Received: 22 June 2020 | Revised: 13 July 2020 | Accepted: 17 July 2020
DOI: 10.1002/sta4.316

WILEY

ORIGINAL ARTICLE

Nonasymptotic support recovery for high-dimensional sparse covariance matrices

Adam B. Kashlak | **Linglong Kong**

Mathematical and Statistical Sciences, University of Alberta, Edmonton, T6G 2G1, Alberta, Canada

Correspondence
Adam B. Kashlak, Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta T6G 2G1, Canada.
Email: kashlak@ualberta.ca

For high-dimensional data, the standard empirical estimator for the covariance matrix is very poor, and thus many methods have been proposed to more accurately estimate the covariance structure of high-dimensional data. In this article, we consider estimation under the assumption of sparsity but regularize with respect to the individual false-positive rate for incorrectly including a matrix entry in the support of the final estimator. The two benefits of this approach are (1) an interpretable regularization parameter removing the need for computationally expensive tuning and (2) extremely fast computation time arising from use of a binary search algorithm implemented to find the best estimator within a carefully constructed operator norm ball. We compare our approach to universal thresholding estimators and lasso-style penalized estimators on both simulated data and data from gene expression for cancerous tumours.

KEYWORDS

concentration inequality, genomics, random matrix, Schatten norm

1 | INTRODUCTION

Covariance matrices and accurate estimators of such objects are of critical importance in statistics. Various standard techniques including principal component analysis and linear and quadratic discriminant analysis rely on an accurate estimate of the covariance structure of multivariate data. Applications can range from genetics and medical imaging data to climate and many others. In the era of high-dimensional data, classical asymptotic estimators perform poorly in applications (Johnstone, 2001; Stein, 1975). Hence, we propose a high-dimensional covariance estimator assuming sparsity on the underlying covariance structure. Our approach aims to control the individual false-positive rate of incorrectly including a null entry in the support of the sparse estimator.

Many sparse estimators for the covariance matrix have been proposed (Pourahmadi, 2011), which is when most of the off-diagonal entries are zero or negligible. Beyond mere theoretical interest, the assumption of sparsity is widely applicable to real data analysis when many of the variable pairings are uncorrelated. Thus, it is desirable to tailor covariance estimation procedures given this assumption of sparsity.

Generally, sparsity implies some bound on the number of non-zero entries in the columns of a covariance matrix. Thus, given $\Sigma \in \mathbb{R}^{d\times d}$ with entries $\sigma_{i,j}$ for $i,j = 1, \dots, d$, there exists some constant $k>0$ such that $\max_{j=1,\dots,d} \sum_{i=1}^{d} [\sigma_{i,j} \neq 0] \leq k$. This can be generalized to "approximate sparsity" as in Rothman et al. (2009) by $\max_{j=1,\dots,d} \sum_{i=1}^{d} |\sigma_{i,j}|^q \leq k$ for some $q \in [0, 1)$. Furthermore, Cai and Liu (2011) define a broader approximately sparse class by bounding the weighted column sums of Σ. In El Karoui (2008), a similar notion referred to as "β-sparsity" is defined. Such classes of sparse covariance matrices allow for good theoretical performance of estimators.

One class of estimators are shrinkage estimators that shrink estimated eigenvalues, eigenvectors, or the matrix itself towards some desired target (Daniels & Kass, 1999, 2001; Dey & Srinivasan, 1985; Haff, 1980; Hoff, 2009; Johnstone & Lu, 2012; Ledoit & Wolf, 2004). Another class of sparse estimators are those that regularize the estimate with lasso-style penalties (Bien & Tibshirani, 2011; Rothman, 2012; Xue et al. 2012). Yet another class consists of thresholding estimators, which map small matrix entries to zero when their magnitude is smaller than some threshold (Bickel & Levina, 2008a, 2008b; Cai & Liu, 2011; Rothman et al. 2009). By banding and tapering, which apply only when the variables are ordered or a notation of proximity exists—for example, spatial, time series, or longitudinal data. As we will not assume such an ordering and construct a methodology that is permutation invariant, these approaches will not be considered. Lastly, there has been substantial work into the estimation of the precision or inverse covariance matrix. This setting is wholly different as an unbiased estimator for the precision matrix does not exist, and thus a different approach is required.

Finite sample guarantees and faster computing time than computationally expensive optimization and cross validation methods were two of the main motivating factors for this research. The recent work of Qiu and Liyanage (2019) similarly proposes an alternative to cross validation-based threshold selection by detailing an elegant analytic equation for a sparse covariance estimator that is optimal in the Frobenius risk. In contrast to this work, we choose an estimator based on the individual false-positive rate for incorrectly including a negligible covariance entry in the support of our estimator. Effectively, the covariance estimation problem becomes one of large-scale hypothesis testing.

Many established methods for sparse estimation make use of a regularization or penalization term incorporated to enforce sparsity (Bien & Tibshirani, 2011; Rothman, 2012). We enforce sparsity via the individual false-positive rate being the probability that we falsely include an off-diagonal entry in the support of the estimator. Note that this is for each entry and not the familywise error rate for all entries at once. Our methodology contains three steps: (i) choose an individual false-positive rate, $0 < \rho \ll 1$; (ii) use ρ to construct a ball about the diagonal of the empirical estimator; (iii) and search that ball for a sparse estimator. The smaller this ball is, the sparser our estimator is forced to be. Thus, ρ acts as an interpretable regularization parameter allowing for greater sparsity as it decreases. Of theoretical note, our proofs in the Appendix use matrix concentration inequalities based on the p-Schatten norms and the spectrum of the covariance, whereas most past results are based on entrywise norms.

In Section 2, the general estimation procedure is outlined, and it is specified for tuning threshold estimators. Section 3 discusses our approach to achieving a certain false-positive rate when attempting to recover the support of the covariance matrix. Lastly, Section 4 details comprehensive simulations comparing our approach to standard techniques such as thresholding and lasso-style penalization. Beyond simulation experiments, a real data set of gene expressions for small round blue-cell tumours (SRBCTs) from the study of Khan et al. (2001) is considered. The Appendix contains auxiliary lemmas and proofs of all the results.

The data that support the findings of this study are openly available at https://bioinf.ucd.ie/people/aedin/R/full_datasets/. Our proposed algorithm is implemented in the function `sparseCov( )` available on CRAN in the R package `sparseMatEst` (Kashlak, 2019). A link to this R package can be found at https://sites.ualberta.ca/~kashlak/kashCode.html.

1.1 | Notation and definitions

The individual false-positive rate ρ is the probability that we incorrectly decide that a given matrix entry is non-zero. For some sparse estimator $\hat{\Sigma}^{\text{sp}}$ with ijth entry $\tilde{\sigma}_{i,j}$ for $i \neq j, \rho = \mathrm{P}(\tilde{\sigma}_{i,j} \neq 0 \,|\, \Sigma_{i,j} = 0)$.

When defining a Banach space of matrices, there are many matrix norms that can be considered. In the article, the main norms of interest are the p-Schatten norms, which will be denoted $\|\cdot\|_{S^p}$ and are defined as follows.

