# Operator norm consistent estimation of large dimensional sparse covariance matrices

Noureddine El Karoui *
*Department of Statistics,*
*University of California, Berkeley*

June 15, 2007

**Abstract**

Estimating covariance matrices is a problem of fundamental importance in multivariate statistics. In practice it is increasingly frequent to work with data matrices $X$ of dimension $n \times p$, where $p$ and $n$ are both large. Results from random matrix theory show very clearly that in this setting, standard estimators like the sample covariance matrix perform in general very poorly.

In this "large $n$, large $p$" setting, it is sometime the case that practitioners are willing to assume that many elements of the population covariance matrix are equal to 0, and hence this matrix is sparse. We develop an estimator to handle this situation. The estimator is shown to be consistent in operator norm, when $p/n \to l \neq 0$, where $l$ is generally finite, as $p \to \infty$. In other words the largest eigenvalue of the difference between the estimator and the population covariance matrix goes to zero. This implies consistency of all the eigenvalues and consistency of eigenspaces associated to isolated eigenvalues.

We also propose a notion of sparsity for matrices that is "compatible" with spectral analysis and is independent of the ordering of the variables.

## 1 Introduction

Estimating covariance matrices is the cornerstone of much of multivariate statistics. Theoretical contributions (see James and Stein (1961), Haff (1980), Anderson (2003), Chap. 7) have been highlighting for a long time the fact that for various loss functions, one could improve on the sample covariance matrix as an estimator of the population covariance matrix, in a non-asymptotic setting.

The "large $n$, large $p$" context, i.e multivariate analysis of datasets for which both the number of observations, $n$ and the number of predictors $p$ are large, is, in a somewhat different setting, highlighting the deficiency of this estimator. To be more precise, when we refer to "large $n$, large $p$" problems, we mean that $p/n$ has a non-zero limit as $n \to \infty$. Results from random matrix theory (Marčenko and Pastur (1967)) make clear that in this asymptotic setting, even at just the level of eigenvalues, the sample covariance matrix will not lead to a consistent estimator. We refer to El Karoui (2006) for a more thorough introduction to these ideas and the consequences of the results for statistical practice.

This is naturally very problematic since this class of results suggest that the sample covariance matrix contains little reliable information about the population covariance. This realization has helped generate a significant amount of work in mathematics, probability and theoretical statistics and the behavior of many hard to analyze quantities are now quite well understood. For instance, under strong distributional assumptions, one can characterize the fluctuation behavior of the largest eigenvalue of sample covariance matrices for quite a large class of population covariance (see e.g El Karoui (2007) for the latest results), or

*Acknowledgements: The author is grateful to Peter Bickel for many very interesting discussions on this and related topics. He would like to thank Elizabeth Purdom for discussions that lead to clarifications at the beginning of this project and Jim Pitman for references. Support from NSF grant DMS-0605169 and hospitality and support from SAMSI in the Fall of 2006 are gratefully acknowledged. **AMS 2000 SC:** 62H12. **Key words and Phrases :** covariance matrices, correlation matrices, adjacency matrices, eigenvalues of covariance matrices, multivariate statistical analysis, high-dimensional inference, random matrix theory, sparsity, $\beta$-sparsity. **Contact :** nkaroui@stat.berkeley.edu

the fluctuation behavior of linear functionals of eigenvalues (see Jonsson (1982), Bai and Silverstein (2004), and Anderson and Zeitouni (2006)). However, until very recently there has been less work in the direction of using these powerful results for the sake of concrete data analysis.

Of course, since this inconsistency phenomenon is now fairly well-known, remedies have been proposed. For instance the interesting paper Ledoit and Wolf (2004) proposes to shrink the sample covariance matrix towards the identity matrix using a shrinkage parameter chosen from the data. In El Karoui (2006), a non-parametric estimator of the spectrum is proposed and shown to be consistent in the sense of weak convergence of distributions. The method in El Karoui (2006) uses convex optimization, random matrix theory (the generalization of Marčenko and Pastur (1967) found in Silverstein (1995)) and ideas from non-parametric function estimation. These estimation methods rely on asymptotic properties of eigenvalues, and as a starting point for estimation of the full covariance matrix, they are essentially trying to get an estimator that is equivariant under the action of the orthogonal (or unitary) group. In other words, the "basis" in which the data is given is not taken advantage of, and the premise of such an analysis is that we should be able to find good estimators in any "basis". While ideally researchers will be able to come up with strategies to solve the estimation problem at this level of generality, it is reasonable to expect that taking advantage of the representation of the data we are given should or might help finding good estimators.

In particular, it is often the case that data analysts are willing to assume that the basis in which the data is given is somewhat nice. Often this translates by the assumption that the population covariance matrix has a particular structure in this basis, which should naturally be taken advantage of. In this situation, it becomes natural to perform certain forms regularization by working directly on the entries of the sample covariance matrix. Various strategies have been proposed (see Huang et al. (2006), Bengtsson and Furrer (2007)) that try to take advantage of the assumed structure. The very interesting paper Bickel and Levina (2007) proposed the idea of "banding" covariance matrices when it is known that the population covariance has small entries far away from the diagonal. The idea is to put to zero all coefficients that are too far away from the diagonal and to keep the other ones unchanged. Remarkably, in Bickel and Levina (2007), the authors show consistency of their estimator in spectral (a.k.a operator) norm, a very nice result. In other words, they show that the largest singular value of the difference between their estimator and the population covariance matrix goes to zero as both dimensions of the matrices go to infinity and for instance when $p/n$ has a finite limit. The requirement of estimating consistently in spectral norm is a very interesting idea (which we adopt in this paper), since then one can deduce easily many results concerning consistency of eigenvalues and eigenspaces. We make this remark more precise in Subsection 3.5, using different arguments than those used by Bickel and Levina.

It might be argued that ideas such as banding essentially assume that one knows a "good" ordering of the variables. As a matter of fact, if we start with a matrix with entries small or zero away from the diagonal and reorder the variables, the new covariance matrix we obtain may not have only small entries away from the diagonal. In some situations, for instance time series analysis, the order of the variables has a statistical/scientific meaning and so using it makes sense. However, in many data analytic problems, there is no canonical ordering of the variables.

Hence to tackle these situations, a natural requirement is to find an estimator which is equivariant under permutations of the variables. We call such estimators permutation-equivariant. Such an estimator would take advantage of the particular nature of the basis in which the data is given, while not requiring the user to find a permutation of the order of the variables that makes the analysis particularly simple. Searching for such a permutation would - in general - be practically infeasible. Note that regularizing the estimator by applying the same function to each of the entries of the matrix leads to permutation-equivariant estimators.

A subject of particular practical interest is the estimation of sparse covariance matrices (see for instance d' Aspremont et al. (2006)) because they are appealing to practitioners for several reasons, including interpretability, presumably ease of estimation and the practically often encountered situation where while many variables are present in the problem, most of them are correlated to only "a few" others.

In this paper we propose to estimate sparse matrices by hard thresholding small entries of the sample covariance matrix and putting them to zero. We propose a combinatorial view of the problem, inspired in part by classical ideas in random matrix theory, going back to Wigner (1955). The notion of sparsity we propose is flexible enough that it makes the proofs manageable and at the same time rich enough that it

captures many practically natural situations.

We show that our estimators are consistent in spectral norm, both in the case of the sample covariance and the sample correlation matrix. No assumptions of normality of the data are required, only the existence of certain moments. As is to be expected, the larger the number of moments available, the easier the task and the larger the class of matrices we can estimate consistently.

## 2 Sparse matrices: concepts and definitions

One conceptual difficulty of this problem is to define a notion of sparsity for matrices that is compatible with spectral analysis. Just as in the case of norms, extending straightforwardly the notions from vectors to matrices can be somewhat unhelpful. In the norm case, the Frobenius norm - the extension of the $\ell_2$ (vector) norm to matrices - is for instance known to not be as good as other matrix norms from a spectral point of view. Similarly here, we will explain that just counting the number of 0's in the matrix - the canonical sparsity notion for vectors - does not yield a "good" notion of sparsity when one investigates the spectral properties of matrices.

Let us illustrate our problem on a concrete example. Consider now two $p \times p$ symmetric matrices with the same number of non-zero coefficients:

$$E_1 = \begin{pmatrix} 1 & \frac{1}{\sqrt{p}} & \frac{1}{\sqrt{p}} & \dots & \frac{1}{\sqrt{p}} \\ \frac{1}{\sqrt{p}} & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \frac{1}{\sqrt{p}} & 0 & 0 & 1 & 0 \\ \frac{1}{\sqrt{p}} & 0 & 0 & \dots & 1 \end{pmatrix} \text{ and } E_2 = \begin{pmatrix} 1 & \frac{1}{\sqrt{p}} & 0 & \dots & 0 \\ \frac{1}{\sqrt{p}} & 1 & \frac{1}{\sqrt{p}} & \dots & 0 \\ 0 & \frac{1}{\sqrt{p}} & 1 & \frac{1}{\sqrt{p}} & \dots \\ \vdots & \ddots & \ddots & \ddots & \frac{1}{\sqrt{p}} \\ 0 & \dots & 0 & \frac{1}{\sqrt{p}} & 1 \end{pmatrix} .$$

Using the Schur decomposition of $E_1$ to compute its characteristic polynomial (see also Subsection 3.3), we see easily that its eigenvalues are $(p-2)$ 1's and $1+\sqrt{p-1}/\sqrt{p}$ and $1-\sqrt{p-1}/\sqrt{p}$. On the other hand, $E_2$ is a well-known matrix, for instance in numerical analysis, and its eigenvalues are $\{1+2\cos(k\pi/(p+1))/\sqrt{p}\}_{k=1}^{p}$. Hence, the extreme eigenvalues of these matrices are very different, but they have the same number of non-zero coefficients and their non-zero coefficients have the same values. This raises the question of trying to come up with an alternative notion of sparsity that while encompassing the canonical notion of "having a large number of zeros" might be better suited for the study and the understanding of spectral properties of matrices.

### 2.1 Matrix sparsity: proposed definition

To describe our proposal, we need to introduce several concepts from graph theory and combinatorics. For the sake of readability we detail them here; they can also be found in for instance Stanley (1986), Section 4.7. To each population covariance matrix, $\Sigma_p$, it is natural to associate an adjacency matrix $A_p(\Sigma_p)$, in the following fashion:

$$A_p(i,j) = 1_{\sigma(i,j)\neq 0} .$$

This matrix $A_p$ can in turn be viewed as the adjacency matrix of a graph $\mathcal{G}_p$, with $p$ vertices, corresponding to the variables in our statistical problem. We call a path (or a walk) on this graph closed if it starts and finishes at the same vertex. The length of a path is the number of edges it traverses. By definition, we call

$$\mathcal{C}_p(k) = \{\text{closed paths of length } k \text{ on the graph with adjacency matrix } A_p\}$$

and

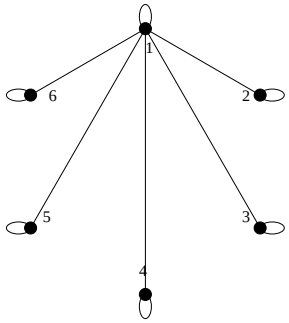
$$\phi_p(k) = \text{Card}\left\{\mathcal{C}_p(k)\right\} .$$

Note that we have

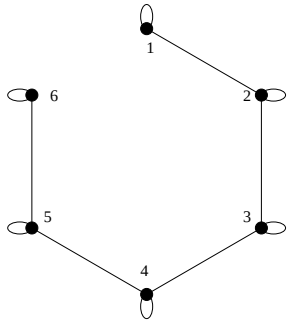
$$\phi_p(k) = \text{trace}\left(A_p^k\right) .$$

The following two drawings show the graphs corresponding to the adjacency matrices of $E_1$ and $E_2$:

Graph corresponding to $E_1$:




Graph corresponding to $E_2$:




**Definition 1.** *We say that a sequence of covariance matrices $\{\Sigma_p\}_{p=1}^{\infty}$ is $\beta$-sparse if the graphs associated to them via $A_p$'s have the property that*

$$\forall k \in 2\mathbb{N}\,, \phi_p(k) \leq f(k) p^{\beta(k-1)+1}$$

*where $f(k) \in \mathbb{R}^+$ is independent of $p$ and $0 \leq \beta \leq 1$.*

*We say that a sequence of matrices is asymptotically $\beta$-sparse if it is $\beta + \epsilon$ sparse for any $\epsilon > 0$.*

*We call $\beta$ an index of sparsity of the sequence of matrices.*

For short, we say that a matrix is $\beta$-sparse instead of saying that a sequence of matrices is $\beta$-sparse when this shortcut does not cause any confusion.

Here are a few simple examples of matrices that are sparse according to our definition.

1. **Diagonal matrices** In the case of diagonal matrices, $A_p = \mathrm{Id}_p$, and $\mathcal{G}_p$ consists only of self-loops at each vertex. Hence $\phi(k) = p$, for all $k$. So a diagonal matrix is 0-sparse.

