Sketched Covariance Testing: A Compression-Statistics Tradeoff

Gautam Dasarathy
Rice University

Parikshit Shah
Yahoo Research

Richard G. Baraniuk
Rice University

***Abstract*—Hypothesis testing of covariance matrices is an important problem in multivariate analysis. Given n data samples and a covariance matrix Σ_0, the goal is to determine whether or not the data is consistent with this matrix. In this paper we introduce a framework that we call *sketched covariance testing*, where the data is provided after being compressed by multiplying by a "sketching" matrix A chosen by the analyst. We propose a statistical test in this setting and quantify an achievable sample complexity as a function of the amount of compression. Our result reveals an intriguing achievable tradeoff between the compression ratio and the statistical information required for reliable hypothesis testing; the sample complexity increases as the fourth power of the amount of compression.**

I. Introduction

Large-scale sparse covariance matrices play a prominent role in a wide range of applications in bioinformatics, climate studies, and economics. Therefore, statistical tasks like inference, structure estimation, and hypothesis testing involving such matrices have been the subject of intensive study in recent years [1]–[6].

In this paper, we focus on the hypothesis-testing problem, which aims to distinguish between two scenarios given data. In the first scenario, or the null-hypothesis, the data appears to be consistent with a zero-mean Gaussian distribution with known covariance $\Sigma_0 \in \mathbb{R}^{p\times p}$. In the second scenario, or the alternate hypothesis, the data appears to be drawn from a zero-mean Gaussian distribution with a covariance $\Sigma \in \mathbb{R}^p$ that is distinct from Σ_0. We are specifically interested in situations where $\Sigma - \Sigma_0$ is *structured*, in particular, when it is "d-distributed sparse", where no more than d entries per row/column of Σ differ from those of Σ_0. Of course, if both Σ_0 and Σ are distributed sparse (see e.g. [2], [7], [8] for various examples to motivate this structural assumption), then naturally their difference is also distributed sparse. However, our main result is more general and requires this assumption only of the difference $\Sigma - \Sigma_0$. Such situations arise naturally, for instance, in anomaly detection or in testing involving protein-protein signaling networks. In the former case, it is of interest to find out if an underlying process was subject to a significant (but structured) statistical change, while in the latter, one might interested in detecting the presence of a (structured) pattern of correlations that corresponds to a particular biological phenomenon.

Importantly, in our setup, the statistical information from the distribution being tested is made available to the analyst through a dimensionality-reduction process known as a *sketch*. Consider n independent samples $X_1, \ldots, X_n$ drawn from a Gaussian distribution $\mathcal{N}(0, \Sigma)$, and let $A \in \mathbb{R}^{m\times p}$ (where typically $m < p$) be a "sketching" matrix that the analyst chooses herself. Instead of directly observing the samples themselves, she observes their sketched or compressed versions Y_i given by $Y_i = AX_i, i = 1, 2, \ldots, n$. The goal then is to conduct the hypothesis test using only these sketched samples. For instance, when conducting hypothesis tests involving protein-protein interactions in a cell, this approach would enable one to dye and record expression levels corresponding to a small number (m) of protein-pools as opposed to doing this exhaustively across the large number (p) of proteins.

Our main result quantifies an achievable sample complexity of the hypothesis test, i.e., the number of samples n required to reliably distinguish between the hypothesis as a function of the compression size m. An interesting feature of this result is the revelation of a novel achievable *compression-statistics tradeoff*; we show that the number of samples n required for the test to succeed is inversely proportional to the fourth power of the size m of the sketch. We believe that this phenomenon holds broadly (i.e., beyond hypothesis testing) and that its exploration is a fruitful avenue for future work.

II. Problem Setup and Main Result

A. Notation, Setup, and Related Work

For $n \in \mathbb{N}$, we write $[n]$ to denote the set $\{1, 2, \ldots, n\}$. We index the elements of the matrix X as either X_{ij} or $[X]_{ij}$, and write $\|X\|_\infty$ to denote the infinity-norm of vec(X), i.e., its absolute maximum element; when the context is clear, we will use the same notation to denote the absolute maximum element of a vector as well.