Definition 1 (p-Schatten Norm)**.** For an arbitrary matrix $\Sigma \in \mathbb{R}^{k \times l}$ and $p \in [1, \infty)$, the p-Schatten norm is $\|\Sigma\|_{S^p} = \mathrm{tr}\{(\Sigma^{\mathsf{T}}\Sigma)^{p/2}\}^{1/p} = \|\nu\|_{\ell^p} = \left(\sum_{i=1}^{\min\{k,l\}} \nu_i^p\right)^{1/p}$, where $\nu = (\nu_1, \ldots, \nu_{\min\{k,l\}})$ is the vector of singular values of Σ and where $\|\cdot\|_{\ell^p}$ is the standard ℓ^p norm in $\mathbb{R}^d$. In the covariance matrix case where $\Sigma \in \mathbb{R}^{d \times d}$ is symmetric and positive semi-definite, $\|\Sigma\|_{S^p} = \mathrm{tr}(\Sigma^p)^{1/p} = \|\lambda\|_{\ell^p}$, where λ is the vector of eigenvalues of Σ. The 1-Schatten norm is referred to as the trace norm and the 2-Schatten norm as the Frobenius norm.

For $p = \infty$, we have the usual operator norm for $\Sigma : \mathbb{R}^l \to \mathbb{R}^k$ with respect to the ℓ^2 norm, $\|\Sigma\|_{S^\infty} = \sup_{\|u\|_{\ell^2}=1} \|\Sigma u\|_{\ell^2} = \|\nu\|_{\ell^\infty} = \max_{i=1,\ldots,\min\{k,l\}} |\nu_i|$, which is the maximal eigenvalue when Σ is symmetric positive semi-definite.

Another family of norms that will be used is the *entrywise* norm, which is written in terms of ℓ^p norms of the entries.

Definition 2 ((p,q)-Entrywise norm)**.** For an arbitrary matrix $\Sigma \in \mathbb{R}^{k \times l}$ with entries $\sigma_{i,j}$ and $p, q \in [1, \infty]$, the (p,q)-entrywise norm is $\|\Sigma\|_{p,q} = \left[\sum_{i=1}^{k} \left(\sum_{j=1}^{l} \sigma_{i,j}^q\right)^{p/q}\right]^{1/p}$ with the usual modification in the case that $p = \infty$ and/or $q = \infty$. When $p = q$, these are the ℓ^p norms of a given matrix treated as a vector in $\mathbb{R}^{kl}$. Note that the 2-Schatten norm, the Frobenius norm, coincides with the $(2,2)$-entrywise norm.

2 | SPARSE ESTIMATION PROCEDURE

Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be a sample of n independent and identically distributed mean zero random vectors with unknown $d \times d$ covariance matrix Σ. Define the empirical estimate of Σ to be $\hat{\Sigma}^{\text{emp}} = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^{\mathsf{T}}$, where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$. Without loss of generality, we assume that all diagonal entries of $\hat{\Sigma}^{\text{emp}}$ are normalized to 1. In practice, we can un-normalize after our algorithm is complete.

The goal of the following procedure is to construct a sparse estimator, $\hat{\Sigma}^{\text{sp}}$, for Σ by first constructing a set containing $\hat{\Sigma}^{\text{sp}}$ centred on $\hat{\Sigma}_0^{\text{emp}}$, the matrix with diagonal entries coinciding with $\hat{\Sigma}^{\text{emp}}$ and off-diagonal entries equal to zero, and then searching this set for the least sparse member—i.e. the entry with as many non-zero entries as possible.

The methodology is as follows:

i. Choose a suitable false-positive rate $\rho \in (0, 1)$, which will typically be close to zero where $\rho = \mathrm{P}(\hat{\Sigma}_{i,j}^{\text{sp}} \neq 0 \,|\, \Sigma_{i,j} = 0)$.
ii. Use the method in Section 2.1 to construct a ball centred at $\hat{\Sigma}_0^{\text{emp}}$ such that the least sparse matrices in that ball have false-positive rate ρ—i.e. those farthest away from the centre $\hat{\Sigma}_0^{\text{emp}}$.
iii. Use the binary search algorithm in Section 2.2 to identify the least sparse element in the above ball denoted $\hat{\Sigma}^{\text{sp}}$.

This algorithm will achieve a desired false-positive rate for our estimator in the large d small n setting as discussed in Section 3. This algorithm is implemented in the function `sparseCov( )` in the R package `sparseMatEst` (Kashlak, 2019).

2.1 | Constructing a sparse ball

To construct a ball centred at $\hat{\Sigma}_0^{\text{emp}}$ as required by Step (ii) above, we implement the following procedure. Theoretical justification is provided in Section 3.

Given a false-positive rate $0 < \rho \leq 0.5$, we construct a ball B_ρ centred on $\hat{\Sigma}_0^{\text{emp}}$ as follows.

i. Find $\eta = 2^a \rho \in (0.5, 1]$ for some $a \in \mathbb{Z}^+$.
ii. Compute λ, the η-quantile of the magnitudes of the off-diagonal entries in $\hat{\Sigma}^{\text{emp}}$. That is, $\lambda > 0$ is the smallest real number such that $([\#\left\{|\hat{\sigma}_{i,j}| > \lambda \,|\, i < j\right\}]/[d(d-1)/2]) \leq \eta$.
iii. Apply hard thresholding to $\hat{\Sigma}^{\text{emp}}$—i.e. given a threshold λ, get $\hat{\Sigma}^{\text{emp}}(\lambda)$ whose entries are $(\hat{\Sigma}^{\text{emp}}(\lambda))_{i,j} = \{\hat{\sigma}_{i,j} \mathbf{1}[|\hat{\sigma}_{i,j}| > \lambda]\}_{i,j}$.
iv. Construct the operator norm ball about $\hat{\Sigma}^{\text{emp}_0}$ of radius $r = 2^{-a}\|\hat{\Sigma}_0^{\text{emp}} - \hat{\Sigma}^{\text{emp}}(\lambda)\|$.

What we have now is $B_\rho = \{\Pi \in \mathbb{R}^{d \times d} : \|\Pi - \hat{\Sigma}_0^{\text{emp}}\|_{S^\infty} \leq r\}$. This set will be searched for its least sparse member using the algorithm in the following section. Note for clarity that increasing the false-positive rate ρ results in an increase in the radius of the ball we will construct, which in turn causes a decrease in the threshold λ.

2.2 | Thresholding within balls

A generalized thresholding operator, as defined in Rothman et al. (2009), is $s_\lambda(\cdot) : \mathbb{R} \to \mathbb{R}$ such that $|s_\lambda(z)| \leq |z|$, $s_\lambda(z) = 0$ for $|z| \leq \lambda$, and $|s_\lambda(z) - z| \leq \lambda$, which shrinks a value z towards zero but by no more than λ. This general definition yields nice theoretical properties and will allow for many types of thresholding including the four threshold operators considered in Section 4. With slight abuse of notation, we will write $s_\lambda(\Sigma)$ to denote application of the thresholding operator to each entry of the matrix Σ. In the past, such an operator is applied to the empirical estimate $\hat{\Sigma}^{\text{emp}}$ for some λ generally chosen via cross validation. Instead of directly choosing a threshold λ, our approach is to find the smallest λ such that $\|s_\lambda(\hat{\Sigma}^{\text{emp}}) - \hat{\Sigma}^{\text{emp}_0}\|_{S^\infty} \leq r$.

