ^2-divergence vanishes, then the total variation also vanishes, which is equivalently to, in view of (8.2), that P cannot be distinguished from Q better than random guessing.

The following lemma due to Ingster and Suslina (see, e.g., [40, p. 97]) gives a formula for the χ^2-divergence of a normal location mixture with respect to the standard normal distribution.

LEMMA 2. *Let P be an arbitrary distribution on $\mathbb{R}^m$. Then*

$$\chi^2(N(0, \mathbf{I}_m) * P \,\|N(0, \mathbf{I}_m)) = \mathbb{E}[\exp(\langle X, \tilde{X}\rangle)] - 1$$

where $*$ denotes convolution and X and $\tilde{X}$ are independently drawn from P.

The proof of the lower bound in Theorem 1 relies on the following lemmas. These results give non-asymptotic both necessary and sufficient conditions for certain moment generating functions involving hypergeometric distributions to be bounded, which show up in the χ^2-divergence calculation. Let $H \sim \mathrm{Hypergeometric}(p, m, m)$, with $\mathbb{P}\{H = i\} = \frac{\binom{m}{i}\binom{p-m}{m-i}}{\binom{p}{m}}, i = 0, \ldots, m$.

LEMMA 3 ([19, Lemma 1]). *Let $p \in \mathbb{N}$ and $m \in [p]$. Let $B_1, \ldots, B_m$ be independently Rademacher distributed. Denote by*

$$G_m \triangleq \sum_{i=1}^{m} B_i$$

the position of a symmetric random walk on $\mathbb{Z}$ starting at 0 after m steps. Then there exist an absolute constant $a_0 > 0$ and function $A : (0, a_0) \mapsto \mathbb{R}_+$ with $A(0+) = 0$, such that if $t = \frac{a}{m} \log \frac{\mathrm{e}p}{m}$ and $a < a_0$, then

$$\mathbb{E}\left[\exp\left(tG_H^2\right)\right] \le A(a). \tag{4.3}$$

LEMMA 4 ([34, Lemma 15, Appendix C]). *Let $p \in \mathbb{N}$ and $m \in [p]$. Then there exist an absolute constant $b_0 > 0$ and function $B : (0, b_0) \mapsto \mathbb{R}_+$ with $B(0+) = 0$, such that if $\lambda = b\left(\frac{1}{m} \log \frac{\mathrm{e}p}{m} \wedge \frac{p^2}{m^4}\right)$ and $b < b_0$, then*

$$\mathbb{E}\left[\exp\left(\lambda H^2\right)\right] \le B(b). \tag{4.4}$$

REMARK 2 (Tightness of Lemmas 3–4). The purpose of Lemma 3 is to seek the largest t, as a function of p and m, such that $\mathbb{E}\left[\exp\left(tG_H^2\right)\right]$ is upper bounded by a constant non-asymptotically. The condition that $t \asymp \frac{1}{m} \log \frac{\mathrm{e}p}{m}$ is in fact both necessary and sufficient. To see the necessity, note that $\mathbb{P}\{G_H = H | H = i\} = 2^{-i}$. Therefore

$$\mathbb{E}\left[\exp\left(tG_H^2\right)\right] \ge \mathbb{E}\left[\exp(tH^2)2^{-H}\right] \ge \exp(tm^2)2^{-m}\,\mathbb{P}\{H = m\} \ge \exp\left(tm^2 - m \log \frac{2p}{m}\right),$$

which cannot be upper bound bounded by an absolute constant unless $t \lesssim \frac{1}{m} \log \frac{\mathrm{e}p}{m}$.

