Detecting Positive Correlations in a Multivariate Sample

Ery Arias-Castro* Sébastien Bubeck† Gábor Lugosi‡

March 29, 2013

Abstract

We consider the problem of testing whether a correlation matrix of a multivariate normal population is the identity matrix. We focus on sparse classes of alternatives where only a few entries are nonzero and, in fact, positive. We derive a general lower bound applicable to various classes and study the performance of some near-optimal tests. We pay special attention to computational feasibility and construct near-optimal tests that can be computed efficiently. Finally, we apply our results to prove new lower bounds for the clique number of high-dimensional random geometric graphs.

Keywords: sparse covariance matrices, sparse detection, high-dimensional data, minimax detection, Bayesian detection, random geometric graphs.

1 Introduction

In multivariate statistics, inference about a covariance (i.e., dispersion) matrix aims at answering questions of dependencies between the variables. This is strictly true when the variables are jointly Gaussian, which is the classical assumption. A basic question is whether the variables are dependent at all. Concretely, consider a simple setting where the components of a random vector are jointly normal, each with zero mean and unit variance. Then the variables are independent if and only if their covariance matrix is the identity matrix. As usual, inference is based on an i.i.d. sample of size m, denoted $X_1, \dots, X_m$ with $X_t = (X_{t,1}, \dots, X_{t,n}) \in \mathbb{R}^n$ for $t = 1, \dots, m$. As stated above, we assume that $\mathbb{E}X_{t,i} = 0$ and $\mathrm{Var}(X_{t,i}) = 1$, and let $\sigma_{i,j} = \mathrm{Cov}(X_{t,i}, X_{t,j})$.

We are interested in testing whether the population covariance matrix is the identity matrix, or not, so the null hypothesis is

$$H_0 : \sigma_{i,j} = 0, \ \forall i \neq j \ .$$

This testing problem is well studied in the classical regime where the dimension n is fixed and the sample size m increases to infinity. See, for example, Muirhead [24, Sec. 8.4], where the generalized likelihood ratio test (GLRT)—against the alternative hypothesis $H_1 : \sigma_{i,j} \neq 0$ for some $i \neq j$—is studied in detail, as well as some unbiased variant. When the dimension is large (i.e.,

*Department of Mathematics, University of California, San Diego
†Department of Operations Research and Financial Engineering, Princeton University
‡ICREA and Department of Economics, Universitat Pompeu Fabra

$n \to \infty$), the GLRT may be degenerate. This is discussed in detail in Ledoit and Wolf [23], where other tests—including a new one—are examined for consistency in this high-dimensional regime. Their ideas are further explored in Srivastava [27], Fisher [19], and Chen, Zhang, and Zhong [13]. All these tests are based on symmetric polynomials of the sample correlation coefficients. We discuss this in Section 3.1. These tests are shown to be consistent when $n/m \to c \in (0, \infty)$ under additional mild conditions.

While these papers focus on general consistency, our focus is on alternatives where the covariance matrix is sparse, meaning that even under the alternative hypothesis, only a few variables are substantially correlated. This sparse setting has been investigated in the last few years, with recent work on the estimation of sparse covariance matrices, see El Karoui [18], Bickel and Levina [8, 9], and Cai, Zhang, and Zhou [12]. To our knowledge, testing for sparse correlation structures in a multivariate sample has not been considered in detail before. This is what we study here.

1.1 Correlation models

We introduce sparse models of correlation matrices to test against. Though many more models are possible, we choose a few emblematic examples that are of interest in a much wider sense within the literature on sparse covariance estimation. In all cases, the null hypothesis is that the observed vector has identity covariance matrix. For the alternative hypothesis, we consider the following prototypical examples:

- **Block model.** The covariance under the alternative hypothesis is the identity matrix except for a $k \times k$ block on the diagonal. Formally, given $\rho > 0$, we assume here that there is a subset of indices of the form $S = \{i, \dots, i + k - 1\}$ modulo n—for aesthetic reasons—such that $\sigma_{i,j} \geq \rho$ if $i, j \in S, i \neq j$. The set S is called the *anomalous set*.

- **Clique model.** This model is defined as the block model with the possible anomalous set S ranging over all the subsets of indices of size k.

- **Perfect matching model.** Suppose n is a perfect square with $n = k^2$. Here the components of the observed vector X correspond to edges of the complete bipartite graph on $2k$ vertices. The alternative hypothesis is that the bipartite graph has a perfect matching such that $\sigma_{i,j} \geq \rho$ for all $i, j \in S, i \neq j$ where S is the anomalous set of indices corresponding to the edges of the perfect matching.

All through, we assume that $k < n/2$. In fact, we will most interested in the regime where $k = o(n)$, where the anomaly only affects a negligible fraction of the variables.

The block model is closely related to the models used in Cai, Zhang, and Zhou [12] to obtain bounds on the minimax risk of estimating sparse matrices. Roughly speaking, Cai, Zhang, and Zhou use the block model with $S = \{1, \dots, k\}$ and place nonzero entries in a (carefully designed) fashion within that block. The fraction of nonzero entries within the block is about one-half. We could also assume that only a fraction of the entries in the block are nonzero and it would only change constants later on. More importantly, to make the detection problem interesting, we need to consider all possible blocks. Note that the block model is parametric. The clique model is a natural generalization of the block model leading to a nonparametric model. The perfect matching model gives an example of a class of sets with a more intricate combinatorial structure which our approach is able to deal with.

1.2 Tests and their risks

As usual, a *test* is a binary-valued function $f : \mathbb{R}^{nm} \to \{0, 1\}$, with $f(X_1, \ldots, X_m) = 1$ meaning that the test rejects the null hypothesis H_0 in favor of the particular alternative hypothesis of interest. We measure the performance of a test based on its *worst-case risk* over the model of interest $\mathcal{M}$, formally defined by

$$R^{\max}(f) = \mathbb{P}_0\{f(X_1, \ldots, X_m) = 1\} + \sup_{M \in \mathcal{M}} \mathbb{P}_M\{f(X_1, \ldots, X_m) = 0\} \,,$$

where $\mathbb{P}_0$ denotes the distribution under the null hypothesis, while $\mathbb{P}_M$ denotes the distribution under the alternative hypothesis associated with a particular covariance structure M. In our setup, $R^{\max}(f)$ depends on n, m, ρ, and the class $\mathcal{C}$ of possible index sets $S \subset \{1, \ldots, n\}$. When all non-zero covariances $\sigma_{i,j}$ are actually equal to the lower bound ρ, then $\mathbb{P}_M$ is determined by S and, with a slight abuse of notation, we write $\mathbb{P}_S$ for $\mathbb{P}_M$. Clearly,

$$R^{\max}(f) \geq \mathbb{P}_0\{f(X_1, \ldots, X_m) = 1\} + \max_{S \in \mathcal{C}} \mathbb{P}_S\{f(X_1, \ldots, X_m) = 0\}$$

and indeed all lower bounds derived in this paper start with this inequality. We will derive upper and lower bounds for the *minimax risk*,

$$R^{\max}_* := \inf_f R^{\max}(f) \,,$$

where the infimum is taken over all measurable functions $f : \mathbb{R}^{nm} \to \{0, 1\}$.

The lower bounds will be obtained by putting a prior on model $\mathcal{C}$ and obtaining a lower bound on the corresponding *Bayesian risk* which never exceeds the worst-case risk. In all cases, we draw the set S uniformly at random within the class $\mathcal{C}$. The upper bounds are obtained by studying the performance of specific tests.

We focus on the case where the dimension n and the sample size m are both large. Of course, such asymptotic statements only make sense if we define sequences of integers $m = m_n$, $k = k_n$, positive reals $\rho = \rho_n$, and classes $\mathcal{C} = \mathcal{C}_n$. This dependency in n will be left implicit. In this asymptotic setting, we say that *reliable* detection is possible (resp. impossible) if $R^{\max}_* \to 0$ (resp. $\to 1$) as $n \to \infty$. Also we say that a sequence of tests (f_n) is asymptotically powerful (resp. powerless) if $R^{\max}(f_n) \to 0$ (resp. $\to 1$).

1.3 A preview of results for the clique model

Among the models we consider, the clique model is perhaps the most compelling because of its relevance in applications and its complexity. Also, for a given value of k, the clique model is the richest possible and therefore for any given ρ, n, m, $R^{\max}_*$ is larger than for any other model. This makes the clique model an important benchmark.

Here we summarize our main findings for this special class. We discover various types of behavior in distinct ranges of the parameters n, m, k, ρ. Roughly speaking, and ignoring logarithmic factors, we arrive at the following conclusions. Two tests are competing for near-optimality. The first one is a 'global' test akin to the classical test Muirhead [24, Sec. 8.4] and the refinements in

Chen, Zhang, and Zhong [13] and Ledoit and Wolf [23]. The second is a 'local' test reminiscent to the generalized likelihood ratio test. The latter dominates the former when

$$\max\left(\frac{k^{3/2}}{n}, \frac{k^2}{n\sqrt{m}}\right)$$

is small, corresponding to smaller values of k.

Our results also uncover an interesting phase-transition phenomenon as the sample size increases. Indeed, we find that, when the sample size m is sub-logarithmic in the dimension n—meaning $m = o(\log n)$—reliable testing is only possible if $\rho \to 1$ or $k^2/n \to \infty$, which is what we found in our previous work [3] when $m = 1$. The situation becomes drastically different when m is at least a sufficiently large constant multiple of $\log n$, where we learn that reliable detection becomes possible in some settings where $\rho \to 0$ and $k = 2$—as long as $\rho \to 0$ sufficiently slowly.

1.3.1 Computational considerations

The "local" test that achieves near-optimal behavior in a large range of the parameters is a scan statistic that requires the computation of a maximum over all $\binom{n}{k}$ subsets of components of size k. In its naive implementation, this test is computationally intractable, unless k is very small. We also believe that computing this test is a fundamentally hard computational problem. We do not have a rigorous argument to prove such a hardness result but it is worth pointing out that the problem is quite similar, in spirit, to the notoriously difficult *hidden clique problem*, see Alon, Krivelevich, and Sudakov [2].

What performance can we achieve with limited computational power? Such questions of trade-off between statistical performance and computational complexity are at the heart of high-dimensional statistics and machine learning. We probe this question and describe a family of tests that balances detection performance and computational complexity.

In particular, in Section 4.5 we design a test that achieves near-optimal performance (similar to that of the scan statistic) and may be computed in polynomial-time in n when $m = O(\log n)$, ρ is a constant, and $k \sim n^a$ for some $a \in (0, 1/2)$. In Section 4.6 we discuss another computationally efficient test based on a convex relaxation of the local test.

1.3.2 An application in the study of random geometric graphs

In Section 7 we apply the lower bound for the optimal risk in the clique model in a perhaps unexpected context and derive a new lower bound for the clique number of a high-dimensional random geometric graph. The setup is as follows.

Consider a random geometric graph on the unit sphere in dimension m. The graph has n vertices, each corresponding to a random point on the unit sphere. Two vertices are connected by an edge if the inner product of the corresponding points is positive. In a recent paper, Devroye, György, Lugosi, and Udina [16] studied the clique number (i.e., the size of the largest clique in the graph) $\omega(n, m)$ of such a graph in various regimes. They showed that when $m \sim c \log n$ for a sufficiently small constant c, $\omega(n, m) = n^{1-o(1)}$ with high probability, while when $m \geq 9 \log^2 n$, $\omega(n, m) = O(\log^3 n)$. However, nothing was known about the behavior of the clique number in between. In particular, it was unclear where exactly the clique number becomes polylogarithmic. In Section 7 we show that the phase transition occurs at $m \asymp \log^2 n$. (We use the notation $a_n \asymp b_n$

when (a_n) and (b_n) are two sequences such that $a_n = O(b_n)$ and $b_n = O(a_n)$.) In particular, we prove that for all $c > 0$, when $m \sim c \log n$, then the median of $\omega(n, m)$ grows as a positive power of n and even for $m \asymp \log^{2-\epsilon} n$, the median of $\omega(n, m)$ grows faster than any power of $\log n$, for all $\epsilon > 0$.