Thus, to locate the least sparse element of the ball B_ρ from the previous section as required by Step (iii) in Section 2, we implement the following procedure.

i. Set $\hat{\Sigma}^{\text{sp}}(0) = (\hat{\Sigma}^{\text{diag}})^{-1/2}(\hat{\Sigma}^{\text{emp}})(\hat{\Sigma}^{\text{diag}})^{-1/2}$ to be the empirical estimator normalized to have a diagonal of ones—i.e. a threshold of 0. Initialize the threshold $\lambda = 0.5$ and write $\hat{\Sigma}^{\text{sp}}(\lambda) = s_\lambda(\hat{\Sigma}^{\text{emp}})$. Let $k = 0$ be the number of steps of the recursion. Choose a false-positive rate ρ and compute r as in the previous section.
ii. Increase $k \leftarrow k + 1$, and then update λ as follows.
    (a) if $\|\hat{\Sigma}^{\text{sp}}(\lambda) - \hat{\Sigma}_0^{\text{emp}}\|_{S^\infty} \leq r$, set $\lambda \leftarrow \lambda - 2^{-k-1}$.
    (b) Otherwise, set $\lambda \leftarrow \lambda + 2^{-k-1}$.
iii. Repeat Step (ii) until k has reached the desired number of iterations. Generally, as few as $k = 10$ will suffice.
iv. The resulting estimator is $\hat{\Sigma}^{\text{sp}} = (\hat{\Sigma}^{\text{diag}})^{1/2}(\hat{\Sigma}^{\text{sp}}(\lambda))(\hat{\Sigma}^{\text{diag}})^{1/2}$, where $\hat{\Sigma}^{\text{sp}}(\lambda)$ is the final matrix resulting from this recursion.

Remark 1 (Binary Search)**.** For Step ii (a) above, we decrease the threshold if the estimator is too sparse. Meanwhile for Step ii (b), we increase the threshold if the estimator has too many non-zero entries.

Remark 2 (Positive definite estimators)**.** Thresholding a covariance matrix often returns a non-positive definite matrix. If $\hat{\Sigma}^{\text{sp}}$ is not positive semi-definite, then it can be projected onto the space of positive semi-definite matrices. A standard past approach is to replace the negative eigenvalues with their absolute values, which maintains the eigen-structure. However, such a projection will have an adverse effect on the support recovery problem as the estimator will no longer be sparse. An alternative is to map $\hat{\Sigma}^{\text{sp}} \to \hat{\Sigma}^{\text{sp}} + \gamma I_d$ for some $\gamma > 0$ large enough to make the result positive definite. This will not affect the recovered support of the matrix.

If we were to replace the operator norm with any metric $d(\cdot, \cdot)$ that is a monotonically decreasing function of the Frobenius norm $\|\hat{\Sigma}^{\text{sp}}(\lambda) - \hat{\Sigma}_0^{\text{emp}}\|_{S^2}$ or another entrywise norm, then the sequence $d(\hat{\Sigma}^{\text{sp}}(\lambda), \hat{\Sigma}_0^{\text{emp}})$ will be decreasing in λ.

Proposition 1. *In the context of the above algorithm, if $\lambda_1 < \lambda_2$, then for any p, q, we have $\|\hat{\Sigma}^{sp}(\lambda_1) - \hat{\Sigma}^{emp}\|_{p,q} \leq \|\hat{\Sigma}^{sp}(\lambda_2) - \hat{\Sigma}^{emp}\|_{p,q}$.*

Proof. As $\lambda_1 < \lambda_2$, the entries of the matrix $\hat{\Sigma}^{\text{sp}}(\lambda_1) - \hat{\Sigma}_0^{\text{emp}}$ are equal to or smaller in absolute value than the entries of $\hat{\Sigma}^{\text{sp}}(\lambda_2) - \hat{\Sigma}_0^{\text{emp}}$. Hence, $\|\hat{\Sigma}^{\text{sp}}(\lambda_1) - \hat{\Sigma}_0^{\text{emp}}\|_{p,q} \leq \|\hat{\Sigma}^{\text{sp}}(\lambda_2) - \hat{\Sigma}_0^{\text{emp}}\|_{p,q}$ by Definition 2. □

This property guarantees that the above algorithm will find the least sparse $\hat{\Sigma}^{\text{sp}}$ in the ball in the sense of having the smallest threshold possible. However, for an arbitrary metric or other p-Schatten norms, this sequence may not necessarily be strictly decreasing in λ. The theoretical results

of Section 3 require the operator norm $\|\hat{\Sigma}^{\text{sp}}(\lambda) - \hat{\Sigma}^{\text{emp}}\|_{S^\infty}$, which does not yield a monotonically decreasing sequence, though this sequence is roughly decreasing in the sense that it is lower bounded by the entrywise $(\infty, 2)$ norm, which is a decreasing sequence. Furthermore, it is upper bounded by the entrywise $(\infty, 1)$ norm, which follows from the Gershgorin circle theorem (Iserles, 2009), and which is also a decreasing sequence.

3 | CONSTRUCTING AN ESTIMATOR WITH A DESIRED FALSE-POSITIVE RATE

To justify the method for constructing the ball from Section 2.1, we will require a class of sparse matrices similar to those from Bickel and Levina ((2008a), (2008b)), Rothman et al. (2008a), and Cai and Liu (2011). Specifically, let

$$\mathcal{U}(\kappa, \delta) = \left\{ \Sigma \in \mathbb{R}^{d\times d} : \max_{i=1,\cdots,d} \sum_{j=1}^{d} \mathbf{1}[\sigma_{i,j} \neq 0] \leq \kappa, \text{if } \sigma_{i,j} \neq 0 \text{ then } |\sigma_{i,j}| \geq \delta > 0 \right\}. \tag{1}$$

For the results regarding the false-positive rate, we are not concerned with the lower bound δ and only with κ, the maximum number of non-zero entries per column or row. As long as κ increases more slowly than the dimension d, which is made specific below, we can achieve a desired false-positive rate.