2. **Matrices with at most $M$ non-zero elements on each line** For these matrices, the corresponding $\mathcal{G}_p$ has at most $M$ edges at each vertex. A simple enumeration shows that $\phi(k) \leq pM^{k-1}$. So these matrices are also 0-sparse.

3. **Matrices with at most $Mp^{\alpha}$ non-zero elements on each line** The same argument shows that $\phi(k) \leq p(Mp)^{\alpha(k-1)}$. So these matrices are $\alpha$ sparse. In particular, full matrices are 1-sparse.

4. **Matrices with at most $M(\log p)^r$ non-zero elements on each line** We have, by simple counting arguments, $\phi_p(k) \leq pM^{k-1}(\log p)^{r(k-1)}$. These matrices are therefore $\beta$-sparse for any $\beta > 0$ and asymptotically 0-sparse.

Given a matrix $S_p$, we can associate to the corresponding $\mathcal{G}_p$ a set of weights on the edges, by simply setting the weight of the edge joining vertices $i$ and $j$ to $S_p(i,j)$. Similarly, for a path, we have

**Definition 2** (Weight of a path)**.** *Given $\gamma$, a closed path of length $k$: $\gamma : \; i_1 \to i_2 \to i_3 \to \ldots \to i_k \to i_{k+1} = i_1$, and a matrix $S_p$, we call $w_\gamma$ the weight of the path $\gamma$. By definition it is*

$$w_\gamma = S_p(i_1, i_2) S_p(i_2, i_3) \ldots S_p(i_k, i_1)\,.$$

We conclude this section by the following simple but important remark:

$$\mathrm{trace}\left(S_p^k\right) = \sum_{\gamma \in \mathcal{C}_p(k)} w_\gamma$$

## 2.2 Remarks on the notion of sparsity proposed

It is clear that if we change the order of the variables in our statistical problem, we do not change the "index of sparsity" of our matrices. This is essentially obvious from the graph representation of the problem. From a more algebraic standpoint, if the permutation that is applied is encoded as a permutation matrix $P$, the covariance in the permuted problem is simply $P'\Sigma_p P$ and the new adjacency matrix is $P' A_p P$

(this matrix is indeed an adjacency matrix). Since $P'P = \mathrm{Id}_p$, we have $\mathrm{trace}\left((P'A_pP)^k\right) = \mathrm{trace}\left(A_p^k\right)$, and hence the sparsity index is unchanged when we permute the variables.

We also note that we could replace the notion of $\beta$-sparsity we use by the requirement that

$$\phi_p(k) \leq f(k)p^{1+\beta k} \ , \forall k \in 2\mathbb{N} \ .$$

This would result in minor differences in the theorems that follow and might be slightly simpler to apply when the only information available concerns the largest eigenvalue of $A_p^2$. From a combinatorial point of view, the notion we use in this paper is more natural and this is what directed our choice.

It is clear that the smaller $\beta$, the sparser the matrix. In particular, if $\beta_0 \leq \beta_1$, a matrix which is $\beta_0$-sparse is also $\beta_1$-sparse. As we will shortly show, the class of $\beta$-sparse matrices is stable by addition, which implies that the sum of a $\beta_0$-sparse and a $\beta_1$-sparse matrix is $(\beta_0 \vee \beta_1)$-sparse.

We conclude this discussion with the proof of the following fact:

**Fact 1.** *The set of $\beta$-sparse matrices is stable by addition. In other words, the sum of two $\beta$-sparse matrices is $\beta$-sparse.*

*Proof of Fact 1.* We call $B_0$ and $B_1$ our "initial" $\beta$-sparse matrices, and $B_2$ their sum. $A_2$, the adjacency matrix of $B_2$ is not the sum of $A_0 + A_1$. In particular, edges that are present in both $A_0$ and $A_1$ may not be present in $A_2$. However, if we add edges to $A_2$, we increase $\phi_p^{(2)}(k)$, the number of closed paths of length $k$ on $A_2$. So in checking the sparsity index of $B_2$, we can work with $\widetilde{A}_2$, which contains all edges in $A_0$ and $A_1$, and contains the graph corresponding to $A_2$ as a subgraph of its own graphical representation. More algebraically, the definition of $\widetilde{A}_2$ is

$$\widetilde{A}_2(i,j) = \min(A_0(i,j) + A_1(i,j), 1) = 1_{A_1(i,j)=1} + 1_{A_0(i,j)=1}1_{A_1(i,j)=0} \ .$$

We can write $\widetilde{A}_2 = \widetilde{A}_0 + A_1$, with $\widetilde{A}_0(i,j) = 1_{A_0(i,j)=1}1_{A_1(i,j)=0}$. Note that $\widetilde{A}_0$ is a symmetric adjacency matrix, may have zeros where $A_0$ has ones, but does not have ones where $A_0$ had zeros. So the graph corresponding to $\widetilde{A}_0$ is a subgraph of the graph corresponding to $A_0$. In particular, $\mathrm{trace}\left(\widetilde{A}_0^{2k}\right) \leq \mathrm{trace}\left(A_0^{2k}\right)$.

The matrices $\widetilde{A}_0$, $A_1$ and $\widetilde{A}_2$ are all symmetric, so when dealing with their eigenvalues we can apply standard results for symmetric matrices. Using Lidskii's theorem (see Bhatia (1997), Corollary III.4.2), we know that

$$\lambda^{\downarrow}(\widetilde{A}_2) \prec \lambda^{\downarrow}(\widetilde{A}_0) + \lambda^{\downarrow}(A_1) \ ,$$

where $\lambda^{\downarrow}(A_1)$ is the vector of decreasing eigenvalues of $A_1$ and the sign $\prec$ means that the left-hand side is majorized by the right hand-side (see Bhatia (1997) p.28 for a definition, if needed). Now the functions $h(x) = x^{2k}$ are convex and we therefore have, using standard results in the theory of majorization (Bhatia (1997), Theorem II.3.1),

$$\mathrm{trace}\left(\widetilde{A}_2^{2k}\right) \leq \sum [\lambda_j(\widetilde{A}_0) + \lambda_j(A_1)]^{2k} \leq 2^{2k-1} \sum \lambda_j(\widetilde{A}_0)^{2k} + \lambda_j(A_1)^{2k} \leq 2^{2k-1} \mathrm{trace}\left(A_0^{2k} + A_1^{2k}\right) \ .$$

Because $A_0$ and $A_1$ are $\beta$-sparse, we see that $\widetilde{A}_2^{2k}$ is. And because we have seen that

$$\mathrm{trace}\left(A_2^{2k}\right) \leq \mathrm{trace}\left(\widetilde{A}_2^{2k}\right) \ ,$$

we conclude that $B_2$ is $\beta$-sparse. □

## 3 Estimation by entry-wise thresholding

To avoid any confusion as to the meaning of the results to be proved, we remind the reader that the spectral norm of a matrix $A$ is defined (see Horn and Johnson (1990), p. 295) as $|||A|||_2 = \max\{\sqrt{\lambda} : \lambda \text{ an eigenvalue of } A^*A\}$; in other words, it is the largest singular value of $A$. When $A$ is a symmetric matrix, $|||A|||_2$ coincides with the spectral radius of $A$: $\rho(A) = \max|\lambda_i(A)|$. In what follows, we use interchangeably the terms spectral norm and operator norm.

When we say that we threshold a variable $x$ at level $t$ we mean that we keep (or replace $x$ by) $x1_{|x|\geq t}$. We also refer to this operation as hard thresholding. Our final remark concerns notation: in what follows, $C$ refers to a generic constant independent of $n$ and $p$. Its value may change from display to display when there is no risk of confusion about the meaning of the statements. If there are, we also use $K$ or $C'$ and they play the same role as $C$.

### 3.1 Estimation of sparse covariance or correlation matrices

We first prove an intermediate result concerning the Gaussian MLE estimator when it is known that the mean of the data is zero (Theorem 1). This is a stepping stone to the more practically relevant results concerning the sample covariance matrix (Theorem 2) and the sample correlation matrix (Theorem 3). The proofs of these later results are essentially the same as that of Theorem 1, but the proof of Theorem 1 is technically a bit less complicated and highlights the key ideas. We refer the reader to Subsection 3.5 for detailed explanations of the consequences of Theorems 1, 2 and 3. Finally, we stress that all of our results are obtained when $p/n$ has a non-zero limit, i.e in the "large $n$, large $p$" setting.

**Theorem 1.** *Suppose $X$ is an $n \times p$ matrix, with $p/n \to l \in (0, \infty)$. Suppose that the rows of $X$ are independent and identically distributed and denote them by $\{X_i\}_{i=1}^n$. Call $\Sigma_p$ the matrix of the vector $X_1$. Suppose $\Sigma_p$ is $\beta$-sparse with $\beta = 1/2 - \eta$ and $\eta > 0$. Suppose that the non-zero coefficients of $\Sigma$ are all greater in absolute value than $Cn^{-\alpha_0}$, with $0 < \alpha_0 = 1/2 - \delta_0 < 1/2$. Suppose further that for all $(i, j)$, $X_{i,j}$ has mean 0 and finite moments of order $4k(\eta)$, with $k(\eta) \geq (1.5 + \epsilon + \eta)/(2\eta)$ and $k(\eta) \in \mathbb{N}$, for some $\epsilon > 0$. Assume that $k(\eta) \geq (2 + \epsilon + \beta)/(2\delta_0)$. Call*

$$S_p = \frac{1}{n} \sum_{i=1}^{n} X_i X_i'$$

*Call $T_\alpha(S_p)$ the matrix obtained from thresholding the entries of $S_p$ at the level $Kn^{-\alpha}$ with $\alpha = 1/2 - \delta > \alpha_0$. Then we have, if we call $\Delta_p = T_\alpha(S_p) - \Sigma_p$,*

$$|||\Delta_p|||_2 = |||T_\alpha(S_p) - \Sigma_p|||_2 \to 0 a.s\,,$$

*where $|||M|||_2$ is the spectral norm of the matrix $M$.*

We postpone a short discussion of the meaning of this theorem to after the statement of Theorem 2, which is arguably more interesting practically.

*Proof of Theorem 1.* We divide the proof into two parts. The first part consist in showing the "oracle" version of the theorem, i.e showing that operator norm consistency happens when one is given the pairs $(i, j)$ for which $\sigma_p(i, j) = 0$. The second part shows that the empirical thresholding does not affect this result.

Let us first remind the reader of a variant of Hölder's inequality. Let $A_1, \dots A_m$ be random variables with finite absolute $m$-th moment. Then we have

$$\left| \mathbf{E}\left( \prod_{i=1}^{m} A_i \right) \right| \leq \prod_{i=1}^{m} \mathbf{E}\left( |A_i|^m \right)^{1/m} .$$

Note that for the case $m = 2$, this is just the Cauchy-Schwarz inequality. So the result is true when $m = 2$. We prove it by induction on $m$. Suppose therefore it is true for all integers less or equal to $m - 1$. Call $B_1 = \prod_{i=2}^{m} A_i$. By Hölder's inequality, we have

$$|\mathbf{E}\left(A_1 B_1\right)| \leq \left(\mathbf{E}\left(|A_1|^m\right)\right)^{1/m} \left[\mathbf{E}\left(|B_1|^{\frac{m}{m-1}}\right)\right]^{\frac{m-1}{m}} .$$

Now, by the induction hypothesis, applied to the random variables $|A_i|^{m/(m-1)}$,

$$\mathbf{E}\left(|B_1|^{\frac{m}{m-1}}\right) = \mathbf{E}\left( \prod_{i=2}^{m} |A_i|^{m/(m-1)} \right) \leq \prod_{i=2}^{m} \left[\mathbf{E}\left(|A_i|^m\right)\right]^{1/(m-1)} .$$

Therefore, $\left[\mathbf{E}\left(|B_1|^{\frac{m}{m-1}}\right)\right]^{\frac{m-1}{m}} \leq \prod_{i=2}^{m} \mathbf{E}\left(|A_i|^m\right)^{1/m}$ and the inequality is verified.

Now given $\gamma(2k)$, a closed path of length $2k$, and the associated matrix $M$ we clearly have

$$|\mathbf{E}\left(w_{\gamma(2k)}\right)| \leq \mathbf{E}\left(\left|w_{\gamma(2k)}\right|\right) \leq \prod_{j=1}^{2k} \left[\mathbf{E}\left(|M(i_j, i_{j+1})|^{2k}\right)\right]^{1/2k} , \tag{1}$$

assuming, for a moment that the relevant moments exist.