Similarly, the condition $\lambda \lesssim \frac{1}{m} \log \frac{\mathrm{e}p}{m} \wedge \frac{p^2}{m^4}$ in Lemma 4 is also necessary. To see this, note that $\mathbb{E}[H] = \frac{m^2}{p}$. By Jensen's inequality, we have $\mathbb{E}\left[\exp\left(\lambda H^2\right)\right] \ge \exp(\frac{\lambda m^4}{p^2})$. Therefore a necessary condition for (4.3) is that $\lambda \le \frac{p^2 \log B}{m^4}$. On the other hand, we have $\mathbb{E}\left[\exp\left(\lambda H^2\right)\right] \ge \exp(\lambda m^2 - m \log \frac{p}{m})$, which implies that $\lambda \lesssim \frac{1}{m} \log \frac{\mathrm{e}p}{m}$.

4.3. *Proof of Theorem 1: lower bound.*

PROOF. *Step 1*: Fix $t > 0$ to be determined later. Recall the random sparse matrix $M = tU_I B U_I$ defined in (4.1), where I is chosen uniformly at random from all subsets of $[p]$ of cardinality k, $u = (u_1, \ldots, u_p)^\top$ consists of independent Rademacher entries, B is a $p \times p$ matrix with i.i.d. Bern($\frac{k}{m}$) entries, and (u, I, B) are independent.

Next we show that the hypothesis $H_0 : X = Z$ versus $H_1 : X = M + Z$ cannot be tested with vanishing probability of error, by showing that the χ^2-divergence is bounded. Let $(\tilde{U}, \tilde{I}, \tilde{B})$ be an independent copy of (U, I, B).

Then $\tilde{M} = \tilde{U}_{\tilde{I}} \tilde{B} \tilde{U}_{\tilde{I}}$ is an independent copy of M. Put $s = t^2$. By Lemma 2, we have

$$\begin{aligned}
\chi^2(P_{X|H_0} \,\|\, P_{X|H_1}) + 1 &= \mathbb{E}\left[\exp\left(\langle M, \tilde{M}\rangle\right)\right] = \mathbb{E}\left[\exp\left(t^2 \langle U_I B U_I, \tilde{U}_{\tilde{I}} \tilde{B} \tilde{U}_{\tilde{I}}\rangle\right)\right] \\
&= \mathbb{E}\left[\exp\left(s \sum_{i \in I \cap \tilde{I}} \sum_{j \in I \cap \tilde{I}} u_i \tilde{u}_i u_j \tilde{u}_j b_{ij} \tilde{b}_{ij}\right)\right] \\
&\overset{\text{(a)}}{=} \mathbb{E}\left[\exp\left(s \sum_{i \in I \cap \tilde{I}} \sum_{j \in I \cap \tilde{I}} u_i u_j a_{ij}\right)\right] \\
&\overset{\text{(b)}}{=} \mathbb{E}\left[\prod_{i \in I \cap \tilde{I}} \prod_{j \in J \cap \tilde{J}} \left(1 + \frac{k^2}{m^2}(e^{s u_i u_j} - 1)\right)\right] \\
&\overset{\text{(c)}}{\le} \mathbb{E}\left[\exp\left\{\frac{k^2}{m^2} \sum_{i \in I \cap \tilde{I}} \sum_{j \in I \cap \tilde{I}} (e^{s u_i u_j} - 1)\right\}\right] \\
&= \mathbb{E}\left[\exp\left\{\frac{k^2}{m^2} \sum_{i \in I \cap \tilde{I}} \sum_{j \in I \cap \tilde{I}} (u_i u_j \sinh(s) + \cosh(s) - 1)\right\}\right] \\
&= \mathbb{E}\left[\exp\left\{\frac{k^2 \sinh(s)}{m^2}\left(\sum_{i \in I \cap \tilde{I}} u_i\right)^2 + \frac{k^2(\cosh(s) - 1)}{m^2} |I \cap \tilde{I}|^2\right\}\right], \qquad (4.5)
\end{aligned}$$