Let $\Sigma_0, \Sigma \in \mathbb{R}^{p\times p}_{\succ 0}$ be two positive definite matrices, and assume that Σ_0 is known. Suppose that $X_1, X_2, \ldots, X_n$ are samples drawn i.i.d. from a zero-mean multivariate normal distribution with covariance matrix Σ; i.e., $X_i \sim \mathcal{N}(0, \Sigma), i \in [n]$. We are interested in the following hypothesis test:

$$H_0 : \Sigma = \Sigma_0 \qquad H_1 : \Sigma \neq \Sigma_0 \tag{1}$$

The focus of this paper is the following twist on this problem. We suppose that we can design a *sketching* matrix $A \in \mathbb{R}^{m\times p}$ with $m < p$ and that, instead of observing the samples $X_i, i \in [n]$, we observe their compressed versions $Y_i = AX_i \in \mathbb{R}^m, i \in [n]$. Our goal is to design and analyze a test statistic that reliably performs the hypothesis test (1) based on this sketched data.

The problem of covariance estimation from full (unsketched) data has been studied extensively (see e.g., [9] and references there). Covariance estimation from compressed samples has begun to receive considerable attention recently. One line of work considers the problem of recovering *low rank* covariance matrices from one-dimensional measurements (see e.g., [10], [11] and the references therein). [12] also considers very low-dimensional measurements, but does not leverage any specific structure during estimation. In the full rank case, [13] and [7] introduce a framework (called *covariance sketching*) which achieves near-optimal compression ratio for *sparse* covariance matrices, but does not come with finite sample guarantees.

Similarly, hypothesis testing of covariance matrices from full data has been a subject of intensive study; see, for instance, [4]–[6] and the references therein. In this paper, we consider hypothesis testing from compressed samples when the underlying difference matrix is sparse, a framework we call *sketched covariance testing*. For this problem, we quantify an intriguing achievable compression-statistics tradeoff, which is novel to the best of our knowledge. It is worth noting that if the covariance matrices themselves are sparse, one might first estimate these matrices (using covariance sketching), and then test the estimates. However, such a procedure would also lack finite sample guarantees, and perhaps more importantly, we expect that this would be sub-optimal in terms of leveraging the statistical information effectively.

B. Choice of the Sketching Matrix A

As in [7], we choose A to be the adjacency matrix of a certain random graph. Such a sketching matrix is an appealing choice since (a) it corresponds to a natural notion of sketching or "pooling" where each row corresponds to a pool, and the support of the row tells us the constituents of this pool, and (b) it has certain combinatorial properties (which may be thought of as a weak notion of small-set vertex expansion) that are critical for our results*.

Definition 1 ($\delta-$left regular random bipartite graph [7])**.** *$G = ([p], [m], E)$ is called a $\delta-$left regular random bipartite graph if the edge set E is generated according to the following process. For each $i \in [p]$, choose δ vertices $j_1, j_2, \ldots, j_\delta$ uniformly and independently at random (without replacement) from $[m]$ and assign these as neighbors to i, i.e., $\{\{i, j_k\}\}_{k=1}^{\delta}$ is added to the edge set E.*

For ease of presentation in [7], each neighbor of $i \in [p]$ was chosen with replacement. However, as noted in [7], all the properties derived in that paper hold when the edges are added without replacement. Note that A is binary and relatively sparse when δ is small. We will prove below that it suffices that $\delta = \mathcal{O}(\log p)$, which can result in significant savings in computational time both when sketching the data and when performing the hypothesis test.

*We refer the reader to [7, Section III-B] for these properties, their proofs, and for more on how these properties are useful.

C. Main Result

Let $\widehat{\Sigma}_Y^{(n)}$ denote the sample covariance matrix of the sketched samples $Y_1, \ldots, Y_n$, that is, $\widehat{\Sigma}_Y^{(n)} = \frac{1}{n}\sum_{k=1}^{n} Y_k Y_k^T$. Note that $\widehat{\Sigma}_Y^{(n)} = A\widehat{\Sigma}^{(n)}A^T$, where $\widehat{\Sigma}^{(n)}$ is the sample covariance matrix corresponding to $\{X_i\}_{i\in[n]}$. This fact follows from the fact that if the random vector $X \sim \mathcal{N}(0, \Sigma)$, then $Y = AX$ is distributed according to $\mathcal{N}(0, A\Sigma A^T)$.

We now define our sketched test statistic as

$$T = \left\| \widehat{\Sigma}_Y^{(n)} - A\Sigma_0 A^T \right\|_\infty \tag{2}$$

and the decision rule as

$$T \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{3}$$

That is, we declare that the alternate hypothesis holds if $T \geq \eta$ for an appropriately chosen η. We are now ready to state the main result of the paper.