• **Oracle part of the proof**

Let us first introduce some notations. We denote by $\sigma_p(i,j)$ the $(i,j)$-th entry of $\Sigma_p$, the population covariance. We call oracle $(S_p)$ the matrix with entries $S_p(i,j)1_{\sigma_p(i,j)\neq 0}$ and

$$\Xi_p = \text{oracle}\,(S_p) - \Sigma_p \, .$$

Note that we have

$$\Xi_p(i,j) = (S_p(i,j) - \sigma_p(i,j))1_{\sigma_p(i,j)\neq 0} \, .$$

In the oracle setting, where we assume we know the patterns of zeros in $\Sigma_p$, so we focus on the matrix $\Xi_p$. Clearly $\Sigma_p$ and $\Xi_p$ have the same patterns of 0's and non-zero, and so if $\Sigma_p$ is $\beta$-sparse, so is $\Xi_p$. Equation (1) shows that if we can control the moments $(\Xi_p(i,j))^{2k}$, we will be able to bound the expected weight of each path. Now we remark that we can write

$$\Xi_p(i,j) = \frac{1}{n}\sum_{m=1}^{n} Z_m$$

where $Z_m$'s are independent, identically distributed and with mean 0, since $S_p$ is unbiased for $\Sigma_p$. By expanding the power, we get that

$$(\Xi_p(i,j))^{2k} = \frac{1}{n^{2k}} \sum_{i_1,\ldots,i_{2k}} Z_{i_1} \ldots Z_{i_{2k}} \, .$$

This last quantity can be rewritten

$$Z_{i_1} \ldots Z_{i_{2k}} = \prod_{i=1}^{n} Z_i^{k_i} \, , \text{ with } \sum_{i=1}^{n} k_i = 2k \, , \text{ and } k_i \geq 0$$

We now remark that if there exists $i_0$ such that $k_{i_0} = 1$, then $\mathbf{E}\left(Z_{i_1} \ldots Z_{i_{2k}}\right) = 0$, by independence and the fact that each of the $Z_i$'s have mean 0. Therefore, in the expansion of $(\Xi_p(i,j))^{2k}$, only the terms for which all non-zero $k_i$'s are greater or equal to 2 will contribute to the expectation. Counting the number of distinct indices appearing in the product above allows us to get a first order estimate of $\mathbf{E}\left(\Xi_p(i,j)^{2k}\right)$. As a matter of fact, the contribution of products with $q$ distinct indices is of order $n^{-2k}n^q$, by simply counting how many such products there are. So we see that to first order, the only products that matter are those for which all the $Z_i$'s raised to a non-zero power are raised to the power 2. Denoting by $n^{[k]}$ the $k$-th factorial moment $n(n-1)\ldots(n-k+1)$, we have, assuming that $\mathbf{E}\left(Z_i^{2k}\right) < \infty$,

$$\mathbf{E}\left(\Xi_p(i,j)\right)^{2k} \leq \frac{n^{[k]}}{n^{2k}} \frac{2k!}{2^k k!} \left[\mathbf{E}\left(Z_i^2\right)\right]^k + \frac{1}{n^{2k}} \mathrm{O}(n^{k-1}) = \mathrm{O}(\frac{1}{n^k}) \, , \text{ if } k \text{ is fixed and } n \to \infty \, .$$

We therefore have

$$\left[\mathbf{E}\left(\Xi_p(i,j)\right)^{2k}\right]^{1/2k} = \mathrm{O}(\frac{1}{\sqrt{n}}) \, .$$

In particular, the weight of a closed path of length $2k$ on the graph with adjacency matrix $A_p(\Sigma_p)$ (or $A_p(\Xi_p)$) and weights $\Xi_p(i,j)$ has the property that

$$\left|\mathbf{E}\left(w_{\gamma(2k)}\right)\right| \leq \mathbf{E}\left(\left|w_{\gamma(2k)}\right|\right) = \mathrm{O}(n^{-k}) \, .$$

Since we have assumed that $\Sigma_p$ and therefore $\Xi_p$ are $\beta$-sparse, we have

$$\mathbf{E}\left(\text{trace}\left(\Xi_p^{2k}\right)\right) = \mathrm{O}(p^{1+\beta(2k-1)}n^{-k}) \ .$$

Since we assume that $p \asymp n$, we see that $\mathbf{E}\left(\text{trace}\left(\Xi_p^{2k}\right)\right) = \mathrm{O}(n^{1/2+\eta-2k\eta})$, where $\eta = 1/2-\beta$. In particular, if $k$ is chosen such that

$$k \geq \frac{1.5+\epsilon+\eta}{2\eta} \ ,$$

we see that

$$\mathbf{E}\left(|||\Xi_p|||_2^{2k}\right) \leq \mathbf{E}\left(\text{trace}\left(\Xi_p^{2k}\right)\right) = \mathrm{O}(n^{-(1+\epsilon)}) \ ,$$

because $\Xi_p$ is a symmetric matrix, so its spectral norm squared is one of its eigenvalues squared. Using Chebyshev's inequality and the first Borel-Cantelli lemma, we conclude that

$$|||\Xi_p|||_2 \to 0 \text{ a.s}$$

Note that $2k > 1 + 1/(2\eta)$ would have guaranteed convergence in probability. The above proof is correct if $Z_m$ has a finite $2k$-th moment. Since $Z_m = X_{m,i}X_{m,j}$, the assumption that the entries of the data matrix $X$ have a $4k$-th moment guarantees the existence of a $2k$-th moment for $Z_m$, using for instance the Cauchy-Schwarz inequality.

We have shown that $|||\text{oracle}\,(S_p) - \Sigma_p|||_2 \to 0$ almost surely, when the conditions of the theorem are satisfied.

• **Non-oracle part of the proof**

We now turn to the non-oracle version of the procedure. It is clear that all we need to do at this point is to show that the thresholding procedure will lead a.s to the right adjacency matrix. Recall the notation $\Delta_p = T_\alpha(S_p) - \Sigma_p$, the difference between our estimator and the population covariance. Call $B_p$ the event $B_p = \{$at least one mistake is made by thresholding$\}$, i.e $A_p(T_\alpha(S_p)) \neq A_p(\Sigma_p)$. Call $E_p$ the event $\{|||\Delta_p|||_2 > \epsilon\}$ and $F_p$ the event $\{|||\Xi_p|||_2 > \epsilon\}$ (we do not index these events by $\epsilon$ to alleviate the notation). Note that

$$E_p = (E_p \cap B_p)\bigcup(E_p \cap B_p^c) \subseteq B_p \bigcup (E_p \cap B_p^c) = B_p \bigcup (F_p \cap B_p^c) \subseteq B_p \bigcup F_p \ .$$

We have already seen that $P(F_p \text{ infinitely often}) = 0$, so if we can show that $P(B_p \text{ i.o}) = 0$, we will have $P(E_p \text{ i.o}) = 0$, as desired.

Call $O_p = \text{oracle}\,(S_p)$ and $S_p^- = S_p - O_p$, where $\text{oracle}\,(S_p)$ is defined above. Note that $S_p^-$ has non-zero entries only where $\Sigma_p$ has entries equal to 0; when that is the case, $S_p^-(i,j) = \widehat{\sigma}(i,j)$. We call $\mathcal{D}_p$ the set of pairs $(i,j)$ such that $\sigma(i,j) = 0$, i.e

$$\mathcal{D}_p = \{(i,j) : \sigma_p(i,j) = 0\} \ .$$

We will first show that the maximal element of $S_p^-$ stays below $n^{-\alpha}$ a.s. Note that in general, for a random matrix $M$ and index sets $I$ and $J$,

$$P(\max_{i\in I, j\in J}|m_{i,j}| > \epsilon) \leq \sum_{i\in I, j\in J} P(|m_{i,j}| > \epsilon) \ .$$

The same moment computations as the ones we made for $\Xi_p$ above show that for the elements of $S_p$ corresponding to $\sigma_p(i,j) = 0$, we have $\mathbf{E}\left(S_p(i,j)^{2k}\right) = \mathrm{O}(n^{-k})$. Therefore,

$$P(\max_{\mathcal{D}_p}|S_p(i,j)| > Cn^{-\alpha}) \leq \sum_{(i,j)\in\mathcal{D}_p} \mathbf{E}\left(S_p(i,j)^{2k}\right)\frac{n^{2k\alpha}}{C^{2k}} = \mathrm{O}(p^2 n^{2k\alpha}n^{-k}) \ .$$

Since we assumed that $p \asymp n$, we see that if $k(1-2\alpha) - 2 \geq 1 + \epsilon$,

$$P(\max_{\mathcal{D}_p}|S_p(i,j)| > Cn^{-\alpha} \text{ i.o}) = 0 \ ,$$

by the first Borel-Cantelli lemma. In other words, if we call $(T_\alpha(S_p))^-$ the thresholded version of the part of $S_p$ that corresponds to indices in $\mathcal{D}_p$, we have that $P((T_\alpha(S_p))^- \neq 0 \text{ i.o}) = 0$.

We now turn our attention to $\mathcal{D}_p^c$, i.e the set of indices for which $\sigma_p(i,j) \neq 0$. Recall that we assumed that these $\sigma_p(i,j)$ satisfied $|\sigma_p(i,j)| \geq Cn^{-\alpha_0}$ and $\alpha_0 < \alpha$. Now note for $(i,j)$ in $\mathcal{D}_p^c$, and $\sigma_p(i,j) \geq 0$, we have $\{|S_p(i,j)| < Cn^{-\alpha}\} \subseteq \{0 \leq \sigma_p(i,j) - Cn^{-\alpha} \leq \sigma_p(i,j) - S_p(i,j)\}$. So, in this case, by using the moment computations made above, and using $C$ to denote a generic constant, we have,

$$P(|S_p(i,j)| < Cn^{-\alpha}) \leq \frac{\mathbf{E}\left(\sigma_p(i,j) - S_p(i,j)\right)^{2k}}{(\sigma_p(i,j) - Cn^{-\alpha})^{2k}} = \mathrm{O}(n^{-k}n^{2k\alpha_0})$$

Similarly, when $\sigma_p(i,j) < 0$, we have

$$P(|S_p(i,j)| < Cn^{-\alpha}) \leq \frac{\mathbf{E}\left(\sigma_p(i,j) - S_p(i,j)\right)^{2k}}{(|\sigma_p(i,j)| - Cn^{-\alpha})^{2k}} = \mathrm{O}(n^{-k}n^{2k\alpha_0})$$

Now note that because $\Sigma_p$ is $\beta$-sparse, there are at most $\mathrm{O}(p^{1+\beta})$ non zero coefficients in $\Sigma_p$: indeed $\phi_p(2)$ counts the number of non-zero coefficients in $\Sigma_p$. From this we conclude that

$$P(\exists (i_0,j_0) \in \mathcal{D}_p^c : |S_p(i_0,j_0)| < Cn^{-\alpha}) \leq \mathrm{O}(n^{k(2\alpha_0-1)}p^{1+\beta}) \ .$$

So if

$$k \geq \frac{2+\epsilon+\beta}{1-2\alpha_0} = \frac{2+\epsilon+\beta}{2\delta_0}$$

then, almost surely, no $S_p(i,j)$ will be wrongly thresholded, if $(i,j) \in \mathcal{D}_p^c$. Combining this result with the one on the indices in $\mathcal{D}_p$, we have

$$P(B_p \text{ i.o}) = 0 \ ,$$

and we have the result announced in the theorem. □

It is however more common practice to use as our estimator of the covariance matrix the sample covariance matrix that differs slightly from the matrix $S_p$ used above, which is the maximum likelihood estimator in the (mean 0) Gaussian case. We now show that for the usual estimator the same strategy works.

**Theorem 2** (**Sample covariance matrix**). *Suppose the assumptions of Theorem 1 are satisfied, but allow now $X_i$ to have a non-zero mean $\mu$. Call*

$$S_p = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})' \ .$$

*Then the result of Theorem 1 holds; namely the thresholded matrix $T(S_p) - \Sigma_p$ converges a.s in spectral norm to 0.*

The previous theorem basically means that if the covariance matrix $\Sigma_p$ is sparse enough, and if the data come from a distribution with enough moments, then thresholding the sample covariance matrix by keeping only terms that are a bit larger than $1/\sqrt{n}$ is a good idea and will lead to an estimator that is consistent in operator norm. This is in stark contrast to simply using the sample covariance matrix, when in the asymptotics considered here, we would not have consistency even at the level of the vector of eigenvalues: in the case of $\Sigma_p = \mathrm{Id}$, this is a consequence of the results of Marčenko and Pastur (1967) or Geman (1980) and we refer to El Karoui (2006) for a thorough discussion.