1.4 More related work

As mentioned before, the literature on *sparse* covariance estimation has become quite extensive. In spite of this surge of interest in sparse high-dimensional models, not much has been done in terms of detection of correlations. We note the work of Verzelen and Villers [30], who consider the task of testing a given dependency structure. Our objective here is admittedly more modest and a more closely related is our own paper [3], which focuses entirely on the case where the sample size is equal to one (i.e., $m = 1$). Our results here are seen to extend those in the one-sample case, with the regimes now partitioned according to the sample size.

Note that our work is different from Butucea and Ingster [11] where the task is the detection of a submatrix with higher per-coordinate mean in a large matrix with i.i.d. Gaussian entries, which is more closely related to the literature on the detection of sparse nonzero entries in the mean of a random vector. Our work has parallels with that literature which, for the clique model, focuses on the "detection-of-means" problem (see Jin [22], Ingster [21], Baraud [5], Donoho and Jin [17], Hall and Jin [20], Arias-Castro, Candès, Helgason, and Zeitouni [4], Addario-Berry, Broutin, Devroye, and Lugosi [1]) defined as follows: Under the null hypothesis, the vectors X_t are i.i.d. standard normal, while under the alternative hypothesis, there is a subset $S \subset \{1, \dots, n\}$ in some class $\mathcal{C}$ of interest such that the X_t are i.i.d. normal with mean $(\mu_1, \dots, \mu_n)^T$ and identity covariance, where $\mu_i \geq \mu$ for $i \in S$ and $\mu_i = 0$ for $i \notin S$. Thus $\mu > 0$ is the minimum (per-coordinate) signal amplitude. Of course, one immediately reduces by sufficiency to the case $m = 1$ by averaging over the sample. This explains why the literature focuses on the case $m = 1$. The connection between the detection-of-means problem with the correlation detection problem studied here was detailed (for $m = 1$) in our previous paper [3], where ρ was found to correspond to μ^2. The connection is based on the following simple representation of equi-correlated normal random variables.

Lemma 1 (Berman [6]) *Let $X_1, \dots, X_k$ be standard normal random variables with $\mathrm{Cov}(X_i, X_j) = \rho > 0$ for $i \neq j$. Then there are independent standard normal random variables $V, Y_1, \dots, Y_k$ such that $X_i = \sqrt{\rho}\, V + \sqrt{1-\rho}\, Y_i$ for all i.*

Thus, given V, the problem becomes that of detecting a subset of variables—here implicitly assumed to be indexed by $S = \{1, \dots, k\}$—with nonzero mean (equal to $\sqrt{\rho}\, V$) and with a variance equal to $1 - \rho$ (instead of 1). This representation was used in [3] to obtain a general lower bound that seemed otherwise out of reach of more standard methods based on the second moment of the likelihood ratio.

This connection with the detection-of-means problem also applies in the case where $m > 1$, but with a twist. Indeed, when detecting correlations one does not average the vectors X_t but their covariances. So a simple reduction to the case $m = 1$ does not apply. However, one may still apply the representation result Lemma 1 to each observation vector X_t, yielding V_t's and $Y_{t,i}$'s that are independent standard normal random variables. By conditioning on $V_1, \dots, V_m$, the problem becomes equivalent to detecting a subset of variables with means $\sqrt{\rho}\, V_t$, $t = 1, \dots, m$. What

makes the situation more complex is that the signs of the V_t's are random. Our approach to finding a general lower bound is based on this representation without which more standard methods seem to fail. The general lower bound, which is the key technical result of this paper, is given in Theorem 1 below.

1.5 Contribution and content of the paper

We obtain a general lower bound in Section 2 akin to, but not a straightforward extension of, the lower bound we obtained in [3]. We then study a number of tests that are near optimal in the sense that they come close to achieving the detection lower bound for various models. This is done in Section 3. We then specialize these general results in Sections 4, 5 and 6, to the three models described in Section 1.1. We also discuss computational issues, particularly in the clique model. In Section 7, we apply our general lower bound to the problem of studying the size of the clique number of a random geometric graph on a high-dimensional sphere. We close the paper with a discussion in Section 8 of possible extensions and challenges.

2 Lower bounds

In this section we derive a general lower bound for the minimax risk $R_*^{\max}$. As mentioned in Section 1.2, the first step is to restrict the supremum in the definition of $R^{\max}(f)$ to covariance matrices in which all the nonzero entries are equal to $\rho > 0$ and then lower bound the maximum by an average. In particular, we have $R_*^{\max} \geq R^*$ where $R^* = \inf_f R(f)$ and

$$R(f) \stackrel{\text{def}}{=} \mathbb{P}_0\{f(X_1, \ldots, X_m) = 1\} + \frac{1}{|\mathcal{C}|} \sum_{S \in \mathcal{C}} \mathbb{P}_S\{f(X_1, \ldots, X_m) = 0\} \, .$$

Note that R^* is just the Bayes risk for the uniform prior on the models $S \in \mathcal{C}$. It is well known that the test f^* that achieves the infimum (i.e., $R(f^*) = R^*$) is the *likelihood ratio test* with critical value 1, and there is a whole machinery that can be used to bound that risk from below.

The following lower bound has a similar flavor as the main result in our previous work [3]. In particular, we make appear some moment of Z, a random variable that represents the size of the overlap of two index sets taken at random from the class $\mathcal{C}$. A straightforward adaptation of the arguments we used in [3] leads to a lower bound in terms of the moment generating function of Z. Here, unfortunately, this quantity is too large to obtain sharp results in most regimes. The key contribution of the following result is to replace the exponential function by the hyperbolic cosine. This allows us to derive much sharper results, essentially because around 0 one has $\exp(x) - 1 \sim x$ while $\cosh(x) - 1 \sim \frac{x^2}{2}$.

Theorem 1 *For any class $\mathcal{C}$, any $\rho \in (0,1)$, and any $a \geq \sqrt{3}$,*

$$R^* \geq \mathbb{P}\left(\chi_m^2 \leq ma^2\right) \left(1 - \frac{1}{2}\sqrt{\mathbb{E}\min\left[\exp\left(m\nu_a Z\right), \cosh^m(\xi_a Z)\right] - 1}\right), \tag{2.1}$$

where

$$\nu_a := \frac{\rho a^2}{1+\rho} - \frac{1}{2}\log(1-\rho^2) \quad \textit{and} \quad \xi_a := \frac{\rho a^2}{1-\rho^2},$$

and where χ^2_m has chi-squared distribution with m degrees of freedom, and $Z = |S \cap S'|$ with S, S' i.i.d. uniform from $\mathcal{C}$.

Proof Following [3] we start the proof by using the representation of the data given by Lemma 1. Under the alternative hypothesis H_1, $X \in \mathbb{R}^{m \times n}$ can be written as

$$X_{t,i} = \begin{cases} Y_{t,i} & \text{if } i \notin S, t \in [m] \\ \sqrt{\rho}\, V_t + \sqrt{1-\rho}\, Y_{t,i} & \text{if } i \in S, t \in [m] \end{cases} \tag{2.2}$$

where, for any positive integer m, $[m] := \{1, \ldots, m\}$, and $(Y_{t,i})_{i \in [n], t \in [m]}$, $(V_t)_{t \in [m]}$ are i.i.d. standard normal random variables.

In our previous paper [3], we conditioned on $V := (V_1, \ldots, V_m)$. Here, instead, we condition on $U := (U_1, \ldots, U_m)$, where $U_t := |V_t|$. Let $\varepsilon_t = \text{sign}(V_t)$, which are i.i.d. Rademacher. We consider now the alternative hypothesis $H_1(u)$, defined as the alternative hypothesis H_1 given $U = u \in \mathbb{R}^m$. Let $R(f)$, L, f^* (resp. $R_u(f)$, L_u, f^*_u) be the risk of a test f, the likelihood ratio, and the optimal (likelihood ratio) test, for H_0 versus H_1 (resp. H_0 versus $H_1(u)$). For any $u \in \mathbb{R}^m$, $R_u(f^*_u) \leq R_u(f^*)$, by the optimality of f^*_u for H_0 vs. $H_1(u)$. Therefore, conditioning on U,

$$R^* = R(f^*) = \mathbb{E}_U R_U(f^*) \geq \mathbb{E}_U R_U(f^*_U) = 1 - \frac{1}{2}\mathbb{E}_U \mathbb{E}_0 |L_U(X) - 1| \,.$$

($\mathbb{E}_U$ is the expectation with respect to U.) Using the fact that $\mathbb{E}_0|L_u(X) - 1| \leq 2$ for all u, we have (with $B(0,a)$ being the euclidean ball of radius a in $\mathbb{R}^m$)

$$\mathbb{E}_U \mathbb{E}_0 |L_U(X) - 1| \leq 2\mathbb{P}\{||U|| > a\sqrt{m}\} + \mathbb{P}\{||U|| \leq a\sqrt{m}\} \max_{u \in B(0, a\sqrt{m})} \mathbb{E}_0 |L_u(X) - 1| \,.$$

Therefore, using the Cauchy-Schwarz inequality,

$$\begin{aligned} 1 - \frac{1}{2}\mathbb{E}_U \mathbb{E}_0 |L_U(X) - 1| &\geq \mathbb{P}\{||U|| \leq a\sqrt{m}\} \left(1 - \frac{1}{2} \max_{u \in B(0, a\sqrt{m})} \mathbb{E}_0 |L_u(X) - 1|\right) \\ &\geq \mathbb{P}\{||U|| \leq a\sqrt{m}\} \left(1 - \frac{1}{2} \max_{u \in B(0, a\sqrt{m})} \sqrt{\mathbb{E}_0 L_u^2(X) - 1}\right) \,. \end{aligned}$$

We turn our attention to bounding $\mathbb{E}_0 L_u^2(X)$ from above. Let $L_{u,\varepsilon,S}(x)$ denote the likelihood ratio when S is anomalous, given u and ε, which is equal to

$$L_{u,\varepsilon,S}(x) = \frac{1}{(1-\rho)^{mk/2}} \exp\left(\sum_{t=1}^m \sum_{i \in S} \frac{x_{t,i}^2}{2} - \frac{(x_{t,i} - \sqrt{\rho}\varepsilon_t u_t)^2}{2(1-\rho)}\right) .$$

Since $L_u(x) = \mathbb{E}_\varepsilon \mathbb{E}_S L_{u,\varepsilon,S}(x)$, by Fubini's theorem, we have

$$\mathbb{E}_0 L_u(X)^2 = \mathbb{E}_{S,S'} \mathbb{E}_{\varepsilon,\varepsilon'} \mathbb{E}_0 L_{u,\varepsilon,S}(X) L_{u,\varepsilon',S'}(X)$$

where $\varepsilon, \varepsilon'$ are i.i.d. Rademacher vectors and S, S' are i.i.d. uniform in the class $\mathcal{C}$. We have

$$L_{u,\varepsilon,S}(x) L_{u,\varepsilon',S'}(x) = (1-\rho)^{-mk} \exp(H_1(x) + H_2(x) + H_3(x)),$$

where

$$H_1(x) := \sum_{t=1}^{m} \sum_{i \in S \cap S'} x_{t,i}^2 - \frac{(x_{t,i} - \sqrt{\rho}\varepsilon_t u_t)^2}{2(1-\rho)} - \frac{(x_{t,i} - \sqrt{\rho}\varepsilon'_t u_t)^2}{2(1-\rho)},$$

$$H_2(x) := \sum_{t=1}^{m} \sum_{i \in S \setminus S'} \frac{x_{t,i}^2}{2} - \frac{(x_{t,i} - \sqrt{\rho}\varepsilon_t u_t)^2}{2(1-\rho)},$$

$$H_3(x) := \sum_{t=1}^{m} \sum_{i \in S' \setminus S} \frac{x_{t,i}^2}{2} - \frac{(x_{t,i} - \sqrt{\rho}\varepsilon'_t u_t)^2}{2(1-\rho)}\ .$$

Let $Z = |S \cap S'|$. We see that $H_1(X), H_2(X), H_3(X)$ are independent of each other under the null hypothesis with

$$\mathbb{E}_0 \exp(H_2(X)) = (1-\rho)^{m|S \setminus S'|/2} = (1-\rho)^{m(k-Z)/2}$$

and similarly for $\mathbb{E}_0 \exp(H_3(X))$, while

$$\mathbb{E}_0 \exp(H_1(X)) = \left(\frac{1-\rho}{1+\rho}\right)^{mZ/2} \exp\left(\frac{\rho Z}{1-\rho^2} \sum_{t=1}^{m} \varepsilon_t \varepsilon'_t u_t^2 - \frac{\rho^2 Z}{1-\rho^2} \|u\|^2\right).$$