Theorem 1. *Under H_1, let d be the maximum number of non-zeros that $\Gamma \triangleq \Sigma - \Sigma_0$ has in any single row or column. There exist constants $c > 0, c_0 \in (0, 0.25), c_1 \in (0, 1)$, and $c_2 > 0$ such that the sketched hypothesis test* (3) *succeeds with probability at least $1 - 4p^{-c}$ provided the following conditions hold†:*

- *The threshold η is set as $\eta = \frac{\Gamma_{\min}}{4}$.*
- *The size of the sketch m satisfies*

$$\sqrt{2c_0\delta^2 dp\frac{\Gamma_{\max}}{\Gamma_{\min}}} \leq m \leq c_1\frac{p}{\log p}. \tag{4}$$

- *The number of samples n satisfies*

$$n \geq c_2\left(\frac{p\log p}{m}\right)^4 \frac{\max_i \Sigma_{ii}^2 \vee \max_i [\Sigma_0]_{ii}^2}{\Gamma_{\min}^2}. \tag{5}$$

Remark 1 (Compression-statistics tradeoff). The above result reveals a graceful tradeoff between the number of samples n and the compression ratio m/p; we see that that n scales like $(p\log p/m)^4$. It is interesting to consider the extremes of this tradeoff. At the low-compression extreme, where $m \sim p/\log p$, n scales as $\mathrm{polylog}(p)$, which may be compared with the results of [1], [5]. At the high-compression extreme, where $m \sim \sqrt{dp}$‡, it suffices if the sample complexity n scales like $\tilde{\mathcal{O}}(p^2)$. More generally, for $m = O(p^\alpha)$ ($\alpha > \frac{1}{2}$), we have an achievable sample complexity of $\tilde{\mathcal{O}}(p^{4(1-\alpha)})$. The upper bound on m in (4) may be an artifact of the analysis, and can potentially be improved.

Remark 2 (Two sample testing). Theorem 1 extends readily to *two-sample testing*, where, given two sets of statistical samples, the goal is to decide whether these samples are drawn from the same distribution or not. It can be verified that if we define T as the maximum absolute deviation of the corresponding sketched sample covariance matrices, then

†$\Gamma_{\max}$ (resp. $\Gamma_{\min}$) is the absolute maximum (resp. minimum) non-zero value in Γ

‡Notice that one can argue, using [7], that $\sqrt{dp}$ (up to log factors) is an information theoretic lower bound for m.

the results of Theorem 1 continue to hold. The "unsketched" version of two sample testing with sparse covariance matrices was studied in [5]. Notice that when presented with sketched data, a naïve application of the framework of [7] would necessitate a statistically expensive full reconstruction of the covariance matrices involved. However, our approach enables the detection of any differences between these matrices directly (and comes with finite sample guarantees).

Remark 3 (Sparsity assumption on Γ). Notice that the smaller m is, the sparser Γ needs to be according to (4). When m is close to its lower bound, i.e., in the high compression regime, the hypothesis test can fail if Γ is dense even in just a single row or column; this follows directly from the argument in [7] that in this case, it is possible to have $A\Gamma A^T = 0$. However, the sparsity requirement on Γ is not strict; indeed Γ can be "approximately sparse" such that it admits a decomposition $\Gamma = \Gamma_s + \Gamma_n$, where Γ_s is sparse with relatively large elements and Γ_n is potentially dense and has low magnitude entries. In this case, theorem 1 continues to hold, with Γ replaced by Γ_s, provided that Γ_n is sufficiently small, i.e., $\|\Gamma_n\|_\infty \leq \left(\frac{m}{p\delta}\right)^2 [\Gamma_s]_{\min}$.

It should also be noted that, intuitively, denser Γ implies Σ and Σ_0 are more different, and hence the hypothesis test should be statistically "easier". This is reflected by the fact that denser Γ requires larger m, and hence n can be much smaller to achieve the same statistical performance.

Remark 4 (Other sketching matrices). It will be interesting to see if the results of this paper and of [7] can be extended to the case of Gaussian or other sketching matrices A, as this might enable better statistical performance; this is an interesting avenue for future work.

III. Proof of Theorem 1

In this section we prove Theorem 1 in two parts. In part (A), we show an upper bound on T that holds with high probability under under H_0, and in part (B) we show a lower bound on T under H_1. Putting these together, and accounting for the probability of violation of these bounds gives us the final result.

