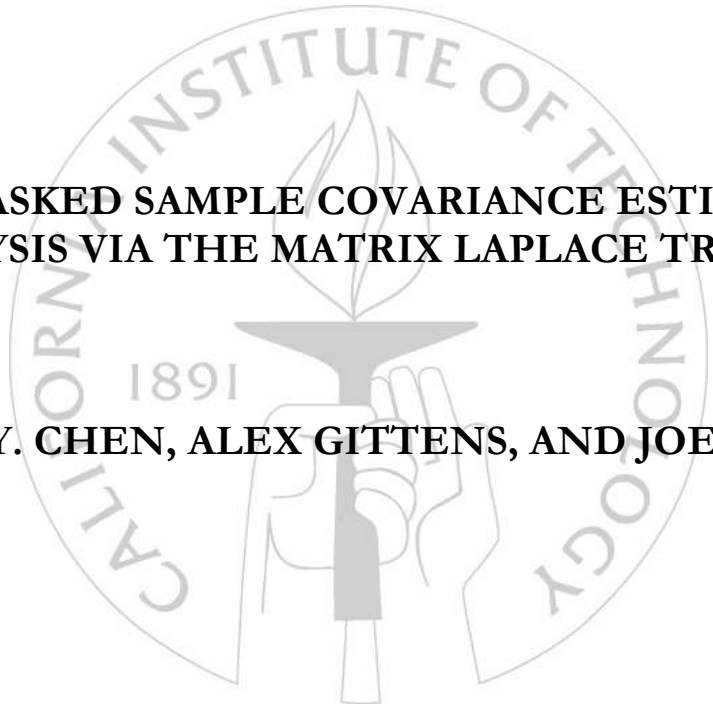

THE MASKED SAMPLE COVARIANCE ESTIMATOR: AN ANALYSIS VIA THE MATRIX LAPLACE TRANSFORM

RICHARD Y. CHEN, ALEX GITTENS, AND JOEL A. TROPP

Technical Report No. 2012-01
February 2012

APPLIED & COMPUTATIONAL MATHEMATICS
CALIFORNIA INSTITUTE OF TECHNOLOGY
mail code 9-94 · pasadena, ca 91125

THE MASKED SAMPLE COVARIANCE ESTIMATOR: AN ANALYSIS VIA THE MATRIX LAPLACE TRANSFORM

RICHARD Y. CHEN, ALEX GITTENS, AND JOEL A. TROPP

Abstract. Covariance estimation becomes challenging in the regime where the number p of variables outstrips the number n of samples available to construct the estimate. One way to circumvent this problem is to assume that the covariance matrix is nearly sparse and to focus on estimating only the significant entries. To analyze this approach, Levina and Vershynin (2011) introduce a formalism called *masked covariance estimation*, where each entry of the sample covariance estimator is reweighed to reflect an *a priori* assessment of its importance.

This paper provides a new analysis of the masked sample covariance estimator based on the *matrix Laplace transform method*. The main result applies to general subgaussian distributions. Specialized to the case of a Gaussian distribution, the theory offers qualitative improvements over earlier work. For example, the new results show that $n = \mathrm{O}(B \log^2 p)$ samples suffice to estimate a banded covariance matrix with bandwidth B up to a relative spectral-norm error, in contrast to the sample complexity $n = \mathrm{O}(B \log^5 p)$ obtained by Levina and Vershynin.

1. Introduction

A fundamental problem in multivariate statistics is to obtain an accurate estimate of the covariance matrix of a multivariate distribution given independent samples from the distribution. This challenge arises whenever we need to understand the spread of the data and its marginals, for example, when we perform regression analysis [Fre05] or principal component analysis [Jol02].

In the classical setting where the number of samples exceeds the number of variables, the behavior of standard covariance estimators is textbook material [JW02, Mui82, MKB80]. The random matrix literature also contains a huge amount of relevant work; we refer to the book [BS10] and the survey [Ver11] for further information.

Modern applications, in contrast, often involve a small number of samples and a large number of variables. The paucity of data makes it impossible to obtain an accurate estimate of a general covariance matrix. As a remedy, we must frame additional assumptions on the model and develop estimators that exploit this extra structure. Over the last few years, a number of papers, including [FB07, BL08b, BL08a, Kar08, RLZ09, CZZ10], have focused on the situation where the covariance matrix is sparse or nearly so. In this case, we imagine that we could limit our attention to the significant entries of the covariance matrix and thereby perform more accurate estimation with fewer samples.

This paper studies a particular technique for the sparse covariance problem that we call the *masked sample covariance estimator*. This approach uses a mask matrix, constructed *a priori*, to specify the importance we place on each entry of the covariance matrix. By reweighting the sample covariance estimate using a mask, we can reduce the influence of entires that we cannot estimate reliably. For instance, if the covariance matrix is approximated by a banded matrix, one sets the entries of the mask to zero outside of the band. This formalism was introduced by Levina and

Date: 6 September 2011. Corrected 28 September 2011.

2010 *Mathematics Subject Classification.* Primary: 60B20.

The authors are with Computing and Mathematical Sciences, MC 305-16, California Inst. Technology, Pasadena, CA 91125. E-mail: ycchen@caltech.edu, gittens@caltech.edu, jtropp@cms.caltech.edu. Research supported by ONR awards N00014-08-1-0883 and N00014-11-1-0025, DARPA award N66001-08-1-2065, AFOSR award FA9550-09-1-0643, and a Sloan Fellowship.

Vershynin [LV11] to provide a unified treatment of earlier methods for sparse covariance estimation; we refer to their paper for a more detailed discussion of prior work.

Levina and Vershynin derive an elegant bound [LV11, Thm. 2.1] for masked covariance estimation of a Gaussian distribution. In this work, we develop a completely different analysis based on the *matrix Laplace transform method* [AW02, Tro11b]. The advantage of this approach is that it applies to general subgaussian distributions and it allows us to obtain more refined information about the quality of the masked sample covariance estimator.

The rest of this Introduction provides an overview of masked covariance estimation and its relationship with classical covariance estimation. In Section 1.6, we present a simplified result for the behavior of the masked sample covariance estimator applied to a Gaussian distribution, and we offer a concrete comparison with the results of Levina and Vershynin [LV11, Thm. 2.1]. More detailed results appear in Section 3.

1.1. **Classical Covariance Estimation.** Consider a random vector

$$\boldsymbol{x} = (X_1, X_2, \dots, X_p)^* \in \mathbb{R}^p.$$

Let $\boldsymbol{x}_1, \dots \boldsymbol{x}_n$ be independent random vectors that follow the same distribution as $\boldsymbol{x}$. For simplicity, we assume that the distribution is known to have zero mean: $\mathbb{E}\,\boldsymbol{x} = \mathbf{0}$. The *covariance matrix* $\boldsymbol{\Sigma}$ is a $p \times p$ matrix that tabulates the second-order statistics of the distribution:

$$\boldsymbol{\Sigma} := \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^*) \tag{1.1}$$

where * denotes the transpose operation. The classical estimator for the covariance matrix is the *sample covariance matrix*, which is obtained from (1.1) by the plug-in principle:

$$\widehat{\boldsymbol{\Sigma}}_n := \frac{1}{n} \sum\nolimits_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^*. \tag{1.2}$$

The sample covariance matrix is an unbiased estimator of the covariance matrix.

Given a tolerance $\varepsilon \in (0, 1)$, we can study how many samples n are typically required to provide an estimate with relative error ε in the spectral norm:

$$\mathbb{E} \left\| \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma} \right\| \leq \varepsilon \left\| \boldsymbol{\Sigma} \right\|. \tag{1.3}$$

This type of spectral-norm error bound is quite powerful. It limits the magnitude of the estimation error for each entry of the covariance matrix; it provides information about the variance of each marginal of the distribution of $\boldsymbol{x}$; it even controls the error in estimating the eigenvalues of the covariance using the eigenvalues of the sample covariance.

Unfortunately, an error bound of the form (1.3) demands a lot of samples. Suppose that the covariance matrix has full rank. Then the number of samples must be at least as large as the number of variables to obtain a nontrivial guarantee. Indeed, when $n < p$, the sample covariance does not even have full rank, so the spectral norm error is bounded away from zero!

Typical positive results on covariance estimation state that we can obtain an accurate estimate for the covariance matrix when the number of samples is proportional to the number of variables, provided that the distribution decays fast enough. For example, assuming that $\boldsymbol{x}$ follows a normal distribution,

$$n \geq \mathrm{C}\, \varepsilon^{-2} p \quad \Longrightarrow \quad \left\| \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma} \right\| \leq \varepsilon \left\| \boldsymbol{\Sigma} \right\| \quad \text{with high probability.} \tag{1.4}$$

We use the analyst's convention that C denotes an absolute constant whose value may change from appearance to appearance. See [Ver11, Thm. 57 et seq.] for details of obtaining the bound (1.4). The work of Srivastava and Vershynin [SV11] contains the most recent news on the classical covariance estimation problem.

1.2. **Motivation for Masked Covariance Estimation.** In the regime $n \ll p$, where we have very few samples, we can never hope to achieve the estimate (1.3). So we must lower our standards. The following example provides some insight on how to proceed.

Example 1.1 (Simultaneous Variance Estimation). Let us how many realizations of a Gaussian random vector we need to accurately estimate the variance of each component.

First, suppose that Z is a zero-mean normal variable with variance v. Given independent copies $Z_1, \dots, Z_n$ of the random variable Z, we can compute the sample variance

$$\widehat{v} := \frac{1}{n} \sum\nolimits_{i=1}^{n} Z_i^2.$$

The estimator $\widehat{v}$ is unbiased, and, up to scaling, it follows a chi-square distribution, so the probability of error satisfies

$$\mathbb{P}\left\{|\widehat{v} - v| \geq \delta v\right\} \leq 2\,\mathrm{e}^{-nt^2/4} \quad \text{for } t \in (0, 1). \tag{1.5}$$

For a clean proof of this type of inequality, see [Bar05, Prop. (2.2)].

Next, suppose that the random vector $\boldsymbol{x}$ follows a zero-mean normal distribution with arbitrary covariance $\mathbf{\Sigma}$, and write σ_{ij} for the (i, j) entry of this matrix. When we use the sample covariance to estimate each of the p diagonal entries of $\mathbf{\Sigma}$, the bound (1.5) implies that

$$\mathbb{P}\left\{\max\nolimits_i |(\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma})_{ii}| \geq (\max\nolimits_i \sigma_{ii}) \cdot t\right\} \leq 2p\,\mathrm{e}^{-nt^2/4}.$$

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

$$n \geq \mathrm{C}\,\varepsilon^{-2} \log p \quad \Longrightarrow \quad \max\nolimits_i |(\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma})_{ii}| \leq \varepsilon \max\nolimits_i \sigma_{ii} \quad \text{with high probability.} \tag{1.6}$$

Since $\max_i \sigma_{ii} \leq \|\mathbf{\Sigma}\|$, the error obtained in (1.6) is smaller than the spectral-norm error in (1.4).