For a covariance estimator $\tilde{\Sigma} \in \mathbb{R}^{d\times d}$, the individual false-positive rate is $\rho(\tilde{\Sigma}) = P(\tilde{\sigma}_{i,j} \neq 0 \,|\, \sigma_{i,j} = 0,\, i \neq j)$, and the empirical false-positive rate is

$$\hat{\rho}(\tilde{\Sigma}) = \frac{\#\{\tilde{\sigma}_{i,j} \neq 0 \,|\, \sigma_{i,j} = 0,\, i > j\}}{\#\{\sigma_{i,j} = 0,\, i > j\}},$$

where $\sigma_{i,j}$ is the ijth entry of the true covariance matrix and $\tilde{\sigma}_{i,j}$ is the ijth entry of the estimator $\tilde{\Sigma}$. For notation, let $\hat{\Sigma}^{\text{emp}}$ be the usual empirical estimate of the covariance matrix. As before, let $\hat{\Sigma}_0^{\text{emp}}$ be the empirical estimator with all off-diagonal entries set to zero thus guaranteeing a false-positive rate of zero. For $\eta \geq 0.5$, let $\hat{\Sigma}_\eta^{\text{emp}}$ be the empirical estimator after application of the hard threshold operator with threshold $M_\eta = \text{quantile}(|\hat{\sigma}_{i,j}|, 1-\eta \,:\, i > j)$, which removes $100(1-\eta)\%$ of the off-diagonal entries, achieving an empirical false-positive rate of approximately η due to the following lemma.

Proposition 2. *Let $\Sigma \in \mathcal{U}(\kappa, \delta)$ from Equation (1) with $\kappa = o(d^\nu)$. Let the $\eta \in [0.5, 1)$ threshold, M_η, be the $1-\eta$ quantile of $|\hat{\sigma}_{i,j}|$ with $i > j$, and let the corresponding thresholded estimator be $\hat{\Sigma}_\eta^{emp} = s_{M_\eta}(\hat{\Sigma}^{emp})$ with the ijth entry denoted $\hat{\sigma}_{i,j}^{(\eta)}$. Then, denoting*

$$\hat{\eta} = \frac{\#\{(i,j) \,|\, i > j,\, \left|\hat{\sigma}_{i,j}^{(\eta)}\right| > 0, \sigma_{i,j} = 0\}}{\#\{\sigma_{i,j} = 0,\, i > j\}},$$

we have that $|\hat{\eta} - \eta| \leq Cd^{\nu-1}$ for some constant $C > 0$.

Remark 3 (Contamination)**.** For this lemma, we want the $(1-\eta)$-quantile of the mean zero entries but have to work with the $(1-\eta)$-quantile of the entire collection, which is contaminated by a small number of elements with non-zero mean. For $\nu < 1$, the error is $O(d^{\nu-1})$ hence for $\eta \approx 0.5$, thresholding based on the η-quantile suffices for large enough d. For small η, say $\eta \approx d^{-1}$, we have to work harder in motivating Theorem 1.

As noted in the remark, we cannot continue to threshold based on the sample quantiles for a very small choice of ρ. However, using the matrices, $\hat{\Sigma}_\eta^{\text{emp}}$ and $\hat{\Sigma}_0^{\text{emp}}$, as reference points, we can interpolate via the following theorem to achieve any desired false-positive rate.

Theorem 1. *Let $\Sigma \in \mathcal{U}(\kappa, \delta)$ from Equation (1) with $\kappa = O(d^\nu)$ for $\nu < 1/2$. Given a desired false-positive rate, $\rho \in (0, 0.5]$, and $\eta = \rho 2^a \in (0.5, 1]$ for some $a \in \mathbb{Z}^+$, let $\hat{\Sigma}_\rho^{emp}$ be the hard thresholded empirical estimator that achieves a false-positive rate of ρ. Then,*

$$\left| \eta \frac{\mathrm{E}\left\|\hat{\Sigma}_\rho^{emp} - \hat{\Sigma}_0^{emp}\right\|_{S^\infty}}{\mathrm{E}\left\|\hat{\Sigma}_\eta^{emp} - \hat{\Sigma}^{emp_0}\right\|_{S^\infty}} - \rho \right| \leq K_1 n \rho^{1/2} d^{-1/4} + K_2 n \rho^{1/4} d^{-1/2} + o(nd^{-1/2}),$$

where K_1, K_2 are universal constants.

Corollary 1. *Given the set-up of Theorem 1, let $\hat{\Sigma}_\rho^{emp}$ have the ijth entry $\tilde{\sigma}_{i,j}$, and let $\sigma_{i,j}$ be the ijth entry of the true covariance matrix Σ. Then, $\mathrm{P}\left(\tilde{\sigma}_{i,j} \neq 0 \,|\, \sigma_{i,j} = 0\right)$ tends towards ρ as $d \to \infty$ and $n = o(d^{1/4})$.*

The corollary follows directly from the theorem. Specifically, when d is large and ρ is near zero, the right-hand side of the expression in Theorem 1 tends to zero, implying that the ratio of the operator norm distances from $\hat{\Sigma}_\rho^{\text{emp}}$ and $\hat{\Sigma}_\eta^{\text{emp}}$ to $\hat{\Sigma}_0^{\text{emp}}$ for ρ and η, respectively,

$$\frac{\mathrm{E}\left\|\hat{\Sigma}_\rho^{\text{emp}} - \hat{\Sigma}_0^{\text{emp}}\right\|_{S^\infty}}{\mathrm{E}\left\|\hat{\Sigma}_\eta^{\text{emp}} - \hat{\Sigma}_0^{\text{emp}}\right\|_{S^\infty}} = \frac{(\text{expected radius for } B_\rho)}{(\text{expected radius for } B_\eta)},$$

is closely proportional to ρ/η. Thus, we can use the known quantities ρ, η, and $\mathrm{E}\left\|\hat{\Sigma}_\eta^{\text{emp}} - \hat{\Sigma}_0^{\text{emp}}\right\|_{S^\infty}$ to approximate $\mathrm{E}\left\|\hat{\Sigma}_\rho^{\text{emp}} - \hat{\Sigma}_0^{\text{emp}}\right\|_{S^\infty}$. Thus, in the methodology of Section 2.1, we contract the η-radius by a factor of $2^{-a} = \rho/\eta$ to get the ρ-radius. This implies that the least sparse matrices in the constructed operator norm ball should have an individual false-positive rate of approximately ρ.

The power of Theorem 1 arises when $d \gg n$. It highlights the interplay between the dimension, sample size, and ρ, the sparseness of the estimator. Furthermore, this result does not require any distributional assumption. It also does not require any assumption on the lower bound δ on the non-zero $|\sigma_{i,j}|$ as it is only concerned with the $\sigma_{i,j}$ that are zero.

4 | NUMERICAL SIMULATIONS

We now apply the methodology in Section 2 to the multivariate Gaussian and Laplace distributions, which will be denoted as *CoM* in the below tables for "concentration of matrices" as the proofs in the Appendix rely on matrix concentration inequalities. We also compare the support recovery of our approach against penalized estimators and application of universal threshold estimators with threshold selection achieved via cross validation.

One alternative is the lasso-style estimator from the R package `PDSCE` (Rothman, 2013), which optimizes

$$\hat{\Sigma}^{\text{PDS}} = \underset{\Sigma \succeq 0}{\arg\min} \left\{ \left\|\Sigma - \hat{\Sigma}^{\text{emp}}\right\|_{S^2} - \tau \log\det(\Sigma) + \lambda \|\Sigma\|_{\ell^1} \right\},$$

with $\tau, \lambda > 0$. Here, the log det term enforces positive definiteness of the final solution, and $\|\cdot\|_{\ell^1}$ is the lasso-style penalty, which enforces sparsity.

