



ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA

ARCHIVIO ISTITUZIONALE  
DELLA RICERCA

## Alma Mater Studiorum Università di Bologna Archivio istituzionale della ricerca

A large covariance matrix estimator under intermediate spikiness regimes

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Farne', M., Montanari, A. (2020). A large covariance matrix estimator under intermediate spikiness regimes. JOURNAL OF MULTIVARIATE ANALYSIS, 176, 1-20 [10.1016/j.jmva.2019.104577].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/744602> since: 2020-03-02

*Published:*

DOI: <http://doi.org/10.1016/j.jmva.2019.104577>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

**Farnè M, Montanari A. A large covariance matrix estimator under intermediate spikiness regimes. *J Multivar Anal.* 2020;176:104577. doi:10.1016/j.jmva.2019.104577**

The final published version is available online at:

<https://doi.org/10.1016/j.jmva.2019.104577>

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

*This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)*

***When citing, please refer to the published version.***

# A large covariance matrix estimator under intermediate spikiness regimes

Matteo Farné<sup>a,\*\*</sup>, Angela Montanari<sup>a,\*</sup>

<sup>a</sup>*Department of Statistical Sciences, University of Bologna, Bologna, Italy*

---

## Abstract

This paper concerns large covariance matrix estimation via composite minimization under the assumption of low rank plus sparse structure. In this approach, the low rank plus sparse decomposition of the covariance matrix is recovered by least squares minimization under nuclear norm plus  $l_1$  norm penalization. The objective is minimized via a singular value thresholding plus soft thresholding algorithm. This paper proposes a new estimator based on an additional least-squares re-optimization step aimed at un-shrinking the eigenvalues of the low rank component estimated in the first step. We prove that such un-shrinkage causes the final estimate to approach the target as closely as possible in spectral and Frobenius norm, while recovering exactly the underlying low rank and sparsity pattern. The error bounds are derived imposing that the latent eigenvalues scale to  $p^\alpha$  and the maximum number of non-zeros per row in the sparse component scales to  $p^\delta$ , where  $p$  is the dimension,  $\alpha \in [0, 1]$ ,  $\delta \in [0, 0.5]$ , and  $\delta < \alpha$ . The sample size  $n$  is imposed to scale at least to  $p^{1.5\delta}$ . The resulting estimator is called UNALCE (UNshrunk ALgebraic Covariance Estimator), and it is shown to outperform state-of-the-art estimators, especially for what concerns fitting properties and sparsity pattern detection. The effectiveness of UNALCE is highlighted by a real example regarding ECB (European Central Bank) banking supervisory data.

*Keywords:* Covariance matrix, Nuclear norm, Penalized least squares, Sparsity, Spiked eigenvalues, Un-shrinkage

*Classification code:* 62J10, 65F35, 93E24, 65F50, 15A18, 62J07

---

## 1. Introduction

Estimation of population covariance matrices from samples of multivariate data is of interest in many high-dimensional inference problems: principal components analysis, classification by discriminant analysis, inferring a graphical model structure, and others. Depending on the goal, the interest is sometimes in inferring the eigen-structure of the covariance matrix (as in principal component analysis) and sometimes in estimating its inverse (as in discriminant analysis or in graphical models). Examples of application areas include gene arrays, functional magnetic resonance imaging, text retrieval, image classification, spectroscopy, climate studies, finance, and macro-economic analysis.

The theory of multivariate analysis for normal variables has been well worked out (see, for example, [2]). However, it soon became apparent that exact expressions were cumbersome, and that multivariate data were rarely Gaussian. The remedy was asymptotic theory for large samples and fixed, relatively small, dimensions. However, in recent years, datasets that do not fit into this framework have become very common, since nowadays the data can be high-dimensional, and sample sizes can be very small relative to dimension.

The traditional covariance estimator, the sample covariance matrix, is known to be dramatically ill-conditioned in a large dimensional context, where the process dimension  $p$  is larger than or close to the sample size  $n$ , even when the population covariance matrix is well-conditioned. Two key properties of the matrix estimation process, i.e. numerical

---

<sup>\*\*</sup>Declarations of interest: none<sup>†</sup>

<sup>\*\*</sup>Corresponding author

*Funding:* This paper is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-17-1-0103.