For the latter, we used the fact that $\varepsilon_t^2 = \varepsilon_t'^2 = 1$, to get

$$\begin{aligned}
& \int_{\mathbb{R}} \exp\left(x_{t,i}^2 - \frac{(x_{t,i} - \sqrt{\rho}\varepsilon_t u_t)^2}{2(1-\rho)} - \frac{(x_{t,i} - \sqrt{\rho}\varepsilon'_t u_t)^2}{2(1-\rho)}\right) \exp(-x_{t,i}^2/2) \frac{dx_{t,i}}{\sqrt{2\pi}} \\
= & \int_{\mathbb{R}} \exp\left(\frac{\rho(\varepsilon_t\varepsilon'_t - \rho)u_t^2}{1-\rho^2} - \frac{1+\rho}{2(1-\rho)} \left(x_{t,i} - \frac{\sqrt{\rho}}{1+\rho}(\varepsilon_t + \varepsilon'_t)u_t\right)^2\right) \frac{dx_{t,i}}{\sqrt{2\pi}} \\
= & \sqrt{\frac{1-\rho}{1+\rho}} \exp\left(\frac{\rho(\varepsilon_t\varepsilon'_t - \rho)u_t^2}{1-\rho^2}\right)\ ,
\end{aligned}$$

where the last line comes from a simple change of variables. Hence,

$$\mathbb{E}_0 L_{u,\varepsilon,S}(X) L_{u,\varepsilon',S'}(X) = (1-\rho^2)^{-mZ/2} \exp\left(\frac{\rho Z}{1-\rho^2} \sum_{t=1}^{m} \varepsilon_t \varepsilon'_t u_t^2 - \frac{\rho^2 Z}{1-\rho^2} \|u\|^2\right).$$

Let $\xi = \xi_1 = \rho/(1-\rho^2)$. Since $(\varepsilon_t\varepsilon'_t : t = 1, \ldots m)$ are i.i.d. Rademacher, we have

$$\mathbb{E}_{\varepsilon,\varepsilon'} \exp\left(\xi Z \sum_{t=1}^{m} \varepsilon_t \varepsilon'_t u_t^2\right) = \prod_{t=1}^{m} \cosh\left(\xi Z u_t^2\right)\ ,$$

so that

$$\mathbb{E}_{\varepsilon,\varepsilon'} \mathbb{E}_0 L_{u,\varepsilon,S}(X) L_{u,\varepsilon',S'}(X) = (1-\rho^2)^{-mZ/2} \prod_{t=1}^{m} \cosh(\xi Z u_t^2) \exp(-\rho \xi Z u_t^2)\ .$$

Holding $Z \geq 1$ fixed, we maximize this over $\|u\|^2 = \sum_t u_t^2 \leq a^2 m$ using Lagrangian multipliers and checking the Karush-Kuhn-Tucker conditions, finding that at a local maximum all u_t^2 must be equal. Hence,

$$\prod_{t=1}^{m} \cosh(\xi Z u_t^2) \exp(-\rho \xi Z u_t^2) \leq \left(\max_{0 \leq c \leq \xi_a Z} \cosh(c) \exp(-\rho c) \right)^m \tag{2.3}$$

$$= \max\left(1,\ \cosh^m(\xi_a Z) \exp(-m\rho\xi_a Z)\right), \tag{2.4}$$

where the last equality comes from the fact that the function $h_\rho(c) := \cosh(c)\exp(-\rho c)$ is decreasing on $(0, \rho)$ and increasing on (ρ, ∞), so that its maximum over $[0, \xi_a Z]$ is either at $c = 0$ or $c = \xi_a Z$. Straightforward calculations lead to

$$h_\rho(c) > 1 \Leftrightarrow g(c) := \frac{1}{c} \log\cosh(c) > \rho\,.$$

Since $g(c) > 1 - \frac{1}{c}\log 2$, the maximum of $h_\rho(c)$ over $c \in [0, \xi_a Z]$ is at $c = \xi_a Z$ when

$$1 - \frac{1}{\xi_a Z} \log 2 \geq \rho \Leftrightarrow a^2 Z \geq \frac{1+\rho}{\rho} \log 2.$$

Since we consider $Z \geq 1$, the last inequality is true if $\rho \geq 1/2$ and $a^2 \geq 3\log(2)$. This inequality is far off when ρ is small, so we need to derive another bound. Noting that g is seen to be strictly increasing on $(0, \infty)$ with range $(0, 1)$, $w(\rho) := g^{-1}(\rho)$ is well-defined and, as a function of ρ, is infinitely differentiable and strictly increasing. Elementary calculations show that

$$w' = \frac{w}{\tanh(w) - \rho},\ w'' = \frac{w\tanh^2(w)}{(\tanh(w)-\rho)^2},$$

implying in particular that w is convex. Also, since $g(c) \geq \frac{1}{c}\log(1+\frac{c^2}{2})$—because $e^c \geq 1+c+c^2/2$ for all $c \geq 0$ and $e^c \geq 1+c$ for all $c \in \mathbb{R}$—we have $g(2) \geq \frac{1}{2}\log(3) > \frac{1}{2}$ and therefore $w(1/2) < 2$. Hence, by convexity of w, we have $w(\rho) \leq w(1/2)\rho < 2\rho$ for all $\rho \leq 1/2$. Now, the maximum in (2.3) at $c = \xi_a Z$ when

$$w(\rho) < \xi_a Z \Leftarrow 2\rho \leq \frac{\rho a^2}{1-\rho^2} Z \Leftrightarrow 2(1-\rho^2) \leq a^2 Z.$$

When $Z \geq 1$, the latter is true when $\rho \leq 1/2$ and $a^2 \geq 3/2$. Hence, given that $a^2 \geq 3$ by assumption, the maximum in (2.3) at $c = \xi_a Z$.

We therefore arrive at

$$\max_{u \in B(0, a\sqrt{m})} \mathbb{E}_0 L_u^2(X) \leq \mathbb{E}\left((1-\rho^2)^{-mZ/2} \cosh^m(\xi_a Z)\exp(-m\rho\xi_a Z)\right)\,.$$

We then have

$$\cosh^m(\xi_a Z)\exp(-m\rho\xi_a Z) \leq \exp(m(1-\rho)\xi_a Z) = \exp\left(\frac{m\rho a^2}{1+\rho} Z\right).$$

And also

$$(1-\rho^2)^{-mZ/2}\exp(-m\rho\xi_a Z) \le \exp\left(-\frac{mZ}{2(1-\rho^2)}\left[2\rho^2+(1-\rho^2)\log(1-\rho^2)\right]\right)$$
$$\le 1,$$

where in the first line we used $a^2 \ge 1$ and in the second line the fact that $s + \frac{1-s}{2}\log(1-s) \ge 0$ for all $s \in (0,1)$. With this, we conclude. ■

In Sections 4, 5, and 6, we specialize Theorem 1 to the different models we described in Section 1.1.

3 Tests

In this section we introduce and briefly discuss two natural tests that will be seen to perform near optimally in various regimes of the parameters. This optimality property will be established in Sections 4, 5, and 6, by comparing simple performance bounds with the implications of Theorem 1.

The first test, that we call "squared-sum test", is based on a global test statistic that does not take the class $\mathcal{C}$ into account at all.

The second test, a "localized" squared-sum test, is based on a simple scan statistic. It may also be interpreted as a simplified version of the generalized likelihood ratio test.

As we will see, one of the two tests above always has a near-optimal performance in all three specific classes we discuss. Thus, the story is essentially complete for the point of view of detection performance. Unfortunately, when the class $\mathcal{C}$ is large—as in the clique model—, the localized squared-sum test is computationally unfeasible, at least in its naive implementation. We discuss two possible substitutes. The first one is a simple "maximum correlation test" that turns out to be nearly optimal for very small values of k. In Section 4.5 we discuss another test in the context of the clique model that is both near-optimal and computationally feasible when the sample size is at most logarithmic in the dimension n. In Section 4.6 a conceptually different computationally efficient alternative is discussed.

All performance bounds derived below are in terms of the average correlation

$$\rho_{\text{ave}} = \frac{1}{k(k-1)} \sum_{i,j\in S: i\neq j} \sigma_{i,j} \ge \rho\,, \tag{3.1}$$

where S is the anomalous set.

3.1 The squared-sum test

Let Σ denote the covariance matrix of the distribution of X_1. When the alternative hypothesis is simply $H_1 : \Sigma \neq I$ (where I is the identity matrix), without any sign restriction on the entries of Σ, one of the simplest tests is that of Nagao [25], which is based on the Frobenius norm of the difference between the sample covariance matrix $\widehat{\Sigma}$ and the identity matrix I. Nagao's test is based on the test statistic

$$\frac{1}{n}\|\widehat{\Sigma} - I\|_F^2 = \frac{1}{n}\,\mathrm{tr}[(\widehat{\Sigma} - I)^2].$$

We also refer to Schott [26], who (like us) assumes that the variables have unit variance under the alternative hypothesis. Ledoit and Wolf [23] show that this test is not always consistent against fixed alternatives. They, and others including Srivastava [27], Fisher [19], and Chen, Zhang, and Zhong [13], suggest variants based on consistent estimates for the Frobenius norm $\mathrm{tr}[(\Sigma - I)^2]$.

Given that we know that the variances are equal to 1 and the correlations are non-negative under the alternative hypothesis, it is more natural to consider the test that rejects for large values of $\sum_{i<j} \widehat{\sigma}_{i,j}$, where $\widehat{\Sigma} = (\widehat{\sigma}_{i,j})$. For simplicity, we consider instead the *squared-sum test* that rejects for large values of the test statistic

$$Y = \sum_{t=1}^{m} \left(\sum_{i=1}^{n} X_{t,i} \right)^2 = m \sum_{i,j=1}^{n} \widehat{\sigma}_{i,j} \; . \tag{3.2}$$

The two tests are thus closely related. In fact, one may easily check that they have similar asymptotic power properties. Our preference for the second test is only for convenience.

The following result gives a simple characterization of the performance of the squared-sum test. Since the test does not use information about the class $\mathcal{C}$, its minimax risk does not depend on the model either.

Proposition 1 *The squared-sum test that rejects H_0 when $Y \geq n(m + a\sqrt{m})$ is asymptotically powerful when $a \to \infty$ such that $a \leq \frac{1}{2}\sqrt{m}$ and $\rho_{\mathrm{ave}}\sqrt{m}\,k^2/n \geq 5a$. The squared-sum test with any threshold value for Y is asymptotically powerless when $\rho_{\mathrm{ave}}\sqrt{m}\,k^2/n \to 0$.*

To interpret this result, note that in a very high-dimensional setting (i.e., when m of smaller order than any positive power of n), and with constant correlation (i.e., ρ_{ave} is a numerical constant), the proposition states that the squared-sum test is powerful in a non-sparse regime where $k \gg \sqrt{n}$. On the other hand in sparse situations with constant correlation, the sample size must be very large for the squared-sum test to be powerful, as one needs $m \gg n^2/k^4$. We will see in the next section that in this sparse regime other tests perform much better.

Proof Under the null hypothesis, $Y \sim n\chi^2_m$, and therefore

$$\mathbb{P}_0(Y > n(m + a\sqrt{m})) \to 0.$$

Under the alternative hypothesis, $Y \sim (1+b)n\chi^2_m$, where $b := \rho_{\mathrm{ave}}\, k(k-1)/n$, and hence

$$\mathbb{P}_1(Y \leq (1+b)n(m - a\sqrt{m})) \to 0, \quad n \to \infty.$$

Using the assumptions on a and b, we have

$$\begin{aligned}
(1+b)n(m - a\sqrt{m}) - n(m + a\sqrt{m}) &\geq n\sqrt{m}\left(b\sqrt{m} - (2+b)a\right) \\
&\geq n\sqrt{m}\left(\frac{1}{2}b\sqrt{m} - 2a\right) \\
&\geq n\sqrt{m}\frac{a}{2} > 0,
\end{aligned}$$

and therefore the test with critical value $n(m + a\sqrt{m})$ is asymptotically powerful.