*Proof.* The proof proceeds as the one of Theorem 1. Since $S_p$ is still unbiased for $\Sigma_p$, the only thing we have to show here is that the $2k-th$ central moments of $S_p(i,j)$ decay in the same fashion as they did in Theorem 1. First let us note that

$$S_p(i,j) = \frac{1}{n-1}\sum_{l=1}^{n}(X_{l,i} - \mu_i)(X_{l,j} - \mu_j) - \frac{n}{n-1}(\bar{X}_i - \mu_i)(\bar{X}_j - \mu_j) \ ,$$

so

$$S_p(i,j) - \sigma_p(i,j) = \frac{1}{n-1}\sum_{l=1}^{n}\left((X_{l,i}-\mu_i)(X_{l,j}-\mu_j) - \sigma_p(i,j)\right)$$
$$-\frac{n}{n-1}\left((\bar{X}_i-\mu_i)(\bar{X}_j-\mu_j) - \frac{1}{n}\sigma_p(i,j)\right) .$$

Now, since $(a+b)^{2k} \leq 2^{2k}(a^{2k}+b^{2k})$, we see that we will have the result we need if we can bound each term in the right-hand side of the previous equation. The technique we used above immediately shows that

$$\mathbf{E}\left(\frac{1}{n-1}\sum_{l=1}^{n}\left[(X_{l,i}-\mu_i)(X_{l,j}-\mu_j) - \sigma_p(i,j)\right]\right)^{2k} = \mathrm{O}\left(\frac{1}{n^k}\right) ,$$

assuming for a moment that all the needed moments exist. For the other part of the equation, the same argument shows that the only thing we need to control is $\mathbf{E}\left((\bar{X}_i-\mu_i)(\bar{X}_j-\mu_j)\right)^{2k}$, since the assumptions we made about the moments of $X_i$ guarantee that $\sigma_p(i,j)$ is bounded in $p$. Using the Cauchy-Schwarz inequality, it is clear that all we need to do is control $\mathbf{E}\left(\bar{X}_i-\mu_i\right)^{4k}$, for all $i$. But $\bar{X}_i - \mu_i$ is a sum of independent mean 0 random variables and the computations we made in the proof of Theorem 1 show that

$$\mathbf{E}\left(\bar{X}_i-\mu_i\right)^{4k} = \mathrm{O}\left(\frac{1}{n^{2k}}\right) .$$

Therefore,

$$\mathbf{E}\left((\bar{X}_i-\mu_i)(\bar{X}_j-\mu_j)\right)^{2k} = \mathrm{O}\left(\frac{1}{n^{2k}}\right) .$$

So we conclude that

$$\mathbf{E}\left(S_p(i,j)-\sigma_p(i,j)\right)^{2k} = \mathrm{O}\left(\frac{1}{n^k}\right) ,$$

just as in the case of the Gaussian MLE estimator. This is all we need to complete the proof of Theorem 2, since the last steps follow exactly from the proof of Theorem 1. The assumptions made guarantee that all the moments used above exist and are finite. □

We note that the distribution of the entries of $X$ can change with $n$ and $p$ as long as the moment conditions are satisfied and the bounds on the moments are uniform in $n$ and $p$. We now turn to the question of estimating correlation matrices.

**Theorem 3** (**Correlation matrices**). *Under the assumptions of Theorem 1, but requiring the boundedness of the $8k(\eta)$-th moments of the $X_{i,j}$'s, if $\Sigma_p$ is now the correlation matrix of the vector $X_i$, and if $S_p$ is now the sample correlation matrix, we have as before*

$$|||T_\alpha(S_p) - \Sigma_p|||_2 \to 0 \ \ a.s .$$

*Proof.* Because of invariance of the problem by centering and scaling, we can assume that the row vector $X_i$ has mean 0, and that the diagonal of its covariance matrix $\Sigma_p$ is full of 1. Then we have $\rho(i,j) = \sigma_p(i,j)$. From the proof of Theorem 1, if it clear that if we can show that $\mathbf{E}\left(S_p(i,j)-\rho(i,j)\right)^{2k} = \mathrm{O}(n^{-k})$ for all $(i,j)$, the same technique as above will lead to the theorem. To show that this is indeed the case, we first make the following elementary remark, which prepares the study of $\widehat{\rho}(i,j) - \rho(i,j)$. Suppose that $F_n$ and $G_n$ are random variables, with $\mathbf{E}\left(F_n-\rho\right)^{2k} = \mathrm{O}(n^{-k})$ for some $\rho \in [-1,1]$, $\mathbf{E}\left(G_n-1\right)^{2k} = \mathrm{O}(n^{-k})$ and

further $|F_n/G_n| \leq 1$. Call $\Omega_n(\epsilon)$ the event $\{\omega : |G_n - 1| < \epsilon\}$. We have

$$\begin{aligned}
\mathbf{E}\left(\frac{F_n}{G_n} - \rho\right)^{2k} &= \mathbf{E}\left(\left[\frac{F_n}{G_n} - \rho\right]^{2k}\left[1_{\Omega_n(\epsilon)} + 1_{\Omega_n^c(\epsilon)}\right]\right) \\
&\leq \mathbf{E}\left(\left[\frac{F_n}{G_n} - \rho\right]^{2k} 1_{\Omega_n(\epsilon)}\right) + \mathbf{E}\left(\left[\frac{F_n}{G_n} - \rho\right]^{2k} 1_{\Omega_n^c(\epsilon)}\right) \\
&\leq \mathbf{E}\left(\left[\frac{F_n - \rho}{G_n} - \rho(1 - \frac{1}{G_n})\right]^{2k} 1_{\Omega_n(\epsilon)}\right) + 2^{2k}\mathbf{E}\left(1_{\Omega_n^c(\epsilon)}\right) \\
&\leq 2^{2k}\mathbf{E}\left(\left[\frac{F_n - \rho}{G_n}\right]^{2k} 1_{\Omega_n(\epsilon)} + \rho^{2k}\left[1 - \frac{1}{G_n}\right]^{2k} 1_{\Omega_n(\epsilon)}\right) + 2^{2k}\mathbf{E}\left(1_{\Omega_n^c(\epsilon)}\right) \\
&\leq \frac{2^{2k}}{(1-\epsilon)^{2k}}\left\{\mathbf{E}\left((F_n - \rho)^{2k} 1_{\Omega_n(\epsilon)}\right) + \rho^{2k}\mathbf{E}\left((G_n - 1)^{2k} 1_{\Omega_n(\epsilon)}\right)\right\} + 2^{2k}\mathbf{E}\left(1_{\Omega_n^c(\epsilon)}\right)
\end{aligned}$$

By Chebyshev's inequality, and our assumptions it is clear that

$$\mathbf{E}\left(\frac{F_n}{G_n} - \rho\right)^{2k} = \mathrm{O}\left(\frac{1}{n^k}\right) .$$

Now we claim that this remark applies in the case of the correlation matrix. We have

$$\widehat{\rho}(i,j) = \frac{F_n(i,j)}{G_n(i,j)} ,$$

where

$$F_n(i,j) = \frac{1}{n-1}\sum_{l=1}^{n}(X_{l,i} - \bar{X}_i)(X_{l,j} - \bar{X}_j) ,$$

and $G_n(i,j) = \sqrt{F_n(i,i)F_n(j,j)}$.

From the moment computations made in the proof of Theorem 2, we see that we have $\mathbf{E}\left(F_n(i,j) - \rho(i,j)\right)^{2k} = \mathrm{O}(n^{-k})$, for all $(i,j)$. Let us denote by $Y_n(i) = F_n(i,i)$; this result implies that

$$\mathbf{E}\left(\sqrt{Y_n(i)} - \sqrt{\rho(i,i)}\right)^{2k} \leq \mathbf{E}\left(Y_n(i) - \rho(i,i)\right)^{2k} / \rho(i,i)^k = \mathbf{E}\left(Y_n(i) - 1\right)^{2k} = \mathrm{O}(n^{-k}) .$$

If we denote by $\alpha_n = \sqrt{Y_n(i)}$ and $\beta_n = \sqrt{Y_n(j)}$, we have, since $\alpha_n\beta_n - 1 = (\alpha_n - 1)\beta_n + \beta_n - 1$,

$$\begin{aligned}
\mathbf{E}\left(\alpha_n\beta_n - 1\right)^{2k} &\leq 2^{2k}\left[\mathbf{E}\left(\beta_n - 1\right)^{2k} + \mathbf{E}\left(\beta_n^{2k}(\alpha_n - 1)^{2k}\right)\right] \\
&\leq 2^{2k}\left[\mathbf{E}\left(\beta_n - 1\right)^{2k} + \sqrt{\mathbf{E}\left(\beta_n^{4k}\right)}\sqrt{\mathbf{E}\left(\alpha_n - 1\right)^{4k}}\right]
\end{aligned}$$

We have already seen that $\mathbf{E}\left(\beta_n - 1\right)^{2k} = \mathrm{O}(n^{-k})$, and since we are assuming the existence of a $8k$-th moment for the $X_{l,j}$, we also have $\sqrt{\mathbf{E}\left(\alpha_n - 1\right)^{4k}} = \mathrm{O}(n^{-k})$. To conclude that $\mathbf{E}\left(\alpha_n\beta_n - 1\right)^{2k} = \mathrm{O}(n^{-k})$, we just need to show that $\mathbf{E}\left(\beta_n\right)^{4k}$ is bounded. But $\beta_n^{4k} = (\beta_n - 1 + 1)^{4k} \leq 2^{4k}((\beta_n - 1)^{4k} + 1)$, from which we conclude that $\mathbf{E}\left(\beta_n\right)^{4k}$ is bounded, since $(\beta_n - 1)^{4k} = \mathrm{O}(n^{-2k})$. We now see that $G_n(i,j) = \sqrt{Y_n(i)Y_n(j)}$ satisfies with $F_n$ the conditions needed to conclude that

$$\mathbf{E}\left(\frac{F_n(i,j)}{G_n(i,j)} - \rho(i,j)\right)^{2k} = \mathbf{E}\left(\widehat{\rho}(i,j) - \rho(i,j)\right)^{2k} = \mathrm{O}(n^{-k}) .$$

□

### 3.2 Approximation of non-sparse matrices by sparse matrices

It is natural to ask whether a thresholding approach can also lead to good results when dealing with matrices which are not sparse per se, but have many coefficients close to zero. In other words, we would like to know when we can approximate non-sparse matrices by sparse matrices obtained by thresholding a sample covariance or correlation matrix. We now present two propositions that relax the sparsity assumptions and still lead to spectral norm convergence. The most general one basically says that if the population covariance matrix can be approximated by a sparse matrix and does not have too many coefficients close to the threshold level $1/\sqrt{n}$, then estimating the (not necessarily sparse) population covariance by thresholding the sample covariance matrix will lead to good results.

Here is our first result in this direction:

**Proposition 1.** *Making the same assumptions as in the theorems above, we now assume that there exists $T_{\alpha_1}(\Sigma_p) = \widetilde{\Sigma}_p$, a version of $\Sigma_p$ thresholded at $n^{-\alpha_1}$, that is $\beta$-sparse. Further we assume that $|||\widetilde{\Sigma}_p - \Sigma_p|||_2 \to 0$. We call $I_{\alpha_1,\alpha_0}$ the set of indices of those $\sigma(i,j)$ for which $Cn^{-\alpha_1} < |\sigma(i,j)| < Cn^{-\alpha_0}$, with $\alpha_0 < \alpha_1 < 1/2 - \delta_0$. Now choose $\alpha \in (\alpha_0, \alpha_1)$. If the adjacency matrix corresponding to $I_{\alpha_1,\alpha_0}$ is $\gamma$-sparse, for some $\gamma \leq \alpha_0 - \zeta_0$, and if the random variables $X_{i,j}$ have moments of order $4k$ ($8k$ in the correlation case), with $k$ satisfying the assumptions put forth in Theorem 1 as well as $k \geq (2+\epsilon-\gamma)/(2(\alpha_0-\gamma))$, for some $\epsilon > 0$, then the conclusions of all the theorems above apply:*

$$|||T_\alpha(S_p) - \Sigma_p|||_2 \to 0 \ a.s\ .$$

While this proposition might appear full of hard to check assumptions, we believe it is useful and not so hard to use when checking whether thresholding is a reasonable idea for a particular estimation problem. We give an example after stating Proposition 2 below. Finally, we note that under the assumptions stated, both $T_{\alpha_0}(\Sigma_p)$ and $T_{\alpha_1}(\Sigma_p)$ are good approximations of $\Sigma_p$ in operator norm.

*Proof.* From the previous proofs, we see that we can divide

$$T_\alpha(S_p) = M_0 + M_1 + M_2\ ,$$

into three parts, $M_0$ corresponding to the indices $(i,j)$ for which $\sigma(i,j)$ is larger (in absolute value) than $Cn^{-\alpha_0}$, $M_1$ corresponding to indices in $I_{\alpha_0,\alpha_1}$, and $M_2$ to those indices for which $\sigma(i,j) < n^{-\alpha_1}$. Similarly, we can write with the same partition of indices,

$$\Sigma_p = T_{\alpha_0}(\Sigma_p) + [T_{\alpha_1}(\Sigma_p) - T_{\alpha_0}(\Sigma_p)] + [\Sigma_p - T_{\alpha_1}(\Sigma_p)] = \Sigma_0 + \Sigma_1 + \Sigma_2\ .$$

With the same notations for the subparts of $\Sigma$, we have from the computations we made in the proofs of the previous theorems that $|||M_0 - \Sigma_0|||_2 \to 0$ a.s (by the oracle part of the proofs), and $|||M_2|||_2 \to 0$ a.s., since the $\widehat{\sigma}_p(i,j)$ corresponding to $|\sigma(i,j)| < n^{-\alpha_1}$ will all be (a.s) thresholded to 0 if the thresholding level is $n^{-\alpha}$, $\alpha < \alpha_1$.