where (a) is due to $(u_m \tilde{u}_m, \ldots, u_m \tilde{u}_m) \overset{\text{(d)}}{=} (u_1, \ldots, u_m)$; (b) follows from $a_{ij} \triangleq b_{ij} \tilde{b}_{ij} \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(\frac{k^2}{m^2})$; (c) follows from the fact that $\log(1 + x) \le x$ for all $x > -1$; (d) is because for $b \in \{\pm 1\}$, we have $\exp(sb) = b \sinh(s) + \cosh(s) - 1$. Recall from Lemma 3 that $\{G_m : m \ge 0\}$ denotes the symmetric random walk on $\mathbb{Z}$. Since $I, \tilde{I}$ are independently and uniformly drawn from all subsets of $[p]$ of cardinality k, we have $H \triangleq |I \cap \tilde{I}| \sim \text{Hypergeometric}(p, m, m)$. Define

$$A(m, s) \triangleq \mathbb{E}\left[\exp\left\{\frac{2k^2 \sinh(s)}{m^2} G_H^2\right\}\right], \qquad (4.6)$$

$$B(m, s) \triangleq \mathbb{E}\left[\exp\left\{\frac{2k^2(\cosh(s) - 1)}{m^2} H^2\right\}\right]. \qquad (4.7)$$

Applying the Cauchy-Schwarz inequality to the right-hand side of (4.5), we obtain

$$\chi^2(P_{X|H_0} \,\|\, P_{X|H_1}) + 1 \le \sqrt{A(m, s) B(m, s)}. \qquad (4.8)$$

Therefore upper bounding the χ^2-divergence boils down to controlling the expectations in (4.6) and (4.7) separately.

Applying Lemma 3 and Lemma 4 to $A(m,s)$ and $B(m,s)$ respectively, we conclude that

$$\frac{k^2(\cosh(s)-1)}{m^2} \le c\left(\frac{1}{m}\log\frac{\mathrm{e}p}{m} \wedge \frac{p^2}{m^4}\right) \;\Rightarrow\; A(m,s) \le C \tag{4.9}$$

$$\frac{k^2\sinh(s)}{m^2} \le \frac{c}{m}\log\frac{\mathrm{e}p}{m} \;\Rightarrow\; B(m,s) \le C \tag{4.10}$$

where c, C are constants so that $C \to 0$ as $c \to 0$. Therefore the best lower bound we get for s is

$$s^* = \max_{k\le m\le p}\left\{(\cosh-1)^{-1}\left(\frac{cm}{k^2}\log\frac{\mathrm{e}p}{m} \wedge \frac{cp^2}{m^2k^2}\right) \wedge \sinh^{-1}\left(\frac{cm}{k^2}\log\frac{\mathrm{e}p}{m}\right)\right\}, \tag{4.11}$$

where the inverses $\sinh^{-1}$ and $(\cosh-1)^{-1}$ are defined with the domain restricted to $\mathbb{R}_+$.

To simplify the maximization in (4.11), we use the following bounds of the hyperbolic functions:

$$\sinh^{-1}(y) \ge \log(2y),\ (\cosh-1)^{-1}(y) \ge \log y, \quad y \ge 0. \tag{4.12}$$

Therefore

$$s^* \ge \log \max_{k\le m\le p}\left(\frac{cm}{k^2}\log\frac{\mathrm{e}p}{m} \wedge \frac{cp^2}{m^2k^2}\right).$$

Choosing $m = \left(\frac{p^2}{\log p}\right)^{\frac{1}{3}}$ yields

$$s^* \gtrsim \log^+\left(\frac{p\log p}{k^3}\right), \tag{4.13}$$

where $\log^+ \triangleq \max\{\log, 0\}$. Note that the above lower bound is vacuous unless $k \le (p\log p)^{\frac{1}{3}}$. To produce a non-trivial lower bound for $k \ge (p\log p)^{\frac{1}{3}}$, note that (4.12) can be improved as follows. If the argument y is restricted to the unit interval, then

$$\sinh^{-1}(y) \ge \sinh^{-1}(1)\, y,\ (\cosh-1)^{-1}(y) \ge \sqrt{y}, \quad y \in [0,1], \tag{4.14}$$

which follows from the Taylor expansion of cosh and the convexity of sinh. Applying (4.14) to (4.11),

$$s^* = \max_{m:\frac{cm}{k^2}\log\frac{\mathrm{e}p}{m}\le 1}\left(\sqrt{\frac{cp^2}{m^2k^2}} \wedge \frac{c\sinh^{-1}(1)m}{k^2}\log\frac{\mathrm{e}p}{m}\right).$$