*Email address:* [matteo.farne2@unibo.it](mailto:matteo.farne2@unibo.it) (Matteo Farné)

stability and identifiability, assume a particular relevance in large dimensions. Both properties are crucial for the theoretical derivation and the practical use of the estimate. A bad conditioned estimate suffers from collinearity and causes its inverse, the precision matrix, to dramatically amplify any error in the data. A large dimension may make it impossible to identify the unknown covariance structure, thus hampering the interpretation of the results.

Regularization approaches to large covariance matrix estimation are therefore being presented and addressed in the literature, both from theoretical and practical perspectives (see [14] for an exhaustive overview). Some authors propose shrinking the sample covariance matrix toward the identity matrix [24], others suggest applying nonlinear transforms to sample eigenvalues [25] or regularizing them by sample splitting [23], while some others consider tapering the sample covariance matrix, i.e. gradually shrinking the off-diagonal elements toward zero [9, 21]. At the same time, a common approach is to encourage sparsity, either by a penalized likelihood approach [20] or by thresholding the sample covariance matrix in different ways: hard-thresholding [6], soft-thresholding [5], generalized thresholding [31], or adaptive thresholding [8]. A consistent bandwidth selection method for all these approaches is described in [30].

A different approach is based on the assumption of a low rank plus sparse structure for the covariance matrix:

$$\boldsymbol{\Sigma}^* = \mathbf{L}^* + \mathbf{S}^*, \quad (1)$$

where  $\mathbf{L}^*$  is low rank with rank  $r < p$ ,  $\mathbf{S}^*$  is positive definite and sparse, with at most  $s$  non-zero off-diagonal elements, and  $\boldsymbol{\Sigma}^*$  is a positive definite matrix. The generic covariance estimator  $\hat{\boldsymbol{\Sigma}}$  can be written as

$$\hat{\boldsymbol{\Sigma}} = \mathbf{L}^* + \mathbf{S}^* + \mathbf{W} = \boldsymbol{\Sigma}^* + \mathbf{W}, \quad (2)$$

where  $\mathbf{W}$  is an error term. The error matrix  $\mathbf{W}$  may be deterministic or stochastic, as explained in [1]. If the data are Gaussian and  $\hat{\boldsymbol{\Sigma}}$  is the unbiased sample covariance matrix  $\boldsymbol{\Sigma}_n$ , then  $\mathbf{W}$  is distributed as a re-centred Wishart random matrix.

In [15], a large covariance matrix estimator, called POET (Principal Orthogonal complEMent Thresholding), is derived under the assumption in (1). POET combines principal component analysis for the recovery of the low rank component and a thresholding algorithm for the recovery of the sparse component. The underlying model assumptions prescribe an approximate factor model with spiked eigenvalues for the data (growing with  $p$ ), thus allowing to reasonably use the first  $r$  principal components of the sample covariance matrix. Furthermore, at the same time, sparsity in the sense of [5] is imposed on the residual matrix. The rank  $r$  of the low rank component is chosen by the information criteria in [4].

Indeed, rank selection represents a relevant issue: when  $p$  is large, setting a large rank would cause the estimate  $\hat{\boldsymbol{\Sigma}}$  to be non-positive definite, while setting a small rank would cause a too relevant variance loss. In the discussion of [15], Yu and Samworth point out that the probability to underestimate the latent rank  $r$  does not asymptotically vanish if the eigenvalues are not really spiked at rate  $O(p)$ . In addition, we note that POET systematically overestimates the proportion of variance explained by the factors (given the true rank) because the eigenvalues of  $\boldsymbol{\Sigma}_n$  are more spiky than the true ones (as shown in [24]).

POET consistency holds given that a number of assumptions is satisfied. The key assumption is the pervasiveness of latent factors, which implies that the principal component analysis of  $\boldsymbol{\Sigma}_n$  asymptotically identifies the eigenvalues and eigenvectors of  $\boldsymbol{\Sigma}^*$  as  $p$  diverges. The results of [15] provide the convergence rates of the relative norm of  $\hat{\boldsymbol{\Sigma}}_{POET} - \boldsymbol{\Sigma}^*$  (defined as  $\|\hat{\boldsymbol{\Sigma}}_{POET} - \boldsymbol{\Sigma}^*\|_{\Sigma} = p^{-1/2} \|\boldsymbol{\Sigma}^{*-1/2} \hat{\boldsymbol{\Sigma}}_{POET} \boldsymbol{\Sigma}^{*-1/2} - \mathbf{I}_p\|_F$ ), the maximum norm of  $\hat{\boldsymbol{\Sigma}}_{POET} - \boldsymbol{\Sigma}^*$ , and the spectral norm of  $\hat{\mathbf{S}}_{POET} - \mathbf{S}^*$ . Under stricter conditions,  $\hat{\mathbf{S}}_{POET}$  and  $\hat{\boldsymbol{\Sigma}}_{POET}$  are proved to be non-singular with probability approaching 1.