Note that $\Sigma_0 + \Sigma_1 = \widetilde{\Sigma}_p$, so $|||\Sigma_2|||_2 \to 0$. To reach the conclusions of the proposition, we need to show that we control $M_1 - \Sigma_1$ in operator norm.

Recall that our assumption is that $\Sigma_1$ is $\gamma$-sparse. We call

$$\Sigma_1 = T_\alpha(\Sigma_1) + R_\alpha(\Sigma_1)\ ,$$

where $T_\alpha(\Sigma_1)$ the version of $\Sigma_1$ thresholded at $n^{-\alpha}$. It is of course also $\gamma$-sparse and so is $R_\alpha(\Sigma_1)$. This implies that

$$|||R_\alpha(\Sigma_1)|||_2^{2k} \leq \text{trace}\left((R_\alpha(\Sigma_1))^{2k}\right) \leq f(k)p^{\gamma(2k-1)}pn^{-2k\alpha}\ ,$$

which goes to 0 if $\gamma \leq \alpha - \epsilon$: we can find $k_0$ an integer, such that the left-hand side goes to 0 as $n$ and $p$ go to infinity.

Using the oracle proof of Theorem 1, we see that if we make no error in thresholding for the indices in $I_{\alpha_0,\alpha_1}$, then $|||\text{oracle}_\alpha(M_1) - T_\alpha(\Sigma_1)|||_2$ tends to 0 a.s. Therefore, all we need to do is check that we control the operator norm of the matrix of possible errors, i.e the difference $M_1 - \text{oracle}_\alpha(M_1)$. Let us call $\Upsilon_1$ this

matrix of potential errors. There are two types of possible errors: either a coefficient is thresholded when it should not have been. Or it is not thresholded when it should have been thresholded. So

$$\Upsilon_1(i,j) = \begin{cases} 0 & \text{if} \quad \text{correctly thresholded } \widehat{\sigma}(i,j) \\ \widehat{\sigma}(i,j) & \text{if} \quad |\sigma_{i,j}| \leq n^{-\alpha} \text{ but did not threshold in } M_1 \\ -\widehat{\sigma}(i,j) & \text{if} \quad |\sigma_{i,j}| > n^{-\alpha} \text{ but did threshold in } M_1 \end{cases}$$

In any case, we conclude that $|\Upsilon_1(i,j)| \leq |\widehat{\sigma}(i,j)| \leq |\widehat{\sigma}(i,j) - \sigma(i,j)| + |\sigma(i,j)|$. Note that all the indices where $\Delta_2$ has potentially non-zero entries are in $I_{\alpha_0,\alpha_1}$, so the corresponding adjacency matrix is $\gamma$-sparse. Since

$$\mathbf{E}\left(\Upsilon_1(i,j)\right)^{2k} \leq 2^{2k}\left(\sigma(i,j)^{2k} + \mathbf{E}\left(\widehat{\sigma}(i,j) - \sigma(i,j)\right)^{2k}\right) = \mathrm{O}(n^{-2\alpha_0 k} + n^{-k}) \,,$$

we conclude as before that the expected weight of a path of length $2k$ is $\mathrm{O}(n^{-2\alpha_0 k})$. Using the assumption of $\gamma$-sparsity of the matrix $\Sigma_1$, we conclude that the total contribution for the expected errors is $\mathrm{O}(p^{\gamma(2k-1)+1} n^{-2\alpha_0 k})$. Therefore, if $k > (2+\epsilon-\gamma)/(2(\alpha_0-\gamma))$, we have a.s convergence. □

The following simple proposition is a clear case of applicability of the ideas of the previous one.

**Fact 2.** *Let $\epsilon > 0$ and suppose that $T_{1+\epsilon}(\Sigma_p)$ is $\beta$-sparse and its non-zero entries are larger than $n^{-\alpha_0}$. Then under the same assumptions as Theorems 1, 2 and 3, we have, for $\alpha_0 < \alpha < 1/2 - delta$,*

$$|||T_\alpha(S_p) - \Sigma_p|||_2 \to 0 \ a.s \,.$$

*Proof of Fact 2.* Take $\alpha_1 = \alpha_0 + \delta$, where $\delta$ is small. In particular, of course, $\alpha_1 < 1/2 < 1+\epsilon$. Here $I_{\alpha_1,\alpha_0}$ is empty so the corresponding matrix is 0-sparse. In particular, that means that in the notation of the proof of Proposition 1, $M_1 = 0$ and similarly for $\Sigma_1$. Since those are in general the only parts that cause problems, So the results on $M_0 - \Sigma_0$ and $M_2$ apply directly and the only thing we have to check is that $|||\Sigma_2|||_2 \to 0$. Note that $\Sigma_2$ contains only entries that of order $n^{-(1+\epsilon)}$ or smaller. Using a Frobenius norm bound, we therefore have

$$|||\Sigma_2|||_2^2 \leq p^2 n^{-(2+2\epsilon)} \ \to 0 \,,$$

so the result is established. □

**Example: a simple (permuted) Toeplitz matrix** We consider a matrix that is often used as an example for estimation: the (Toeplitz) covariance matrix $\Sigma_p$, with $\Sigma(i,j) = \rho^{|i-j|}$, $|\rho| < 1$. Of course, we can also consider the same matrix where the variables have been randomly permuted and hence the Toeplitz structure destroyed. However, on any given line, the entries are still a (possibly random) permutation of the $\rho^{|i-j|}$. We apply Proposition 1. To do so, we just need to count how many coefficients on each row are between $n^{-\alpha_1}$ and $n^{-\alpha_0}$, for $\alpha_0$ and $\alpha_1$ to be chosen later. Note that $|\rho|^k \leq n^{-\alpha_1}$ is equivalent to $k \geq \log(n)\alpha_1/\log(1/|\rho|)$. So $T_{\alpha_1}(\Sigma_p)$ is asymptotically 0 sparse, as it contains only $\mathrm{O}(\log(n))$ non-zero terms on each row. Similarly, the adjacency matrix corresponding to $I(\alpha_0,\alpha_1)$ is also asymptotically 0 sparse as there are at most $\mathrm{O}(\log(n))$ terms on each of its row. Finally, we need to check that the thresholded $\Sigma_p$ is a good approximation of $\Sigma_p$. Recall that for a symmetric matrix $M$, $|||M|||_2 \leq max_i(\sum_j |m_{i,j}|)$. (See for instance Bickel and Levina (2007) or Stewart and Sun (1990), p. 70.) Now, $\sum_{k\geq k_0} \rho^k = \rho^{k_0}/(1-\rho)$, so $|||\Sigma_p - T_{\alpha_1}(\Sigma_p)|||_2 \leq n^{-\alpha_1}/(1-|\rho|)$, which tends to 0 as $n$ goes to infinity. So we conclude that Proposition 1 applies and thresholding the sample covariance (resp. correlation) matrix corresponding to this population covariance will yield an operator norm consistent estimator, a.s., provided the moment conditions are satisfied. In this situation, the moment conditions translate simply into $k \geq 2+\epsilon$ for some $\epsilon$, because $\alpha_0$ can be chosen arbitrarily close to $1/2$ and $\gamma$ arbitrarily close to 0.

Finally, we have the following corollaries that apply to all the theorems and proposition above.

**Corollary 1** (Infinitely many moments)**.** *Suppose that the entries of $X$ have infinitely moments. Then all the above results hold with only the sparsity conditions having to be checked.*

**Corollary 2** (Asymptotic $\beta$-sparsity)**.** *Suppose that the sequence $\Sigma_p$ is asymptotically $\beta$-sparse. Then all the above results apply, with the modification that $\beta$ be replaced by $\widetilde{\beta}_\epsilon = \beta + \epsilon$ for $\epsilon > 0$ but arbitrarily small. In particular, moment conditions need only to be satisfied and checked with $\widetilde{\beta}_\epsilon$. In the situation where one has infinitely many moments, one therefore only needs to check that the sparsity conditions are satisfied by a $\widetilde{\beta}_\epsilon$.*

### 3.3 About $1/2$-sparse matrices

The previous computations are very clearly limited to the case where $\beta < 1/2$. A natural question is therefore to ask if this limitation is inherent to the problem, or if it is a consequence of the bounds we use in the mathematical analysis. We now want to highlight the problems that occur in the case $\beta = 1/2$ and show that our result is "sharp": at the level of generality at which we are working, (at least some) $1/2$-sparse matrices are not estimable in operator norm. To show this, we will produce a $1/2$-sparse matrix that cannot be consistently estimated in operator norm even at the oracle level.

To do so, we consider estimating a matrix $A$ of the following form:

$$\Sigma_p = \begin{pmatrix} 1 & \alpha_2 & \alpha_3 & \dots & \alpha_p \\ \alpha_2 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \alpha_{p-1} & 0 & 0 & 1 & 0 \\ \alpha_p & 0 & 0 & \dots & 1 \end{pmatrix} .$$

To simplify the problem, we assume that the data is multivariate Gaussian, with mean 0, and that we know that the diagonal is composed only of 1's. We estimate $\Sigma_p$ using the sample covariance matrix, putting to 1 the main diagonal, and using the oracle information to put to 0 all other terms except the first row and columns. We call $\widehat{\Sigma}_p$ the corresponding estimator. Note that

$$\Sigma_p - \widehat{\Sigma}_p = \begin{pmatrix} 0 & \alpha_2 - \widehat{\alpha}_2 & \alpha_3 - \widehat{\alpha}_3 & \dots & \alpha_p - \widehat{\alpha}_p \\ \alpha_2 - \widehat{\alpha}_2 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \alpha_{p-1} - \widehat{\alpha}_{p-1} & 0 & 0 & 0 & 0 \\ \alpha_p - \widehat{\alpha}_p & 0 & 0 & \dots & 0 \end{pmatrix} .$$

Using the Schur complement formula for determinants (see for instance Horn and Johnson (1990), p.22), we see that the characteristic polynomial of this matrix is

$$p(\lambda) = \lambda^{p-2}(\lambda^2 - \sum_{i=2}^{p}(\alpha_i - \widehat{\alpha}_i)^2) \text{ , and therefore}$$

$$|||\widehat{\Sigma}_p - \Sigma_p|||_2 = \sqrt{\sum_{i=2}^{p}(\alpha_i - \widehat{\alpha}_i)^2} .$$

Note that the computation holds for the corresponding adjacency matrix, giving that $\phi_p(2k) = \text{trace}\left(A_p^{2k}\right) = 2(p-1)^k$. So this matrix is $1/2$-sparse.

Now, since we assume the data is Gaussian, it is clear that $\lambda_1 = |||\widehat{\Sigma}_p - \Sigma_p|||_2$ has infinitely many moments, using for instance Frobenius norm as a bound on $\lambda_1$. Also, $\mathbf{E}\left(\lambda_1^2\right) = \sum_{i=2}^{p}\mathbf{E}\left((\alpha_i - \widehat{\alpha}_i)^2\right)$. The covariance of elements of the sample covariance matrix is well-known in the Wishart case; see for instance Anderson (2003), Theorem 3.4.4 p. 87. In our context, we see that $\mathbf{E}\left((\alpha_i - \widehat{\alpha}_i)^2\right) = (1 + \alpha_i^2)/(n-1) = \nu_i/(n-1)$. In particular,

$$\mathbf{E}\left(\lambda_1^2\right) = \frac{p - 1 + \sum_{i=2}^{p}\alpha_i^2}{n-1} \geq \frac{p-1}{n-1} \to l > 0 .$$

We now turn to showing that $\lambda_1^2$ actually converges in probability to this limit.