A. Analysis under H_0

In what follows we produce an upper bound on T under H_0 that holds with high probability. Let $\phi = \delta/(m-\delta+1)$.

Proposition 1. *Under H_0, the test statistic T is bounded from above by ε_2 with probability at least $1 - m\exp\left(-\frac{p\phi\varepsilon_1^2}{2\phi+2\varepsilon_1/3}\right) - 4m^2\exp\left(\frac{-n\varepsilon_2^2}{3200p^4\phi^4(1+\varepsilon_1)^4(\max_{i\in[p]}[\Sigma_0]_{ii})^2}\right)$.*

Proof: In order to bound T, we use Lemma 3 from Appendix A on the concentration of sample covariance matrices. To use this, we need an upper bound on the variance of the pooled random vector Y. Notice that there are two sources of randomness here, the randomness due to the sampling and the randomness in the generation of the sketching matrix A. We first control the variance for a fixed compression matrix A.

Lemma 1. *Let us fix A. For $r \in [m]$ the following upper bound holds on the $r-$th diagonal element of $\Sigma_Y = A\Sigma A^T$: $[\Sigma_Y]_{rr} \leq \left(\sum_{i=1}^p A_{ri}\right)^2 \max_{i\in[p]} \Sigma_{ii}$.*

Proof: Observe that $[\Sigma_Y]_{rr} = \mathrm{Var}[Y_r]$, where $Y_r = \sum_{i=1}^p A_{ri}X_i$ is the $r-$th coordinate of Y. Also, Y_r is $0-$mean, which implies that we have the following expression holds $[\Sigma_Y]_{rr} = \mathrm{Var}[Y_r] = \mathbb{E}\left[\left(\sum_{i=1}^p A_{ri}X_i\right)^2\right]$. Next, notice that

$$\mathbb{E}\sum_{i,j=1}^p A_{ri}A_{rj}X_iX_j \leq \left(\sum_{i=1}^p A_{ri}\right)^2 \max_{i\in[p]}\Sigma_{ii},$$

where we have used the fact that the A_{ri}'s are always non-negative, and that $\Sigma_{ij} \leq \sqrt{\Sigma_{ii}\Sigma_{jj}}$ by Hadamard's inequality [15]. ∎

The next step is to control the sum $\sum_{i=1}^p A_{ri}$ for all $r \in [m]$. The following result gives us a high probability upper bound.

Lemma 2. *When A is chosen according to the $\delta-$left regular bipartite random graph model of Definition 1, the following upper bound holds with probability greater than $1 - m\exp\left(-\frac{p\phi\varepsilon_1^2}{2\phi+2\varepsilon_1/3}\right)$: $\max_{r\in[m]}\sum_{i=1}^p A_{ri} \leq p\phi(1+\varepsilon_1)$*

Proof: Note that the A_{ri}'s are mutually independent, since the choice of neighbors for $i \in [p]$ is independent of the choice of neighbors for any other $j \in [p]$. Next, observe that $A_{ri} = 1$ iff r is one of the δ neighbors chosen for the left vertex i. Therefore, $\mathbb{P}[A_{ri}=1] = \binom{m}{\delta-1}/\binom{m}{\delta} = \delta/(m-\delta+1) \triangleq \phi(\delta,m)$. We suppress the dependence on δ and m.

As alluded to earlier, we are interested in regimes where δ is significantly smaller than p (and therefore m). This implies that the random variables that make up the sum have low variance. And, to leverage this, we use Bernstein's inequality (which we reproduce as Lemma 4 in Appendix A). In this case, we have the following: $|A_{ri}-\phi| \leq 1-\phi$, and $\mathrm{Var}[A_{ri}] = \phi(1-\phi) \leq \phi$. Applying Lemma 4 to the A_{ri}'s,

$$\mathbb{P}\left[\sum_{i=1}^p A_{ri} > p\phi(1+\varepsilon_1)\right] \leq \exp\left(-p\phi\varepsilon_1^2/2\phi + 2\varepsilon_1/3\right).$$

A union bound over $r \in [m]$ gives us the desired result. ∎

Now, we notice that under H_0, $\Sigma = \Sigma_0$. Therefore, $T = \left\|A\widehat{\Sigma}_0^{(n)}A^T - A\Sigma_0A^T\right\|_\infty$. Recall that our goal is to produce a probabilistic upper bound on T. And towards this end, letting $\mathcal{E}_1$ denote the event that the implication of Lemma 2 holds, we estimate the probability that T is larger than ε_2.