When the covariance $\mathbf{\Sigma} = \mathbf{I}$, it can be shown that at least $\log p$ samples are *required* to achieve the bound (1.6).

Example 1.1 suggests an intriguing possibility. Although we need at least p samples to estimate the entire covariance matrix, roughly $\log p$ samples suffice to estimate the diagonal. It turns out that this phenomenon is generic: *If we estimate only a small portion of the covariance matrix, then we can reduce the number of samples dramatically.* This observation is widely applicable because there are many problems where we do not need to know all of the second-order statistics.

Partitioning Variables: Suppose that we divide the stock market into disjoint sectors, and we would like to study the interactions among the monthly returns for stocks within each sector. The list of returns for all the stocks can be treated as a random vector. We block the covariance matrix of this random vector to conform with the market sectors, and we estimate only the entries in the diagonal blocks.

Spatial or Temporal Localization: A simple random model for grayscale images treats the intensity of each pixel as a random variable. Nearby pixels tend to be bright or dark together, while distant pixels are usually uncorrelated. Thus, we might limit our attention to the interactions between a pixel and the pixels directly adjacent to it. This model suggests that we estimate the entries of the covariance that lie within a (generalized) band about the diagonal.

Graph Structures: Consider a stochastic model for the spread of an epidemic through a social network. At each time instant, we label an individual with a random variable that measures how sick he is. Since transmission only occurs along links in the network, neighbors are likely to be sick or well together. As a result, we might want to focus on estimating the covariance for individuals separated by one degree. In this case, the adjacency matrix of the graph determines which pairs to estimate.

1.3. **The Mask Matrix.** We can treat all the examples from Section 1.2 using a formalism that was introduced by Levina and Vershynin [LV11]. Let $\boldsymbol{M}$ be a fixed $p \times p$ symmetric matrix with real entries, which we call the *mask matrix*. The basic idea is to construct a mask that guides our attention to specific parts of the covariance matrix.

In the simplest case, the mask has 0–1 values that indicate which entries of the covariance we must attend to. The presence of a unit entry $m_{ij} = 1$ tells us to estimate the interaction between the ith and jth variable; a zero entry $m_{ij} = 0$ means that we abdicate from making any estimate of this interaction. In Example 1.1, we are only interested in the diagonal entries of the covariance, so we are using the mask $\boldsymbol{M}_{\mathrm{diag}} = \mathbf{I}$. Here are some other basic examples:

$$\boldsymbol{M}_{\mathrm{group}} := \begin{bmatrix} 1 & 1 & & & & \\ 1 & 1 & & & & \\ & & 1 & 1 & & \\ & & 1 & 1 & & \\ & & & & & 1 \end{bmatrix}; \quad \boldsymbol{M}_{\mathrm{band}} := \begin{bmatrix} 1 & 1 & & & & \\ 1 & 1 & 1 & & & \\ & 1 & 1 & 1 & & \\ & & 1 & 1 & 1 & \\ & & & 1 & 1 & \end{bmatrix}; \quad \boldsymbol{M}_{\mathrm{graph}} := \begin{bmatrix} 1 & & & 1 & 1 & \\ & 1 & & & & 1 \\ & & 1 & & & 1 \\ 1 & & & 1 & & \\ 1 & & & & 1 & \\ & 1 & & & & 1 \end{bmatrix}.$$

The matrix $\boldsymbol{M}_{\mathrm{group}}$ corresponds to the case where we partition variables into three subgroups, and we make estimates only within subgroups. Masks such as $\boldsymbol{M}_{\mathrm{band}}$ arise from banded covariance estimation, which occurs for spatially localized random fields. The mask $\boldsymbol{M}_{\mathrm{graph}}$ might occur when the variables exhibit a graphical dependency structure.

In more complicated situations, we can allow the mask to take arbitrary nonnegative values and then interpret the magnitude of each entry as a requirement on the precision of the estimate. When m_{ij} is large, we must study the interaction between the ith and jth variable carefully. When m_{ij} is small, we are less vigilant about how well we estimate the (i, j) entry of the covariance matrix. An example of a mask with general entries is the Kac matrix

$$\boldsymbol{M}_{\mathrm{Kac}} := \begin{bmatrix} 1 & \varphi & \varphi^2 & \varphi^3 & \varphi^4 \\ \varphi & 1 & \varphi & \varphi^2 & \varphi^3 \\ \varphi^2 & \varphi & 1 & \varphi & \varphi^2 \\ \varphi^3 & \varphi^2 & \varphi & 1 & \varphi \\ \varphi^4 & \varphi^3 & \varphi^2 & \varphi & 1 \end{bmatrix} \quad \text{where } \varphi \in (0, 1).$$

The mask $\boldsymbol{M}_{\mathrm{Kac}}$ tapers the covariances exponentially depending on the distance $|i - j|$ between the variables. This type of example might be relevant for the study of spatially localized processes.

Most of the regularization techniques for sparse covariance estimation studied in the literature, such as [BL08b, FB07, CZZ10], can be described using mask matrices. The initial works focus on specific cases, such as banded masks and tapered masks, whereas we have followed Levina and Vershynin [LV11] by allowing an arbitrary symmetric matrix $\boldsymbol{M}$. We refer to the papers cited in this paragraph for further background and references.

Remark 1.2. Let us emphasize that the entries of the mask can take both positive and negative values, but it is harder to find a clear interpretation of a mask that has negative entries.

1.4. **The Masked Sample Covariance Estimator.** Suppose that we have specified a symmetric $p \times p$ mask $\boldsymbol{M}$ with real entries. The *masked covariance* and the *masked sample covariance estimator* are the two matrices

$$\boldsymbol{M} \odot \boldsymbol{\Sigma} \quad \text{and} \quad \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n,$$

where the symbol $\odot$ denotes the componentwise (i.e., Schur or Hadamard) product. The goal of this work is to study the error incurred when we estimate the masked covariance matrix using the masked sample covariance:

$$\left\| \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma} \right\|. \tag{1.7}$$

As noted by Levina and Vershynin [LV11, Sec. 1], control on the error (1.7) also delivers information about how well we estimate the full covariance because

$$\big\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\big\| \le \big\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\big\| + \big\|\boldsymbol{M} \odot \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\big\|. \tag{1.8}$$

The first term in (1.8) reflects the variance of the estimator about its mean value, while the second term represents the bias in the estimate owing to the presence of the mask. It is important to select a mask $\boldsymbol{M}$ that simultaneously controls both the variance and the bias. Understanding the variance term requires an excursion into random matrix theory, and it comprises the main subject of this work. Studying the bias term involves only a deterministic analysis, which should be undertaken with a specific application in mind.

When the error (1.7) is small, the masked sample covariance yields accurate estimates for each component of the covariance where the corresponding entry of $\boldsymbol{M}$ is large, as well as the variance of some specially chosen marginals. When the error (1.8) is also small, the masked sample covariance provides additional information about the variance of all marginals of the distribution of $\boldsymbol{x}$, as well as estimates for the eigenvalues of the covariance.

1.5. **The Complexity of a Mask.** The number of samples we need to control (1.7) depends on "how much" of the covariance matrix we are attempting to estimate. We quantify the complexity of the mask using two separate metrics. First, define the square of the maximum column norm of the mask matrix:

$$\|\boldsymbol{M}\|_{1\to 2}^2 := \max_j \left(\sum\nolimits_i m_{ij}^2\right).$$

Roughly, the parenthesis reflects the number of interactions we want to estimate that involve the variable j, and the maximum computes a bound over all p variables. The second metric is the spectral norm $\|\boldsymbol{M}\|$ of the mask matrix, which provides a more global view of the complexity of the interactions that we estimate.

Some examples may illuminate how these metrics reflect the properties of the mask. First, suppose that we estimate the entire covariance matrix, so the mask is the matrix of ones:

$$\boldsymbol{M} = \text{matrix of ones} \quad\Longrightarrow\quad \|\boldsymbol{M}\|_{1\to 2}^2 = p \quad\text{and}\quad \|\boldsymbol{M}\| = p.$$

We will see that the value p here corresponds with the factor p in the sample complexity bound (1.4). Next, consider the mask that arises in banded covariance estimation:

$$\boldsymbol{M} = \text{0–1 matrix, bandwidth } B \quad\Longrightarrow\quad \|\boldsymbol{M}\|_{1\to 2}^2 \le B \quad\text{and}\quad \|\boldsymbol{M}\| \le B$$

because there are at most B ones in each row and column. When $B \ll p$, the banded mask is much less complex than the matrix of ones, and estimation is commensurately easier. Third, assuming the mask is a Kac matrix, we have

$$\boldsymbol{M} = \text{Kac matrix, parameter } \varphi \quad\Longrightarrow\quad \|\boldsymbol{M}\|_{1\to 2}^2 \le \frac{1}{1-\varphi^2} \quad\text{and}\quad \|\boldsymbol{M}\| \le \frac{1}{1-\varphi}.$$

For a fixed value of φ, neither quantity depends on the total number of variables, so covariance estimation with this mask should require very few samples.

Remark 1.3. In each example above, the two metrics take very similar values, but this coincidence does not always occur. Although the spectral norm dominates the maximum column norm, the *square* of the maximum column norm can be substantially larger or substantially smaller than the spectral norm. We have omitted examples to support this point because they do not seem to arise naturally in the setting of masked covariance estimation.

1.6. **Masked Covariance Estimation for Gaussian Distributions.** This paper develops a bound for the estimation error (1.7) when the random vector $\boldsymbol{x}$ follows a subgaussian distribution with zero mean. For illustrative purposes, this section focuses on the simpler case where the random vector has a normal distribution. The general results appear in Section 3.

Theorem 1.4 (Masked Covariance Estimation for Gaussian Distributions)**.** *Fix a $p \times p$ symmetric mask matrix $\boldsymbol{M}$. Suppose that $\boldsymbol{x}$ is a Gaussian random vector in $\mathbb{R}^p$ with mean zero. Define the covariance matrix $\boldsymbol{\Sigma}$ and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ as in* (1.1) *and* (1.2)*. Then the expected estimation error satisfies*

$$\mathbb{E}\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \leq 8\left[\left(\frac{\|\boldsymbol{M}\|_{1\to 2}^2 \log(6p)}{n}\right)^{1/2} + \frac{\|\boldsymbol{M}\| \log^2(6np)}{n}\right]\|\boldsymbol{\Sigma}\|. \tag{1.9}$$

Theorem 1.4 is a simplified version of Corollary 3.3. The reader is encouraged to examine the full result, which includes several substantial refinements.