The similar method from the R package `spcov` (Bien & Tibshirani, 2012), which uses a majorize–minimize algorithm to determine

$$\hat{\Sigma}^{\text{MMA}} = \underset{\Sigma \succeq 0}{\arg\min} \left\{ \mathrm{tr}\left(\hat{\Sigma}^{\text{emp}}\Sigma^{-1}\right) - \log\det(\Sigma^{-1}) + \lambda \|\Sigma\|_{\ell^1} \right\}$$

for some penalization $\lambda > 0$, was also considered but proved to run too slowly on high-dimensional matrices—that is, $d \geq 200$—to be included in the numerical experiments.

We also compare our method against the four universal thresholding estimators applied to the empirical covariance matrix from (Rothman et al. 2009), Hard, Soft, SCAD, and Adaptive LASSO:

$$\hat{\Sigma}_\lambda^{\text{Hard}} = \{\hat{\sigma}_{i,j} \mathbf{1}\left[|\hat{\sigma}_{i,j}| > \lambda\right]\}_{i,j}, \qquad \hat{\Sigma}_\lambda^{\text{SCAD}} = \begin{cases} \hat{\sigma}_{i,j}^{\text{Soft}} & \text{for } |\hat{\sigma}_{i,j}| \leq 2\lambda \\ \frac{a-1}{a-2}(|\hat{\sigma}_{i,j}| - 2\lambda) + \lambda & \text{for } 2\lambda < |\hat{\sigma}_{i,j}| \leq a\lambda \\ \hat{\sigma}_{i,j}^{\text{Hard}} & \text{for } |\hat{\sigma}_{i,j}| > a\lambda \end{cases},$$

$$\hat{\Sigma}_\lambda^{\text{Soft}} = \{\text{sign}(\hat{\sigma}_{i,j})(|\hat{\sigma}_{i,j}| - \lambda)_+\}_{i,j}, \qquad \hat{\Sigma}_\lambda^{\text{Adpt}} = \{\text{sign}(\hat{\sigma}_{i,j})(|\hat{\sigma}_{i,j}| - \lambda^{\xi+1}|\hat{\sigma}_{i,j}|^{-\xi})_+\}_{i,j},$$

where $\hat{\sigma}_{i,j}$ is the (i,j)th entry of the empirical covariance estimate, $a = 3.7$, and $\xi = 1$. The parameter $\lambda > 0$ is the threshold, which is chosen in practice via cross validation with respect to the Frobenius norm. Briefly, the data are split into half, and two empirical estimators are formed, one is thresholded and λ is selected to minimize the Frobenius distance between the one empirical estimate and the other thresholded estimate.

4.1 | Multivariate Gaussian data

Table 1 displays false-positive and true-positive percentages for seven sparse estimators computed over 100 replications of a random sample of size $n = 50$ of $d = 50, 100, 200, 500$ dimensional multivariate Gaussian data with a tri-diagonal covariance matrix Σ whose diagonal entries are 1 and whose off-diagonal entries are 0.3. We can clearly see that the concentration-based estimator approaches the desired false-positive rate—either 1% or 5%—as the dimension increases. In contrast, the thresholding estimators with threshold λ chosen via cross validation generally start with higher false-positive percentages, which tend to zero as the dimension increases. As noted in previous work, hard thresholding is overly aggressive. The PDS method is very stable across changes in the dimension and maintains a constant 3.4% false-positive rate and 50% true-positive rate. A receiver operating characteristic (ROC) curve for this simulation is displayed on the left of Figure 1 achieved by varying the false-positive rate from 0% to 5%. Thus, our method achieves the same empirical true-positive rate as each of the cross-validated threshold methods does with the added benefit of having an interpretable penalization parameter.

TABLE 1 Percentage of false and true positives for multivariate Gaussian data and Σ tri-diagonal with diagonal entries 1 and off-diagonal entries 0.3

	False positive %				True positive %			
Dimension	50	100	200	500	50	100	200	500
CoM 1%	0.0	0.1	0.3	1.0	0.0	7.7	20.7	32.0
CoM 5%	1.0	2.2	3.5	4.7	33.1	42.9	51.5	56.0
PDS	3.4	3.4	3.4	3.4	50.0	50.0	51.5	50.6
Hard	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0
Soft	2.0	0.7	0.2	0.0	38.5	25.4	16.2	7.5
SCAD	2.1	0.7	0.3	0.0	39.0	26.0	16.4	7.5
Adpt	0.3	0.1	0.0	0.0	17.4	10.0	5.8	2.0

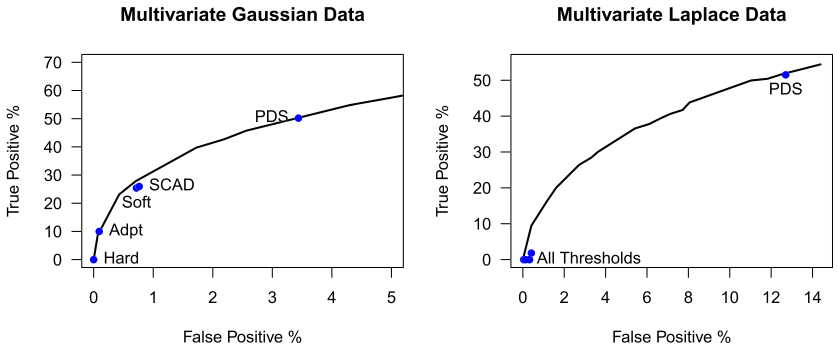

FIGURE 1 A line demarcating the trade-off between false- and true-positive recoveries for multivariate Gaussian (left) and Laplace (right) data from 100 replications of sample size $n = 50$ and dimension $d = 100$

TABLE 2 Percentage of false and true positives for multivariate Laplace data and Σ tri-diagonal with diagonal entries 1 and off-diagonal entries 0.3

	False positive %				True positive %			
Dimension	50	100	200	500	50	100	200	500
CoM 1%	0.2	0.4	0.7	1.1	4.5	9.2	13.0	17.2
CoM 5%	2.2	3.3	4.1	4.7	22.8	29.3	32.1	34.1
PDS	12.4	12.7	12.2	12.2	51.0	51.5	51.0	51.2
Hard	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Soft	1.2	0.4	0.2	0.0	11.3	1.8	0.0	0.0
SCAD	0.8	0.3	0.2	0.0	8.6	0.0	0.0	0.0
Adpt	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0

4.2 | Multivariate Laplace data

There are many possible ways to extend the univariate Laplace distribution onto $\mathbb{R}^d$. For the following simulation study, we chose the extension detailed in Eltoft et al. (2006). Namely, let $Z \sim \mathcal{N}(0, \sigma^2)$, and let $V \sim$ Exponential (1). Then, $X = \sqrt{V}Z \sim$ Laplace $\left(\sigma/\sqrt{2}\right)$, which has pdf $f(x) = \sqrt{2}\sigma^{-1}\exp\left(-\sqrt{2}\,|x|\,/\sigma\right)$ and variance $\mathrm{Var}(X) = \sigma^2$. For the multivariate setting, now let $Z \in \mathbb{R}^d$ be multivariate Gaussian with zero mean and covariance Σ, and, once again, let $V \sim$ Exponential (1). Then, we declare $X = \sqrt{V}Z$ to have a multivariate Laplace distribution with zero mean and covariance Σ.