Suppose that $b\sqrt{m} \to 0$. We still have that $Y/n \sim \chi^2_m$ under H_0 while $Y/n \sim (1+b)\chi^2_m$ under H_1. If m is fixed, Y/n is asymptotically χ^2_m under the alternative hypothesis since $b \to 0$ in this

case. If $m \to \infty$, $(Y/n - m)/\sqrt{2m}$ is asymptotically standard normal under both the null and the alternative hypotheses, since under H_1,

$$\frac{Y/n - m}{\sqrt{2m}} = \frac{Z - m}{\sqrt{2m}} + \frac{bZ}{\sqrt{2m}} = \frac{Z - m}{\sqrt{2m}} + O_P(b\sqrt{m}),$$

where $Z = Y/(n(1+b)) \sim \chi^2_m = O_P(m)$. We may apply Slutsky's theorem to conclude. ■

3.2 A localized squared-sum test

When k is smaller, global tests such as the squared-sum test are not very powerful. The generalized likelihood ratio test "scans" over all subsets S in the class $\mathcal{C}$. Instead of studying the generalized likelihood ratio test, we consider a localized version of the squared-sum test that has similar power and is a little easier to analyze. The *localized squared-sum test* rejects the null hypothesis for large values of the test statistic

$$Y_{\text{scan}} = \max_{S \in \mathcal{C}} Y_S, \quad \text{where} \quad Y_S = \sum_{t=1}^{m} \left(\sum_{i \in S} X_{t,i} \right)^2 = m \sum_{i,j \in S} \widehat{\sigma}_{i,j} \ .$$

The following result gives sufficient conditions for the test to be asymptotically powerful. The conditions are in terms of the cardinality of the class $\mathcal{C}$. Sharper bounds that take into account the fine metric structure of $\mathcal{C}$ are also possible by more careful bounding of the distribution of Y_{scan} under the null hypothesis. However, as we will see below, this bound is already quite sharp for the specific classes considered in this paper, and to preserve relative simplicity of the arguments we do not pursue sharper bounds here.

Proposition 2 *The localized squared-sum test that rejects the null hypothesis if $Y_{\text{scan}} > bkm$, is asymptotically powerful when $\rho_{\text{ave}} k \geq 3(b-1)$ and either*

$$\frac{\log |\mathcal{C}|}{m} \to 0 \quad \textit{and} \quad b \geq 1 + \sqrt{\frac{5 \log |\mathcal{C}|}{m}} \ ,$$

or

$$\frac{\log |\mathcal{C}|}{m} \to \infty \quad \textit{and} \quad b \geq \frac{3 \log |\mathcal{C}|}{m} \ .$$

Note that the proposition makes a distinction between a *ultra-high* dimensional setting where $m \ll \log |\mathcal{C}|$, and a (potentially) high dimensional setting where $m \gg \log |\mathcal{C}|$. In the former case, the result states that k needs to be large enough for the localized squared-sum test to be powerful. On the other hand in the other less extreme regime the correlation ρ_{ave} can be as small as $\frac{1}{k}\sqrt{\frac{\log |\mathcal{C}|}{m}}$.

In particular in the sparse regime where k is a constant the result above ensures that asymptotic power is achieved for ρ_{ave} as small as $\sqrt{\frac{\log n}{m}}$. In the non-sparse regime, for large classes $\mathcal{C}$ (i.e., such that $\log |\mathcal{C}|$ is roughly $k \log n$), asymptotic power is achieved in the ultra-high dimensional setting (that is $m \ll k \log n$) for ρ_{ave} as small as $\frac{\log n}{m}$, while in the high dimensional setting (that is $m \gg k \log n$) it is achieved for ρ_{ave} as small as $\sqrt{\frac{\log n}{km}}$.

Other consequences of this proposition will be discussed in the next sections.

Proof Observe that under the null hypothesis $Y_S \sim k\chi^2_m$ for all $S \in \mathcal{C}$. By a simple Chernoff bound for the chi-square distribution, for all $b > 1$,

$$\mathbb{P}\left(\chi^2_m > bm\right) \le \exp\left(-\frac{m}{2}H(b)\right), \tag{3.3}$$

where $H(b) := b - 1 - \log b$ for $b > 1$. Hence, by the union bound,

$$\mathbb{P}_0(Y_{\text{scan}} > bkm) \le |\mathcal{C}| \exp\left(-\frac{m}{2}H(b)\right). \tag{3.4}$$

When $(\log|\mathcal{C}|)/m \to 0$, using the fact that $H(b) \sim (b-1)^2/2$ when $b \to 1$, we see that the right-hand side in (3.4) tends to zero when $b \ge 1 + \sqrt{5\log|\mathcal{C}|/m}$. When $(\log|\mathcal{C}|)/m \to \infty$, using the fact that $H(b) \sim b$ when $b \to \infty$, so the right-hand side in (3.4) tends to zero when $b \ge 3\log|\mathcal{C}|/m$.

Under the alternative hypothesis where S is anomalous—that is, under $\mathbb{P}_S$—, $Y_S \sim (k + \rho_{\text{ave}}\, k(k-1))\chi^2_m$, and therefore

$$Y_{\text{scan}} \ge Y_S > km(1 + \rho_{\text{ave}}(k-1))(1 - O_P(1/\sqrt{m})). \tag{3.5}$$

Hence, the test is asymptotically powerful when $(1 + \rho_{\text{ave}}(k-1))(1 - O_P(1/\sqrt{m})) > b$, and we can check that $\rho_{\text{ave}}k \ge 3(b-1)$ implies this in both regimes. ■

The case when $\log|\mathcal{C}|/m \asymp 1$ can be dealt with in the same way, yielding that, with a proper choice of threshold, the localized squared-sum test is asymptotically powerful when

$$\rho_{\text{ave}}k \ge A \max\left(\sqrt{\frac{\log|\mathcal{C}|}{m}}, \frac{\log|\mathcal{C}|}{m}\right), \tag{3.6}$$

for a sufficiently large constant A.

When the class $\mathcal{C}$ is large (i.e., has size exponential in k), the test statistic Y_{scan} may be difficult to compute as it involves solving a nontrivial combinatorial optimization problem. This is the case for the clique model (unless k is very small) and the matching model.

3.3 Maximum correlation test

Finally, we mention the possibly simplest test that one would think of when confronted with testing H_0 in the sparse regime. This is the test that rejects for large values of the maximum pairwise empirical correlation

$$Y_{\max} = \max_{i \ne j} \sum_{t=1}^{m} X_{t,i}X_{t,j}\,.$$

In fact, this test does have some power in the sparse regime, and is actually near-optimal when k is fixed as the following result shows. However, one cannot expect a good performance of this test for large values of k. An advantage of this test is that it may be computed efficiently in a straightforward manner.

Proposition 3 *The maximum correlation test that rejects H_0 when $Y_{\max} > \sqrt{5m \log n}$ is asymptotically powerful when*

$$\rho_{\text{ave}} \geq \sqrt{5(\log n)/m} \, . \tag{3.7}$$

Note that the performance described in this proposition matches the one of the localized squared-sum test for the sparse regime where k is a constant. Moreover the performance is better than the squared-sum test for k is small enough, precisely when $k^2 \ll \sqrt{\log n}/n$.

Proof Assume that $m \geq 5 \log n$ for otherwise the condition for ρ_{ave} is vacuous. For $i \neq j$ fixed, under the null hypothesis, $X_{t,i}X_{t,j}, t = 1, \dots, m$ are i.i.d. with zero mean, unit variance, and finite moment generating function in a neighborhood of the origin. In fact, it is equal to $(1-\lambda^2)^{-1/2}$ for $\lambda \in (-1, 1)$. Hence, by a standard result on moderate deviations, see, e.g., Dembo and Zeitouni [15, Th 3.7.1],

$$\limsup_{m \to \infty} \frac{m}{b_m^2} \log \mathbb{P}_0 \left\{ \sum_{t=1}^{m} X_{t,i}X_{t,j} > b_m \right\} \leq -\frac{1}{2}$$

for any sequence (b_m) such that $\sqrt{m} = o(b_m)$ and $b_m/m \to 0$. We choose $b_m = \sqrt{5m \log n}$ and use the union bound, to get

$$\mathbb{P}_0\{Y_{\max} > \sqrt{5m \log n}\} \leq \binom{n}{2} \mathbb{P}_0 \left\{ \sum_{t=1}^{m} X_{t,i}X_{t,j} > b_m \right\} \to 0 \, .$$

Under the alternative hypothesis when $S \subset [n]$ is anomalous, pick $i \neq j$ in S such that $X_{t,i}X_{t,j}, t = 1, \dots, m$ are i.i.d. with mean larger than ρ_{ave} and variance smaller than 2, and by Chebyshev's inequality,

$$\sum_{t=1}^{m} X_{t,i}X_{t,j} = m\rho_{\text{ave}} + O_P(\sqrt{m}) \, .$$

From this, the result follows immediately. ■

4 Clique model

In this section we discuss the implications of the general results of the previous sections for the clique model. We derive a lower bound based on Theorem 1 in various ranges of the parameters and compare it with the performance bounds for the squared-sum test and the scan statistics-based test considered in Section 3. We also propose a goodness-of-fit test for the case where $\rho \to 1$, and consider two alternative tests that take computational considerations into account. A digest is provided at the end of the section.

4.1 Lower bounds for the clique model

In order to apply Theorem 1, note that in the clique model, Z has hypergeometric distribution with parameters (n, k, k), which is stochastically bounded by the binomial distribution with parameters

(k,p), where $p := k/(n-k)$. In particular, for all $\xi \geq 0$,

$$\mathbb{E}\exp(\xi Z) \leq (1-p+pe^{\xi})^k, \tag{4.1}$$

by Bennett's inequality, for all $t > 1$,

$$\mathbb{P}\left(Z \geq tkp\right) \leq \exp\left(-\frac{kp}{1-p}H(t)\right), \tag{4.2}$$

where $H(t) := t(\log t - 1) + 1$. Note that $H(t) \sim t^2/2$ when $t \to 0$ and $H(t) \sim t\log t$ when $t \to \infty$.

We distinguish various regimes of the parameters in which the minimax risk and the optimal test behave differently. Note that Theorem 1 implies that reliable detection is impossible (i.e., $R^* \to 1$) whenever one can choose a such that $a \to \infty$ and either $\limsup \mathbb{E}\left[\cosh^m(\xi_a Z)\right] \leq 1$ or $\limsup \mathbb{E}\exp(m\nu_a Z) \leq 1$. This is what we do in all cases listed below.

Case 1: large k. Suppose that k is so large and ρ is so small that

$$\frac{k}{n} \to 0 \quad \text{and} \quad \frac{k^2}{n} \to \infty \quad \text{and} \quad \rho\sqrt{m}\,\frac{k^2}{n} \to 0. \tag{4.3}$$

We first note that these conditions imply that $\rho \to 0$. Let $\zeta = \rho\sqrt{m}k^2/n$ and choose $a, b \to \infty$ such that $a^2 b\zeta \to 0$. When $Z \leq bk^2/n$, we use the fact that $\rho a^2 bZ \to 0$ and $\cosh(x) \leq 1 + x^2$ when $x \in (0,1)$ to get that for all sufficiently large n,

$$\cosh^m(\xi_a Z) \leq \left(1 + (\xi_a Z)^2\right)^m \leq \exp(m\xi_a^2 Z^2) = 1 + o(1),$$

since $\sqrt{m}\xi_a Z \leq \sqrt{m}\frac{\rho a^2}{1-\rho^2}\frac{bk^2}{n} = O(\zeta a^2 b) = o(1)$.

We now show that, if, in addition to (4.3), we have either $\rho m \to 0$ or $\rho^2 mk \to 0$, then

$$\mathbb{E}\left[\cosh^m(\xi_a Z)\mathbb{1}_{\{Z > bk^2/n\}}\right] = o(1).$$

This implies that reliable detection is impossible in this range of the parameters.