Remark 1.5. In the actual practice of covariance estimation, we center each sample empirically by subtracting the sample mean $\bar{\boldsymbol{x}} = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i$. The sample covariance (1.2) is computed using the centered samples $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}}$ instead of the original samples $\boldsymbol{x}_i$. The theory in this paper can be extended to cover the masked covariance estimator formed with centered samples; see [LV11, Rem. 4] for the details of the argument.

1.6.1. *Sample Complexity Bound.* Theorem 1.4 allows us to develop conditions on the number n of samples that we need to control the estimation error with high probability. Markov's inequality can be used to convert (1.9) into an error bound that holds in probability. For example, with probability at least 99%,

$$\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \leq \mathrm{C}\left[\left(\frac{\|\boldsymbol{M}\|_{1\to 2}^2 \log p}{n}\right)^{1/2} + \frac{\|\boldsymbol{M}\| \log^2(np)}{n}\right]\|\boldsymbol{\Sigma}\|. \tag{1.10}$$

For stronger exponential error bounds, we refer to Corollary 3.3. To obtain the sample complexity, assume that $n \leq p$, and let $\varepsilon \in (0, 1)$. Then (1.10) yields the statement

$$n \geq \mathrm{C}\left[\varepsilon^{-2}\|\boldsymbol{M}\|_{1\to 2}^2 \log p + \varepsilon^{-1}\|\boldsymbol{M}\| \log^2 p\right] \quad \Longrightarrow \quad \left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \leq \varepsilon\|\boldsymbol{\Sigma}\| \tag{1.11}$$

with probability at least 99%.

1.6.2. *Is this Sample Complexity Bound Optimal?* Levina and Vershynin show that the sample complexity of masked covariance estimation must exhibit a logarithmic dependence on the number p of variables [LV11, Rem. 3]. They also argue that there should be a linear dependence on the maximum number of interactions that involve a single variable [LV11, Eqn. (1.4) et seq.]; this term appears in (1.11) in the guise of $\|\boldsymbol{M}\|_{1\to 2}^2$. As a consequence of these observations, it seems plausible that the first summand in the sample bound (1.11) has the optimal form. On the other hand, we believe that the factor $\log^2 p$ in the second summand could probably be reduced to $\log p$.

The discussion in Example 1.1 suggests that it may be possible to improve the dependence of the sample complexity bound (1.11) on the spectral norm $\|\boldsymbol{\Sigma}\|$ of the covariance. Indeed, we have obtained a refinement of this type. See Corollary 3.3 for details.

1.6.3. *Application Example.* Consider the banded covariance estimation problem, with the mask

$$\boldsymbol{M} = 0\text{–}1 \text{ matrix with bandwidth } B.$$

See the matrix $\boldsymbol{M}_{\mathrm{band}}$ displayed on page 4 for an instance with $B = 3$ and $p = 5$. The sample complexity bound (1.11) and the norm calculations from Section 1.5 demonstrate that

$$n \geq \mathrm{C}\left[\varepsilon^{-2} B \log p + \varepsilon^{-1} B \log^2 p\right] \tag{1.12}$$

is sufficient to provide a relative estimation error ε in spectral norm with 99% probability. For comparison, recall the sufficient condition (1.4) that the sample complexity for estimating the entire covariance with relative error ε satisfies

$$n \geq \mathrm{C}\, \varepsilon^{-2} p.$$

When the bandwidth is much smaller than the number of variables ($B \ll p$), the masked covariance estimator outperforms the classical covariance estimator. On the other hand, when the bandwidth is comparable with the number of variables, the analysis of the masked covariance estimator gives a sample complexity bound (1.12) that is worse by a polylogarithmic factor.

We remark that, when ε is constant, the second summand in (1.12) always dominates the first as $p \to \infty$. On the other hand, the first summand is larger when $\varepsilon \leq \log^{-1} p$. In other words, the excess logarithm in the second term of (1.12) does not have an impact on the sample complexity when we are seeking highly accurate covariance estimates.

1.6.4. *Comparison with Bounds of Levina and Vershynin.* Theorem 1.4 should be compared with the main result of Levina and Vershynin [LV11, Thm. 2.1], which states that

$$\mathbb{E} \left\| \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma} \right\| \leq \mathrm{C} \left[\frac{\|\boldsymbol{M}\|_{1\to 2} \log^{5/2} p}{\sqrt{n}} + \frac{\|\boldsymbol{M}\| \log^3 p}{n} \right] \|\boldsymbol{\Sigma}\|.$$

The associated sample complexity bound is

$$n \geq \mathrm{C} \left[\varepsilon^{-2} \|\boldsymbol{M}\|_{1\to 2}^2 \log^5 p + \varepsilon^{-1} \|\boldsymbol{M}\| \log^3 p \right]. \tag{1.13}$$

Our sample complexity bound (1.11) has exactly the same structure as (1.13), but we have managed to remove a moderate number of logarithms.

We do not feel that chopping down logs is an interesting pursuit *per se*. Instead, the value of this work stems from the fact that we have applied an argument that is completely different from previous work on masked covariance estimation. Our approach provides some qualitative refinements over Levina and Vershynin's bound in the Gaussian setting (Corollary 3.3), and it also extends to the general subgaussian distributions (Theorem 3.2).

1.6.5. *Proof Techniques.* The argument in this paper is based on a recent set of ideas, collectively known as the *matrix Laplace transform method*. This approach can be regarded as a generalization of the classical technique, attributed to Bernstein, that develops probability inequalities for a random variable in terms of bounds for its cumulant generating function. Tropp [Tro11b], building on work of Ahlswede and Winter [AW02], demonstrates that the scalar approach admits a tight analogy in the matrix setting. See Section 2.4 for an overview of this technique.

The matrix Laplace transform method is particularly well suited for studying sums of independent random matrices. To apply these techniques, we express the error as a sum of i.i.d. random matrices, each with zero mean:

$$\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma} = \frac{1}{n} \sum\nolimits_{i=1}^{n} \boldsymbol{M} \odot (\boldsymbol{x}_i \boldsymbol{x}_i^* - \mathbb{E}\, \boldsymbol{x}\boldsymbol{x}^*).$$

The main challenge is to study the matrix cumulant generating function of each summand:

$$\log \mathbb{E} \exp\left(\theta \boldsymbol{M} \odot (\boldsymbol{x}_i \boldsymbol{x}_i^* - \mathbb{E}\, \boldsymbol{x}\boldsymbol{x}^*)\right) \quad \text{for } \theta > 0. \tag{1.14}$$

The key technical result of this paper is a semidefinite upper bound for the matrix cgf (1.14). This estimate requires a number of substantial new ideas, including a symmetrization argument, a careful analysis of the variance of the random matrix in the exponent of (1.14), and a delicate truncation bound.

1.7. **Organization of the paper.** The rest of the paper is organized as follows. Section 2 introduces our notation and some preliminaries. Section 3 presents the main result for zero-mean subgaussian distributions, together with its proof and the proof of Theorem 1.4. In Section 4, we deal with the technical challenge of estimating the matrix cumulant generating function (1.14).

2. Preliminaries

This section sets out the background material we require for the proof. The argument depends on a very recent set of ideas, collectively known as the *matrix Laplace transform method*. We introduce the main results from this theory in Section 2.4, and we provide references to the primary sources. The rest of the material here is more or less standard. Section 2.1 states our notational conventions, Section 2.2 describes some basic properties of the Schur product, and Section 2.3 includes key facts about subgaussian random variables.

2.1. **Notation and Conventions.** In this paper, we work exclusively with real numbers. Plain italic letters always refer to scalars. Bold italic lowercase letters, such as $\boldsymbol{a}$, refer to column vectors. Bold italic uppercase letters, such as $\boldsymbol{A}$, denote matrices. All matrices in this work are square; the dimensions are determined by context. We write $\mathbf{0}$ for the zero matrix and $\mathbf{I}$ for the identity matrix. The matrix unit $\mathbf{E}_{ij}$ has a unit entry in the (i, j) position and zeros elsewhere.

The symbol * denotes the transpose operation on vectors and matrices. We use the term *self-adjoint* to refer to a matrix that satisfies $\boldsymbol{A} = \boldsymbol{A}^*$ to avoid confusion between symmetric matrices and symmetric random variables. Curly inequalities refer to the positive-semidefinite partial ordering on self-adjoint matrices: $\boldsymbol{A} \preccurlyeq \boldsymbol{B}$ if and only if $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite.

The function $\operatorname{diag}(\cdot)$ maps a vector $\boldsymbol{a}$ to a matrix whose diagonal entries correspond with the entries of $\boldsymbol{a}$. We write $\operatorname{tr}(\cdot)$ for the trace of a matrix. The symbol $\odot$ denotes the componentwise (i.e., Schur or Hadamard) product of two matrices.

We write $\|\cdot\|$ for both the ℓ_2 vector norm and the associated operator norm, which is usually called the *spectral norm*. The norm $\|\cdot\|_\infty$ returns the absolute maximum entry of a vector. For clarity, we use a separate notation $\|\cdot\|_{\max}$ for the absolute maximum entry of a matrix. The maximum column norm $\|\cdot\|_{1\to 2}$ is defined as

$$\|\boldsymbol{A}\|_{1\to 2} := \max_j \left(\sum\nolimits_i |a_{ij}|^2\right)^{1/2}.$$

The notation reflects the fact that this is the natural norm for linear maps from ℓ_1 into ℓ_2.

We reserve the symbol ε for a *Rademacher random variable*, which takes the two values ± 1 with equal probability. We also assume that all random variables are sufficiently regular that we are justified in computing expectations, interchanging limits, and so forth.

2.2. **Facts about the Schur Product.** The proof depends on some basic properties of Schur products. The first result is a simple but useful algebraic identity. For each square matrix $\boldsymbol{A}$ and each conforming vector $\boldsymbol{x}$,

$$\boldsymbol{A} \odot \boldsymbol{x}\boldsymbol{x}^* = \operatorname{diag}(\boldsymbol{x})\boldsymbol{A}\operatorname{diag}(\boldsymbol{x}). \tag{2.1}$$

The second result states that the Schur product with a positive-semidefinite matrix is order preserving. That is, for a fixed positive-semidefinite matrix $\boldsymbol{A}$,

$$\boldsymbol{B}_1 \preccurlyeq \boldsymbol{B}_2 \quad \text{implies} \quad \boldsymbol{A} \odot \boldsymbol{B}_1 \preccurlyeq \boldsymbol{A} \odot \boldsymbol{B}_2. \tag{2.2}$$

This property follows from Schur's theorem [HJ94, Thm. 7.5.3], which demonstrates that the Schur product of two positive-semidefinite matrices remains positive semidefinite.