$$\mathbb{P}[T \geq \varepsilon_2] \leq m\exp\left(-\frac{p\phi\varepsilon_1^2}{2\phi+2\varepsilon_1/3}\right) + \mathbb{P}[T \geq \varepsilon_2|\mathcal{E}_1]. \quad (6)$$

The second term above is bounded from above as follows

$$\begin{aligned}
&\mathbb{P}[T \geq \varepsilon_2|\mathcal{E}_1] \\
&\leq \sum_{r,s\in[m]} \mathbb{P}\left[\left|[A\widehat{\Sigma}_0^{(n)}A^T]_{rs} - [A\Sigma_0A^T]_{rs}\right| > \varepsilon_2\right] \\
&\overset{(a)}{\leq} 4m^2\exp\left(\frac{-n\varepsilon_2^2}{3200\left(\max_r\sum_{i=1}^p A_{ri}\right)^4\max_i([\Sigma_0]_{ii})^2}\right) \\
&\overset{(b)}{\leq} 4m^2\exp\left(\frac{-n\varepsilon_2^2}{3200p^4\phi^4(1+\varepsilon_1)^4(\max_{i\in[p]}[\Sigma_0]_{ii})^2}\right). \quad (7)
\end{aligned}$$

(a) follows from the concentration of the sample covariance matrix (Lemma 3) along with an application of Lemma 1, and (b) follows from Lemma 2 (or more precisely, since we are conditioning on the event $\mathcal{E}_1^c$ holds). Putting equations (6) and (7) together, we get the desired result. ∎

B. Analysis under H_1

In this section, we prove a probabilistic lower bound on T under H_1.

Proposition 2. *There exists constants $c_0 \in (0, 0.25), c > 0$ such that with probability at least*

$$1-p^{-c} - m\exp\left(-\frac{p\phi^2\varepsilon_1^2}{2\phi+2\varepsilon_1/3}\right) - 4m^2\exp\left(\frac{-n\varepsilon_3^2}{3200p^4\phi^4(1+\varepsilon_1)^4(\max_{i\in[p]}\Sigma_{ii})^2}\right)$$

the test statistic satisfies $T \geq \frac{\Gamma_{\min}}{2} - \varepsilon_3$ under H_1, provided the compression size m satisfies $m \geq \sqrt{2c_0\delta^2 dp\frac{\Gamma_{\max}}{\Gamma_{\min}}}$.

Proof: We begin the proof by first observing that

$$T = \left\|\widehat{\Sigma}_Y^{(n)} - A\Sigma_0 A^T\right\|_\infty \geq \left\|\Sigma_Y - A\Sigma_0 A^T\right\|_\infty - \left\|\widehat{\Sigma}_Y^{(n)} - \Sigma_Y\right\|_\infty \triangleq T_1 - T_2. \quad (8)$$

Therefore, to obtain a lower bound on T, we first obtain an upper bound on T_2 that holds with high probability. Notice that this is essentially the same statement as in Proposition 1 with a different covariance matrix. Therefore, it follows that

$$\mathbb{P}\left[T_2 > \varepsilon_3\right] \leq m\exp\left(-\frac{p\phi\varepsilon_1^2}{2\phi+2\varepsilon_1/3}\right) + 4m^2\exp\left(\frac{-n\varepsilon_3^2}{3200p^4\phi^4(1+\varepsilon_1)^4(\max_{i\in[p]}\Sigma_{ii})^2}\right). \quad (9)$$

Next, we bound T_1 from below. Towards this end, as before, we let $\Gamma = \Sigma - \Sigma_0$. That is, with this notation, we can write $T_1 = \left\|A\Gamma A^T\right\|_\infty$. For the sake of this proof, and for the ease of relating our results back to those in [7], here we will work with the "tensor product form" of the quantity $A\Gamma A^T$. Let γ denote vec(Γ) $\in \mathbb{R}^{p^2}$. To keep presentation simple, we write γ_{ij} to index γ where $i, j \in [p]$, that is, $\gamma_{ij} = \Gamma_{ij}$. Also, we let $\mathcal{A} \in \mathbb{R}^{m^2\times p^2}$ denote the matrix $A \otimes A$, the tensor (or Kronecker) product of A with itself. And, as above, we index this matrix as $\mathcal{A}_{rs,ij}$, where $r, s \in [m]$ and $i, j \in [p]$. Notice that from the definition of the tensor product, $\mathcal{A}_{rs,ij} = A_{ri}A_{sj}$. With this notation in place, observe that the following holds for any $r, s \in [m]$

$$[\Sigma_Y - A\Sigma_0 A^T]_{rs} = \left[A\Gamma A^T\right]_{rs} = [\mathcal{A}\gamma]_{rs} = \sum_{i,j\in[p]} \mathcal{A}_{rs,ij}\gamma_{ij}.$$