A standard result in Gaussian multivariate analysis (see Anderson (2003), Theorem 3.3.2) states that we can write $\widehat{\alpha}_i - \alpha_i = (\sum_{k=1}^{n-1} Z_k)/(n-1)$, where the $Z_k$'s are i.i.d and mean 0. Hence we get that

$$\mathbf{E}\left(((\widehat{\alpha}_i - \alpha_i)^2 - \nu_i/(n-1))^2\right) = \frac{1}{(n-1)^4}\mathbf{E}\left(\sum_{k_1,k_2,k_3,k_4} Z_{k_1}Z_{k_2}Z_{k_3}Z_{k_4}\right) .$$

In the above sum, the terms that contain an index repeated only once contribute zero to the expectation. After elementary computations, we see that to first order this expectation is $\mathrm{O}(2\nu_i^2/n^2)$. Using the same ideas (see Appendix), we get that, for $i \neq j$,

$$\mathbf{E}\left(((\widehat{\alpha}_i - \alpha_i)^2 - \nu_i/(n-1))((\widehat{\alpha}_j - \alpha_j)^2 - \nu_j/(n-1))\right) = \mathrm{O}(\frac{2}{n^2}\alpha_j^2\alpha_i^2 \vee \frac{1}{n^3}) \ .$$

Hence we have that

$$\mathrm{var}\left(\lambda_1^2\right) = \mathrm{O}\left(\sum_{i=2}^{p} \frac{2\nu_i^2}{n^2} + \sum_{i\neq j} \frac{2\alpha_j^2\alpha_i^2}{n^2} \vee \frac{1}{n^3}\right) = \mathrm{O}\left(\sum_{i=2}^{p} \frac{2\nu_i^2}{n^2} + \frac{2}{n^2}(\sum_{i=2}^{p}\alpha_i^2)^2 \vee \frac{1}{n}\right)$$

Therefore, if for instance, $\alpha_i = \frac{1}{\sqrt{p}}$, $\mathrm{var}\left(\lambda_1^2\right) = \mathrm{O}(\frac{p}{n^2} + \frac{1}{n}) \to 0$ and

$$\lambda_1^2 - \frac{p - 1 + \sum_{i\geq 2}\alpha_i^2}{n} \to 0 \text{ in probability} \ ,$$

and therefore

$$\lambda_1^2 \geq \frac{p-1}{n} \text{ in probability} \ .$$

Note that the same result would apply if $\alpha_i = p^{-1/2+\epsilon}$, with $\epsilon > 0$, a situation where thresholding would work for $\beta$-sparse graphs, with $\beta \leq 1/2 - \gamma$. (The case of $\alpha_i = \frac{1}{\sqrt{p}}$ is not covered in our theorems whereas the case $\alpha_i = p^{-1/2+\epsilon}$ is.)

Note also that if we had tried to estimate $\Sigma_p$ using oracle information about the location of the non-zero coefficient but nothing about the fact the diagonal was equal to 1, we would have encountered the same problem. As a matter of fact, if we call $M_p$ the diagonal matrix with entries $\widehat{\sigma}(i,i)$, we have from previous results in the paper (our moment computations and the 0-sparsity of this matrix) that $|||M_p - \mathrm{Id}_p|||_2 \to 0$ a.s. Note that because $\widehat{\Sigma}_p$ had 1's on its diagonal,

$$\widehat{\Sigma}_p + M_p - \mathrm{Id}_p = \begin{pmatrix} \widehat{\sigma}(1,1) & \widehat{\alpha}_2 & \widehat{\alpha}_3 & \dots & \widehat{\alpha}_p \\ \widehat{\alpha_2} & \widehat{\sigma}(2,2) & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \widehat{\alpha}_{p-1} & 0 & 0 & \widehat{\sigma}(p-1,p-1) & 0 \\ \widehat{\alpha}_p & 0 & 0 & \dots & \widehat{\sigma}(p,p) \end{pmatrix} \ .$$

So for the oracle estimator that uses only information about the location of the non-zero coefficients, we have

$$|||\widehat{\Sigma}_p + (M_p - \mathrm{Id}_p) - \Sigma_p|||_2 \geq |||\widehat{\Sigma}_p - \Sigma_p|||_2 - |||M_p - \mathrm{Id}_p|||_2 > \frac{p-1}{2n} \text{ a.s} \ .$$

This example shows that even using oracle information for estimation of the $\Sigma_p$ pointed out above does not lead to an operator norm consistent estimator, in the presence of this simple 1/2-sparse graph. This suggests that for these graphs, simple thresholding might not be a good method. It also suggests that the conditions of our theorems have more to do with the method we propose than with unrefined mathematical details in its analysis.

### 3.3.1 Complement on this example

In what follows we investigate in more details the case where $\alpha_i = 1/\sqrt{p}$. One might ask whether, despite the fact that $|||\Sigma_p - \widehat{\Sigma}_p|||_2$ does not go to zero, $\widehat{\Sigma}_p$ does not have some good characteristics as an estimator of $\Sigma_p$ anyway. In what follows, we show that both for the eigenvalues and eigenvectors, this is not the case.

The previous computations essentially show that

$$\mathbf{E}\left(\lambda_1^2(\widehat{\Sigma}_p - \mathrm{Id}_p)\right) = \sum_{i\geq 2}\alpha_i^2 + \frac{1+\alpha_i^2}{n-1} = (\lambda_1(\Sigma_p - \mathrm{Id}))^2 + \frac{p-1}{n-1} + \frac{p-1}{p(n-1)} \ ,$$

so at the level of eigenvalues, the answer is negative. Note that the eigenvectors of $\Sigma_p - \mathrm{Id}_p$ and therefore of $\Sigma_p$ are known. The ones corresponding to the non-zero eigenvalues are, calling $\lambda_+ = \sqrt{\sum_{i\geq 2} \alpha_i^2}$

$$u_+ = \frac{1}{\sqrt{2}\lambda_+} \begin{pmatrix} \lambda_+ \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} \quad \text{and } u_- = \frac{1}{\sqrt{2}\lambda_+} \begin{pmatrix} \lambda_+ \\ -\alpha_2 \\ \vdots \\ -\alpha_p \end{pmatrix} .$$

We call $\widehat{u}_+$ the eigenvector corresponding to the positive eigenvalue of $\widehat{\Sigma}_p - \mathrm{Id}_p$. When $\alpha_i = 1/\sqrt{p}$, $\mathrm{cov}\,(\widehat{\alpha}_\mathrm{i}, \widehat{\alpha}_\mathrm{j}) = (1_{i=j} + 1/p)/(n-1)$ and $\lambda_+ = \sqrt{(p-1)/p}$. Using the expression above for the eigenvectors, we have

$$2\lambda_+ \widehat{\lambda}_+ \langle u_+, \widehat{u}_+ \rangle = \lambda_+ \widehat{\lambda}_+ + \frac{1}{\sqrt{p}} \sum_{i\geq 2} \widehat{\alpha}_i \, .$$

Now $\mathrm{var}\left(\sum_{i\geq 2} \widehat{\alpha}_i\right) = (p-1)(1+1/p)/(n-1) + (p-1)(p-2)/(p(n-1))$, and $\mathbf{E}\left(\sum_{i\geq 2} \widehat{\alpha}_i\right) = (p-1)/\sqrt{p}$, from which we conclude that

$$\frac{1}{\sqrt{p}} \left( \sum_{i\geq 2} \widehat{\alpha}_i \right) - \left(1 - \frac{1}{p}\right) \to 0 \quad \text{in probability.}$$

Since when all $\alpha_i = 1/\sqrt{p}$,

$$\mathrm{var}\left( \sum_{i\geq 2} \widehat{\alpha}_i^2 \right) \leq 2 \left( \mathrm{var}\left( \sum_{i\geq 2} (\widehat{\alpha}_i - \alpha_i)^2 \right) + \frac{4}{p} \mathrm{var}\left( \sum_{i\geq 2} \widehat{\alpha}_i \right) \right) \, ,$$

the above computations show that $\widehat{\lambda}_+ \to \sqrt{1+p/n}$ in probability and therefore, using Slutsky's lemma, we get that

$$\langle u_+, \widehat{u}_+ \rangle \to \frac{1}{2} \left( 1 + \frac{1}{\sqrt{1+p/n}} \right) \text{ in probability .}$$

So when $p/n$ has a non-zero limit, the angle between these two vectors has a finite non-zero limit (in probability), showing that the eigenvectors are not consistently estimated.

### 3.4 Discussion

In the following, we call $\widehat{\Sigma}_p$ our (final) estimator of $\Sigma_p$, which is obtained from the standard estimator $S_p$. As above, we denote by $\Delta_p = \widehat{\Sigma}_p - \Sigma_p$, $\Xi_p = \mathrm{oracle}(S_p) - \Sigma_p$, where $\mathrm{oracle}(S_p)$ the oracle version of $S_p$, and $D_p = S_p - \Sigma_p$.

#### 3.4.1 Finite dimensional character and sharpening of the bounds

As is clear from the proofs, all the bounds we derive are valid at $n$ and $p$ fixed. Essentially, we get bounds on the probability of deviation of the largest eigenvalue of the matrix $\Delta_p$ from 0. These bounds are polynomial in nature since we used Chebyshev's inequality and worked with moments.

Note that in particular cases, such as when the entries of the data matrix are bounded or satisfy certain tail conditions, these bounds can be sharpened by using (exponential or gaussian) concentration inequalities for the difference $d_{i,j} = \hat{\sigma}(i,j) - \sigma(i,j)$. If the entries of $X$ are bounded in absolute value by a constant $C$, in the setting of Theorem 1, Hoeffding's inequality (see Hoeffding (1963)) would for instance give that

$$P(|d_{i,j}| > t) = P(|\hat{\sigma}(i,j) - \sigma(i,j)| > t) \leq 2\exp(-nt^2/(2C^4)) \, .$$

This is a simple consequence of the fact that $\hat{\sigma}(i,j)$ is a sum of i.i.d random variables and their mean is $\sigma(i,j)$. (Of course, a slight adjustment is needed when dealing with sample covariance matrices, but

it does not change the exponential character of the bounds. We give the argument in the simplest case where $S_p = X^*X/n$, the Gaussian MLE when we know the mean is zero.) Suppose that the non-zero coefficients of $\Sigma_p$ are bounded below, in absolute value by $C_1 n^{-1/2+b}$. If we call $B_p$ the event $B_p = \{$at least one mistake is made by the thresholding procedure$\}$, and if we decide to refine our thresholding to a $(\log(n))^a/\sqrt{n}$ threshold, we see, using a simple union bound, that

$$P(B_p) \leq 2p^2(\exp(-(\log n)^{2a}/(2C^4)) + \exp(-((\log n)^a - C_1 n^b)^2/(2C^4)) \ .$$

Therefore, by adding assumptions to our problem, we are able to get sharper bounds on the probability of making a mistake by thresholding.

We can also get better bounds on the probability that $|||\Xi_p|||_2 > \epsilon$ and $|||\Delta_p|||_2 > \epsilon$. We assume that $\Sigma_p$ is $\beta$-sparse and use the corresponding notations. Of course, the event $|||\Xi_p|||_2 > \epsilon$ is contained in the event $\text{trace}\left(\Xi_p^{2k}\right) > \epsilon^{2k}$, which is contained in the event $\max|w_\gamma(2k)| > \epsilon^{2k}/(f(k)p^{1+\beta(2k-1)})$ which is contained in the event $\max|d_{i,j}| > \epsilon/(f(k)^{1/2k}p^{1/2k+\beta(1-1/2k)})$. Hence, by using Hoeffding's inequality, we get

$$P(|||\Xi_p|||_2 > \epsilon) \leq 2p^2 \exp(-n\epsilon^2 p^{-2\beta} p^{(\beta-1)/k}/(2C^4 f(k)^{1/k})) \ .$$

Finally, using the fact that $\{|||\Delta_p|||_2 > \epsilon\} \subseteq \left(\{|||\Xi_p|||_2 > \epsilon\} \cap B_p^c\right) \cup B_p$, we see that

$$P(|||\Delta_p|||_2 > \epsilon) \leq P(B_p) + P(|||\Xi_p|||_2 > \epsilon) \ ,$$

for which we just derived bounds. Similar types of bounds can be obtained in the context of Theorem 2, when, for instance, Hoeffding's inequality applies.

Though these results are sharper than the ones announced in the theorems above, they are less general. Because one of our concern was maximal distributional generality, we decided to give the theorems in general form with less sharp bounds.

### 3.4.2 Beyond the finite $p/n$ limit

A close look at the proofs of the theorems and the bounds above reveal that the assumption that $p/n$ has a finite limit can be relaxed. As a matter of fact, our bounds on expected values of traces are generically of the form $\mathrm{O}(p^\gamma n^{-\lambda})$, and all we require is that this quantity goes to zero fast enough. If we focus on the oracle version of the theorems we see that the bounds are of the form

$$\mathbf{E}\left(\text{trace}\left(\Xi_p^{2k}\right)\right) = \mathrm{O}(n^{-k}p^{1+\beta(2k-1)}) \ .$$

If $p = \mathrm{O}(n^\nu)$, we see that the exponent in $n$ becomes of the form $k(2\beta\nu - 1) + \nu(1-\beta)$. If this quantity is less than $-(1+\epsilon)$ for some $\epsilon > 0$ and $k = k_0$, then we will have a.s convergence of $\Xi_p$ to zero in operator norm. This condition is satisfied if

$$\nu \leq \frac{k - (1+\epsilon)}{1 + \beta(2k-1)} \ .$$

So in particular, if we are working with random variables with infinitely many moments, the oracle results will hold almost surely for a $\beta$-sparse matrix when

$$p = \mathrm{O}(n^{1/(2\beta)-\eta}) \ , \text{ for some } \eta \text{ arbitrarily small } .$$

As a matter of fact, all we need to do is pick a finite number $k_1$ such that

$$\frac{k_1 - (1+\epsilon)}{1 + \beta(2k_1 - 1)} > 1/(2\beta) - \eta$$

and carry out the analysis for $\mathbf{E}\left(\text{trace}\left(\Xi_p^{2k_1}\right)\right)$. $k_1$ exists (and is finite), because $\frac{k-(1+\epsilon)}{1+\beta(2k-1)} \to 1/(2\beta)$, as $k$ goes to infinity. If there are only $4k_1$ moments the results will hold, too.