2.3. **Subgaussian Random Variables.** There are several different ways to formalize the concept of a random variable that decays faster than a Gaussian random variable [Ver11]. For the purposes of this paper, the following definition is most convenient.

Definition 2.1 (Subgaussian random variable)**.** *A random variable X is* subgaussian *if there exists a positive constant K such that*

$$\mathbb{P}\{|X| > t\} \le 2\,\mathrm{e}^{-t^2/K^2} \quad \text{for all } t \ge 0.$$

The subgaussian coefficient $\kappa(X)$ *is defined to be the infimal K for which this inequality holds.*

We can bound all the moments of a subgaussian random variable X in terms of its subgaussian coefficient:

$$\mathbb{E}\,|X|^q = \int_0^\infty qt^{q-1}\,\mathbb{P}\{|X| > t\}\,\mathrm{d}t \le \int_0^\infty qt^{q-1}\cdot 2\,\mathrm{e}^{-t^2/\kappa(X)^2}\,\mathrm{d}t = 2\kappa(X)^q\,\Gamma(q/2+1).$$

In particular, the raw fourth moment of X satisfies

$$\mathbb{E}\,|X|^4 \le 4\kappa(X)^4. \tag{2.3}$$

2.4. **The Matrix Laplace Transform Method.** In classical probability, the Laplace transform method is a powerful tool for obtaining tail bounds for a sum of independent random variables. In their influential paper [AW02], Ahlswede and Winter describe a generalization of the Laplace transform method that applies to a sum of independent random matrices. Subsequent papers by Oliveira [Oli10a, Oli10b], by Tropp [Tro11b, Tro11a], and by Hsu et al. [HKZ11] all contain substantial refinements and extensions of the original idea. Altogether, these tools are easy to use, remarkably effective, and widely applicable.

In analogy with the scalar case, we study large deviations using a matrix version of the moment generating function (mgf) and the cumulant generating function (cgf). Let $\boldsymbol{Z}$ be a self-adjoint random matrix. Using the matrix exponential, we define the matrix mgf and matrix cgf, respectively, to be

$$\boldsymbol{M}_{\boldsymbol{Z}}(\theta) := \mathbb{E}\,\mathrm{e}^{\theta\boldsymbol{Z}} \quad \text{and} \quad \boldsymbol{\Xi}_{\boldsymbol{Z}}(\theta) := \log \mathbb{E}\,\mathrm{e}^{\theta\boldsymbol{Z}} \quad \text{for } \theta \in \mathbb{R}.$$

Note that these expectations may not exist for all values of θ. The matrix cgf can be interpreted as an *exponential mean*, an average that emphasizes large deviations of the spectrum with the same sign as the parameter θ.

The matrix mgf contains valuable information about the behavior of the maximum eigenvalue of a symmetric random matrix. The following result is a matrix analog of the classical approach to large deviations, which is attributed to Bernstein.

Proposition 2.2 (Matrix Laplace transform bound)**.** *Let $\boldsymbol{Z}$ be a random, self-adjoint matrix. For each $t \in \mathbb{R}$,*

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Z}) \ge t\} \le \inf_{\theta>0}\left\{\mathrm{e}^{-\theta t}\cdot \mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta\boldsymbol{Z}}\right\}. \tag{2.4}$$

In this form, Proposition 2.2 is due to Oliveira [Oli10b, Sec. 3], but the main idea goes back to the paper [AW02] of Ahlswede and Winter. See [Tro11b, Prop. 3.1] for a succinct proof.

In our application, the random matrix $\boldsymbol{Z}$ can be expressed as a sum of i.i.d. zero-mean random, self-adjoint matrices. The argument relies on a symmetrization procedure, which introduces additional randomness into the series.

Proposition 2.3 (Symmetrization bound)**.** *Consider a sequence $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\}$ of independent, random, self-adjoint matrices. For each $\theta \in \mathbb{R}$,*

$$\mathbb{E}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^n \theta(\boldsymbol{Y}_i - \mathbb{E}\,\boldsymbol{Y}_i)\right) \le \mathbb{E}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^n 2\theta\varepsilon_i\boldsymbol{Y}_i\right),$$

where $\{\varepsilon_i\}$ are independent Rademacher random variables that are also independent from $\{\boldsymbol{Y}_i\}$.

The proof of Proposition 2.3 is essentially identical with the proof of Lemma 7.6 in [Tro11b], so we omit the argument.

The matrix Laplace transform method derives its power from a deep technical result that allows us to bound the mgf of a sum of independent random matrices in terms of the cgfs of the summands. We state a simplified version of this fact that suits our needs.

Proposition 2.4 (Subadditivity of cgfs)**.** *Let $\boldsymbol{Y}$ be a random, self-adjoint matrix. Consider a finite sequence $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n\}$ of independent copies of $\boldsymbol{Y}$. For each $\theta \in \mathbb{R}$,*

$$\mathbb{E}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^{n} \theta \boldsymbol{Y}_i\right) \leq \mathrm{tr}\exp\left(n \log \mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{Y}}\right).$$

Proposition 2.4 is due to Tropp [Tro11b, Lem. 3.4]. The main ingredient in the proof is a celebrated concavity theorem established by Lieb [Lie73, Thm. 6].

We use these techniques to develop a matrix Bernstein inequality that is adapted for partial covariance estimation. The final ingredient in our argument is a matrix mgf bound that parallels the classical mgf bound underlying Bernstein's inequality.

Proposition 2.5 (Bernstein matrix mgf bound)**.** *Let $\boldsymbol{Y}$ be a random, self-adjoint matrix that satisfies*

$$\mathbb{E}\,\boldsymbol{Y} = \mathbf{0} \quad \text{and} \quad \lambda_{\max}(\boldsymbol{Y}) \leq R \quad \text{almost surely.}$$

When $\theta \in (0, R^{-1})$,

$$\mathbb{E}\,\mathrm{e}^{\theta \boldsymbol{Y}} \preccurlyeq \mathbf{I} + \frac{\theta^2}{2(1-\theta R)} \cdot \mathbb{E}(\boldsymbol{Y}^2).$$

Proposition 2.5 follows immediately from [Tro11b, Lem. 6.7] and the classical inequality

$$\frac{\mathrm{e}^{\theta R} - \theta R - 1}{R^2} \leq \frac{\theta^2}{2(1-\theta R)} \quad \text{valid for } \theta \in (0, R^{-1}).$$

We can verify this bound by comparing derivatives. The constants in this inequality can be improved, but we have chosen the version here to streamline other aspects of the argument.

3. Masked Covariance Estimation for a Subgaussian Distribution

In this section, we state and prove our main error estimates for masked covariance estimation. Section 3.1 defines two concentration parameters that measure the spread of the distribution. We present the main theorem for subgaussian distributions in Section 3.2, and we specialize to Gaussian distributions in Section 3.3. Section 3.4 shows how to derive the result for Gaussian matrices from the main theorem. Finally, we establish the main result in Section 3.5.

3.1. Concentration Parameters.

The effectiveness of the masked sample covariance estimator depends on the concentration properties of the distribution of $\boldsymbol{x}$. Let us introduce two quantities that measure different facets of the variation of the random vector.

The *subgaussian coefficient* $\kappa(\boldsymbol{x})$ of the distribution is defined to be the maximum subgaussian coefficient of a single component of the vector:

$$\kappa(\boldsymbol{x}) := \max_i \kappa(X_i). \tag{3.1}$$

In other words, we assume that each component of the distribution exhibit subgaussian decay with variance controlled by $\kappa(\boldsymbol{x})^2$.

We do not need every marginal of the distribution to be subgaussian with controlled variance, but we do require some information on the spread of the distribution in other directions. Define the *uniform fourth moment* $\nu(\boldsymbol{x})$ by the formula

$$\nu(\boldsymbol{x}) := \sup_{\|\boldsymbol{u}\|=1} (\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4)^{1/4}. \tag{3.2}$$

The uniform fourth moment measures how much the worst marginal varies.

Note that both $\kappa(\boldsymbol{x})$ and $\nu(\boldsymbol{x})$ have the same homogeneity as the random vector $\boldsymbol{x}$. (This property is sometimes expressed by saying that the quantities have the same dimension, the same units, or the same scaling.) As a consequence, $\kappa^2(\boldsymbol{x})$ and $\nu^2(\boldsymbol{x})$ have the same homogeneity as the covariance matrix $\boldsymbol{\Sigma}$.

In the sequel, we abbreviate $\kappa := \kappa(\boldsymbol{x})$ and $\nu := \nu(\boldsymbol{x})$ whenever the distribution of the random vector $\boldsymbol{x}$ is clear.

Remark 3.1. For Gaussian distributions, the uniform fourth moment ν always dominates the subgaussian coefficient κ. In the worst case, ν can be much larger than κ. Indeed, suppose that X is a standard normal random variable, and consider the random vector $\boldsymbol{x} = (X, X, \dots, X)^* \in \mathbb{R}^p$. Although the subgaussian coefficient $\kappa(\boldsymbol{x}) = \sqrt{2}$, the directional fourth moment $\nu(\boldsymbol{x}) = 12^{1/4}\sqrt{p}$.

For other kinds of distributions, the subgaussian coefficient κ may be substantially larger than the uniform fourth moment ν. Examples of this phenomenon already emerge in the univariate case.

3.2. **Main Result for Masked Covariance Estimation.** The following theorem provides detailed information about the expectation and tail behavior of the error in the masked sample covariance estimator for a zero-mean subgaussian distribution.