And, with this notation $T_1 = \max_{r,s\in[m]}\left|\sum_{i,j\in[p]}\mathcal{A}_{rs,ij}\gamma_{ij}\right|$. As in [7], we may think of $\mathcal{A} = A\otimes A$ as the adjacency matrix of a *tensor product graph* $G\otimes G$, where G is drawn randomly according to Definition 1. Notice that $G\otimes G$ is also a bipartite graph with left and right sets given by $[p]\times[p]$ and $[m]\times[m]$ respectively. The edge set of $G\otimes G$ (denoted as $E\otimes E$) is such that $(r,i),(s,j)\in E \Leftrightarrow (rs,ij)\in E\otimes E$.

Following along the lines of [7, Section IV], we begin by considering an arbitrary ordering of the set $[p]\times[p]$ and we order the edges in $E\otimes E$ lexicographically based on this ordering, i.e., the first δ^2 edges $e_1,\ldots,e_{\delta^2}$ in $E_1\otimes E_2$ are those that correspond to the first element as per the ordering on $[p]\times[p]$ and so on. One can imagine that the graph $G\otimes G$ is formed by including these edges sequentially as per the ordering on the edges. This enables us to partition the edge set into the set E_{coll}^c of edges that do not collide with any of the previous edges as per the ordering, and the set $E_{\text{coll}} := (E\otimes E)\setminus E_{\text{coll}}^c$. (A similar proof technique appears in [16]). Let $\Omega\subset[p]\times[p]$ denote the support of Γ, and also we write $\mathbb{1}_E^{ijrs}$ to mean $\mathbb{1}\{(ij,rs)\in E\}$.

$$\max_{r,s\in[m]}\left|\sum_{i,j\in[p]}\mathcal{A}_{rs,ij}\gamma_{ij}\right| = \max_{r,s\in[m]}\left|\sum_{(ij,rs)\in E\otimes E}\gamma_{ij}\right| = \max_{r,s\in[m]}\left|\sum_{(ij,rs)\in E_{\text{coll}}^c}\gamma_{ij} + \sum_{(ij,rs)\in E_{\text{coll}}}\gamma_{ij}\right| \geq \max_{r,s\in[m]}\left(\left|\sum_{(i,j)\in\Omega}\mathbb{1}_{E_{\text{coll}}^c}^{ijrs}\gamma_{ij}\right| - \left|\sum_{(i,j)\in\Omega}\mathbb{1}_{E_{\text{coll}}}^{ijrs}\gamma_{ij}\right|\right). \quad (10)$$

Before we proceed, observe that for a fixed $r, s \in [m]$, the first sum above has only one term in it by the definition of the set E_{coll}^c. Therefore, the first term above is bounded from below by $\Gamma_{\min}$. This, along with a triangle inequality on the second term above, implies that:

$$\begin{aligned} T_1 &\geq \max_{r,s\in[m]}\left(\Gamma_{\min} - \sum_{(i,j)\in\Omega}\mathbb{1}_{E_{\text{coll}}}^{ijrs}\left|\gamma_{ij}\right|\right) \\ &= \Gamma_{\min} - \min_{r,s\in[m]}\sum_{(i,j)\in\Omega}\mathbb{1}_{E_{\text{coll}}}^{ijrs}\left|\gamma_{ij}\right| \\ &\geq \Gamma_{\min} - \frac{1}{m^2}\sum_{r,s\in[m]}\sum_{(i,j)\in\Omega}\mathbb{1}_{E_{\text{coll}}}^{ijrs}\left|\gamma_{ij}\right|, \end{aligned} \quad (11)$$