Choosing $m = \sqrt{\frac{pk}{4c^2 \log \frac{ep}{k}}}$ yields $\frac{cm}{k^2} \log \frac{ep}{k} \le 1$. We then obtain

$$s^* \gtrsim \sqrt{\frac{p}{k^3} \log \frac{ep}{k}}. \tag{4.15}$$

Step 2: To conclude kt as a valid lower bound for λ with $t = \sqrt{s^*}$ given in (4.13) and (4.15), we invoke Lemma 5 in the supplementary material [20, Section 8]. To this end, we need to show that with high probability, M is $O(k)$-sparse and $\|M\|_2 = \Omega(kt)$. Define events

$$E_1 = \{M \in \mathcal{M}(p, 2k)\}, \quad E_2 = \{\|M\|_2 \ge kt/2\}.$$

It remains to show that both are high-probability events. Since I is independent of B, we shall assume, without loss of generality, that $I = [m]$. For the event E_1, by the union bound and Hoeffding's inequality, we have

$$\mathbb{P}\{E_1^{\mathrm{c}}\} = \mathbb{P}\{B_{II} \notin \Theta(m, 2k)\} \le m^2 \mathbb{P}\left\{\sum_{i=1}^{m} b_{i1} \ge 2k\right\} \le m^2 \exp(-mk^2) = o(1), \tag{4.16}$$

where $b_{i1} \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(\frac{k}{m})$. For the event E_2, again by Hoeffding's inequality,

$$\begin{aligned}
\mathbb{P}\{E_2\} &= \mathbb{P}\{\|B_{II}\|_2 \ge k/2\} \ge \mathbb{P}\left\{\|M\mathbf{1}_I\|_2 \ge \frac{k}{2}\|\mathbf{1}_I\|_2\right\} \\
&\ge \mathbb{P}\left\{\sum_{j=1}^{m} b_{ij} \ge \frac{k}{2}, \forall i \in [m]\right\} \ge 1 - m\,\mathbb{P}\left\{\sum_{j=1}^{m} b_{1j} < \frac{k}{2}\right\} \\
&\ge 1 - m\exp(-mk^2/4) = 1 - o(1).
\end{aligned} \tag{4.17}$$

The desired lower bound now follows from Lemma 5 in the supplementary material [20, Section 8].

Finally, we note that the lower bound continues to hold up to constant factors even if M is constrained to be symmetric. Indeed, we can replace M with the symmetrized version $M' = \left[\begin{smallmatrix} 0 & M \\ M^\top & 0 \end{smallmatrix}\right]$ and note that the bound on χ^2-divergence remains valid since $\langle M', \tilde{M}' \rangle = 2\langle M, \tilde{M} \rangle$. □

5. Detecting sparse covariance matrices. In this section we describe the test procedures for the the covariance model and prove the upper bound part of Theorem 2. The lower bound proof is given in the supplementary material [20, Section 9].

Let $X_1, \dots, X_n$ be independently sampled from $N(0, \Sigma)$. Define the sample covariance matrix as

$$S = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top, \tag{5.1}$$

which is a sufficient statistic for Σ.

The following result is the counterpart of Theorem 4 for entrywise thresholding that is optimal in the highly sparse regime:

THEOREM 6. *Let C, C' be constants that only depend on τ. Let $\epsilon \in (0,1)$. Define $\hat{\Sigma} = (S_{ij}\mathbf{1}\{|S_{ij}| \geq t\})$, where $\tau = \sqrt{C \log \frac{p}{\epsilon}}$. Assume that $n \geq C' \log p$. If $\lambda\sqrt{n} > 2kt$, then the test $\psi(S) = \mathbf{1}\left\{\|\hat{\Sigma}\|_2 \geq \lambda\right\}$ satisfies*

$$\mathbb{P}_{\mathbf{I}}(\psi = 1) + \sup_{\Sigma \in \Xi(p,k,\lambda,\tau)} \mathbb{P}_{\Sigma}(\psi = 0) \leq \epsilon$$

for all $1 \leq k \leq p$.