Theorem 3.2 (Masked Covariance Estimation for Subgaussian Distributions)**.** *Fix a $p \times p$ symmetric mask matrix $\boldsymbol{M}$. Suppose that $\boldsymbol{x}$ is a subgaussian random vector in $\mathbb{R}^p$ with mean zero. Define the covariance matrix $\boldsymbol{\Sigma}$ and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ as in* (1.1) *and* (1.2)*. Then the expected estimation error satisfies*

$$\mathbb{E}\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \le \left[\frac{16\kappa^2\nu^2 \left\|\boldsymbol{M}\right\|_{1\to 2}^2 \log(2ep)}{n}\right]^{1/2} + \frac{4\kappa^2 \left\|\boldsymbol{M}\right\| \log^2(2enp)}{n}. \tag{3.3}$$

Furthermore, for each $t > 0$, the estimation error satisfies the tail bound

$$\mathbb{P}\left\{\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \ge t\right\} \le 2ep \cdot \exp\left(\frac{-nt^2/2}{8\kappa^2\nu^2 \left\|\boldsymbol{M}\right\|_{1\to 2}^2 + 4\kappa^2 \left\|\boldsymbol{M}\right\| \log(4np) \cdot t}\right). \tag{3.4}$$

The subgaussian coefficient κ and the uniform fourth moment ν are defined in (3.1) *and* (3.2)*.*

The proof of Theorem 3.2 appears in Section 3.5. We can extend this result to the case where we center the observations using the sample mean before computing the sample covariance; the argument is identical with the one described by Levina and Vershynin [LV11, Rem. 4] for the Gaussian case.

3.2.1. *Interpretation and Consequences.* Let us take a moment to discuss Theorem 3.2. First, we note that the error in the masked sample covariance estimator can be expressed as

$$\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \boldsymbol{M} \odot (\boldsymbol{x}_i\boldsymbol{x}_i^* - \mathbb{E}\,\boldsymbol{x}\boldsymbol{x}^*), \tag{3.5}$$

using the definitions (1.1) and (1.2) of the covariance and sample covariance. For each i, the parenthesis in (3.5) has subexponential tails because the random vector $\boldsymbol{x}_i$ is subgaussian. Therefore, the formula (3.5) expresses the error as an average of subexponential random variables.

Consequently, we expect the estimation error to obey a probability inequality just like (3.4). For moderate values of t, the error (3.4) exhibits subgaussian decay, an intimation of the normal profile that emerges when the number of samples tends to infinity. For large values of t, the error has subexponential decay, owing to the heavier tails of the summands in (3.5). Likewise, the two terms in the expected error bound (3.3) correspond with the two regimes in the tail bound. The first term reflects the subgaussian decay, while the second term comes from the subexponential decay.

The scale for subgaussian decay is controlled by a measure of the variance σ^2 of each summand:

$$\sigma^2 = 8\kappa^2\nu^2 \|\boldsymbol{M}\|_{1\to 2}^2 .$$

We see that moderate deviations depend on the local properties of the mask, as encapsulated in $\|\boldsymbol{M}\|_{1\to 2}^2$. The appearance of the subgaussian coefficient κ in σ^2 reflects the variance of each component of the random vector. The presence of the uniform fourth moment ν shows that there is also a role for the spread of the random vector in every direction.

The scale for subexponential decay is controlled by a second quantity,

$$R = 4\kappa^2 \|\boldsymbol{M}\| \log(4np).$$

Large deviations reflect more global properties of the mask owing to the presence of $\|\boldsymbol{M}\|$. The subgaussian coefficient κ arises here because the tails of the distribution drive the tails of the error. Note that the large-deviation behavior only depends on the individual components of the random vector being subgaussian; we attribute this fact to the basis-dependent nature of the Schur product. The logarithmic factor in R emerges from a truncation argument, and we believe it is parasitic.

We can obtain a sample complexity bound directly from the probability inequality (3.4) in Theorem 3.2. Assume that $n \le p$ and that $\varepsilon \in (0, 1)$. Then

$$n \ge \mathrm{C} \cdot \frac{\kappa^2}{\nu^2}\left[\frac{\|\boldsymbol{M}\|_{1\to 2}^2 \log p}{\varepsilon^2} + \frac{\|\boldsymbol{M}\| \log^2 p}{\varepsilon}\right] \quad \Longrightarrow \quad \left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \le \varepsilon\nu^2 \tag{3.6}$$

with high probability. The square ν^2 of the uniform fourth moment has the same homogeneity as the covariance matrix, so (3.6) is a type of relative error bound. As before, the first summand reflects the subgaussian part of the tail, while the second summand comes from the subexponential part. A novel feature of the sample bound (3.6) is the presence of the ratio κ^2/ν^2, which is a dimensionless measure of the shape of the distribution. This ratio can be very large or very small, so it should be assessed within the scope of a particular application.

3.3. **Specialization to Gaussian Distributions.** It is natural to apply Theorem 3.2 to study the performance of masked covariance estimation for a zero-mean Gaussian random vector. In this case, the covariance matrix determines the distribution completely, so we can obtain a more transparent statement that does not involve the concentration parameters κ and ν.

Corollary 3.3 (Masked Covariance Estimation for Gaussian Distributions)**.** *Fix a $p \times p$ symmetric mask matrix $\boldsymbol{M}$. Suppose that $\boldsymbol{x}$ is a Gaussian random vector in $\mathbb{R}^p$ with mean zero. Define the covariance matrix $\boldsymbol{\Sigma}$ and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ as in* (1.1) *and* (1.2)*. Then the expected estimation error satisfies*

$$\mathbb{E}\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \le \sqrt{\frac{56 \|\boldsymbol{\Sigma}\|_{\max} \|\boldsymbol{\Sigma}\| \|\boldsymbol{M}\|_{1\to 2}^2 \log(6p)}{n}} + \frac{8 \|\boldsymbol{\Sigma}\|_{\max} \|\boldsymbol{M}\| \log^2(6np)}{n}.$$

Furthermore, for each $t > 0$, the estimation error satisfies the tail bound

$$\mathbb{P}\left\{\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \ge t\right\} \le 6p \cdot \exp\left(\frac{-nt^2}{56 \|\boldsymbol{\Sigma}\|_{\max} \|\boldsymbol{\Sigma}\| \|\boldsymbol{M}\|_{1\to 2}^2 + 16 \|\boldsymbol{\Sigma}\|_{\max} \|\boldsymbol{M}\| \log(4np) \cdot t}\right).$$

The proof of Corollary 3.3 appears below in Section 3.4. Theorem 1.4 of the Introduction follows quickly from this result when we apply the inequality $\|\boldsymbol{\Sigma}\|_{\max} \le \|\boldsymbol{\Sigma}\|$ and complete some numerical estimates.

It is fruitful to compare Corollary 3.3 directly with earlier work on masked covariance estimation for a Gaussian distribution. Assume that $n \le p$ and $\varepsilon \in (0, 1)$. Then Corollary 3.3 delivers a sample complexity bound of the form

$$n \ge \mathrm{C} \cdot \frac{\|\boldsymbol{\Sigma}\|_{\max}}{\|\boldsymbol{\Sigma}\|} \left[\frac{\|\boldsymbol{M}\|_{1\to 2}^2 \log p}{\varepsilon^2} + \frac{\|\boldsymbol{M}\| \log^2 p}{\varepsilon} \right] \quad \Longrightarrow \quad \left\| \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma} \right\| \le \varepsilon \|\boldsymbol{\Sigma}\| \tag{3.7}$$

with high probability. The bound (3.7) is similar with the results of Levina and Vershynin [LV11], stated in (1.13), but two improvements are worth mentioning.

First, recall that the sample complexity bound (1.6) we present in Example 1.1 depends on the absolute maximum entry of the covariance matrix, rather than its spectral norm. A similar refinement appears in the bound (3.7) on account of the ratio of the two norms. This ratio never exceeds one, and it can be as small as p^{-1} for particular choices of the covariance matrix. We interpret this term as saying that covariance estimation is easier when the variables are highly correlated with each other. This represents a new phenomenon that previous authors have not identified.

The second improvement over (1.13), which has less conceptual significance, is the reduction of the number of logarithmic factors.

3.4. **Proof of Corollary 3.3 from Theorem 3.2.** The result for Gaussian distributions is a direct consequence of the main theorem because the covariance matrix $\boldsymbol{\Sigma}$ of a zero-mean Gaussian vector $\boldsymbol{x}$ characterizes the distribution completely. As a consequence, we just need to estimate the concentration parameters $\kappa(\boldsymbol{x})$ and $\nu(\boldsymbol{x})$ in terms of $\boldsymbol{\Sigma}$.

First, we compute the subgaussian coefficient $\kappa(\boldsymbol{x})$. Observe that the ith component X_i of the vector $\boldsymbol{x}$ is a Gaussian random variable with variance σ_{ii}, where σ_{ii} denotes the ith diagonal entry of $\boldsymbol{\Sigma}$. The usual Gaussian tail bound demonstrates that

$$\mathbb{P}\{|X_i| > t\} \le 2\,\mathrm{e}^{-t^2/2\sigma_{ii}}.$$

According to Definition 2.1, the subgaussian coefficient $\kappa(X_i)^2 \le 2\sigma_{ii}$, and so the subgaussian coefficient of the vector satisfies

$$\kappa(\boldsymbol{x})^2 \le \max_i 2\sigma_{ii} = 2\,\|\boldsymbol{\Sigma}\|_{\max}.$$

The latter equality holds because the absolute maximum entry of a positive-definite matrix occurs on its diagonal.

Next, we compute the uniform fourth moment $\nu(\boldsymbol{x})$. Fix a unit vector $\boldsymbol{u}$. The distribution of the marginal $\boldsymbol{u}^*\boldsymbol{x}$ is Gaussian with mean zero. To compute the variance $\sigma_{\boldsymbol{u}}^2$ of the marginal, we write $\boldsymbol{x} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{g}$, where $\boldsymbol{g}$ is a standard Gaussian vector. Then

$$\sigma_{\boldsymbol{u}}^2 = \mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^2 = \mathbb{E}\,|\boldsymbol{u}^*(\boldsymbol{\Sigma}^{1/2}\boldsymbol{g})|^2 = \boldsymbol{u}^*\boldsymbol{\Sigma}^{1/2}(\mathbb{E}\,\boldsymbol{g}\boldsymbol{g}^*)\boldsymbol{\Sigma}^{1/2}\boldsymbol{u} = \boldsymbol{u}^*\boldsymbol{\Sigma}\boldsymbol{u} \le \|\boldsymbol{\Sigma}\|.$$

The fourth moment of a Gaussian variable equals three times its squared variance, so

$$\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4 = 3\sigma_{\boldsymbol{u}}^4 \le 3\,\|\boldsymbol{\Sigma}\|^2.$$

We conclude that the uniform fourth moment satisfies

$$\nu(\boldsymbol{x}) = \sup_{\|\boldsymbol{u}\|=1} (\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4)^{1/4} \le 3^{1/4}\,\|\boldsymbol{\Sigma}\|^{1/2}.$$

To complete the argument, substitute the estimates for $\kappa(\boldsymbol{x})$ and $\nu(\boldsymbol{x})$ into Theorem 3.2 and make some numerical estimates.

3.5. **Proof of Theorem 3.2.** The argument follows the same lines as the classical Laplace transform technique. For clarity, we break the presentation into discrete steps.