where in the last step we use the average to bound the minimum. Next, we observe that $\sum_{r,s\in[m]}\sum_{i,j\in\Omega}\mathbb{1}_{E_{\text{coll}}}^{ijrs}\left|\gamma_{ij}\right| \leq \sum_{i,j\in[p]} c_{ij}\left|\gamma_{ij}\right|$, where c_{ij} is the number of collisions involving the edges emanating from (i,j). Now, we use a key result from [7], which we paraphrase as Lemma 5 (or more specifically Corollary 1) in Appendix A, that guarantees that there is a constant $c_0 \in (0, 0.25)$ such that $c_{ij} \leq c_0\delta^2$ for all $(i,j) \in \Omega$ with probability at least $1 - p^{-c}$ provided $\delta \in \mathcal{O}(\log p)$. Therefore, w.h.p, we may bound (11) as follows

$$T_1 \geq \Gamma_{\min} - \frac{c_0\delta^2}{m^2}\sum_{(i,j)\in\Omega}\left|\gamma_{ij}\right| \geq \Gamma_{\min} - \frac{c_0\delta^2\Gamma_{\max}dp}{m^2}. \quad (12)$$

This means that T_1 is bounded from below by $\Gamma_{\min}/2$ with probability at least $1 - p^{-c}$, provided $m \geq \sqrt{2c_0\delta^2 dp\frac{\Gamma_{\max}}{\Gamma_{\min}}}$.

Putting this together with (9) in (8), we obtain the desired result.

■

Finally, we combine the results of Proposition 1 and 2 to prove Theorem 1

C. Combining Propositions 1 and 2

Notice that if we set $\varepsilon_3 = \varepsilon_2 = \eta = \frac{\Gamma_{\min}}{4}$, then, by a union bound, we have a valid test except with probability at most

$$m \exp\left(-\frac{p\phi\varepsilon_1^2}{2\phi + 2\varepsilon_1/3}\right) + p^{-c} + 4m^2 \exp\left(\frac{-n\Gamma_{\min}^2}{16 \cdot 3200 p^4 \phi^4 (1+\varepsilon_1)^4 (\max_{i\in[p]}[\Sigma_0]_{ii})^2}\right) + 4m^2 \exp\left(\frac{-n\Gamma_{\min}^2}{51200 p^4 \phi^4 (1+\varepsilon_1)^4 (\max_{i\in[p]}\Sigma_{ii})^2}\right). \quad (13)$$

First, we observe that since $\phi = \frac{\delta}{m-\delta+1}$, there exists a constant $c_1 \in (0,1)$ such that provided $m < c_1 \frac{p}{\log p}$, we have that $m \exp\left(-\frac{p\phi\varepsilon_1^2}{2\phi+2\varepsilon_1/3}\right) \leq p^{-c}$. Next, we use the fact that $\delta \in \mathcal{O}(\log p)$ to deduce that there is a constant $c_2 > 0$ such that if

$$n \geq c_2 \left(\frac{p \log p}{m}\right)^4 \frac{\max_i \Sigma_{ii}^2 \vee \max_i [\Sigma_0]_{ii}^2}{\Gamma_{\min}^2}, \quad (14)$$

the third and fourth term in (13) are no greater than p^{-c}. This gives us the proof of Theorem 1.

REFERENCES

[1] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *The Annals of Statistics*, vol. 36, pp. 2577–2604, Dec. 2008.

[2] J. Bien and R. J. Tibshirani, "Sparse estimation of a covariance matrix," *Biometrika*, vol. 98, pp. 807–820, Nov. 2011.

[3] T. T. Cai and H. H. Zhou, "Optimal rates of convergence for sparse covariance matrix estimation," *The Annals of Statistics*, vol. 40, pp. 2389–2420, Oct. 2012.

[4] T. T. Cai and Z. Ma, "Optimal hypothesis testing for high dimensional covariance matrices," *Bernoulli*, vol. 19, pp. 2359–2388, Nov. 2013.

[5] T. Cai, W. Liu, and Y. Xia, "Two-Sample Covariance Matrix Testing and Support Recovery in High-Dimensional and Sparse Settings," *Journal of the American Statistical Association*, vol. 108, pp. 265–277, Jan. 2013.

[6] W. Li and Y. Qin, "Hypothesis testing for high-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 128, pp. 108–119, July 2014.

[7] G. Dasarathy, P. Shah, B. N. Bhaskar, and R. D. Nowak, "Sketching Sparse Matrices, Covariances, and Graphs via Tensor Products," *IEEE Transactions on Information Theory*, vol. 61, no. 3, pp. 1373–1388, 2015.