Case 1(a). In addition to (4.3), assume

$$\rho m \to 0\,. \tag{4.4}$$

We choose a such that $a^2\rho m \to 0$. We use the bound $\cosh(x) \leq \exp(x)$ and (4.2), to get

$$\mathbb{E}\left[\cosh^m(\xi_a Z)\mathbb{1}_{\{Z > bk^2/n\}}\right] \leq \sum_{z > bk^2/n} \exp\left(m\xi_a z - \frac{kp}{1-p}H\left(\frac{z}{kp}\right)\right).$$

where $p = k/(n-k)$. We have $m\xi_a \sim m\rho a^2 = o(1)$ and $\frac{kp}{1-p}H\left(\frac{z}{kp}\right) \sim z\log\left(\frac{z}{k^2/n}\right) \geq z\log(b)$ uniformly over $z > bk^2/n$. Hence, eventually,

$$\begin{aligned}\mathbb{E}\left[\cosh^m(\xi_a Z)\mathbb{1}_{\{Z > bk^2/n\}}\right] &\leq \sum_{z > bk^2/n} \exp\left(-\frac{1}{2}z\log(b)\right) \\ &\sim \exp\left(-\frac{1}{2}b(k^2/n)\log(b)\right) = o(1)\,.\end{aligned}$$

Case 1(b). In addition to (4.3), assume

$$\rho^2 mk \to 0\,. \tag{4.5}$$

We may assume that $k \le m$ for otherwise $\rho m \to 0$, which we already covered. We choose a such that $a^2\rho\sqrt{mk} \to 0$—which implies in particular that $a^2\rho k \to 0$. We use the bounds $\cosh(x) \le 1+x^2$ for $x \in [0,1]$, the fact that $Z \le k$—since $Z = |S \cap S'|$ with $|S| = |S'| = k$—and the fact that $a^2\rho k \to 0$, to get

$$\cosh^m(\xi_a Z) \le \left(1 + (\xi_a Z)^2\right)^m \le \exp(m\xi_a^2 Z^2) \le \exp(2a^4\rho^2 m Z^2) \le \exp(2a^4\rho^2 mkZ),$$

eventually. Since $a^4\rho^2 mk = o(1)$, we apply (4.2) and proceed exactly as before, reaching the same conclusion.

Case 2: small k, moderate m. Suppose

$$\frac{k^2}{n} \to 0 \quad \text{and} \quad \rho\sqrt{\frac{mk}{\log(n/k^2)}} \to 0 \quad \text{and} \quad \frac{k\log(n/k^2)}{m} \to 0. \tag{4.6}$$

Let $\zeta = \rho\sqrt{mk/\log(n/k^2)}$ and choose $a \to \infty$ such that $a\zeta \to 0$ and $a^2\rho k \to 0$. The latter is possible because (4.6) implies that $\rho k \to 0$. Then, as in Case 1(b),

$$\cosh^m(\xi_a Z) \le \exp(2a^4\rho^2 mkZ)\,.$$

We then use (4.1) to get

$$\begin{aligned}
\mathbb{E}\cosh^m(\xi_a Z) &\le \left(1 + \frac{k}{n-k} e^{2a^4 m\rho^2 k}\right)^k \\
&\le \exp\left(2\frac{k^2}{n} e^{2a^4 m\rho^2 k}\right) \\
&= \exp\left(2\exp\left(2a^4 m\rho^2 k - \log(n/k^2)\right)\right) \\
&= 1 + o(1)
\end{aligned}$$

and again, reliable detection is impossible by Theorem 1.

Case 3: small k, small m. Suppose

$$\frac{k^2}{n} \to 0 \quad \text{and} \quad \rho\frac{m}{\log(n/k^2)} \to 0 \quad \text{and} \quad \limsup \rho < 1\,. \tag{4.7}$$

Hence, there is some $\varepsilon > 0$ fixed such that $\rho < 1-\varepsilon$. Let $\zeta = \rho m/\log(n/k^2)$ and choose $a \to \infty$ such that $a^2\zeta \to 0$. We use the fact that $\cosh(x) \le \exp(x)$, and use the same bound on the moment generating function of Z, to get (eventually)

$$\begin{aligned}
\mathbb{E}\cosh^m(\xi_a Z) &\le \mathbb{E}\exp(m\xi_a Z) \\
&\le \left(1 + \frac{k}{n-k} e^{ma^2\rho/\varepsilon}\right)^k \\
&= \exp\left(2\exp\left(a^2 m\rho/\varepsilon - \log(n/k^2)\right)\right) \\
&= 1 + o(1),
\end{aligned}$$

implying that reliable detection is impossible.

Case 4: very large ρ. Suppose

$$(1-\rho)^{1/2}\left(\frac{n}{k^2}\right)^{1/m} \to \infty \quad \text{and} \quad \rho \to 1. \tag{4.8}$$

This is the only situation where we bound

$$\begin{aligned}\mathbb{E}\exp(m\nu_a Z) &\le \left(1+\frac{k}{n-k}e^{m\nu_a}\right)^k \\ &\le \exp\left(2\frac{k^2}{n}e^{m\nu_a}\right) \\ &\le \exp\left(2\exp\left(m(a^2-\zeta)\right)\right)\ ,\end{aligned}$$

where $\zeta := \frac{1}{2}\log(1-\rho^2) + \frac{1}{m}\log(n/k^2) \to \infty$ by (4.8). Therefore, it suffices to choose $a \to \infty$ such that $a^2 = o(\zeta)$, to have the last expression on the right-hand side tend to zero.

The discussion of these various regimes leads to the following.

Corollary 1 *In the clique model, under either* (4.3) *with* (4.4) *or* (4.5)*,* (4.6)*,* (4.7)*, or* (4.8)*,* $R^* \to 1$.

We will see below that (4.3), combined with either (4.4) or (4.5), is tight up to logarithmic factors. This is also the case of (4.6) and (4.7), unless $k^2/n \to 0$ as a negative power of n. Note also that the result is silent in the regime when $k^2/n \to 0$ and $\rho\sqrt{m}\,k^2/n \to 0$. However, it is covered by (4.6) when $\log(n/k^2)/k \to 0$, and by (4.7) when $\log(n/k^2)/\sqrt{m} \to 0$, so again it is a matter of logarithmic factors. The bound (4.8) is also tight up to log factors when $m = o(\log n)$. We mention that the typical exposition in the detection-of-means literature, for example in Donoho and Jin [17], avoids the discussion of such fine details by assuming that $k = n^\alpha$ for some $\alpha \in (0,1)$.

4.2 Localized squared-sum test

Next we take a closer look at the performance of the localized squared-sum test for the clique model. In this case we have $|\mathcal{C}| = \binom{n}{k}$ so $\log|\mathcal{C}| \sim k\log(n/k)$. Plugging this into (3.6),we see that the local squared-sum test is asymptotically powerful when

$$\rho_{\text{ave}} \ge A\max\left(\frac{\log(n/k)}{m}, \sqrt{\frac{\log(n/k)}{km}}\right)\ , \tag{4.9}$$

and the constant A is large enough. Based on this and Corollary 1, we conclude that the test is near-optimal in regimes (4.6) and (4.7), though only up to a logarithmic factor if $k^2/n \to 0$ slower than any power of n. It is also near-optimal up to a logarithmic factor in regime (4.3) when neither (4.4) nor (4.5) is satisfied.

However, we do not have such a guarantee in the regime (4.3) (with either (4.4) or (4.5)). In this range of parameters, it is the squared-sum test that yields an optimal performance up to a

logarithmic factor. Also, comparing Proposition 1 and Proposition 2, we see that the local test dominates when $\max\left(1, (k/m)^{1/2}\right)\ k^{3/2}/n$ tends to zero faster than $1/\log(n/k)$.

As we mentioned earlier, computing the scan statistic Y_{scan}, or even approximating with enough precision, seems to be fundamentally hard. In fact, we conjecture that computing *any* test with a near-optimal performance is fundamentally hard in some range of the parameters. We see this is as a challenging and important research problem.

We make some progress in this direction in two ways. In Section 4.5, we suggest a test that has good performance and that is efficiently computable if m is only logarithmic in n. In Section 4.6, inspired by recent work of Berthet and Rigollet [7], we consider a convex relaxation of the problem following d'Aspremont, El Ghaoui, Jordan, and Lanckriet[14].

4.3 The case of ρ constant

Assume that $\rho < 1$ is a constant, independent of n. From our previous work [3] in the case of $m = 1$, we know that $R^* \to 1$ unless $k^2/n \to \infty$, in which case the squared-sum test is asymptotically powerful. Now we learn from Corollary 1 (4.7) that $R^* \to 1$ when $k^2/n \to 0$ and $m = o(\log(n/k^2))$. Hence, in the case of ρ constant and n/k^2 a positive power of n, a sample size m sub-logarithmic in the dimension n is not enough for reliable detection, and is qualitatively on par with the case of $m = 1$.

The situation changes dramatically when the sample size m becomes at least logarithmic in the dimension n. Indeed, even for $k = 2$, both the localized squared-sum test and the maximum correlation test have a vanishing risk for any constant value of ρ when $\log(n)/m \to 0$. This reveals an interesting "phase transition" occurring when the sample size is about logarithmic in the dimension.

4.4 The case of ρ tending to 1

The regime in (4.8) does not have a match in either the squared-sum test or the localized squared-sum test. It is instead met by a goodness-of-fit test which is a variant of test proposed and analyzed in our previous work [3] in the same regime with $m = 1$, although the construction here is slightly different.

Here we assume that $\rho \to 1$ so fast that

$$\frac{k}{m \log n} \to \infty \ \text{ and } \ \frac{k^2}{n} \to 0 \ \text{ and } \ m \le n \ \text{ and } \ (1-\rho)^{1/2} \left(\frac{n}{k^2} \log n\right)^{1/m} \sqrt{m} \to 0. \tag{4.10}$$

Let ζ denote the last term tending to zero in (4.10), and choose $a \to \infty$ such that $a \max(\zeta, 1/\log n) \to 0$.

The test we propose is based on the idea that the variables that are positively correlated are closer together than the other variables that are independent of each other. Take $\eta \to 0$ such that

$$\eta = \max\left(((m+1)\log(n)/n)^{1/m},\ 2a(1-\rho)^{1/2}\sqrt{m}\right)\ , \tag{4.11}$$

which is possible by (4.10). Let $w_1, \ldots, w_\ell \in \mathbb{R}^m$ be an η-covering of $B(0, a\sqrt{m})$ with $\ell \le (a\sqrt{m}/\eta)^m$, and let $R_s = B(w_s, 2\eta)$. We count the number of data points in each R_s, yielding $B_s = \#\{i : X_i \in R_s\}$, and consider the test that rejects for large values of $\max\{B_s : s = 1, \ldots, \ell\}$.

Under the null hypothesis, $B_s \sim \mathrm{Bin}(n, p_s)$ where

$$p_s := (2\pi)^{-m/2} \int_{R_s} e^{-\|x\|^2/2} \mathrm{d}x \leq p := \eta^m.$$

A simple combination of Bernstein's inequality and the union bound gives

$$\max_s B_s \leq np + \sqrt{3np \log \ell},$$

when $\log \ell = o(np)$, which is the case when $m \log(a\sqrt{m}/\eta) = o(n\eta^m)$ and is implied under (4.11), since under our assumptions

$$\begin{aligned} m \log(a\sqrt{m}/\eta) - n\eta^m &\leq m \log(a\sqrt{m}) + \log n - (m+1) \log n \\ &= m \left(\log a + \frac{1}{2} \log m - \log n \right) \to -\infty. \end{aligned}$$

Under the alternative hypothesis where S is anomalous, we use the representation (2.2), to get

$$\|X_i - \sqrt{\rho} V\| = (1-\rho)^{1/2} \|Y_i\| \ ,$$

with

$$\mathbb{P}\left(\|Y_i\| \leq a\sqrt{m} \right) \geq 1 - \frac{1}{a^2} \ ,$$

by Markov's inequality. Hence, by another application of Markov's inequality,

$$\mathbb{P}\left\{ \#\left\{ i \in S : \|Y_i\| \leq a\sqrt{m} \right\} \geq \frac{k}{2} \right\} \geq 1 - \frac{2}{a^2} \to 1 \ .$$

Let s be such that $V \in B(w_s, \eta)$, so by the triangle inequality, the region R_s contains at least $k/2$ anomalous vectors X_i with high probability. Reasoning as in [3], we deduce that, in that case, $B_s \geq np + k/2 - O_P(\sqrt{np})$.