To extend the test (3.8) to covariance model, we need a test statistic for $\|\Sigma - \mathbf{I}\|_{\mathrm{F}}^2$. Consider the following U-statistic proposed in [17, 24]:

$$Q(S) = p + \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \langle X_i, X_j \rangle^2 - \langle X_i, X_i \rangle - \langle X_j, X_j \rangle, \tag{5.2}$$

Then $Q(S)$ is a unbiased estimator of $\|\Sigma - \mathbf{I}\|_{\mathrm{F}}^2$. We have the following result for the moderately sparse regime:

THEOREM 7. *Let $m = C\sqrt{\frac{kp}{\log \frac{ep}{k}}}$, where C is the universal constant from Theorem 3. Define the following test statistic*

$$T_m(S) = \max\{\|S_{II}\|_2 : I \subset [p], |I| = m\} \tag{5.3}$$

and the test

$$\psi(S) = \mathbf{1}\{Q(S) \geq s\} \vee \mathbf{1}\{T_m(X) \geq t\} \tag{5.4}$$

where

$$s \triangleq 2\log\frac{1}{\epsilon} + 2p\sqrt{\log\frac{1}{\epsilon}}, \quad t \triangleq 2\sqrt{m} + 4\sqrt{m \log \frac{ep}{m}}. \tag{5.5}$$

There exists a universal constant C_0 such that the following holds. For any $\epsilon \in (0, 1/2)$, if

$$\lambda \geq C_0 \left\{ kp \log \frac{1}{\epsilon} \log \left(\frac{p}{k} \log \frac{1}{\epsilon} \right) \right\}^{\frac{1}{4}}, \tag{5.6}$$

then the test (5.4) satisfies

$$\mathbb{P}_0(\psi = 1) + \sup_{M \in \Theta(p,k,\lambda)} \mathbb{P}_M(\psi = 0) \leq \epsilon$$

for all $1 \leq k \leq p$.

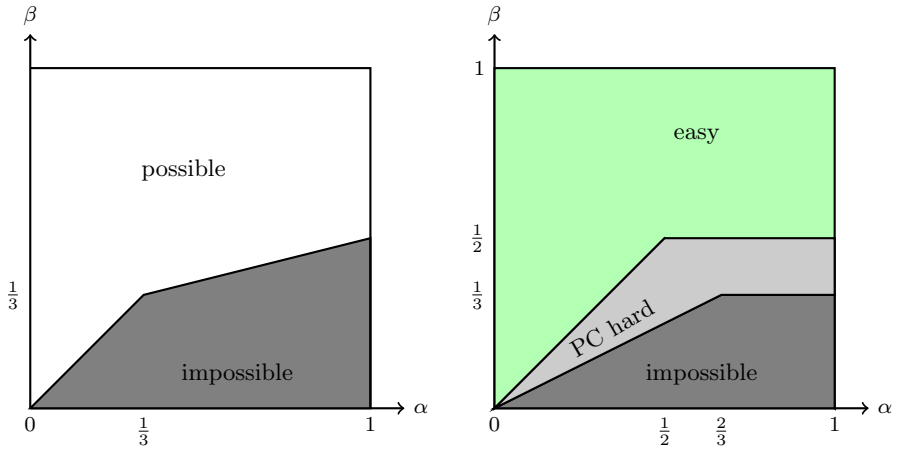
The proofs of Theorems 6 and 7 parallel those of Theorems 4 and 5. Next we point out the main distinction. For Theorem 6, the only difference is the Gaussian tail is replaced by the concentration inequality $\mathbb{P}\{|S_{ij} - \Sigma_{ij}| \geq a\} \leq c_0 \exp(-c_1 n t^2)$ for all $|t| \leq c_2$, where c_i's are constants depending only on τ [22, Eq. (26)]. For Theorem 7, let $\tilde{S} \triangleq \Sigma^{-\frac{1}{2}} S \Sigma^{-\frac{1}{2}}$, which is a $k \times k$ standard Wishart matrix with n degrees of freedom. Applying the deviation inequality in [18, Proposition 4], we have $\mathbb{E}[\|\tilde{S} - I_k\|_2^2] \lesssim \frac{k}{n} + \frac{k^2}{n^2}$. Since $\|S - \Sigma\|_2 \leq \|\Sigma\|_2 \|\tilde{S} - I_k\|_2$, we have $\mathbb{E}\left[\|S - \Sigma\|_2^2\right] \lesssim \lambda^2 \left(\frac{k}{n} + \frac{k^2}{n^2}\right)$.