On the other hand, the non-oracle results will be satisfied in the context of Theorem 1 as soon as

$$\nu \leq \frac{k(1-2\alpha_0) - (1+\epsilon)}{2} \ ,$$

a constraint much less restrictive than the previous one in general. Finally, we note that Proposition 1 would apply if, assuming the other constraints had been satisfied, we also had

$$\nu \leq \frac{2\alpha_0 k - (1+\epsilon)}{1 + \gamma(2k-1)} \ .$$

## 3.5 Consequences of spectral norm convergence

### 3.5.1 Convergence of eigenvalues

We recall some classical facts from matrix analysis. First, if $A$ and $B$ are two symmetric matrices, and if $\lambda_i$ is their $i$-th eigenvalue, where the eigenvalues are sorted in decreasing order, we have, by Weyl's Theorem (Theorem 4.3.1 in Horn and Johnson (1990))

$$|\lambda_i(A) - \lambda_i(B)| \leq |||A - B|||_2 \, .$$

Because the matrix $S_p$ is symmetric, the thresholded version of it is symmetric, too. Therefore the operator norm convergence we showed implies the following:

**Fact 3.** *When the thresholded estimator $\widehat{\Sigma}_p$ is a spectral norm consistent estimator of the population covariance or correlation matrix $\Sigma_p$, all the eigenvalues of $\widehat{\Sigma}_p$ are consistent estimators of the population eigenvalues.*

### 3.5.2 Convergence of eigenvectors

Perhaps even more interestingly, controlling the spectral norm allows us to get very good control on the angles between the eigenspaces of the population and sample covariance matrix, through the use of the classical $\sin(\theta)$ theorems of Davis and Kahan (Davis and Kahan (1970), Section 2 and Stewart and Sun (1990), Section V.3). For the sake of completeness we quote a version of this important result (Theorem 2 in Davis and Kahan (1970)) and show how to exploit it in our context.

**Theorem 4** ($\sin(\theta)$ Theorem)**.** *Suppose $\Sigma_p$ has the spectral resolution*

$$\begin{pmatrix} X_1' \\ X_2' \end{pmatrix} \Sigma_p (X_1 X_2) = diag(L_1, L_2)$$

*with $(X_1 X_2)$ an orthogonal matrix, $X_1$ being a $p \times k$ matrix. Suppose $Z$ is a $p \times k$ matrix with orthogonal columns, and for any hermitian matrix $M$ of order $k$, call $R = \Sigma_p Z - ZM$. Suppose the eigenvalues of $M$ are contained in an interval $[\alpha, \beta]$ and that for some $\delta > 0$, the eigenvalues of $L_2$ are contained in $\mathbb{R} \backslash [\alpha - \delta, \beta + \delta]$. Then for any unitarily invariant norm,*

$$\|\sin \Theta[\mathcal{R}(X_1), \mathcal{R}(Z)]\| \leq \frac{\|R\|}{\delta} \, ,$$

*where $\Theta[\mathcal{R}(X_1), \mathcal{R}(Z)]$ stands for the canonical angles between the column space of $X_1$ and that of $Z$.*

These angles are closely connected to canonical correlation analysis: their cosines are the canonical correlations for the "data matrices" $X_1$ and $Z$.

We therefore have the following corollaries to Theorems 2 and 3:

**Corollary 3** (Consistency of eigenspaces)**.** *Suppose $\Sigma_p$ has a group of eigenvalues contained in an interval and separated from the other eigenvalues by $\delta > 0$. Call the set of their indices (after say ordering them) $J$. Then the canonical angles between the column space of the corresponding eigenvectors and the column space of the eigenvectors of $\widehat{\Sigma}_p$ (our thresholding estimator) corresponding to the eigenvalues of $\widehat{\Sigma}_p$ with index set $J$ goes to zero a.s.*

*Proof.* Call $\widehat{\lambda}_j$ the eigenvalues of $\widehat{\Sigma}_p$ with index set $J$. Let $M$ be the diagonal matrix with diagonal entries the $\{\widehat{\lambda}_j\}$. Call $L_2$ the set consisting of the other eigenvalues of $\Sigma_p$. Note that the convergence of eigenvalues guarantees that the $\{\widehat{\lambda}_j\}_{j \in J}$ will a.s stay away from that of $L_2$, by a distance at least $\delta_2 > 0$. Call $Z_j$ the eigenvectors corresponding to $\widehat{\lambda}_j$ and $Z$ the matrix with columns $Z_j$ (if some eigenvalues have multiplicity higher than 1, then we pick a set of such eigenvectors). We can write $\Sigma_p = \widehat{\Sigma}_p - \Delta_p$ with $|||\Delta_p|||_2 \to 0$ a.s. Note that $\widehat{\Sigma}_p Z = ZM$, so $\Sigma_p Z = ZM - \Delta_p Z$. Therefore $R = -\Delta_p Z$ and because $|||\cdot|||_2$ is matrix norm and the columns of $Z$ are orthonormal, $|||R|||_2 \leq |||\Delta_p|||_2$. Applying Theorem 4 with these inputs gives the result. □

### 3.6 Practical considerations

The theoretical part of this paper essentially says $\beta$-sparse matrices with $\beta < 1/2$ are asymptotically estimable, in the strong notion of estimability induced by the spectral norm. However, it does not give much information about how to choose the thresholding parameter.

In practice, covariance matrices are estimated for a purpose other than simply estimating them. So in concrete applications, users would most likely be able to find a penalty function that incorporates a measure of performance of a certain estimator and mitigates it with how sparse the corresponding matrix is. Then cross-validation or resampling techniques might be used to assess the performance of different estimators and choose the threshold from the data. Note also, that in Bickel and Levina (2007), Section 5, the authors propose a technique for choosing a banding parameter from the data, which is shown empirically to work quite well. Such technique is transferable in our context, through some fairly straightforward steps.

However, a shortcoming of resampling techniques is their heavy computational cost. Thresholding methods are appealing because they are easily "parallelizable" and can be used on very large dimensional datasets. Therefore having an a priori method that works reasonably well and is not too computationally expensive is also worthwhile. Of course there is a clear link between thresholding and testing the hypothesis that a certain parameter is 0. As a practical ansatz, one method that can be tried is the following: get a $p$-value for the hypothesis $\sigma(i,j) = 0$ for all $i > j$. Such a $p$-value can be obtained by bootstrap methods and since we are dealing with means those reduce to a simple $z$-test. Then perform the Benjamini-Hochberg procedure (see Benjamini and Hochberg (1995)) for these $p$-values, using the FDR parameter $1/\sqrt{p}$. Though the theoretical part of Benjamini and Hochberg (1995) does not apply, we found in the practical examples we ran (limited to Gaussian simulations and relatively simple population covariance matrices) that this worked reasonably well. We include some pictures illustrating our simulations below (see Appendix A.1). If speed is the most important issue, not using the FDR but testing each entry at level $\alpha/\sqrt{p}$ seems also to yield reasonable results.

We note that it is possible that our estimators will not be positive definite: thresholding entry-wise the sample covariance or correlation matrix does not guarantee positive-definiteness of the resulting estimator. Our theorems however say that if the population matrices have a smallest eigenvalue bounded away from zero (uniformly in $p$), then asymptotically our estimators will yield positive-definite matrices (in that case, the theorems also imply spectral norm consistency of $\widehat{\Sigma}_p^{-1}$ for $\Sigma_p^{-1}$). If, in practice, one encounters a non-positive definite estimator, it is clear that the problem at hand should dictate the strategy to remedy this flaw. Two general ideas can nevertheless be applied: one might think of "projecting" the estimator on the cone of positive-semidefinite matrices, using semi-definite programming and probably a sparseness penalty. The feasibility of this idea depends of course of the dimensionality of the problem and it is unlikely to work well (at this point in time) in truly high-dimension. Another idea would be to do a singular value decomposition of the estimator, which is possible even in high-dimension, since the estimator is by construction sparse, and hence falls within the reach of several fast algorithms in numerical linear algebra. Then one could keep a smaller rank approximation of $\widehat{\Sigma}_p$ as the final estimator, $\widehat{\Sigma}_f$, by putting for instance all the negative eigenvalues of $\widehat{\Sigma}_p$ to zero (or instead of 0 a real $g(p)$, with $g(p) \to 0$). Note that $\widehat{\Sigma}_f$ can also be shown to be a consistent estimator of the population covariance, in spectral norm, since $|||\widehat{\Sigma}_f - \widehat{\Sigma}_p|||_2 \to 0$ because the negative eigenvalues of $\widehat{\Sigma}_p$ have to converge to zero (otherwise $|||\widehat{\Sigma}_p - \Sigma_p|||_2$ would not tend to 0). The main drawback of such a solution to the positive definiteness problem is that we may lose the sparsity of the estimator, a feature that is in general desirable. However, its spectral characteristics would be quite easy to obtain, even in high-dimension.

Finally, we note that the results of this paper suggest that acting entry-wise on the sample covariance matrix is a way to create good estimators of $\Sigma_p$. In particular, when other issues such as robustness or contamination by heavy-tailed data arise, using (entry-wise) more robust estimators than the sample covariance is likely to give improved results.

## 4 Conclusion

In this paper we have investigated the theoretical properties of the idea of thresholding the entries of a sample covariance (or correlation) matrix to better estimate the population covariance, when it is

assumed (or known) to be sparse. We have shown that the natural notion of sparsity, coming from problems concerning random vectors is not appropriate when one is concerned with estimating matrices. By contrast, we propose an alternative notion of sparsity, based on properties of the graph corresponding to the adjacency matrix of the population covariance. We have shown that our notion of sparsity divides sharply classes of matrices that are estimable through hard thresholding and those that are not, an appealing property. The notion of sparsity we propose is invariant under permutation of the order of the variables and hence is well-suited for the analysis of problems where there is no canonical ordering of the variables.

We show that $\beta$-sparse matrices, with $\beta < 1/2$ are consistently estimable in operator (a.k.a spectral) norm, a very strong notion of convergence that implies consistency of all eigenvalues and eigenspaces corresponding to eigenvalues separated from the rest of the spectrum (see Subsection 3.5). Practically, the results of simulations are maybe not as striking as one may have hoped for, but lead to great improvement over the sample covariance (or correlation) matrix.

We also show that certain non-sparse matrices are estimable by sparse matrices through the thresholding method we analyzed. Numerically, this method has many advantages in terms of implementation. It is easy to implement, and leads to sparse matrices, which have the desirable property that their eigenvalues and eigenvectors can be numerically computed efficiently, even in high-dimension. Also, since the method acts in an entrywise fashion, the corresponding algorithm is easily parallelizable and in general produces results quickly.

Statistically, our results mean that under the assumption of $\beta$-sparsity, $\beta < 1/2$, applying the natural practical idea of thresholding the entries of a sparse matrix leads to excellent convergence properties. However, we also show that in situations that are not inconceivable in practice, i.e $\beta \geq 1/2$, this strategy may sometime fail to give an estimator as good as what we required. More sophisticated approaches may be needed in these more difficult cases, though, as noted above, the simple thresholding approach has even then many practical virtues.

# Appendix

## A 1/2-sparse matrices: details of computations

In what follows, we use the notation $N$ for the quantity $n-1$ (so $N = n-1$) in an effort to alleviate the notation. The computations that follow are used in Subsection 3.3 and the notations are defined there. Recall that $\nu_i = 1 + \alpha_i^2$ and $i \geq 2$. We give a detailed explanation of our estimate of

$$\mathbf{E}\left(((\widehat{\alpha}_i - \alpha_i)^2 - \nu_i/N)((\widehat{\alpha}_j - \alpha_j)^2 - \nu_j/N)\right) .$$

Clearly, the only thing we need to control is $\mathbf{E}\left((\widehat{\alpha}_i - \alpha_i)^2(\widehat{\alpha}_j - \alpha_j)^2\right)$, since $\nu_i/N$ and $\nu_j/N$ are the means of $(\widehat{\alpha}_i - \alpha_i)^2$ and $(\widehat{\alpha}_j - \alpha_j)^2$. Note that we can write $(\widehat{\alpha}_i - \alpha_i) = \sum_{k=1}^N Z_k(i)/N$, where the $Z_k(i)$'s are i.i.d and mean 0. Similarly, we can write $(\widehat{\alpha}_j - \alpha_j) = \sum_{k=1}^N Y_k(j)/N$. Note that $Y_k(j)$ is independent of $Z_l$ if $k$ is different from $l$. Therefore,

$$\mathbf{E}\left((\widehat{\alpha}_i - \alpha_i)^2(\widehat{\alpha}_j - \alpha_j)^2\right) = \frac{1}{N^4}\mathbf{E}\left(\sum Z_{k_1}(i)Z_{k_2}(i)Y_{k_3}(j)Y_{k_4}(j)\right) .$$

In the previous sum if an index appears only once in the product, the expectation is zero. So only terms where each index appears an even number of times will matter.