3.5.1. *The Matrix Laplace Transform Method.* We begin with the proof of the probability inequality (3.4). First, split the tail bound for the spectral norm into two pieces:

$$\mathbb{P}\left\{\left\|\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma}\right\|\geq t\right\}\\ \leq\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma})\geq t\right\}+\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{M}\odot\boldsymbol{\Sigma}-\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n)\geq t\right\}. \quad (3.8)$$

This inequality depends on the fact $\|\boldsymbol{A}\|=\max\{\lambda_{\max}(\boldsymbol{A}),\lambda_{\max}(-\boldsymbol{A})\}$, valid for each self-adjoint matrix $\boldsymbol{A}$, and an invocation of the union bound. We develop an estimate for the first term on the right-hand side of (3.8); an essentially identical argument applies to the second term.

The matrix Laplace transform bound, Proposition 2.2, allows us to control the first term on the right-hand side of (3.8) in terms of a matrix mgf.

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma})\geq t\right\}=\mathbb{P}\left\{\lambda_{\max}\left(n(\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma})\right)\geq nt\right\}\\ \leq\inf_{\theta>0}\left\{\mathrm{e}^{-\theta nt}\cdot\mathbb{E}\,\mathrm{tr}\exp\left(\theta n(\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma})\right)\right\}. \quad (3.9)$$

In the first line of (3.9), we have rescaled both sides of the event and applied the positive homogeneity of the maximum eigenvalue. Let us introduce notation for the trace of the matrix mgf:

$$E(\theta):=\mathbb{E}\,\mathrm{tr}\exp\left(\theta n(\boldsymbol{M}\odot\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{M}\odot\boldsymbol{\Sigma})\right). \quad (3.10)$$

Our main task is to obtain a suitable bound for $E(\theta)$.

3.5.2. *Symmetrizing the Random Sum.* The random matrix appearing in (3.10) admits a natural expression as a sum of centered, independent random matrices. To see why, substitute the definitions (1.1) and (1.2) of the population covariance matrix $\boldsymbol{\Sigma}$ and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ to obtain

$$E(\theta)=\mathbb{E}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^{n}\theta\big(\boldsymbol{M}\odot\boldsymbol{x}_i\boldsymbol{x}_i^*-\mathbb{E}\,\boldsymbol{M}\odot\boldsymbol{x}_i\boldsymbol{x}_i^*\big)\right).$$

The samples $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ are statistically independent, so the summands are independent, centered random matrices. Therefore, we may apply the symmetrization lemma, Proposition 2.3, to reach

$$E(\theta)\leq\mathbb{E}\,\mathrm{tr}\exp\left(\sum\nolimits_{i=1}^{n}2\theta\varepsilon_i(\boldsymbol{M}\odot\boldsymbol{x}_i\boldsymbol{x}_i^*)\right), \quad (3.11)$$

where $\{\varepsilon_i\}$ is a sequence of independent Rademacher random variables that is also independent from the sequence $\{\boldsymbol{x}_i\}$ of samples. The benefit of the estimate (3.11) is that each Schur product involves a rank-one matrix, which greatly simplifies our computations.

3.5.3. *Matrix cgf Bound for the Matrix mgf.* The summands on the right-hand side of (3.11) are i.i.d., so we can apply Proposition 2.4 on the subadditivity of matrix cgfs to see that

$$E(\theta)\leq\mathrm{tr}\exp(n\cdot\log\mathbb{E}\exp(2\theta\varepsilon\boldsymbol{M}\odot\boldsymbol{x}\boldsymbol{x}^*)). \quad (3.12)$$

The chief technical contribution of this paper consists in the following matrix cgf bound:

$$\log\mathbb{E}\exp(2\theta\varepsilon\boldsymbol{M}\odot\boldsymbol{x}\boldsymbol{x}^*)\preccurlyeq\frac{\theta^2\sigma^2}{2(1-\theta R)}\cdot\mathbf{I}+\frac{1}{n}\cdot\mathbf{I}\quad\text{when }\theta\in(0,R^{-1}), \quad (3.13)$$

where

$$\sigma^2:=8\kappa^2\nu^2\,\|\boldsymbol{M}\|_{1\to2}^2\quad\text{and}\quad R:=4\kappa^2\,\|\boldsymbol{M}\|\log(4np). \quad (3.14)$$

The concentration parameters κ and ν that characterize $\boldsymbol{x}$ are defined as in (3.1) and (3.2). The calculation underlying (3.13) requires several pages and some substantial new ideas. We encapsulate the details in Lemma 4.1, which is the subject of Section 4.

The trace exponential is monotone with respect to the semidefinite order [Pet94, Prop. 1], so we can substitute the cgf bound (3.13) into our estimate (3.12) for $E(\theta)$. Thus,

$$E(\theta) \le \operatorname{tr} \exp\left(\frac{\theta^2\sigma^2 n}{2(1-\theta R)} \cdot \mathbf{I} + \mathbf{I}\right) = ep \cdot \exp\left(\frac{\theta^2\sigma^2 n}{2(1-\theta R)}\right). \tag{3.15}$$

The second relation depends on the fact that the identity matrix has dimension p. The inequality (3.15) is just what we need to establish the probability inequality and the expectation bound that constitute the conclusions of Theorem 3.2.

3.5.4. *Probability Bound for the Estimation Error.* We are now prepared to complete our bound for the tail probability, initiated in (3.8). Substitute the estimate (3.15) for the matrix mgf into the Laplace transform bound (3.9) to discover that

$$\mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right) \ge t\right\} \le ep \cdot \inf_{\theta>0} \exp\left(-\theta n t + \frac{\theta^2\sigma^2 n}{2(1-\theta R)}\right).$$

Select the classical value for the parameter: $\theta = t/(\sigma^2 + Rt)$. This choice yields an upper bound for the first term on the right-hand side of (3.8):

$$\mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right) \ge t\right\} \le ep \cdot \exp\left(\frac{-nt^2}{2(\sigma^2 + Rt)}\right). \tag{3.16}$$

The second term on the right-hand side of (3.8) admits the same upper bound:

$$\mathbb{P}\left\{\lambda_{\max}\left(\boldsymbol{M} \odot \boldsymbol{\Sigma} - \boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n\right) \ge t\right\} \le ep \cdot \exp\left(\frac{-nt^2}{2(\sigma^2 + Rt)}\right). \tag{3.17}$$

The proof of (3.17) is essentially identical with the proof of (3.16), so we omit the details.

Finally, recall the definition (3.14) for the quantities σ^2 and R. Then introduce the relations (3.16) and (3.17) into the probability inequality (3.8) to establish the tail bound (3.4) stated in Theorem 3.2.

3.5.5. *Bound for the Expected Estimation Error.* Although it is possible to control the expected error by integrating the tail bound (3.4), we obtain somewhat better results through a direct application of the estimate (3.15) for the matrix mgf $E(\theta)$.

The argument is based on the following inequality, of independent interest, which provides a way to bound the expected spectral norm of a matrix in terms of its mgf. Let $\boldsymbol{Z}$ be a random, self-adjoint matrix, and fix a positive number θ. We have the following chain of relations:

$$\begin{aligned}
\mathbb{E}\,\|\boldsymbol{Z}\| &\le \theta^{-1} \log \mathbb{E}\, \mathrm{e}^{\theta\|\boldsymbol{Z}\|} \\
&= \theta^{-1} \log \mathbb{E}\, \mathrm{e}^{\max\{\lambda_{\max}(\theta\boldsymbol{Z}),\ \lambda_{\max}(-\theta\boldsymbol{Z})\}} \\
&= \theta^{-1} \log \mathbb{E} \max\left\{\lambda_{\max}(\mathrm{e}^{\theta\boldsymbol{Z}}), \lambda_{\max}(\mathrm{e}^{-\theta\boldsymbol{Z}})\right\} \\
&\le \theta^{-1} \log\left(\mathbb{E} \operatorname{tr} \mathrm{e}^{\theta\boldsymbol{Z}} + \mathbb{E} \operatorname{tr} \mathrm{e}^{-\theta\boldsymbol{Z}}\right).
\end{aligned} \tag{3.18}$$

For the first inequality, multiply and divide by θ; then invoke Jensen's inequality to bound the expectation by an exponential mean. The second relation expresses the spectral norm of a symmetric matrix in terms of eigenvalues. In the third line, we pull the maximum through the exponential and then apply the spectral mapping theorem to draw out the eigenvalue maps. Finally, replace the maximum by a sum, and bound the maximum eigenvalue of the matrix exponential, which is positive definite, by the trace.

We intend to apply (3.18) to the random matrix

$$\boldsymbol{Z} = n(\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}).$$

According to the definition (3.10) of the function $E(\theta)$, the trace of the mgf of the matrix $\boldsymbol{Z}$ coincides with $E(\theta)$. Therefore, when the parameter $\theta \in (0, R^{-1})$, our upper bound (3.15) for $E(\theta)$ demonstrates that

$$\mathbb{E}\operatorname{tr} \mathrm{e}^{\theta \boldsymbol{Z}} = E(\theta) \leq \mathrm{e}p \cdot \exp\left(\frac{\theta^2\sigma^2 n}{2(1-\theta R)}\right). \tag{3.19}$$

The argument underlying the bound (3.15) for the trace mgf of $\boldsymbol{Z}$ also applies to $-\boldsymbol{Z}$, whereby

$$\mathbb{E}\operatorname{tr} \mathrm{e}^{-\theta \boldsymbol{Z}} \leq \mathrm{e}p \cdot \exp\left(\frac{\theta^2\sigma^2 n}{2(1-\theta R)}\right). \tag{3.20}$$

Introduce (3.19) and (3.20) into the norm bound (3.18) to reach

$$n \cdot \mathbb{E}\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \leq \theta^{-1}\left(\log(2\mathrm{e}p) + \frac{\theta^2\sigma^2 n}{2(1-\theta R)}\right).$$

Minimize the right-hand side over admissible values of θ, ideally with a computer algebra system. This computation yields

$$n \cdot \mathbb{E}\left\|\boldsymbol{M} \odot \widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{M} \odot \boldsymbol{\Sigma}\right\| \leq \sqrt{2\sigma^2 n \log(2\mathrm{e}p)} + R\log(2\mathrm{e}p).$$

Divide through by n and recall the definition (3.14) of the quantities σ^2 and R. Combine the two logarithms in the second term to complete the proof of the expected error bound (3.3) from Theorem 3.2.

4. The Matrix cgf of a Schur Product

In this section, we work out the details of the matrix cgf bound (3.13) that stands at the center of Theorem 3.2. The following lemma contains a complete statement of the result.