[8] H. Liu, L. Wang, and T. Zhao, "Sparse Covariance Matrix Estimation With Eigenvalue Constraints," *Journal of Computational and Graphical Statistics*, vol. 23, pp. 439–459, Apr. 2014.

[9] T. Cai, C.-H. Zhang, and H. H. Zhou, "Optimal rates of convergence for covariance matrix estimation," *The Annals of Statistics*, vol. 38, pp. 2118–2144, Aug. 2010.

[10] T. Cai and A. Zhang, "ROP: Matrix recovery via rank-one projections," *The Annals of Statistics*, vol. 43, pp. 102–138, Feb. 2015.

[11] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and Stable Covariance Estimation From Quadratic Sampling via Convex Programming," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 4034–4059.

[12] M. Azizyan, A. Krishnamurthy, and A. Singh, "Extreme Compressive Sampling for Covariance Estimation," June 2015.

[13] G. Dasarathy, P. Shah, B. N. Bhaskar, and R. D. Nowak, "Covariance sketching," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1026–1033, IEEE, 2012.

[14] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing $\ell 1$-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[15] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[16] R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff, and M. J. Strauss, "Combining geometry and combinatorics: A unified approach to sparse signal recovery," in *Allerton Conference on Communication, Control, and Computing*, pp. 798–805, IEEE, 2008.

[17] S. Bernstein, "Theory of probability," (Russian) 1927.

APPENDIX A
AUXILIARY RESULTS

We will need the following result on the concentration of the elements of the sample covariance matrix about their true values; this is a specialization of [14, Lemma 1].

Lemma 3 (Sample covariance Concentration [14])**.** *Let $Y_1, \ldots, Y_n$ be i.i.d samples from $\mathcal{N}(0, \Sigma_Y)$. The sample covariance matrix $\widehat{\Sigma}_Y^{(n)} = \frac{1}{n}\sum_{k\in[n]} Y_k Y_k^T$ satisfies the following deviation inequalities for any fixed $(r,s) \in [m] \times [m]$:*

$$\mathbb{P}\left[\left|[\widehat{\Sigma}_Y^{(n)}]_{rs} - [\Sigma_Y]_{rs}\right| > \varepsilon\right] \leq 4\exp\left\{\frac{-n\varepsilon^2}{3200\left(\max_r [\Sigma_Y]_{rr}\right)^2}\right\}.$$

We will also need the Bernstein's inequality, which we state here [17].

Lemma 4 (Bernstein's Inequality)**.** *Let $R_1, \ldots, R_p$ be independent $0-$mean random variables such that $|R_i| \leq \zeta$ a.s. Let $\sigma^2 \geq \frac{1}{p}\sum_{i=1}^p \mathrm{Var}[R_i]$. Then for any $a > 0$, we have $\mathbb{P}[\sum_{i=1}^p R_i \geq pa] \leq \exp\left(-pa^2/2\sigma^2 + 2\zeta a/3\right)$.*

Finally, we will state the following lemma from [7, Lemma 1] that states a crucial property of the random tensor product graphs that we are considering here. Letting $N(S)$ denote the neighborhood of a left set $S \subset [p]$, we have:

Lemma 5. *Suppose $G = ([p],[m],E)$ is a random $\delta-$left regular bipartite graph (cf. Definition 1) chosen such that $\delta = \mathcal{O}(\log p)$ and $m \in \Omega(\sqrt{dp}\log p)$. Let $\Omega \in [p] \times [p]$ denote the support of (a $d-$distributed sparse) Γ. Then there exists a constant $c_0 \in \left(0, \frac{1}{4}\right)$ such that $G_1 \otimes G_2$ has the following properties with probability exceeding $1 - p^{-c}$, for some $c > 0$: (a) $|N(\Omega)| \geq p\delta^2(1-c_0)$. (b) For any $(i,i') \in ([p] \times [p]) \setminus \Omega$ we have $|N(i,i') \cap N(\Omega)| \leq c_0\delta^2$. (c) For any $(i,i') \in \Omega$, $|N(i,i') \cap N(\Omega \setminus (i,i'))| \leq c_0\delta^2$.*

In particular, we will need the following corollary, which follows from part (c) in Lemma 5.

Corollary 1. *Under the conditions of Lemma 5, if we let $c_{ij}, (i,j) \in \Omega$ be as in the proof of Proposition 2, then we have that $c_{ij} \leq c_0\delta^2$ for all $(i,j) \in \Omega$ with probability at least $1 - p^{-c}$.*