Hence, the test is asymptotically powerful when $\frac{k}{2} \geq \sqrt{4np \log \ell}$, which is the case when $n\eta^m m \log(a\sqrt{m}/\eta) = o(k^2)$. This is seen easily because when $\eta = 2(\log(n)/n)^{1/m}$, we have

$$\begin{aligned} n\eta^m m \log(a\sqrt{m}/\eta) &\leq (m+1)^2 (\log n) \left(\log(a\sqrt{m}) + \log n \right) \\ &= O(m \log n)^2 = o(k^2 = \ . \end{aligned}$$

Finally, when $\eta = 2a(1-\rho)^{1/2}\sqrt{m} \geq 2(\log(n)/n)^{1/m}$, we have

$$\begin{aligned} n\eta^m m \log(a\sqrt{m}/\eta) &\leq n \left(2a(1-\rho)^{1/2}\sqrt{m} \right)^m m \left(\log(a\sqrt{m}) + \log n \right) \\ &\leq k^2 (4a\zeta)^m = o(k^2) \ . \end{aligned}$$

Remark. In (4.10) we really have in mind a setting where $\rho \to 1$, $k = n^\alpha$ with $\alpha < 1/2$ fixed, and $m = o(\log n)$, since the maximum-correlation test is asymptotically powerful at ρ constant when $m \geq C \log n$ and C is sufficiently large. In that case, comparing with (4.8), this goodness-of-fit test is optimal up to a sub-logarithmic factor. We note that another goodness-of-fit test based on $H_i := X_i / \|X_i\|$ achieves a similar bound, but with ζ defined as

$$\zeta = (1-\rho)^{1/2} \left(\frac{n}{k^2} \right)^{1/(m-2)} \ .$$

4.5 Balancing detection ability and running time

Given the often enormous size of data sets that statisticians need to handle as an every-day practice, it is of great interest to design computationally efficient, yet near-optimal tests. In the case of the clique model, this is a highly non-trivial task, because the class $\mathcal{C}$ has size exponential in k and computing the localized squared-sum test (or other versions of the generalized likelihood ratio test and scan statistics) involves a non-trivial optimization problem over all $\binom{n}{k}$ elements of $\mathcal{C}$. In fact, often it seems that small testing risk and computational efficiency are contradicting terms. In this section we show that in at least one non-trivial instance, it is possible to design a computationally efficient (i.e., computable in time quadratic in n) test that has near optimal risk.

This is the case when the sample size m is (at most) logarithmic in n and $k \sim n^a$ for some $a \in (0,1)$. (Recall from Section 4.3 that this is a quite interesting range of parameters.)

To introduce a family of tests that balance detection performance and computational complexity, let $\ell \in \{1,\dots,m\}$ and define

$$Y(\ell) = \max_{S:|S|=k} \max_{T:|T|=\ell} \sum_{t\in T}\sum_{i\in S} X_{t,i}\ .$$

Since

$$Y(\ell) = \max_{T:|T|=\ell} \sum_{i=n-k+1}^{n} X_{T,(i)}\ , \quad X_{T,i} := \sum_{t\in T} X_{t,i},$$

where $X_{T,(1)} \le \cdots \le X_{T,(n)}$ are the ordered $X_{T,i}$'s, the statistic $Y(\ell)$ can be computed in $O(\binom{m}{\ell}(n\log(n)k + \ell\log(m)))$ time by first sorting $(X_{T,i} : i = 1,\dots,n)$ and summing the largest k, for all subsets T of size ℓ, and then maximizing over these.

For example, when $m \le \log_2 n$, then $\binom{m}{\ell} \le 2^m \le n$ and the test may be computed in time $O(n^2 \log n)$. Even when $m \sim C\log n$ for some constant $C > 0$, we may choose $\ell \sim \gamma m$ such that $C\gamma\log(1/\gamma) \le 1$. In that case $\binom{m}{\ell} \le 2^{\ell\log_2(m/\ell)} \le n$ and again the test may be computed in time $O(n^2\log n)$. The next proposition bounds the risk of the test.

Proposition 4 *Take $\ell \le m/7$. The test that rejects H_0 when*

$$Y(\ell) > a := \sqrt{2\ell k(\ell\log(em/\ell) + k\log(en/k))},$$

is asymptotically powerful in the clique model when

$$\rho_{\text{ave}} \ge 3\left(\frac{\log(n/k)}{\ell} + \frac{\log(m/\ell)}{k}\right)\ .$$

Proof Since under the null hypothesis $\sum_{t\in T}\sum_{i\in S} X_{t,i} \sim \mathcal{N}(0,\ell k)$, by a standard bound for the maximum of a finite set of Gaussian variables,

$$Y(\ell) \le \sqrt{2\ell k \log\binom{m}{\ell}\binom{n}{k}} \le a,$$

with probability converging to 1.

Under the alternative hypothesis where S is anomalous, we have

$$Y(\ell) \geq \sqrt{\rho_{\text{ave}}}\, k \left(Z_{(m-\ell+1)} + \cdots + Z_{(m)} \right) ,$$

where $Z_{(1)} \leq \cdots \leq Z_{(m)}$ are the ordered values of

$$Z_t := \left(k + \rho_{\text{ave}}\, k(k-1) \right)^{-1/2} \sum_{i \in S} X_{t,i} \, ,$$

which are i.i.d. standard normal. Since we assume that $m \to \infty$, $\mathbb{P}(Z_{(m-\ell+1)} \geq 1) \to 1$ when $\ell/m \leq 1/7 < \mathbb{P}\left(\mathcal{N}(0,1) > 1\right)$. Hence, $Y(\ell) \geq \sqrt{\rho_{\text{ave}}}\, k\ell$ with probability tending to one under the alternative hypothesis. Therefore, the test is asymptotically powerful when $\sqrt{\rho_{\text{ave}}}\, k\ell > a$, which follows from the assumptions when $k, \ell \to \infty$. ■

In the regime of (4.7) with $m \sim C \log n$, we see that the test is optimal up to a constant factor in ρ when $k \sim n^a$ for some $a < 1/2$. In this range of parameters, it seems hopeless to compute (or even approximate) the local squared-sum test.

However, when m is much larger than logarithmic in n, this test also requires super-polynomial computational time and therefore it is not useful in practice. In such cases one may have to resort to sub-optimal tests such as the maximum correlation test described in Section 3.3. It is an important and difficult challenge to find out the possibilities and limitations of powerful detection taking computational constraints into account.

4.6 A convex relaxation

In parallel to our work, Berthet and Rigollet [7] study a related problem of detecting a sparse principal component. The setting there is the same, except for the alternative hypothesis, where the covariance matrix is of the form $\Sigma = I + \theta v v^\top$, with $\theta > 0$ and v a unit vector with at most k non-zero components. They study a test based on $\lambda_k^{\max}(\widehat{\Sigma})$, the largest k-sparse eigenvalue of $\widehat{\Sigma}$, defined as

$$\lambda_k^{\max}(A) = \max_{|S|=k} \lambda^{\max}(A_S),$$

where A_S denotes the principal submatrix of A indexed by S and $\lambda^{\max}(A)$ the largest eigenvalue of A. This test is, in fact, intimately related to our localized squared-sum test, as we shall see in the analysis below. Berthet and Rigollet [7] prove that this test—with a proper choice of critical value—is near-optimal. However, just like our localized squared-sum test, the test of Berthet and Rigollet [7] is also computationally unfeasible due to the maximization over $\binom{n}{k}$ sets. For computational reasons, [7] turn to the convex relaxation of d'Aspremont, El Ghaoui, Jordan, and Lanckriet [14], for which they also establish a performance bound.

Berthet and Rigollet [7] show that, when $n, m \to \infty$, with high probability under the null hypothesis,

$$\lambda_k^{\max}(\widehat{\Sigma}) \leq 1 + C \max\left(\frac{k \log(n/k)}{m}, \sqrt{\frac{k \log(n/k)}{m}} \right) ,$$

for a universal constant C. Under the alternative hypothesis where S is anomalous, we have

$$\lambda_k^{\max}(\widehat{\Sigma}) \geq \lambda^{\max}(\widehat{\Sigma}_S) \geq \frac{1}{k} \sum_{i,j \in S} \widehat{\sigma}_{ij} = \frac{1}{km} Y_S \, ,$$

and in (3.5), we saw that $Y_S \geq km(1+\rho_{\text{ave}}(k-1))(1-O_P(1/\sqrt{m}))$. Hence, this test is asymptotically powerful when

$$\rho_{\text{ave}} \geq 3C \max\left(\frac{\log(n/k)}{m}, \sqrt{\frac{\log(n/k)}{km}}\right) , \tag{4.12}$$

which matches the performance of the localized squared-sum test (4.9) up to a multiplicative constant.

The semidefinite relaxation of d'Aspremont, El Ghaoui, Jordan, and Lanckriet [14] for $\lambda_k^{\max}$ is

$$\text{SDP}_k(A) = \max \text{tr}(AZ) \quad \text{subject to } Z \succeq 0,\ \text{tr}(Z) = 1,\ |Z|_1 \leq k ,$$

where the maximum is over all positive semidefinite matrices $Z = (Z_{st}) \in \mathbb{R}^{m\times m}$ and $|Z|_1$ denotes $\sum_{s,t}|Z_{st}|$. The quantity $\text{SDP}_k(A)$ can be computed efficiently as it is a semidefinite program. Berthet and Rigollet [7] suggest to use the test statistic $\text{SDP}_k(\widehat{\Sigma})$. They show that, when $n, m \to \infty$, with high probability under the null hypothesis,

$$\text{SDP}_k(\widehat{\Sigma}) \leq 1 + Ck \max\left(\frac{\log(n/k)}{m}, \sqrt{\frac{\log(n/k)}{m}}\right) ,$$

for a universal constant C, while under the alternative hypothesis,

$$\text{SDP}_k(\widehat{\Sigma}) \geq \lambda_k^{\max}(\widehat{\Sigma}) \geq (1+\rho_{\text{ave}}(k-1))(1-O_P(1/\sqrt{m})) .$$

Hence, the test based on the statistic $\text{SDP}_k(\widehat{\Sigma})$ is asymptotically powerful when

$$\rho_{\text{ave}} \geq 3C \max\left(\frac{\log(n/k)}{m}, \sqrt{\frac{\log(n/k)}{m}}\right) . \tag{4.13}$$

This rate matches (4.12) when $(k/m)\log(n/k) \to \infty$, and is otherwise comparable to what the maximum correlation test achieves (3.7). Thus, the relaxed test of Berthet and Rigollet is computationally efficient and near-optimal when the sample size is of smaller order that $k\log(n/k)$. Note that this allows one to handle larger values of m than for the test introduced in Section 4.5 where m had to be at most a constant multiple of $\log n$.

Interestingly, Berthet and Rigollet [7] also show that their analysis of the relaxed test is optimal in the following sense. If one can improve the rate given in (4.13) for the relaxed test, then one obtains an algorithm that improves by an order of magnitude upon known results for the *hidden clique problem*, see Alon, Krivelevich, and Sudakov [2]. We refer to [7] for a more precise statement.

4.7 Digest

In this section we briefly summarize our findings for the case of the clique model in simplified regimes. We consider combinations of the following settings:

- The sparse regime corresponds to $k \ll n^{1/2-\epsilon}$ for some $\epsilon > 0$. The non-sparse regime corresponds to $k \gg \sqrt{n}$.

- The ultra-high dimensional setting corresponds to $m \ll k \log n$ [29], while the (potentially) high dimensional setting corresponds to $m \gg k \log n$.

Our results can be summarized as follows:

- In the sparse high dimensional setting, detection is impossible when $\rho \ll \sqrt{(\log n)/(mk)}$ (see (4.6)). On the other hand when $\rho \gg \sqrt{(\log n)/(mk)}$, the localized squared-sum test is asymptotically powerful. Furthermore in the extremely sparse case where k is a constant, the same performance is achieved by the maximum correlation test.

- In the sparse ultra-high dimensional setting, detection is impossible when $\rho \ll (\log n)/m$ (see (4.7)). This rate is matched by the localized squared-sum test, and by the computationally efficient test of Berthet and Rigollet [7] based on SDP_k (see Section 4.6). Furthermore when $m \sim C \log n$ the test of Section 4.5 can also be computed in polynomial time (in n) and is asymptotically powerful for $\rho \gg (\log n)/m + 1/k$.