Table 2 displays false-positive and true-positive percentages for seven sparse estimators computed over 100 replications of a random sample of size $n = 50$ of $d = 50,100,200,500$ dimensional multivariate Laplace data with a tri-diagonal covariance matrix Σ whose diagonal entries are 1 and whose off-diagonal entries are 0.3. Similarly to the previous setting, the concentration-based estimator approaches the desired false-positive rate—either 1% or 5%—as the dimension increases. All universal thresholding estimators set most of the entries in the matrix to zero when threshold λ is chosen via cross validation. The PDS method is still stable across changes in the dimension but fixates on a much higher false-positive rate around 12.5% and a similar true-positive rate of 51%. A ROC curve for this simulation is displayed on the right of Figure 1 achieved by varying the false-positive rate from 0% to 14%.

4.3 | Small round blue-cell tumour data

Following the same analysis performed in Rothman et al. (2009), Cai and Liu (2011), and Qiu and Liyanage (2019), we will consider the data set resulting from the SRBCT microarray experiment (Khan et al. 2001). The data set consists of a training set of 64 vectors containing 2,308 gene

Non-zero (%)	CoM 10%	CoM 5%	CoM 1%	PDS
Informative	30.3%	25.6%	8.5%	47.3%
Uninformative	5.4%	2.7%	0.4%	15.6%
	Hard	Soft	SCAD	Adpt
Informative	6.0%	24.7%	21.3%	9.9%
Uninformative	0.3%	2.3%	1.8%	0.7%

TABLE 3 The percentages of non-zero off-diagonal entries in the six covariance estimates partitioned into two parts: the informative 40×40 block of the highest scoring genes; the uninformative remaining matrix entries

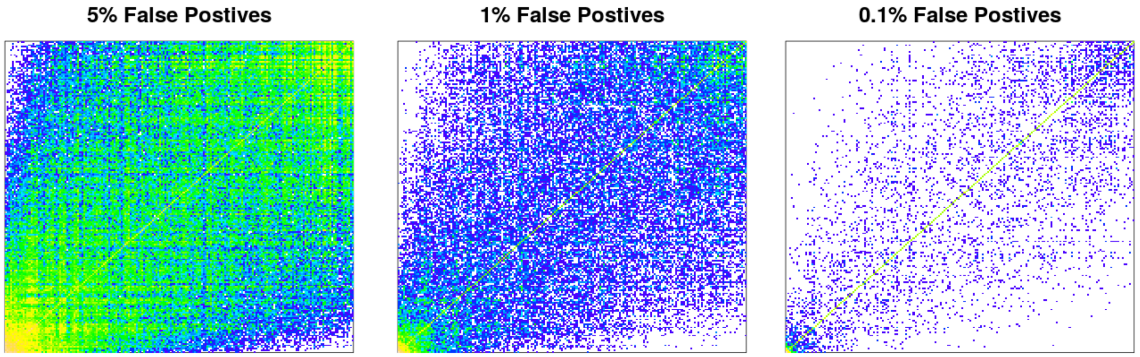

FIGURE 2 A density plot of the number of non-zero entries in $\hat{\Sigma}^{sp} \in \mathbb{R}^{2,308 \times 2,308}$ partitioned into 12×12 blocks for false-positive rates of 5%, 1%, and 0.1%

expressions. The data contains four types of tumours denoted as EWS, BL-NHL, NB, and RMS. As performed in the previous papers, the genes are ranked by their respective amount of discriminative information according to their *F*-statistic

$$F = \frac{\frac{1}{k-1}\sum_{m=1}^{k} n_m(\bar{x}_m - \bar{x})^2}{\frac{1}{n-k}\sum_{m=1}^{k}(n_m - 1)\hat{\sigma}_m^2},$$

where $\bar{x}$ is the sample mean, $k = 4$ is the number of classes, $n = 64$ is the sample size, n_m is the sample size of class m, and likewise, $\bar{x}_m$ and $\hat{\sigma}_m^2$ are, respectively, the sample mean and variance of class m. The top 40 and bottom 160 scoring genes were selected to provide a mix of the most and least informative genes.

Table 3 displays the results of applying the four threshold estimators with cross validation, the PDS method, and our concentration-based thresholding with the sub-Gaussian formula and with false-positive rates of 10%, 5%, and 1%. The percentage of matrix entries that are retained for the most informative 40×40 block and the least informative block is tabulated. Depending on the chosen false-positive rate, our concentration-based estimators give similar results to Soft and SCAD thresholding. PDS is the least conservative of the methods as it keeps the most entries. Hard and Adaptive LASSO thresholding are the most aggressive methods.

Note also in Table 3 that ρ is decreased from 10% to 5% to 1%, which is by a factor of 2 and then by a factor of 5. Empirically, we observe relative decreases in the number of uninformative entries kept of $5.4/2.7 = 2.0$ and $2.7/0.4 = 6.75$, which tracks with decrease in ρ.

Our method is also computationally efficient enough to consider the entire $2,308 \times 2,308$ matrix at once. It took only 131.3 s to compute $\hat{\Sigma}^{sp}$ on an Intel i7-7567U CPU, 3.50 GHz. In contrast, the PDS method, which still has significantly faster run times than cross validating the threshold estimators, took over 101 min to finish. False-positive rates of 5%, 1%, and 0.1% were tested. The fraction of non-zero entries in $\hat{\Sigma}^{sp}$ was 8.6%, 2.0%, and 0.22%, respectively. For comparison, the fraction of non-zero entries retained by PDS was 17.7%. If such an analysis is meant to lead to follow-up research on specific gene pairings, then culling as many false positives as possible is of critical importance. The sparse covariance estimator was partitioned into 12×12 blocks, and the number of non-zero entries was tabulated for each. The results are displayed in Figure 2.

ORCID

Adam B. Kashlak https://orcid.org/0000-0002-4050-7784

REFERENCES

Bickel, P. J., & Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, *36*(6), 2577–2604.

Bickel, P. J., & Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, *36*(1), 199–227.

Bien, J., & Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, *98*(4), 807–820.

Bien, J., & Tibshirani, R. (2012). spcov: Sparse estimation of a covariance matrix. Received from https://CRAN.R-project.org/package=spcov R package version 1.01.