At the same time, a number of non-asymptotic methods has been presented. In [11], the exact recovery of the covariance matrix in the noiseless context is first proved. The result is achieved by minimizing a specific convex non-smooth objective, which is the sum of the nuclear norm of the low rank component and the  $l_1$  norm of the sparse component. In [10], which is an extension of [11], the exact recovery of the inverse covariance matrix in the noisy graphical model setting is provided. The authors prove that, in the worst case, the number of necessary samples in order to ensure consistency is  $n = O(p^3/r^2)$ , even if the required condition for the positive definiteness of the estimate is  $p \leq 2n$ .

An approximate solution to the recovery and identifiability of the covariance matrix in the noisy context is described in [1]. Even there, the condition  $p \leq n$  is unavoidable for standard results on large deviations and non-asymptotic random matrix theory. An exact solution to the same problem, based on the results in [10], is then shown

in [27]. The resulting estimator is called LOREC (LOW Rank and sparse Covariance estimator) and is proved to be both algebraically and parametrically consistent in the sense of [10].

In [10], algebraic consistency is defined as follows.

**Definition 1.** A pair of symmetric matrices  $(\mathbf{S}, \mathbf{L})$  with  $\mathbf{S}, \mathbf{L} \in \mathbb{R}^{p \times p}$  is an algebraically consistent estimate of the low rank plus sparse decomposition (1) for the covariance matrix  $\Sigma^*$  if the following conditions hold:

- (i) The sign pattern of  $\mathbf{S}$  is the same as that of  $\mathbf{S}^*$ :  $\text{sign}(\mathbf{S}_{ij}) = \text{sign}((\mathbf{S}^*)_{i,j})$ ,  $\forall i, j$ . Here, we assume that  $\text{sign}(0) = 0$ ;
- (ii) The rank of  $\mathbf{L}$  is the same as the rank of  $\mathbf{L}^*$ ;
- (iii) Matrices  $\mathbf{L} + \mathbf{S}$ ,  $\mathbf{S}$ , and  $\mathbf{L}$  are such that  $\mathbf{L} + \mathbf{S}$  and  $\mathbf{S}$  are positive definite and  $\mathbf{L}$  is positive semidefinite.

Parametric consistency holds if the estimates  $(\mathbf{S}, \mathbf{L})$  are close to  $(\mathbf{S}^*, \mathbf{L}^*)$  in some norm with probability approaching 1. In [10], it is defined as follows.

**Definition 2.** A pair of symmetric matrices  $(\mathbf{S}, \mathbf{L})$  with  $\mathbf{S}, \mathbf{L} \in \mathbb{R}^{p \times p}$  is a parametrically consistent estimate of the low rank plus sparse decomposition (1) for the covariance matrix  $\Sigma^*$  if the norm  $g_\gamma = \max(\|\mathbf{S} - \mathbf{S}^*\|_\infty / \gamma, \|\mathbf{L} - \mathbf{L}^*\|_2)$ , where  $\|\cdot\|_\infty$  denotes the maximum norm, converges to 0 with probability approaching 1.

LOREC shows several advantages compared to POET. The most important is that the estimates are both algebraically and parametrically consistent, while POET provides only parametric consistency. Nevertheless, LOREC suffers from some drawbacks, especially concerning the estimated latent eigenvalues. Moreover, the strict condition  $p \leq n$  is required, while POET allows for  $p \ln(p) \gg n$ .

For these reasons, we propose a new estimator, UNALCE (UNshrunk ALgebraic Covariance Estimator), based on the ‘unshrinkage’ (the technical meaning will be clarified in Section 4) of the estimated eigenvalues of the low rank component, which allows to improve the fitting properties of LOREC systematically. We assume that the non-zero eigenvalues of  $\mathbf{L}^*$  are proportional to  $p^\alpha$ ,  $\alpha \in [0, 1]$  (the so called generalized spikiness context). Under the assumption that the maximum number of non-zeros per row in  $\mathbf{S}^*$ , called ‘maximum degree’, scales to  $p^\delta$  (with  $\delta \in [0, 0.5]$  and  $\delta < \alpha$ ), we prove that our estimator possesses a non-asymptotic error bound allowing  $n$  to be as small as  $p^{1.5\delta}$ . We derive absolute bounds depending on  $\alpha$  for the low rank, the sparse component, and the overall estimate. We also identify the conditions for positive definiteness and invertibility and for rank and sparsity pattern recovery. In this way, we provide a unique framework for covariance estimation via composite minimization under the low rank plus sparse assumption.