- In the non-sparse regime, the squared sum test is asymptotically powerful for $\rho \gg k^2/(n\sqrt{m})$. This rate is optimal (see (4.3)) if either m or k is not too large (that is either (4.4) or (4.5) is satisfied). In case neither (4.4) nor (4.5) is satisfied, the localized squared-sum test is asymptotically powerful.

5 Block model

Next we discuss the consequences of our main results for the block model which serves as a prototypical example of a "small" or "parametric" class. We focus on the case where ρ is bounded away from 1. Specifically, we assume that $\rho \le \rho_0 < 1$, and define $C_0 = (1 - \rho_0^2)^{-1}$.

In this model, to apply Theorem 1, we may use the obvious bound $Z \le k \mathbb{1}_{\{S \cap S' \neq \emptyset\}}$. Noting that $\mathbb{P}(S \cap S' \neq \emptyset) \le 2k/n$, we have

$$\mathbb{E} \cosh^m(\xi_a Z) \le 1 + \frac{2k}{n} \cosh^m(\xi_a k) \, .$$

We distinguish between two main regimes and we show that $R^* \to 1$ in both cases.

Case 1: moderate m. Suppose

$$\rho k \sqrt{\frac{m}{\log(n/k)}} \to 0 \quad \text{and} \quad \frac{k}{n} \to 0 \quad \text{and} \quad \frac{\log(n/k)}{m} \to 0 \, . \tag{5.1}$$

Let $\zeta = \rho k \sqrt{m/\log(n/k)}$ and choose $a \to \infty$ such that $a^2 \zeta \to 0$ and $a^2 \rho k \to 0$. The latter is possible because (5.1) implies that $\rho k \to 0$. We use the bound $\cosh(x) \le 1 + x^2$ for $x \in (0, 1)$, to get

$$\cosh^m(k\xi_a) = \big(1 + (k\xi_a)^2\big)^m \le \exp(C_0^2 a^4 \rho^2 k^2 m) \, ,$$

for n sufficiently large. Then

$$\frac{2k}{n} \exp(C_0^2 a^4 \rho^2 k^2 m) = 2 \exp(-\log(n/k)(1 - C_0^2 a^4 \zeta^2)) \to 0 \, ,$$

by our assumptions. Theorem 2.1 now implies that reliable detection is impossible in this range of the parameters.

Case 2: small m. Suppose

$$\rho k \frac{m}{\log(n/k)} \to 0 \quad \text{and} \quad \frac{k}{n} \to 0 \,. \tag{5.2}$$

Let $\zeta = \rho k m / \log(n/k)$ and choose $a \to \infty$ such that $a^2\zeta \to 0$. We use the bound $\cosh(x) \le \exp(x)$ to get

$$\frac{2k}{n} \cosh^m(k\xi_a) \le 2 \exp\left(-\log(n/k)(1 - C_0 a^2 \zeta)\right) \to 0 \,.$$

This discussion leads to the following.

Corollary 2 *In the block model, under either* (5.1) *or* (5.2)*,* $R^* \to 1$.

In view of Corollary 2, the squared-sum test is near-optimal for the block model only when $k \asymp n$. However, the localized squared-sum test has a much better performance. We have $|\mathcal{C}| = n$, and plugging this into (3.6), we see that the localized squared-sum test is asymptotically powerful when

$$\rho_{\text{ave}} k \ge A \max\left(\sqrt{(\log n)/m},\ (\log n)/m\right) \,,$$

for a large enough constant $A > 0$. With Corollary 1, we conclude that the test is near-optimal except in the case where $k/n \to 0$ slower than any negative power of n, where the test is optimal up to a logarithmic factor.

6 Perfect matching model

Here we work out the corollaries of our main results for the perfect matching model. This model illustrates how one may proceed when the model in question has a non-trivial combinatorial structure. In order to use Theorem 1, one needs to use the specific properties of the class. We focus on the case where ρ is bounded away from 1. Specifically, we assume that $\rho \le \rho_0 < 1$, and define $C_0 = (1 - \rho_0^2)^{-1}$.

In the perfect matching model, Z is distributed as the number of fixed points in a random permutation over $\{1, \dots, k\}$. It is well known that

$$\mathbb{P}\{Z = z\} = \frac{1}{z!} \sum_{s=0}^{k-z} \frac{(-1)^s}{s!} \le \frac{1}{z!} \left(\frac{1}{e} + \frac{1}{(k-z+1)!}\right) \,, \quad \forall z \in \{0, \dots, k\}. \tag{6.1}$$

We prove that $R^* \to 1$ in two main regimes with the help of Theorem 1. To simplify notation, we assume that k is even and recall that $n = k^2$ in this model.

Case 1: small m. Suppose

$$\rho \sqrt{k \max(k, m)} \to 0 \,. \tag{6.2}$$

We choose $a \to \infty$ such that $a^2\rho\sqrt{k\max(k,m)} \to 0$. We use the bounds $\cosh(x) \le 1 + x^2$ for $x \in (0,1)$ and $Z \le k$, and the fact that $a^2\rho k \to 0$, to get, for n sufficiently large,

$$\cosh^m(\xi_a Z) \le \exp(C_0^2 a^4 \rho^2 m Z^2) \le \exp(C_0^2 a^4 \rho^2 m k Z).$$

Now let $c = C_0^2 a^4 \rho^2 m k$. Using (6.1), one obtains

$$\begin{aligned}
\mathbb{E}\cosh^m(\xi_a Z) &\le \mathbb{E}\exp(cZ) \\
&\le \sum_{z=0}^{k} \frac{1}{z!}\left(\frac{1}{e} + \frac{1}{(k-z+1)!}\right)\exp(cz) \\
&\le \exp(\exp(c)-1) + \frac{k+1}{(k/2+1)!}\exp(ck) \\
&\le 1 + o(1),
\end{aligned}$$

because $c \to 0$ and $\log[(k/2+1)!] \sim (k/2)\log k$ as $k \to \infty$.

Case 2: moderate m. Suppose

$$\frac{\rho m}{\log(\min(k,m))} \to 0. \tag{6.3}$$

We choose $a \to \infty$ such that $a^2\rho m/\log(\min(k,m)) \to 0$. Using (6.1) one obtains

$$\begin{aligned}
\mathbb{E}\cosh^m(\xi_a Z) &\le \mathbb{E}\cosh^m(\xi_a Z)\mathbb{1}_{\{Z<k/2\}} + \mathbb{P}\{Z \ge k/2\}\cosh^m(\xi_a k) \\
&\le \sum_{z=0}^{k/2-1} \frac{1}{z!}\left(\frac{1}{e} + \frac{1}{(k/2)!}\right)\cosh^m(\xi_a z) + \mathbb{P}\{Z \ge k/2\}\exp(\xi_a m k) \\
&\le \frac{1}{e}\sum_{z=0}^{+\infty} \frac{1}{z!}\cosh^m(\xi_a z) + \left(\frac{k}{(k/2)!} + \mathbb{P}\{Z \ge k/2\}\right)\exp(\xi_a m k)\,.
\end{aligned}$$

Now we take care separately of these last two terms. First note that

$$\frac{k}{(k/2)!} + \mathbb{P}\{Z \ge k/2\} \le \frac{3k}{(k/2)!} \le \exp((k/3)\log k)$$

when k is large enough, and since $\xi_a m k = O(a^2\rho m k) = o(k\log k)$ by our choice of a, we obtain

$$\left(\frac{k}{(k/2)!} + \mathbb{P}\{Z \ge k/2\}\right)\exp(\xi_a m k) \to 0\,.$$

For the other term the situation is slightly more subtle. Let Y be a sum of m independent Rademacher random variables. Using the binomial identity it is easy to prove that

$$\cosh^m(\xi_a z) = \mathbb{E}\exp(\xi_a z Y)\,,$$

and thus we have

$$\frac{1}{e}\sum_{z=0}^{+\infty} \frac{1}{z!}\cosh^m(\xi_a z) = \mathbb{E}\left[\exp(\exp(\xi_a Y) - 1)\right]\,.$$

Now thanks to Hoeffding's inequality, we obtain for any $t > 0$,

$$\mathbb{E}\left[\exp(\exp(\xi_a Y) - 1\right] \leq \exp(\exp(\xi_a t) - 1) + \exp(-t^2)\exp(\exp(\xi_a m) - 1)\ .$$

In particular with $t = m/\log m$, using that $\xi_a m = O(a^2 \rho m) = o(\log m)$ by our choice of a, this shows that

$$\mathbb{E}(\exp(\exp(\xi_a Y) - 1) = 1 + o(1).$$

This discussion leads to the following.

Corollary 3 *Consider the class of perfect matchings on the complete bipartite graph. Under either of* (6.2)*, or* (6.3)*,* $R^* \to 1$.

It is easy to derive upper bounds for the performance of the localized squared-sum test in this model. All we need to observe is that $|\mathcal{C}| = k!$ and therefore $\log|\mathcal{C}| \sim k\log k$ when $k \to \infty$. Plugging this into (3.6), we see that the local squared-sum test is asymptotically powerful when

$$\rho_{\text{ave}} k \geq A \max\left(\sqrt{k\log(k)/m},\ k\log(k)/m\right)\ ,$$

Thus ignoring logarithmic factors, the requirement is that $\rho_{\text{ave}}\sqrt{m\min(k,m)}$ be large. Looking at Corollary 3, the complement of (6.2) *or* (6.3) corresponds (roughly) to $\rho\sqrt{k\max(k,m)} \to \infty$ *and* $\rho m/\log\min(k,m) \to \infty$, which is the same requirement if we ignore logarithmic factors. Thus the local squared-sum test is near-optimal.

7 The clique number of random geometric graphs

In this section we describe a, perhaps unexpected, application of Theorem 1. We use this theorem to derive a lower bound for the clique number of random geometric graphs on high-dimensional spheres.

To describe the problem, let $p \in (0,1)$ and let $Z_1, \ldots, Z_n$ be independent random vectors, uniformly distributed on the unit sphere $S_{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$. A random geometric graph $G(n,d,p)$ is defined by vertex set $V = \{1,\ldots,n\}$ and vertex i and vertex j are connected by an edge if and only if $(Z_i, Z_j) \geq t_{p,d}$ where the threshold value $t_{p,d}$ is such that

$$\mathbb{P}\{(Z_1, Z_2) \geq t_{p,d}\} = p$$

(i.e., the probability that an edge is present equals p). The *clique number* $\omega(n,d,p)$ is the size of the largest clique of $G(n,d,p)$ (i.e., the largest fully connected subset of vertices). In Devroye, György, Lugosi, and Udina [16] the behavior of the random variable $\omega(n,d,p)$ is studied for fixed values of p when n is large and $d = d_n$ grows as a function of n. The rate of growth of $\omega(n,d,p)$ is shown to depend in a crucial way of how fast d_n increases with n. Specifically, the following results are established (and hold with probability converging to 1 as $n \to \infty$):

$$\begin{array}{lcl}
d_n = O(1) & \Longrightarrow & \omega(n,d,p) = \Omega(n) \\
d_n \to \infty & \Longrightarrow & \omega(n,d,p) = o(n) \\
d_n = o(\log n) & \Longrightarrow & \mathbb{E}\omega(n,d,p) = \Omega(n^{1-\epsilon}) \text{ for all } \epsilon > 0 \\
d_n \geq 9\log^2 n & \Longrightarrow & \omega(n,d,p) = O(\log^3 n) \\
d_n/\log^3 n \to \infty & \Longrightarrow & \omega(n,d,p) = (2 + o(1))\log_p n
\end{array}$$

where $a_n = \Omega(b_n)$ means that $b_n = O(a_n)$. We see that the clique number behaves in drastically different ways between $d_n = o(\log n)$ — when $\omega(n, d, p)$ grows almost linearly—and $d_n \sim \log^2 n$ — when $\omega(n, d, p)$ has a poly-logarithmic growth at most.