6. Computational limits. In this section we address the computational aspects of detecting sparse matrices in both the Gaussian noise and the covariance model.

Gaussian noise model. The computational hardness of the red region (reducibility from planted clique) in Figure 1 follows from that of submatrix detection in Gaussian noise [14, 47], which is a special case of the model considered here. The statistical and computational boundary of submatrix detection is shown in Figure 2(b), in terms of the tradeoff between the sparsity $k = p^\alpha$ and the spectral norm of the signal $\lambda = p^\beta$. Below we explain how Figure 2(b) follows from the results in [47].

The setting in [47] also deals with the additive Gaussian noise model (1.2), where, under the alternative, the entries of the mean matrix M is at least θ on a $k \times k$ submatrix and zero elsewhere, with $k = p^\alpha$ and $\theta = p^{-\gamma}$. Since $\|M\|_2 \geq k\theta$, this instance is included in the alternative hypothesis in (1.3) with $\lambda = p^\beta$ and $\beta = \alpha - \gamma$. It is shown that (see [47, Theorem 2 and Fig. 1]) detection is computationally at least as hard as solving the planted clique problem when $\gamma > 0 \vee (2\alpha - 1)$, i.e., $\beta < \alpha \wedge (1 - \alpha)$. Note that this bound is not monotone in α, which can be readily improved to $\beta < \alpha \wedge \frac{1}{2}$, corresponding to the computational limit in Figure 2(b). Similarly, detection

is statistically impossible when $\gamma > \frac{\alpha}{2} \vee (2\alpha - 1)$, i.e., $\beta < \frac{\alpha}{2} \wedge (1-\alpha)$. Taking the monotone upper envelope leads to $\beta < \frac{\alpha}{2} \wedge \frac{1}{3}$, yielding the statistical limit in Figure 2(a). Finally, Figure 1 can be obtained by superimposing the statistical-computational limits in Figure 2(a) on top of the statistical limit obtained in the present paper as plotted in Figure 2(b).



(a) Statistical boundary for detecting sparse matrices (this paper).

(b) Statistical-computational boundary for detecting submatrices [47].

FIG 2. *Detection boundary for k-sparse matrices and $k \times k$ submatrices M in noise, where $k = p^\alpha$ and $\|M\|_2 = \lambda = p^\beta$.*

Sparse covariance model. For the problem of detecting sparse covariance matrices, which is defined by the 4-tuple (n, p, k, λ), the picture is less complete than the additive-noise counterpart; this is mainly due to the extra parameter n. Indeed, the statistical lower bound in Theorem 2 holds under the extra assumptions (2.10) and (2.11) that the sample size is sufficiently large, while the current computational lower bound for sparse PCA in the literature [14, 32, 54] also requires a number of conditions including the assumption of $n \le p$. Nevertheless, if we still let $k = p^\alpha$ and $\lambda\sqrt{n} = p^\beta$ and focus on the tradeoff between the (α, β) pair, the statistical and computational limits in Figure 1 continue to hold. Next we explain how to deduce the computational hardness of the red region from that of sparse PCA in the spiked Gaussian covariance model [32].