The remainder of the paper is organized as follows. In Section 2, we establish the notation, set up the model, briefly recall definitions and key properties of LOREC approach, and outline our novel contributions. In Section 3, we define a new estimator, that we call ALCE (ALgebraic Covariance Estimator), and we state the necessary assumptions for algebraic and parametric consistency. In Section 4, we then define the UNALCE (UNshrunk ALCE) estimator, proving that the unshrinkage of thresholded eigenvalues of the low rank component is the key to improve fitting properties as much as possible given a finite sample, preserving algebraic consistency. In Section 5, we propose a new model selection criterion specifically tailored to our model setting. In Section 6, we provide a real Euro Area banking data example which clarifies the effectiveness of our approach (a thorough simulation study is presented in the supplementary material, Section 2). Finally, in Section 7, we draw conclusions and discuss the most relevant findings. The proofs of all theorems and corollaries are reported in Appendix A.

## 2. Numerical estimation and spiked eigenvalues

### 2.1. Notation

Let us define a  $p \times p$  symmetric positive-definite matrix  $\mathbf{M}$ . We denote by  $\lambda_i(\mathbf{M})$ ,  $i \in \{1, \dots, p\}$ , the eigenvalues of  $\mathbf{M}$  in descending order. Then, we recall the following norm definitions:

- (i) Element-wise:

(a)  $L_0$  norm:  $\|\mathbf{M}\|_0 = \sum_{i=1}^p \sum_{j=1}^p \mathbb{1}(\mathbf{M}_{ij} \neq 0)$ , which is the total number of non-zeros;

(b)  $L_1$  norm:  $\|\mathbf{M}\|_1 = \sum_{i=1}^p \sum_{j=1}^p |\mathbf{M}_{ij}|$ ;

(c) Frobenius norm:  $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^p \mathbf{M}_{ij}^2}$ ;

110 (d) Maximum norm:  $\|\mathbf{M}\|_\infty = \max_{i \leq p, j \leq p} |\mathbf{M}_{ij}|$ .

(ii) Induced by vector:

(a)  $\|\mathbf{M}\|_{0,v} = \max_{i \leq p} \sum_{j \leq p} \mathbb{1}(\mathbf{M}_{ij} \neq 0)$ , which is the maximum number of non-zeros per column, defined as the maximum ‘degree’ of  $\mathbf{M}$ ;

(b)  $\|\mathbf{M}\|_{1,v} = \max_{i \leq p} \sum_{j \leq p} |\mathbf{M}_{ij}|$ ;

115 (c) Spectral norm:  $\|\mathbf{M}\|_2 = \lambda_1(\mathbf{M})$ .

(iii) Schatten:

(a) Nuclear norm of  $\mathbf{M}$ , here defined as the sum of the eigenvalues of  $\mathbf{M}$ :  $\|\mathbf{M}\|_* = \sum_{i=1}^p \lambda_i(\mathbf{M})$ .

## 2.2. Model setup

120 Let us suppose that the population covariance matrix of our data is the sum of a low rank and a sparse component. A  $p$ -dimensional random vector  $\mathbf{x}$  is said to have a low rank plus sparse structure if its covariance matrix  $\Sigma^*$  satisfies relationship (1):

$$\Sigma^* = \mathbf{L}^* + \mathbf{S}^*,$$

where:

1.  $\mathbf{L}^*$  is a positive semidefinite symmetric  $p \times p$  matrix with at most rank  $r \ll p$ ;

125 2.  $\mathbf{S}^*$  is a positive definite  $p \times p$  sparse matrix with at most  $s \ll p(p-1)/2$  non-zero off-diagonal elements and maximum degree  $s'$ .