Lemma 4.1 (Matrix cgf Bound for a Schur Product)**.** *Fix a self-adjoint matrix $\boldsymbol{M}$. Let $\boldsymbol{x} = (X_1, \dots, X_p)^*$ be a random vector, and let ε be a Rademacher variable, independent from $\boldsymbol{x}$. For each positive integer n,*

$$\log \mathbb{E}\exp(2\theta\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*) \preccurlyeq \frac{\theta^2\sigma^2}{2(1-\theta R)} \cdot \mathbf{I} + \frac{1}{n} \cdot \mathbf{I} \quad \text{when } \theta \in (0, R^{-1}),$$

where

$$\sigma^2 := 8\kappa^2\nu^2 \left\|\boldsymbol{M}\right\|_{1\to 2}^2 \quad \text{and} \quad R := 4\kappa^2 \left\|\boldsymbol{M}\right\| \log(4np).$$

The concentration parameters κ and ν associated with $\boldsymbol{x}$ are defined as in (3.1) *and* (3.2)*.*

To prove Lemma 4.1, we would like to invoke the Bernstein mgf bound, Proposition 2.5, but several obstacles stand in the way. First, estimating the variance of the random matrix $2\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*$ involves a surprisingly delicate calculation. Second, this random matrix is typically unbounded, which requires us to develop a new type of truncation argument. We address ourselves to these tasks in the next two subsections.

4.1. Computing the Variance.

The Bernstein mgf bound demands that we compute the variance of the random matrix $2\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*$. The following lemma contains this estimate. Our key insight is that the monotonicity (2.2) of the Schur product allows us to replace one factor in the product by a scalar matrix. This act of diagonalization simplifies the estimate tremendously because we erase the off-diagonal entries when we take the Schur product with an identity matrix.

Lemma 4.2 (Semidefinite variance bound)**.** *Under the assumptions of Lemma 4.1, it holds that*

$$\mathbb{E}(2\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 \preccurlyeq 8\kappa^2\nu^2 \left\|\boldsymbol{M}\right\|_{1\to 2}^2 \cdot \mathbf{I}.$$

Proof. First, we treat the leading constant and the Rademacher random variable.

$$\mathbb{E}(2\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 = 4\,\mathbb{E}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2. \tag{4.1}$$

The expectation with respect to $\boldsymbol{x}$ is not so easy to handle. To begin, we perform some algebraic manipulations to consolidate the remaining randomness. The Schur product identity (2.1) implies that

$$\begin{aligned}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 &= (\operatorname{diag}(\boldsymbol{x})\boldsymbol{M}\operatorname{diag}(\boldsymbol{x}))^2 \\ &= \operatorname{diag}(\boldsymbol{x})(\boldsymbol{M}\operatorname{diag}(\boldsymbol{x})^2\boldsymbol{M})\operatorname{diag}(\boldsymbol{x}) = (\boldsymbol{M}\operatorname{diag}(\boldsymbol{x})^2\boldsymbol{M}) \odot \boldsymbol{x}\boldsymbol{x}^*.\end{aligned}$$

Rewrite the diagonal matrix as a linear combination of matrix units: $\operatorname{diag}(\boldsymbol{x})^2 = \sum_i X_i^2\,\mathbf{E}_{ii}$. The bilinearity of the Schur product now yields

$$(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 = \Big[\boldsymbol{M}\Big(\sum\nolimits_i X_i^2\,\mathbf{E}_{ii}\Big)\,\boldsymbol{M}\Big] \odot \boldsymbol{x}\boldsymbol{x}^* = \sum\nolimits_i (\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot (X_i^2\,\boldsymbol{x}\boldsymbol{x}^*).$$

Take the expectation of this expression to reach

$$\mathbb{E}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 = \sum\nolimits_i (\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot [\mathbb{E}(X_i^2\,\boldsymbol{x}\boldsymbol{x}^*)]. \tag{4.2}$$

Next, we invoke the monotonicity (2.2) of the Schur product to make a diagonal estimate for each summand in (4.2):

$$(\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot [\mathbb{E}(X_i^2\boldsymbol{x}\boldsymbol{x}^*)] \preccurlyeq \lambda_{\max}(\mathbb{E}(X_i^2\boldsymbol{x}\boldsymbol{x}^*)) \cdot (\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot \mathbf{I}.$$

The Rayleigh–Ritz variational formula [Bha97, Cor. III.1.2] allows us to write the maximum eigenvalue as a supremum. Thus,

$$\begin{aligned}\lambda_{\max}(\mathbb{E}(X_i^2\boldsymbol{x}\boldsymbol{x}^*)) &= \sup_{\|\boldsymbol{u}\|=1} \boldsymbol{u}^*\,[\mathbb{E}(X_i^2\,\boldsymbol{x}\boldsymbol{x}^*)]\,\boldsymbol{u} = \sup_{\|\boldsymbol{u}\|=1} \mathbb{E}\left[X_i^2\,|\boldsymbol{u}^*\boldsymbol{x}|^2\right] \\ &\le \sup_{\|\boldsymbol{u}\|=1} (\mathbb{E}\,X_i^4)^{1/2}\,(\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4)^{1/2} \le 2\kappa(X_i)^2 \sup_{\|\boldsymbol{u}\|=1} (\mathbb{E}\,|\boldsymbol{u}^*\boldsymbol{x}|^4)^{1/2} \le 2\kappa^2\nu^2.\end{aligned}$$

The first inequality is Cauchy–Schwarz. For the second inequality, we apply (2.3) to bound the fourth moment of X_i in terms of the subgaussian coefficient. The final inequality follows from the definitions (3.1) and (3.2) of the concentration parameters. Combine the last two displays to obtain

$$(\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot [\mathbb{E}(X_i^2\boldsymbol{x}\boldsymbol{x}^*)] \preccurlyeq 2\kappa^2\nu^2 \cdot (\boldsymbol{M}\mathbf{E}_{ii}\boldsymbol{M}) \odot \mathbf{I}. \tag{4.3}$$

To complete our bound for the variance, we introduce (4.3) into (4.2), which delivers

$$\mathbb{E}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 \preccurlyeq 2\kappa^2\nu^2 \cdot \boldsymbol{M}^2 \odot \mathbf{I}$$

The remaining matrix is diagonal, so we can control it using only its maximum entry:

$$\mathbb{E}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2 \preccurlyeq 2\kappa^2\nu^2 \max\nolimits_i(\boldsymbol{M}^2)_{ii} \cdot \mathbf{I} = 2\kappa^2\nu^2\,\|\boldsymbol{M}\|_{1\to 2}^2 \cdot \mathbf{I}$$

The second relation follows from the fact that the diagonal entries of $\boldsymbol{M}^2$ list the squared norms of the columns of $\boldsymbol{M}$, and $\|\boldsymbol{M}\|_{1\to 2}$ computes the maximum column norm of $\boldsymbol{M}$. Substitute the latter expression into (4.1) to conclude. □

4.2. **Proof of Lemma 4.1.** This subsection contains the main steps in the proof of Lemma 4.1. We begin by explaining the motivation behind our approach.

We would like to invoke the Bernstein matrix mgf inequality, Proposition 2.5, to control the mgf of $2\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*$. This proposition requires the maximum eigenvalue of the random matrix to satisfy an almost sure bound. Using the Schur product identity (2.1), we can develop a simple estimate for the maximum eigenvalue:

$$\lambda_{\max}(2\varepsilon\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*) \le 2\,\|\operatorname{diag}(\boldsymbol{x})\boldsymbol{M}\operatorname{diag}(\boldsymbol{x})\| \le 2\,\|\boldsymbol{M}\|\,\|\operatorname{diag}(\boldsymbol{x})\|^2 = 2\,\|\boldsymbol{M}\|\,\|\boldsymbol{x}\|_\infty^2\,. \tag{4.4}$$

Unfortunately, the random variable $\|\boldsymbol{x}\|_\infty$ is typically unbounded, which suggests that we cannot apply the Bernstein approach directly.

To tackle this problem, we develop a truncation argument in Section 4.2.1, which splits the distribution of the random matrix $2\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*$ into two pieces, depending on the size of $\|\boldsymbol{x}\|_\infty$. This technique allows us to apply the Bernstein estimate to the bounded part of the random matrix (Section 4.2.2). To handle the unbounded part, we use the inequality (4.4) to develop a coarse tail estimate that we can integrate directly (Section 4.2.3). Section 4.2.4 combines these results to complete the argument.

4.2.1. *The Truncation Argument.* As we have explained, we intend to decompose the random matrix $2\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*$ based on the magnitude of the random variable $\|\boldsymbol{x}\|_\infty$. To that end, define the event

$$\mathscr{A} := \{\|\boldsymbol{x}\|_\infty^2 \le B\}, \tag{4.5}$$

where we determine a suitable truncation level B later.

Now, let us split the matrix mgf into expectations over $\mathscr{A}$ and $\mathscr{A}^c$:

$$\begin{aligned}\mathbb{E}\exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*) &= \mathbb{E}\left[\exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathscr{A}}\right] + \mathbb{E}\left[\exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathscr{A}^c}\right] \\ &\preccurlyeq \mathbb{E}\exp((2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathscr{A}}) + \mathbb{E}\left[\exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathscr{A}^c}\right]. \end{aligned} \tag{4.6}$$

The first identity follows because the two indicators form a partition of unity. In the second line, notice that the first term can only increase in the semidefinite order when we draw the indicator $\mathbb{1}_{\mathscr{A}}$ into the exponential.

4.2.2. *Bernstein Estimate for the Bounded Part of the Random Matrix.* We can interpret the first term on the right-hand side of (4.6) as the mgf of a random matrix whose maximum eigenvalue is bounded; this matrix mgf admits a Bernstein-type estimate.

We must verify that the truncated matrix $(2\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathscr{A}}$ satisfies the hypotheses of Proposition 2.5. First, note that

$$\mathbb{E}[(2\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathscr{A}}] = \mathbf{0}$$

because the Rademacher variable ε is independent from $\boldsymbol{x}$ and, hence, from $\mathscr{A}$. Second, continuing the calculation (4.4), we determine that the maximum eigenvalue is bounded.

$$\lambda_{\max}((2\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathscr{A}}) \le 2\|\boldsymbol{M}\|\|\boldsymbol{x}\|_\infty^2 \cdot \mathbb{1}_{\mathscr{A}} \le 2B\|\boldsymbol{M}\|. \tag{4.7}$$

The second inequality in (4.7) relies on the definition (4.5) of the truncation event. Third, we apply Lemma 4.2 to obtain a semidefinite bound for the variance.

$$\mathbb{E}[(2\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathscr{A}}]^2 \preccurlyeq \mathbb{E}[(2\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)^2] \preccurlyeq 8\nu^2\kappa^2\|\boldsymbol{M}\|_{1\to 2} \cdot \mathbf{I}. \tag{4.8}$$

Of course, discarding the indicator in (4.8) only increases the semidefinite order.