The above-mentioned results leave open the question of where exactly the "phase transition" occurs, and whether the upper bound in the regime $d_n \sim \log^2 n$ is sharp. In this section we are able to answer both of these questions. Below we establish a general lower bound for the clique number which implies that, perhaps surprisingly, the phase transition occurs around $\log^2 n$ and that the upper bounds above cannot be improved in an essential way. We show that the median of the clique number $\omega(n, d, p)$ is bounded from below by $\exp(\kappa \log^2 n/d)$ where κ is a positive constant that depends on p only. This implies, for example, that if $d \sim c \log n$ for some $c > 0$, then $\omega(n, d, p)$ grows as a positive power of n. On the other hand, even when $d \sim \log^{2-\epsilon} n$ for any fixed $\epsilon > 0$, then $\omega(n, d, p)$ is much larger than any power of $\log n$. For the sake of simplicity, we only state the result for the case of $p = 1/2$. The argument is identical for other values of p.

Theorem 2 *There exist universal constants $c_1, c_2, c_3, c_4 > 0$ such that for all n, d such that $d \geq c_1 \log(c_2 n)$, the median of the clique number $\omega(n, d, 1/2)$ satisfies*

$$\operatorname{med}(\omega(n, d, 1/2)) \geq c_3 \exp\left(\frac{c_4 \log^2 (c_2 n)}{d}\right) .$$

One may take $c_1 = 7/16$, $c_2 = 16 \log 2$, $c_3 = 1/16$, and $c_4 = 49/5120$. In particular,

$$d \leq c_4 \log^{2-\varepsilon} n \quad \textit{implies} \qquad \operatorname{med}(\omega(n, d, 1/2)) = \Omega(\exp(\log^{\varepsilon} n)) .$$

Proof The basic idea of the proof is to define a test that works well whenever the median clique number is small. But then the lower bound of Theorem 1 implies that the clique number cannot be small.

Let $\omega_0 = \operatorname{med}(\omega(n, d, 1/2))$ for short. Consider the clique model with $m = d$, all nonzero correlations equal to ρ and $k = 16\omega_0$. For $i = 1, \ldots, n$, let $X^{(i)} = (X_{i,1}, \ldots, X_{i,d}) \in \mathbb{R}^d$, and define the random geometric graph G on the normalized vectors $Z_i = X^{(i)}/\|X^{(i)}\|$, connecting points Z_i and Z_j whenever $(Z_i, Z_j) \geq 0$. The test statistic we consider is the clique number of the resulting graph, denoted by ω. (This test was suggested and analyzed in [16]. Here we combine their analysis with Theorem 1 to derive a lower bound for the median clique number.)

Under the null hypothesis (when $\rho = 0$), the Z_i's are i.i.d. uniform on the sphere S_{d-1} implying that $G \sim G(n, d, 1/2)$ and, consequently, $\omega \sim \omega(n, d, 1/2)$. Devroye, György, Lugosi, and Udina [16] show that, under the alternative hypothesis, with probability at least $7/8$, the graph contains a clique of size k whenever

$$\binom{k}{2} < (1/8) e^{d\rho^2/10}. \tag{7.1}$$

When this is the case, the test that accepts the null hypothesis when $\omega < k$ has a probability of type II error bounded by $1/8$. To bound the probability of type I error of this test, we first prove that $\mathbb{E}_0\omega < 2\omega_0$ for any d and n sufficiently large. We start with

$$\mathbb{E}_0\omega \geq 2\omega_0 \quad \Leftrightarrow \quad \mathbb{E}_0\omega - \omega_0 \geq \frac{1}{2}\mathbb{E}_0\omega \quad \Rightarrow \quad \frac{1}{2}\mathbb{E}_0\omega \leq \mathbb{E}_0\omega - \omega_0 \leq \sqrt{\operatorname{var}(\omega)},$$

where in the last step we used the well-known fact that the difference between the mean and the median of any random variable is bounded by its standard deviation. Now observe that ω, as a function of the independent random variables $Z_1, \ldots, Z_n$, is a configuration function in the sense of Talagrand [28] which implies that $\mathrm{var}(\omega) \leq \mathbb{E}_0\omega$, (see [10, Corollary 3.8]. We arrive at

$$\mathbb{E}_0\omega \geq 2\omega_0 \quad \Rightarrow \quad \frac{1}{2}\mathbb{E}_0\omega \leq \sqrt{\mathbb{E}_0\omega} \quad \Leftrightarrow \quad \mathbb{E}_0\omega \leq 4.$$

However, it is a simple matter to show that $\mathbb{E}_0\omega > 4$ for all d if n is sufficiently large. (To see this it suffices to show that the probability that 5 random points form a clique is bounded away from zero. This follows from the arguments of [16].) We then bound the probability of type I error as follows

$$\mathbb{P}_0\{\omega \geq k\} = \mathbb{P}_0\{\omega \geq 16\omega_0\} \leq \mathbb{P}_0\{\omega \geq 8\mathbb{E}_0\omega\} \leq \frac{1}{8},$$

where we used Markov's inequality in the last line.

Combining the bounds on the probabilities of type I and type II errors, we conclude that $R^* \leq 1/4$. Put it another way,

$$R^* > 1/4 \Longrightarrow \binom{16\omega_0}{2} \geq (1/8)e^{d\rho^2/10}.$$

Now, by Theorem 1, we see that

$$(16\omega_0)^2 < 4(\ln 2)ne^{-16\rho d/7} \Longrightarrow R^* > 1/4.$$

We conclude that, for any $\rho \in (0, 1)$,

$$(16\omega_0)^2 < 4(\ln 2)ne^{-16\rho d/7} \Longrightarrow (16\omega_0)^2 \geq (1/4)e^{d\rho^2/10}.$$

Therefore, if ρ is such that $4(\ln 2)ne^{-16\rho d/7} > (1/4)e^{d\rho^2/10}$, then $(16\omega_0)^2 \geq (1/4)e^{d\rho^2/10}$. Choosing $\rho = (7/(16d))\log((16\log 2)n)$—which is possible since $d \geq (7/16)\log((16\log 2)n)$—clearly satisfies the required inequality and this choice gives rise to the announced lower bound. ■

8 Discussion

We close this paper by discusing some open problems and directions of further research.

Sharper bounds. The cornerstone of our analysis is the lower bound stated in Theorem 1. It is powerful enough that we can deduce useful bounds in many different models, which are seen to be optimal up to constant or logarithmic factors. While a considerable effort has been devoted in the related detection-of-means problem for finding the right constants, one wonders if it is possible to obtain results that fine here, at least in some regimes. One possible avenue is via the truncated second moment approach, which underlies the lower bounds in Ingster [21], Donoho and Jin [17], Hall and Jin [20], Butucea and Ingster [11]. The computations are rather daunting in the setup of this paper and we decided not to take this route. Note that the second moment approach (without truncation) has limited applicability, though it is a little more useful here than it is in the case where $m = 1$.

Comparison with the detection-of-means setting. Our results reveal that the dependence on the sample size is different here. In the detection-of-means setting, one reduces by sufficiency to the case where $m = 1$ by simply averaging the multiple observations and working with $\bar{X}_1, \ldots, \bar{X}_n$, where $\bar{X}_i = \frac{1}{m}\sum_t X_{t,i}$. Therefore, if initially $X_{t,i} \sim \mathcal{N}(\mu, 1)$ when i is anomalous, we now have $\bar{X}_i \sim \mathcal{N}(\mu, 1/m)$. Therefore, we reduce the problem to where $m = 1$ and μ is replaced by $\sqrt{m}\mu$. From this, we know that reliable detection is possible if either

$$\frac{k^2}{n} \to \infty \quad \text{and} \quad \mu^2 m \frac{k^2}{n} \to \infty \,,$$

or

$$\frac{k^2}{n} \to 0 \quad \text{and} \quad \mu^2 m / \log(n) \geq C \,,$$

where C is a large enough constant. In our previous work, we argued that, at least when ρ is bounded away from 1, the parameter ρ in the correlation-detection problem played a similar role as μ^2 in the detection-of-means setting. The case when the sample size $m \to \infty$ is, however, quite different both in the 'dense' regime $\rho\sqrt{m} \gg n/k^2$, and in the 'intermediary' regime $\rho\sqrt{m} \succ \sqrt{\log(n/k)/k}$. We also note that this regime does seem to have an equivalent in the detection-of-means setting.

Computational considerations. The localized squared-sum test, although near-optimal in some regimes, is not computationally tractable in the clique model. Following the footsteps of Berthet and Rigollet [7], we find that the convex relaxation of d'Aspremont, El Ghaoui, Jordan, and Lanckriet [14] leads to a computationally-tractable test of comparable performance in the 'sparsest' regime where $m = O(k \log(n))$. The best performance achievable by tests that can be computed in polynomial time is yet unknown and remains an intriguing open problem.

General correlations. More generally, the problem of detecting correlations of arbitrary sign—not just positive correlations like we do here—remains open. Even though one can design natural tests akin to our squared-sum and local squared-sum tests for that situation, the challenge is in deriving tight lower bounds. We mention that our approach to obtaining a lower bound in Section 2 does not apply here, since the representation (2.2) is not valid when the correlations may be negative.

Acknowledgements

We would like to thank the anonymous referees for helpful comments and suggestions. We also thank Quentin Berthet and Philippe Rigollet for shedding some light on their results at the *Non-parametric and High-dimensional Statistics* conference, held in December of 2012, in Luminy, France. EAC was partially supported by NSF grant DMS 1120888 and ONR grant N00014-09-1-0258. GL was supported by the Spanish Ministry of Science and Technology grant MTM2009-09063 and PASCAL2 Network of Excellence under EC grant no. 216886.

References

[1] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi. On combinatorial testing problems. *Ann. Statist.*, 38(5):3063–3092, 2010.

[2] N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. *Random Structures and Algorithms*, 13:457–466, 1999.

[3] E. Arias-Castro, S. Bubeck, and G. Lugosi. Detection of correlations. *Ann. Statist.*, 2012. To appear.

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

[5] Y. Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606, 2002.

[6] S. M. Berman. Equally correlated random variables. *Sankhyā Ser. A*, 24:155–156, 1962.

[7] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. Available online at http://arXiv.org/abs/1202.5070, 2012.

[8] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604, 2008.

[9] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227, 2008.

[10] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities; A nonasymptotic theory of independence*. Oxford University Press, Oxford, 2013.

[11] C. Butucea and Y. I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. Available online http://arxiv.org/abs/1109.0898, 2011.

[12] T. T. Cai, C.-H. Zhang, and H. H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38(4):2118–2144, 2010.

[13] S. X. Chen, L.-X. Zhang, and P.-S. Zhong. Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.*, 105(490):810–819, 2010.

[14] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007.

[15] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2010. Corrected reprint of the second (1998) edition.

[16] L. Devroye, A. György, G. Lugosi, and F. Udina. High-dimensional random geometric graphs and their clique number. *Electron. J. Probab.*, 16:2481–2508, 2011.

[17] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 2004.

[18] N. El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.*, 36(6):2717–2756, 2008.

[19] T. J. Fisher. On testing for an identity covariance matrix when the dimensionality equals or exceeds the sample size. *Journal of Statistical Planning and Inference*, 142(1):312 – 326, 2012.

[20] P. Hall and J. Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.*, 38(3):1686–1732, 2010.

[21] Y. I. Ingster. Minimax detection of a signal for ℓ_n^l balls. *Math. Methods Statist.*, 7:401–428, 1999.

[22] J. Jin. *Detecting and Estimating Sparse Mixtures*. PhD thesis, Stanford University, 2003.

[23] O. Ledoit and M. Wolf. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.*, 30(4):1081–1102, 2002.

[24] R. J. Muirhead. *Aspects of multivariate statistical theory*. John Wiley & Sons Inc., New York, 1982. Wiley Series in Probability and Mathematical Statistics.

[25] H. Nagao. On some test criteria for covariance matrix. *Ann. Statist.*, 1:700–709, 1973.

[26] J. R. Schott. Testing for complete independence in high dimensions. *Biometrika*, 92(4):951–956, 2005.

[27] M. S. Srivastava. Some tests concerning the covariance matrix in high dimensional data. *J. Japan Statist. Soc.*, 35(2):251–272, 2005.

[28] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.*, 81:73–205, 1995.

[29] N. Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat.*, 6:38–90, 2012.

[30] N. Verzelen and F. Villers. Goodness-of-fit tests for high-dimensional Gaussian linear models. *Ann. Statist.*, 38(2):704–752, 2010.