To this end, due to monotonicity, it suffices to demonstrate a "hard instance", i.e., a sequence of triples (n, λ, k) indexed by p, for every (α, β) such that $\frac{1}{3} < \alpha < \frac{1}{2}$ and $\beta < 1$. Given samples $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$, the

computational aspect of testing

$$H_0 : \Sigma = \mathbf{I}, \quad \text{versus} \quad H_1 : \Sigma = \mathbf{I} + \lambda u u^\top, \tag{6.1}$$

where the eigenvector u is both k-sparse and unit-norm, has been studied in [32]. Fix $\alpha \in (\frac{1}{3}, \frac{1}{2})$. Let $n = p^\eta$, $k = p^\alpha$ and $\lambda = \frac{ck^2}{n \log^2 n}$, so that $\beta = 2\alpha - \eta$, and let $\frac{1}{a} \le \eta \le 1$ to be chosen later; here $a > 1$ and $c > 0$ are absolute constants from [32, Theorem 5.4]. By assumption, $(2\alpha, 4\alpha) \cap (\frac{1}{a}, 1) \neq \emptyset$; pick any η therein. Then we have $\lambda \ll 1$ and (6.1) is indeed an instance of (2.6). By the choice of the parameters, the conditions of [32, Theorem 5.4] are fulfilled, namely, $\beta < \alpha$ and $\alpha > \frac{\eta}{4}$, and the detection problem (6.1) and hence (2.6) are at least as hard as the planted clique problem.

References.

[1] N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie. Testing k-wise and almost k-wise independence. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 496–505. ACM, 2007.

[2] N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. *Random Structures and Algorithms*, 13(3-4):457–466, 1998.

[3] E. Arias-Castro, S. Bubeck, and G. Lugosi. Detection of correlations. *The Annals of Statistics*, 40(1):412–435, 2012.

[4] E. Arias-Castro, S. Bubeck, and G. Lugosi. Detecting positive correlations in a multivariate sample. *Bernoulli*, 21(1):209–241, 2015.

[5] E. Arias-Castro, E. J. Candés, H. Helgason, and O. Zeitouni. Searching for a trail of evidence in a maze. *The Annals of Statistics*, 36(4):1726–1757, 2008.

[6] E. Arias-Castro, E. J. Candès, and Y. Plan. Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011.

[7] E. Arias-Castro, D. L. Donoho, and X. Huo. Near-optimal detection of geometric objects by fast multiscale methods. 51(7):2402–2425, 2005.

[8] E. Arias-Castro and N. Verzelen. Community detection in dense random networks. *Ann. Statist.*, 42(3):940–969, 06 2014.

[9] S. Balakrishnan, M. Kolar, A. Rinaldo, A. Singh, and L. Wasserman. Statistical and computational tradeoffs in biclustering. In *NIPS 2011 Workshop on Computational Trade-offs in Statistical Learning*, 2011.

[10] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 30:1046–1066, 2013.

[11] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

[12] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

[13] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[14] C. Butucea and Y. I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 2013.

[15] T. T. Cai, J. X. Jeng, and J. Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 73(5):629 – 662, 2011.

[16] T. T. Cai, T. Liang, and A. Rakhlin. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *The Annals of Statistics*, 45:1403–1430, 2017.

[17] T. T. Cai and Z. Ma. Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli*, 19(5B):2359–2388, 2013.

[18] T. T. Cai, Z. Ma, and Y. Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074 – 3110, 2013.

[19] T. T. Cai, Z. Ma, and Y. Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, 161(3-4):781–815, Apr. 2015.

[20] T. T. Cai and Y. Wu. Supplement to "Statistical and computational limits for sparse matrix detection". Jan 2019.

[21] T. T. Cai and M. Yuan. Minimax rate optimal detection of very short signal segments. *arXiv preprint arXiv:1407.2812*, 2014.

[22] T. T. Cai and H. H. Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40:2389–2420, 2012.

[23] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[24] S. X. Chen, L.-X. Zhang, and P.-S. Zhong. Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, 105(490):810–819, 2010.