According to the spectral theorem, we can write  $\mathbf{L}^* = \mathbf{U}_L \mathbf{D} \mathbf{U}_L^\top = \mathbf{B} \mathbf{B}^\top$ , where  $\mathbf{B} = \mathbf{U}_L \mathbf{D}^{1/2}$ ,  $\mathbf{U}_L$  is a  $p \times r$  semi-orthogonal matrix,  $\mathbf{D}$  is a  $r \times r$  diagonal matrix, with  $\mathbf{D}_{jj} > 0$ ,  $j \in \{1, \dots, r\}$ . Let us suppose that the  $p \times 1$  random vector  $\mathbf{x}$  is generated according to the following model:

$$\mathbf{x} = \mathbf{B} \mathbf{f} + \epsilon,$$

130 where  $\mathbf{f}$  is a  $r \times 1$  random vector with  $E(\mathbf{f}) = \mathbf{0}_r$ ,  $V(\mathbf{f}) = \mathbf{I}_r$ , and  $\epsilon$  is a  $p \times 1$  random vector with  $E(\epsilon) = \mathbf{0}_p$ ,  $V(\epsilon) = \mathbf{S}^*$ . The random vector  $\mathbf{x}$  is thus assumed to be zero mean, without loss of generality. Given a sample  $\mathbf{x}_k$ ,  $k \in \{1, \dots, n\}$ ,  $\Sigma_n = (n-1)^{-1} \sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^\top$  is the  $p \times p$  sample covariance matrix.

It is easy to observe that  $\mathbf{x}$  follows a low rank plus sparse structure:

$$E(\mathbf{x} \mathbf{x}^\top) = E\{(\mathbf{B} \mathbf{f} + \epsilon)(\mathbf{B} \mathbf{f} + \epsilon)^\top\} = E(\mathbf{B}^\top \mathbf{f} \mathbf{f}^\top \mathbf{B}) + E(\mathbf{B} \mathbf{f} \epsilon^\top) + E(\epsilon \mathbf{B}^\top \mathbf{f}^\top) + E(\epsilon \epsilon^\top) = \mathbf{B} \mathbf{B}^\top + \mathbf{S}^* = \mathbf{L}^* + \mathbf{S}^* = \Sigma^*$$

135 under the usual assumption that  $\text{cov}(\mathbf{f}, \epsilon) = E(\mathbf{f} \epsilon^\top) = \mathbf{0}_{r \times p}$  ( $r \times p$  null matrix). Assuming  $p$  fixed, it is also useful to recall that  $\Sigma_n$  is asymptotically strongly consistent (see [19]). If we assume a normal distribution for  $\mathbf{f}$  and  $\epsilon$ , then  $\Sigma_n$  is unbiased for any fixed  $n$  (see [2]), and the matrix  $\mathbf{W} := \Sigma_n - (\mathbf{L}^* + \mathbf{S}^*)$  is distributed as a re-centred Wishart random matrix. In any case, the normality assumption is not essential for our setting.

### 2.3. Nuclear norm plus $l_1$ norm heuristics

Under the assumption in (1), the need arises to develop a method that can consistently estimate the covariance matrix  $\Sigma^*$  as well as determine the sparsity pattern of  $\mathbf{S}^*$  and the spikiness pattern of the eigenvalues of  $\mathbf{L}^*$  simultaneously. Such an estimation problem is stated as

$$\min_{\mathbf{L}, \mathbf{S}} \frac{1}{2} \|(\mathbf{L} + \mathbf{S}) - \Sigma_n\|_F^2 + \psi \text{rank}(\mathbf{L}) + \rho \|\mathbf{S}\|_{0,off}, \quad (3)$$

where  $\|\mathbf{S}\|_{0,off} = \sum_{i=1}^{p-1} \sum_{j=i+1}^p \mathbb{1}(\mathbf{S}_{ij}^* \neq 0)$  (because the diagonal of  $\mathbf{S}$  is preserved as in [15]). This is a combinatorial problem, which is known to be NP-hard, since both  $\text{rank}(\mathbf{L})$  and  $\|\mathbf{S}\|_{0,off}$  are non-convex.

The tightest convex relaxation of problem (3), as shown in [17], is

$$\min_{\mathbf{L}, \mathbf{S}} \frac{1}{2} \|(\mathbf{L} + \mathbf{S}) - \Sigma_n\|_F^2 + \psi \|\mathbf{L}\|_* + \rho \|\mathbf{S}\|_{1,off}, \quad (4)$$

where  $\psi$  and  $\rho$  are non-negative threshold parameters, and  $\|\mathbf{S}\|_{1,off} = \sum_{i=1}^{p-1} \sum_{j=i+1}^p |\mathbf{S}_{ij}^*|$ . The use of nuclear norm for covariance matrix estimation was introduced in [18]. The feasible set of (4) is the set of all  $p \times p$  symmetric positive definite matrices  $\mathbf{S}$  and all  $p \times p$  symmetric positive semi-definite matrices  $\mathbf{L}$ .