In view of (4.7) and (4.8), we define a variance parameter and a uniform bound parameter

$$\sigma^2 := 8\kappa^2\nu^2\|\boldsymbol{M}\|_{1\to 2}^2 \quad \text{and} \quad R := 2B\|\boldsymbol{M}\|. \tag{4.9}$$

Finally, we apply Proposition 2.5 and the variance estimate (4.8) to achieve

$$\mathbb{E}\exp((2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathscr{A}}) \preccurlyeq \mathbf{I} + \frac{\theta^2\sigma^2}{2(1-\theta R)} \cdot \mathbf{I}. \tag{4.10}$$

The relation (4.10) is valid for all $\theta \in (0, R^{-1})$.

4.2.3. *Controlling the Unbounded Part of the Random Matrix.* We treat the second term on the right-hand side of (4.6) by making a rough bound that we can integrate directly. First, observe that

$$\exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*) \preccurlyeq \exp(2\theta \cdot \lambda_{\max}(\boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)) \cdot \mathbf{I} \preccurlyeq \exp(2\theta \, \|\boldsymbol{M}\| \, \|\boldsymbol{x}\|_\infty^2) \cdot \mathbf{I}.$$

We have applied the semidefinite relation $\mathrm{e}^{\boldsymbol{A}} \preccurlyeq \mathrm{e}^{\lambda_{\max}(\boldsymbol{A})} \cdot \mathbf{I}$, valid for each self-adjoint matrix $\boldsymbol{A}$, followed by the eigenvalue bound (4.4). Multiply both sides by the indicator $\mathbb{1}_{\mathscr{A}^c}$, and take the expectation to reach

$$\begin{aligned}\mathbb{E}[\exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathscr{A}^c}] &\preccurlyeq \mathbb{E}[\exp(2\theta \, \|\boldsymbol{M}\| \, \|\boldsymbol{x}\|_\infty^2)\mathbb{1}_{\mathscr{A}^c}] \cdot \mathbf{I} \\ &=: \mathbb{E}[\mathrm{e}^{\alpha W}\mathbb{1}_{\mathscr{A}^c}] \cdot \mathbf{I}. \end{aligned} \tag{4.11}$$

In the expression (4.11), we have abbreviated

$$\alpha := 2\theta \, \|\boldsymbol{M}\| \quad \text{and} \quad W := \|\boldsymbol{x}\|_\infty^2 . \tag{4.12}$$

We apply classical techniques to bound the remaining expectation.

Observe that we can control the tail probability of W using the subgaussian coefficient κ. Indeed,

$$\begin{aligned}\mathbb{P}\{W > w\} &= \mathbb{P}\left\{\max\nolimits_i |X_i|^2 > w\right\} \\ &\le \sum\nolimits_{i=1}^p \mathbb{P}\left\{|X_i|^2 > w\right\} \le \sum\nolimits_{i=1}^p 2\,\mathrm{e}^{-w/\kappa(X_i)^2} \le 2p\,\mathrm{e}^{-w/\kappa^2}. \end{aligned} \tag{4.13}$$

The second relation is the union bound. The third follows from Definition 2.1 of the subgaussian coefficient $\kappa(X)$ of a random variable X, while the last depends on the definition (3.1) of the subgaussian coefficient κ of the random vector $\boldsymbol{x}$.

Next, we invoke a standard integration-by-parts argument [Bil79, Eqn. (21.10)] to study the expectation in (4.11). Since $\mathscr{A}^c = \{W > B\}$,

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{\alpha W}\mathbb{1}_{\mathscr{A}^c}] &= \mathrm{e}^{\alpha B} \cdot \mathbb{P}\{W > B\} + \alpha \int_B^\infty \mathrm{e}^{\alpha w} \cdot \mathbb{P}\{W > w\} \, \mathrm{d}w \\ &\le \mathrm{e}^{\alpha B} \cdot 2p\,\mathrm{e}^{-B/\kappa^2} + \alpha \int_B^\infty \mathrm{e}^{\alpha w} \cdot 2p\,\mathrm{e}^{-w/\kappa^2} \, \mathrm{d}w \\ &= 2p\left[1 + \frac{\alpha}{1/\kappa^2 - \alpha}\right] \mathrm{e}^{-(1/\kappa^2 - \alpha)B}. \end{aligned} \tag{4.14}$$

We have used the tail bound (4.13) twice to obtain the inequality in the second line. The third line follows when we evaluate the definite integral under the assumption that $\alpha < 1/\kappa^2$.

To continue the bound on the right-hand side of (4.14), we need to make a careful estimate. Owing to definition (4.12) of α, the condition

$$\theta \le \frac{1}{4\kappa^2 \, \|\boldsymbol{M}\|} \quad \Longrightarrow \quad \alpha \le \frac{1}{2\kappa^2}. \tag{4.15}$$

Assume that θ satisfies the hypothesis of (4.15). Now, observe that the right-hand side of the inequality (4.14) is an increasing function of α. Therefore, we may increase α to $1/2\kappa^2$ on the right-hand side of (4.14) and then set the truncation level

$$B = 2\kappa^2 \log(4np) \tag{4.16}$$

to obtain the bound

$$\mathbb{E}[\mathrm{e}^{\alpha W}\mathbb{1}_{\mathscr{A}^c}] \le 4p\,\mathrm{e}^{-B/2\kappa^2} = \frac{1}{n}.$$

Introduce this expression into (4.11) to conclude that

$$\mathbb{E}[\exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*)\mathbb{1}_{\mathscr{A}^c}] \preccurlyeq \frac{1}{n} \cdot \mathbf{I}. \tag{4.17}$$

Finally, we verify that the truncation level B forces the parameter θ to satisfy the hypothesis of (4.15). Recall the definition (4.9) of the bound parameter and the definition (4.16) of the truncation level to see that

$$R = 2B \|\boldsymbol{M}\| = 4\kappa^2 \|\boldsymbol{M}\| \log(4np).$$

We have already assumed that $\theta < R^{-1}$. It follows that

$$\theta < \frac{1}{R} = \frac{1}{\log(4np)} \cdot \frac{1}{4\kappa^2 \|\boldsymbol{M}\|} \leq \frac{1}{4\kappa^2 \|\boldsymbol{M}\|}.$$

This observation completes the tail estimate.

4.2.4. *Combining the Results.* We have obtained estimates for the two terms in our truncation bound (4.6). Introduce (4.10) and (4.17) into (4.6) to reach

$$\mathbb{E} \exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*) \preccurlyeq \mathbf{I} + \frac{\theta^2\sigma^2}{2(1-\theta R)} \cdot \mathbf{I} + \frac{1}{n} \cdot \mathbf{I},$$

where σ^2 and R are defined in (4.9). We have also assumed that $\theta \in (0, R^{-1})$. The logarithm is operator monotone [Bha07, Exer. 4.2.5], so

$$\log \mathbb{E} \exp(2\theta\varepsilon \boldsymbol{M} \odot \boldsymbol{x}\boldsymbol{x}^*) \preccurlyeq \log \left[\mathbf{I} + \frac{\theta^2\sigma^2}{2(1-R)} \cdot \mathbf{I} + \frac{1}{n} \cdot \mathbf{I} \right].$$

To complete the proof of Lemma 4.1, we invoke the semidefinite relation $\log(\mathbf{I} + \boldsymbol{A}) \preccurlyeq \boldsymbol{A}$, which holds for each positive semidefinite matrix $\boldsymbol{A}$.

References

[AW02] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, 2002.

[Bar05] A. Barvinok. Measure concentration. Available at `http://www.math.lsa.umich.edu/~barvinok/total710.pdf`, 2005.

[Bha97] R. Bhatia. *Matrix Analysis*. Springer, New York, NY, 1997.

[Bha07] R. Bhatia. *Positive Definite Matrices*. Princeton Univ. Press, Princeton, NJ, 2007.

[Bil79] P. Billingsley. *Probability and Measure*. Wiley, New York, NY, 1979.

[BL08a] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604, 2008.

[BL08b] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227, 2008.

[BS10] Z. D. Bai and J. W. Silverstein. *Spectral Analysis of Large-Dimensional Random Matrices*. Springer, New York, NY, 2010.

[CZZ10] 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.

[FB07] R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivar. Anal.*, 98(2):227–255, 2007.

[Fre05] D. A. Freedman. *Statistical Models: Theory and Practice*. Cambridge Univ. Press, Cambridge, 2005.

[HJ94] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge Univ. Press, Cambridge, 1994.

[HKZ11] D. Hsu, S. M. Kakade, and T. Zhang. Dimension-free tail inequalities for sums of random matrices. Available at `arXiv:1104.1672`, Apr. 2011.

[Jol02] I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, NY, 2002.

[JW02] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, Englewood Cliffs, NJ, 6th edition, 2002.

[Kar08] N. El Karoui. Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.*, 36(6):2717–2756, 2008.

[Lie73] E. H. Lieb. Convex trace functions and the Wigner–Yanase–Dyson conjecture. *Adv. Math.*, pages 11:267–288, 1973.

[LV11] E. Levina and R. Vershynin. Partial estimation of covariance matrices. *Probab. Theory Related Fields*, 2011.

[MKB80] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1980.

[Mui82] R. J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley, New York, NY, 1982.

[Oli10a] R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Available from `arxiv:0911.0600`, 2010.

[Oli10b] R. I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electron. Commun. Probab.*, 15:203–212, 2010.

[Pet94] D. Petz. A survey of certain trace inequalities. *Functional Analysis and Operator Theory, Banach Center Publications*, 30:287–298, 1994.

[RLZ09] A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.*, 104(485):177–186, 2009.

[SV11] N. Srivastava and R. Vershynin. Covariance estimation for distributions with $2 + \varepsilon$ moments. Available at `arXiv:1106.2275`, June 2011.

[Tro11a] J. A. Tropp. Freedman's inequality for matrix martingales. *Electron. Commun. Probab*, 16:262–270, Aug. 2011.

[Tro11b] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, Aug. 2011.

[Ver11] R. Vershynin. *Compressed Sensing: Theory and Applications*, chapter Introduction to the non-asymptotic analysis of random matrices. Cambridge Univ. Press, Cambridge, 2011. To appear. Available at `http://www-personal.umich.edu/~romanv/papers/non-asymptotic-rmt-plain.pdf`.