[25] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Publ. Math. Inst. Hungar. Acad. Sci., Ser. A*, 8:85–108, 1963.

[26] K. Davidson and S. Szarek. Local operator theory, random matrices and Banach spaces. In W. Johnson and J. Lindenstrauss, editors, *Handbook on the Geometry of Banach Spaces*, volume 1, pages 317–366. Elsevier Science, 2001.

[27] D. L. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

[28] J. Fan, P. Rigollet, and W. Wang. Estimation of functionals of sparse covariance matrices. *The Annals of statistics*, 43(6):2706, 2015.

[29] A. A. Fedotov, P. Harremoës, and F. Topsøe. Refinements of Pinsker's inequality. *Information Theory, IEEE Transactions on*, 49(6):1491–1498, 2003.

[30] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical Algorithms and a Lower Bound for Planted Clique. *ArXiv e-prints*, Jan. 2012.

[31] W. Feller. *An Introduction to Probability Theory and Its Applications: Volume I.* John Wiley & Sons New York, 1968.

[32] C. Gao, Z. Ma, and H. H. Zhou. Sparse CCA: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017.

[33] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, USA, 3rd edition, 1996.

[34] B. Hajek, Y. Wu, and J. Xu. Computational lower bounds for community detection on random graphs. *Conference on Learning Theory (COLT)*, 2015. arxiv:1406.6625.

[35] B. Hajek, Y. Wu, and J. Xu. Computational lower bounds for community detection on random graphs. In *Proceedings of COLT 2015*, pages 899–928, June 2015.

[36] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016.

[37] E. Hazan and R. Krauthgamer. How hard is it to approximate the best nash equilibrium? *SIAM Journal on Computing*, 40(1):79–91, 2011.
[38] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
[39] Y. I. Ingster. On some problems of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics*, 6(1):47 – 69, 1997.
[40] Y. I. Ingster and I. A. Suslina. *Nonparametric Goodness-of-fit Testing under Gaussian Models*. Springer, New York, NY, 2003.
[41] M. Jerrum. Large cliques elude the Metropolis process. *Random Structures & Algorithms*, 3(4):347–359, 1992.
[42] N. E. Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, pages 2717–2756, 2008.
[43] O. Klopp and A. B. Tsybakov. Estimation of matrices with row sparsity. *Problems of Information Transmission*, 51(4):335–348, 2015.
[44] W. Kozakiewicz. On the convergence of sequences of moment generating functions. *Annals of Mathematical Statistics*, 18(1):61–69, 1947.
[45] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York, NY, 1986.
[46] K. Lounici, M. Pontil, S. Van De Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.
[47] Z. Ma and Y. Wu. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116, 2015.
[48] Z. Ma and Y. Wu. Volume ratio, sparsity, and minimaxity under unitarily invariant norms. *IEEE Transactions on Information Theory*, 61(12):6939 – 6956, Dec 2015.
[49] A. Onatski, M. J. Moreira, and M. Hallin. Asymptotic power of sphericity tests for high-dimensional data. *The Annals of Statistics*, 41(3):1204–1231, 2013.
[50] A. Onatski, M. J. Moreira, and M. Hallin. Signal detection in high dimension: The multispiked case. *The Annals of Statistics*, 42(1):225–254, 2014.
[51] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 54(4):21, 2007.
[52] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY, 2009.
[53] R. Vershynin. High-dimensional probability: An introduction with applications in data science. Book draft, available at https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf.
[54] T. Wang, Q. Berthet, and R. J. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5):1896–1930, 2016.
[55] D. Yang, Z. Ma, and A. Buja. Rate optimal denoising of simultaneously sparse and low rank matrices. *The Journal of Machine Learning Research*, 17(1):3163–3189, 2016.

SUPPLEMENTARY MATERIAL

Supplementary material for "Statistical and Computational Limits for Sparse Matrix Detection"
(; .pdf). Due to space constraints, some proofs are deferred to the supplementary document [20].