We first focus on terms where we have two distinct indices: the contribution of such terms is

$$\frac{N(N-1)}{N^4}\mathbf{E}\left(Z_1^2Y_2^2 + Z_1Z_2Y_1Y_2 + Z_1Z_2Y_2Y_1\right) .$$

We can limit our investigations to the terms with two distinct indices since there are only $N$ terms of the form $Z_1^2Y_1^2$, so their contribution will be asymptotically negligible. Now, $\mathbf{E}\left(Z_1^2Y_2^2\right) = \nu_i\nu_j$, by independence and definition. Also, if $X$ is multivariate Gaussian vector with covariance $\Sigma_p$,

$$\begin{aligned}\mathbf{E}\left(Z_1(i)Y_1(j)\right) &= \mathbf{E}\left((X_1X_i - \alpha_i)(X_1X_j - \alpha_j)\right) = \mathbf{E}\left(X_1^2X_iX_j - \alpha_i\alpha_j\right)\\ &= \sigma(1,1)\sigma(i,j) + \alpha_i\alpha_j + \alpha_j\alpha_i - \alpha_i\alpha_j = \alpha_i\alpha_j + 1_{i=j} ,\end{aligned}$$

by using the fact that we are working with Gaussian random variables. Therefore, if $i \neq j$,

$$\begin{aligned}\mathbf{E}\left(\{(\widehat{\alpha}_i-\alpha_i)^2-\frac{\nu_i}{N}\}\{(\widehat{\alpha}_j-\alpha_j)^2-\frac{\nu_j}{N}\}\right) &= \left(\frac{1}{N^2}-\frac{1}{N^3}\right)\left[2(\alpha_i\alpha_j)^2+\nu_i\nu_j\right]+\frac{\mathbf{E}\left(Z_1^2Y_1^2\right)}{N^3}-\frac{\nu_i\nu_j}{N^2}\,,\\ &= 2(\alpha_i\alpha_j)^2\left(\frac{1}{N^2}-\frac{1}{N^3}\right)+\frac{\mathbf{E}\left(Z_1^2Y_1^2\right)-\nu_i\nu_j}{N^3}\,,\\ &= \mathrm{O}\left(\frac{(\alpha_i\alpha_j)^2}{N^2}\vee\frac{1}{N^3}\right)\end{aligned}$$

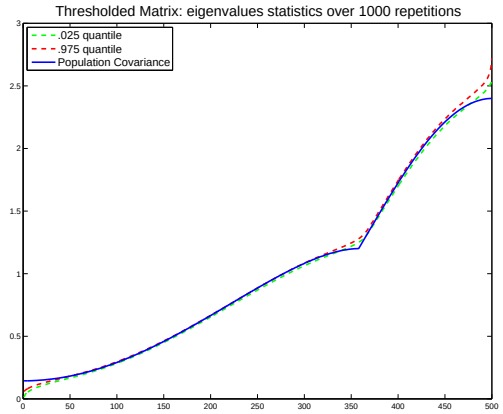
In the case where $\alpha_i\alpha_j = \mathrm{o}(1/\sqrt{N})$, we see that this term is of order $1/n^3$. In general,

## A.1 Performance of estimator: graphical illustration

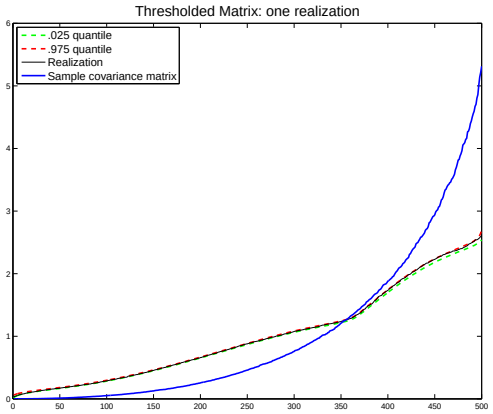
The images of this subsection illustrate the performance of the estimator, assessing visually its variability and comparing it to the sample covariance matrix. All simulations were done with Gaussian data; the thresholding was made according to the FDR rule - in connection with $z$-tests - with FDR parameter $1/\sqrt{p}$. Our illustrations focus on the properties of eigenvalues because they are easier to visualize.

All matrices investigated are (symmetric) Toeplitz matrices, because of the ease with which they can be simulated. We did not randomly permute the "variables" because this would have had no effect on the performance of the estimator; in particular, the eigenvalues would be exactly the same. These matrices can be summarized by their first row, which is what we refer to when speaking of "coefficients" below.

**Case of a Toeplitz matrix, with $n = p = 500$, and coefficients** $(1, 0.3, 0.4, 0, \ldots, 0)$ This situation should be fairly easy since the non-zero coefficients are quite large compared to the variance of $\widehat{\sigma}(i,j)$'s for those $(i,j)$ for which $\sigma(i,j) = 0$.



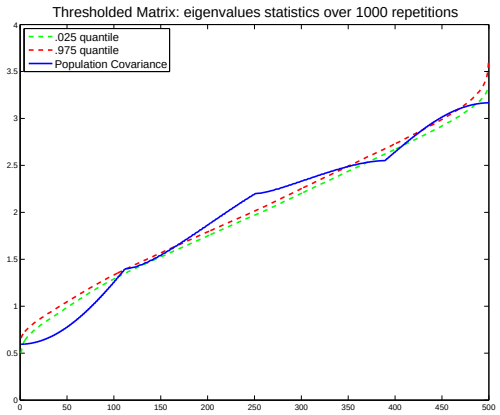

(a) Variability of estimator and population spectrum: scree plot of population and corresponding confidence bounds for ordered eigenvalues of our estimator



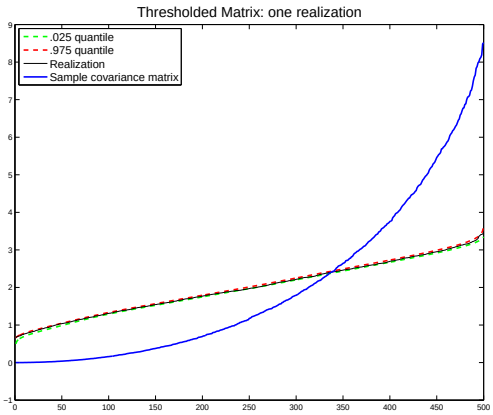

(b) Comparison between scree plot of our estimator (aka "Realization": the continuous line between the two dashed ones) and that of the sample covariance matrix on one realization, picked at random from our 1,000 repetitions

Figure 1: Case of a Toeplitz $(1, 0.3, 0.4, 0, \ldots, 0)$ population covariance matrix $\Sigma_p$, $n = p = 500$. The dashed lines correspond to the .025 and .975 quantiles of the empirical distribution of the $k$-th eigenvalue, for $k = 1$ to $p$. The data was $\mathcal{N}(0, \Sigma_p)$ and the experiment was repeated 1,000 times. As we can see the estimator is very stable. It does well, especially "far" from the edges of the spectrum. For this particular $\Sigma_p$, it can be explained by the fact that the non-zero coefficients in the matrix are easily detectable, when $n = 500$. The improvement over the sample covariance matrix is quite dramatic.

**Case of a Toeplitz matrix, with $n = p = 500$, and coefficients $(2, .2, .3, 0, -.4, 0, \ldots, 0)$** This situation is a bit harder than the one above a priori, as the non-zero coefficients are not as large compared to the variance of $\widehat{\sigma}(i,j)$'s for those $(i,j)$ for which $\sigma(i,j) = 0$ as they are in the previous example.




(a) Variability of estimator and population spectrum: scree plot of population and corresponding confidence bounds for ordered eigenvalues of our estimator




(b) Comparison between scree plot of our estimator (aka "Realization": the continuous line between the two dashed ones) and that of the sample covariance matrix on one realization, picked at random from our 1,000 repetitions

Figure 2: Case of a Toeplitz $(2, .2, .3, 0, -.4, 0, \ldots, 0)$ population covariance matrix $\Sigma_p$, $n = p = 500$. The dashed lines correspond to the .025 and .975 quantiles of the empirical distribution of the $k$-th eigenvalue, for $k = 1$ to $p$. The data was $\mathcal{N}(0, \Sigma_p)$ and the experiment was repeated 1,000 times. As we can see the estimator is very stable. It does capture the support of the spectrum fairly accurately, but is not as good in the capturing the fine details of the bulk. For this particular $\Sigma_p$, there is (compared to the previous example of Figure 1) a certain lack of accuracy when estimating the adjacency matrix $A_p$ of $\Sigma_p$, when $n = 500$. The improvement over the sample covariance matrix is quite dramatic.

**Case of a non-sparse Toeplitz matrix, with $n = 500, p = 100$, and coefficients $\{.3^k\}_{k=0}^{p-1}$** This situation illustrates the approximation of a non-sparse matrix by a sparse matrix. As seen above, this population covariance can be approximated in spectral norm by a 0-sparse matrix. In these type of situations, it is possible that thresholding might be a bit "harsh" and "smoother" regularization approaches might lead to better empirical results.

## References

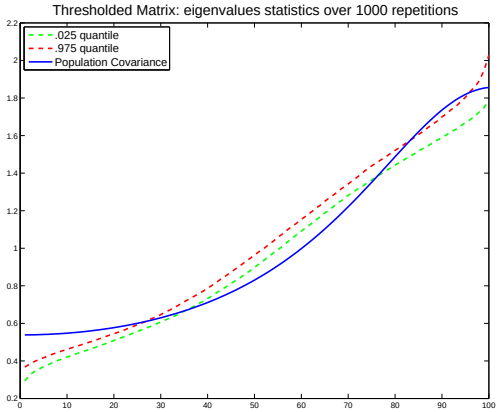

Anderson, G. W. and Zeitouni, O. (2006). A CLT for a band matrix model. *Probab. Theory Related Fields* **134**, 283–338.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis.* Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition.

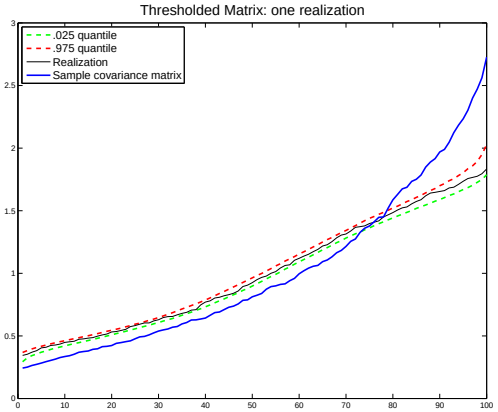
Bai, Z. D. and Silverstein, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32**, 553–605.

Bengtsson, T. and Furrer, R. (2007). Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis* **98**, 227–255.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289–300.




(a) Variability of estimator and population spectrum: scree plot of population and corresponding confidence bounds for ordered eigenvalues of our estimator




(b) Comparison between scree plot of our estimator (aka "Realization": the continuous line between the two dashed ones) and that of the sample covariance matrix on one realization, picked at random

Figure 3: Case of a Toeplitz $\{0.3^k\}_{k=0}^{p-1}$ population covariance matrix $\Sigma_p$, $n = 500, p = 100$. The dashed lines correspond to the .025 and .975 quantiles of the empirical distribution of the $k$-th eigenvalue, for $k = 1$ to $p$. The data was $\mathcal{N}(0, \Sigma_p)$ and the experiment was repeated 1,000 times. As we can see the estimator is very stable. The problem is harder for the thresholding technique than the one illustrated in Figure 1, and it is possible that less "harsh" regularizations might perform slightly better. The improvement over the sample covariance matrix is still quite dramatic.


Bhatia, R. (1997). *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York.

Bickel, P. J. and Levina, E. (2007). Regularized estimation of large covariance matrices. *The Annals of Statistics* To Appear.

d' Aspremont, A., Banerjee, O., and El Ghaoui, L. (2006). First-order methods for sparse covariance selection Available at `math.OC/0609812`.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7**, 1–46.

El Karoui, N. (2006). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Submitted* See `math.ST/0609418`.

El Karoui, N. (2007). Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability* **35**, 663–714.

Geman, S. (1980). A limit theorem for the norm of random matrices. *Ann. Probab.* **8**, 252–261.

Haff, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.* **8**, 586–597.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30.

Horn, R. and Johnson, C. (1990). *Matrix Analysis*. Cambridge University Press.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85–98.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pp. 361–379. Univ. California Press, Berkeley, Calif.

Jonsson, D. (1982). Some limit theorems for the eigenvalues of a sample covariance matrix. *J. Multivariate Anal.* **12**, 1–38.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88**, 365–411.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)* **72 (114)**, 507–536.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55**, 331–339.

Stanley, R. P. (1986). *Enumerative combinatorics. Vol. I.* The Wadsworth & Brooks/Cole Mathematics Series. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA. With a foreword by Gian-Carlo Rota.

Stewart, G. W. and Sun, J. G. (1990). *Matrix perturbation theory.* Computer Science and Scientific Computing. Academic Press Inc., Boston, MA.

Wigner, E. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math. (2)* **62**, 548–564.