Cai, T., & Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, *106*(494), 672–684.
Daniels, M. J., & Kass, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, *94*(448), 1254–1263.
Daniels, M. J., & Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, *57*(4), 1173–1184.
Dey, D. K., & Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *The Annals of Statistics*, *13*(4), 1581–1591.
El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, *36*(6), 2717–2756.
Eltoft, T., Kim, T., & Lee, T.-W. (2006). On the multivariate Laplace distribution. *IEEE Signal Processing Letters*, *13*(5), 300–303.
Haff, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, *8*(3), 586–597.
Hoff, P. D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(5), 971–992.
Iserles, A. (2009). *A first course in the numerical analysis of differential equations*. Cambridge, United Kingdom: Cambridge University Press.
Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, *29*(2), 295–327.
Johnstone, I. M., & Lu, A. Y. (2012). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, *104*(486), 682–693.
Kashlak, A. B. (2019). sparsematest: Sparse matrix estimation and inference. https://CRAN.R-project.org/package=sparseMatEst R package version 1.0.0.
Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., ..., & Peterson, C. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, *7*(6), 673–679.
Latała, R. (2005). Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, *133*(5), 1273–1282.
Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411.
Pourahmadi, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statistical Science*, *26*(3), 369–387.
Qiu, Y., & Liyanage, J. S. S. (2019). Threshold selection for covariance estimation. *Biometrics*, *75*(3), 895–905.
Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika*, *99*(3), 733–740.
Rothman, A. J. (2013). PDSCE: Positive definite sparse covariance estimators. https://CRAN.R-project.org/package=PDSCE R package version 1.2.
Rothman, A. J., Levina, E., & Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, *104*(485), 177–186.
Stein, C. (1975). Estimation of a covariance matrix. *Rietz Lecture*, *39th IMS Annual Meeting*, Atlanta, Georgia.
Tao, T. (2012). *Topics in random matrix theory* (Vol. *132*). Providence, Rhode Island, USA: American Mathematical Soc..
Xue, L., Ma, S., & Zou, H. (2012). Positive-definite L1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, *107*(500), 1480–1491.

How to cite this article: Kashlak AB, Kong L. Nonasymptotic support recovery for high-dimensional sparse covariance matrices. *Stat*. 2021;10:e316. https://doi.org/10.1002/sta4.316

APPENDIX A: PROOFS

Proof of Proposition 2. We begin with the collection of $N = d(d-1)/2$ random variables $\hat{\sigma}_{i,j} = n^{-1}\sum_{k=1}^{n} X_{k,i}X_{k,j}$, which we will denote as $Z_1, \ldots, Z_N$. Without loss of generality, assume that $Z_1, \ldots, Z_{N_0}$ have mean zero and $Z_{N_0+1}, \ldots, Z_{N_1+N_0}$ have non-zero mean and $N = N_0 + N_1$. To achieve η false positives, we would find the index k_0 corresponding to the $\lfloor(1-\eta)N_0\rfloor$ order statistic of the $Z_1, \ldots, Z_{N_0}$, and set all entries $|Z_i| \leq |Z_{k_0}|$ to zero. Instead, we find the index $\hat{k}$ corresponding to the $\lfloor(1-\eta)N\rfloor$ order statistic of all the Z_i.

Given that $\Sigma \in \mathcal{U}(\kappa, \delta)$, we have $\left|k_0 - \hat{k}\right| \leq \kappa d$. Thus, when considering the achieved false-positive rate $\#\{|Z_i| < |Z_{\hat{k}}| \,|\, i \leq N_0\}/N_0$ to the target rate $\#\{|Z_i| < |Z_{k_0}| \,|\, i \leq N_0\}/N_0$, we have $|\eta - \hat{\eta}| \leq \frac{\kappa d}{N_0} = \frac{2\kappa}{d-1-2\kappa} = O(d^{\nu-1})$. □

The proof of Theorem 1 from Section 3 relies on the following lemma involving symmetrization of random covariance matrices. In turn, the proof of this lemma relies mainly on Latała (2005) also discussed in Tao (2012).

Lemma 1. *Let $R \in \mathbb{R}^{d\times d}$ be a real valued symmetric random matrix with zero diagonal and mean zero entries bounded by 1 and not necessarily iid, and let $B \in \{0,1\}^{d\times d}$ be an iid symmetric Bernoulli random matrix with entries $b_{i,j} = b_{j,i} \sim$ Bernoulli(ρ) for $\rho \in (0,1)$. Denoting the entrywise or Hadamard product by $\circ$, let $A = R \circ B$. Let $\varepsilon \in \{-1,1\}^{d\times d}$ be a symmetric random matrix with iid Rademacher entries $\varepsilon_{i,j}$ for $j < i$ and $\varepsilon_{i,j} = \varepsilon_{j,i}$. Then, $\mathrm{E}\|A\circ\varepsilon\|_{S^\infty} \leq K_1 d^{1/2}\rho^{1/4} + K_2 d^{3/4}\rho^{1/2}$, where K_1, K_2 are universal constants.*

Proof. This proof follows from the result of Latała (2005) Theorem 2—also found in Theorem 2.3.8 of Tao (2012)—without the assumption of iid entries in the random matrix but with many entries equal to zero.

We first apply the expectation with respect to ε and use the result from Latała (2005).

$$\mathrm{E}\|A\circ\varepsilon\|_{S^\infty} = \mathrm{E}_A\mathrm{E}_\varepsilon\|A\circ\varepsilon\|_{S^\infty} \leq \left[\mathrm{E}_A\mathrm{E}_\varepsilon\|A\circ\varepsilon\|_{S^\infty}^2\right]^{1/2} \leq \left[K_1\mathrm{E}\max_{i=1,\ldots,d}\left(\sum_{j=1}^{d}a_{i,j}^2\right) + K_2\mathrm{E}\left(\sum_{i,j=1}^{d}a_{i,j}^4\right)^{1/2}\right]^{1/2}$$

with K_1, K_2 universal constants. For the second term in the above equation, we have via Jensen's inequality and the fact that $|a_{i,j}| \leq 1$ that

$$\mathsf{E}\left(\sum_{i,j=1}^{d} a_{i,j}^4\right)^{1/2} \leq \left(\sum_{i,j=1}^{d} \mathsf{E}a_{i,j}^4\right)^{1/2} \leq (d^2\rho)^{1/2} = d\rho^{1/2}.$$

For the first term in the above equation, we make use of the fact that $|a_{i,j}| \leq 1$ and that only ρ are non-zero resulting in

$$\mathsf{E}\max_{i=1,\dots,d}\left(\sum_{j=1}^{d} a_{i,j}^2\right) \leq \mathsf{E}\left(\sum_{i=1}^{d}\left(\sum_{j=1}^{d} a_{i,j}^2\right)^2\right)^{1/2} \leq \mathsf{E}\left(\sum_{i,j}^{d} a_{i,j}^2 + \sum_{i,j\neq k=1}^{d} a_{i,j}a_{i,k}\right)^{1/2} \leq \left(d^2\rho + (d^3-d^2)\rho^2\right)^{1/2} \leq d\rho^{1/2} + d^{3/2}\rho.$$

Combining the above results and updating the constants K_1, K_2 as necessary give the desired result that $\mathsf{E}\|A\circ\varepsilon\|_{S^\infty} \leq \left[K_1 d\rho^{1/2} + K_2 d^{3/2}\rho\right]^{1/2} \leq K_1 d^{1/2}\rho^{1/4} + K_2 d^{3/4}\rho^{1/2}$. □

Proof of Theorem. Without loss of generality, we can normalize $\hat{\Sigma}^{\text{emp}}$ such that the diagonal entries are 1. Thus $\hat{\Sigma}_0^{\text{emp}} = I_d$, the d dimensional identity matrix, and the off-diagonal entries of all matrices considered will be bounded in absolute value by one.

For the empirical covariance estimator, $\left\|\hat{\Sigma}^{\text{emp}} - \hat{\Sigma}_0^{\text{emp}}\right\|_{S^\infty} = \left\|\hat{\Sigma}^{\text{emp}}\right\|_{S^\infty} - 1$. We can decompose $\hat{\Sigma}^{\text{emp}}$ into three parts: the diagonal of ones; the off-diagonal terms corresponding to $\sigma_{i,j} \neq 0$; and the off-diagonal terms corresponding to $\sigma_{i,j} = 0$. The number of non-zero off-diagonal terms is bounded in each row/column by κ. Hence,

$$\left\|\hat{\Sigma}^{\text{emp}} - \hat{\Sigma}_0^{\text{emp}}\right\|_{S^\infty} \leq \left\|\hat{\Sigma}_{\neq 0}^{\text{emp}}\right\|_{S^\infty} + \left\|\hat{\Sigma}_{=0}^{\text{emp}}\right\|_{S^\infty} \leq \kappa + \left\|\hat{\Sigma}_{=0}^{\text{emp}}\right\|_{S^\infty},$$

where $\hat{\Sigma}_{=0}^{\text{emp}}$ has entries $\hat{\sigma}_{i,j}$ such that $\mathsf{E}\hat{\sigma}_{i,j} = 0$.

Let the entrywise or Hadamard product of two similar matrices A and B be $A \circ B$ with the ijth entry $(a_{i,j}b_{i,j})_{i,j}$. For ease of notation, we denote as $\Pi_0 = \hat{\Sigma}_{=0}^{\text{emp}}$. Let Π_1 be the result of randomly removing one half of the entries from Π_0, which is $\Pi_1 = \Pi_0 \circ B$ where $B \in \{0,1\}^{d\times d}$ is a symmetric random matrix with iid Bernoulli (1/2) entries. Considering the corresponding symmetric Rademacher random matrix, $\varepsilon = 2B - 1$, we then have that $\mathsf{E}\|\Pi_1\|_{S^\infty} = \mathsf{E}\|\Pi_0\circ B\|_{S^\infty} = \frac{1}{2}\mathsf{E}\|\Pi_0 \pm \Pi_0\circ\varepsilon\|_{S^\infty}$. where the $\pm$ comes from the symmetry of ε. Thus, $\left|\mathsf{E}\|\Pi_1\|_{S^\infty} - \frac{1}{2}\mathsf{E}\|\Pi_0\|_{S^\infty}\right| \leq \frac{1}{2}\mathsf{E}\|\Pi_0\circ\varepsilon\|_{S^\infty}$.

This idea can be iterated. Let $\Pi_m = \Pi_0 \circ B_1 \circ \dots \circ B_m$ with the B_i iid copies of B from before. Then, similarly,

$$\begin{aligned}\mathsf{E}\,\|\Pi_m\|_{S^\infty} &\leq \frac{1}{2}\mathsf{E}\,\|\Pi_{m-1}\|_{S^\infty} + \frac{1}{2}\,\|\Pi_{m-1}\circ\varepsilon_m\|_{S^\infty},\\ \mathsf{E}\|\Pi_m\|_{S^\infty} &\geq \frac{1}{2}\mathsf{E}\|\Pi_{m-1}\|_{S^\infty} - \frac{1}{2}\|\Pi_{m-1}\circ\varepsilon_m\|_{S^\infty}.\end{aligned}$$

Moreover, $|\mathsf{E}\|\Pi_m\|_\infty - 2^{-m}\mathsf{E}\|\Pi_0\|_{S^\infty}| \leq \sum_{j=0}^{m-1} 2^{-m+j}\mathsf{E}\|\Pi_j\circ\varepsilon\|_{S^\infty}$. Applying Lemma 1 m times and updating universal constants K_1, K_2 as necessary result in

$$\begin{aligned}\left|\mathsf{E}\|\Pi_m\|_\infty - 2^{-m}\mathsf{E}\|\Pi_0\|_{S^\infty}\right| &\leq \sum_{j=0}^{m-1} 2^{-m+j}\left[K_1 d^{1/2}2^{-j/4} + K_2 d^{3/4}2^{-j/2}\right]\\ &\leq K_1 d^{1/2}2^{-m/4} + K_2 d^{3/4}2^{-m/2}.\end{aligned}$$

Thus, for $\rho = 2^{-m}$, we have $|\mathsf{E}\|\Pi_m\|_\infty - \rho\mathsf{E}\|\Pi_0\|_{S^\infty}| \leq K_1 d^{1/2}\rho^{1/4} + K_2 d^{3/4}\rho^{1/2}$.

We want to replace the Π_m with $\hat{\Sigma}_\rho^{\text{emp}} - \hat{\Sigma}_0^{\text{emp}}$ and similarly for Π_0. The off-diagonal entries such that $\sigma_{i,j} \neq 0$ can contribute at most $\kappa = o(d^\nu)$, $\nu < 1$, to the operator norm. Hence,

$$\left|\mathsf{E}\left\|\hat{\Sigma}_\rho^{\text{emp}} - \hat{\Sigma}_0^{\text{emp}}\right\|_{S^\infty} - \rho\mathsf{E}\left\|\hat{\Sigma}^{\text{emp}} - \hat{\Sigma}_0^{\text{emp}}\right\|_{S^\infty}\right| \leq K_1 d^{1/2}\rho^{1/4} + K_2 d^{3/4}\rho^{1/2} + (1+\rho)o(d^\nu).$$

We lastly apply the crude—but effective in the nonasymptotic setting—bound $\left\|\hat{\Sigma}^{\text{emp}}\right\|_{S^\infty} \geq d/n$ almost surely. Dividing by $\mathsf{E}\left\|\hat{\Sigma}^{\text{emp}} - \hat{\Sigma}_0^{\text{emp}}\right\|_{S^\infty}$ results in

$$\left|\frac{\mathsf{E}\left\|\hat{\Sigma}_\rho^{\text{emp}} - \hat{\Sigma}_0^{\text{emp}}\right\|_{S^\infty}}{\mathsf{E}\left\|\hat{\Sigma}^{\text{emp}} - \hat{\Sigma}_0^{\text{emp}}\right\|_{S^\infty}} - \rho\right| \leq K_1 n d^{-1/2}\rho^{1/4} + K_2 n d^{-1/4}\rho^{1/2} + o(nd^{\nu-1}).$$

Thus, we require $\nu < 1/2$ to make the final term negligible for large d with respect to the others.

We can extend this result to arbitrary $\rho \in (0, 0.5]$ by using the simple observation that given such a ρ, there exists an $a \in \mathbb{Z}^+$ such that $2^a\rho \in [0.5, 1)$. Therefore, setting $\eta = 2^a\rho$ and replacing $\hat{\Sigma}^{\text{emp}}$ with the corresponding matrix $\hat{\Sigma}_\eta^{\text{emp}}$ from Proposition 2 allow us to proceed as above. □