The objective (4) is minimized according to an alternate thresholding algorithm, composed of a singular value thresholding (SVT, [7]) and a soft thresholding step [12]. In order to speed up convergence, Nesterov's acceleration scheme for composite gradient mapping minimization problems [28] is applied. Details of the algorithm are reported in the supplementary material, Section 2.

From a statistical viewpoint, (4) is a penalized least squares heuristics, composed of a smooth least squares term ( $0.5\|(\mathbf{L} + \mathbf{S}) - \Sigma_n\|_F^2$ ) and a non-smooth composite penalty ( $\psi\|\mathbf{L}\|_* + \rho\|\mathbf{S}\|_{1,off}$ ). The choice of (4) allows to lower the condition number of the estimates and the parameter space dimensionality simultaneously. In principle, different losses could be used, like Stein's one [13]. However, the classical Frobenius loss does not require normality and is computationally appealing.

From an algebraic viewpoint, (4) is an algebraic matrix variety recovery problem. In the covariance matrix setting described in equation (1), matrices  $\mathbf{L}^*$  and  $\mathbf{S}^*$  are assumed to come from the following sets of matrices:

$$\mathcal{B}(r) = \{\mathbf{L} \in \mathbb{R}^{p \times p} \mid \mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{U}^\top, \mathbf{U} \in \mathbb{R}^{p \times r} \text{ semi-orthogonal}, \mathbf{D} \in \mathbb{R}^{r \times r} \text{ diagonal}\}, \quad (5)$$

$$\mathcal{A}(s) = \{\mathbf{S} \in \mathbb{R}^{p \times p} \mid |\text{support}(\mathbf{S})| \leq s\}. \quad (6)$$

$\mathcal{B}(r)$  is the variety of matrices with at most rank  $r$ .  $\mathcal{A}(s)$  is the variety of (element-wise) sparse matrices with at most  $s$  non-zero elements, where  $\text{support}(\mathbf{S})$  is the orthogonal complement of  $\ker(\mathbf{S})$  and  $|\text{support}(\mathbf{S})|$  denotes its dimension.

In [11], a notion of rank-sparsity incoherence is developed. It is expressed as an uncertainty principle between the sparsity pattern of a matrix and its row/column space. This notion has been introduced because  $\mathbf{L}^*$  cannot be identified if it is nearly sparse, and  $\mathbf{S}^*$  cannot be identified if it is nearly low rank. Denoting by  $T(\mathbf{L}^*)$  and  $\Omega(\mathbf{S}^*)$  the tangent spaces to  $\mathcal{B}(r)$  and  $\mathcal{A}(s)$ , respectively, the following rank-sparsity incoherence measures between  $T(\mathbf{L}^*)$  and  $\Omega(\mathbf{S}^*)$  are defined:

$$\xi(T(\mathbf{L}^*)) = \max_{\mathbf{M} \in T(\mathbf{L}^*), \|\mathbf{M}\|_2 \leq 1} \|\mathbf{M}\|_\infty, \quad (7)$$

$$\mu(\Omega(\mathbf{S}^*)) = \max_{\mathbf{M} \in \Omega(\mathbf{S}^*), \|\mathbf{M}\|_\infty \leq 1} \|\mathbf{M}\|_2. \quad (8)$$

Quantities (7) and (8) control the identifiability of  $\mathbf{L}^*$  and  $\mathbf{S}^*$  in (1). In fact, a necessary and sufficient condition for identifiability is that  $T(\mathbf{L}^*)$  and  $\Omega(\mathbf{S}^*)$  have a transverse intersection, i.e. they intersect only at the origin. In [11], it is proved that  $T(\mathbf{L}^*)$  and  $\Omega(\mathbf{S}^*)$  are transverse if and only if (7) and (8) are small. Therefore, the product  $\mu(\Omega(\mathbf{S}^*))\xi(T(\mathbf{L}^*))$  is a rank-sparsity incoherence measure and bounding it controls both for identification and recovery.

The described approach was first used for deriving the LOREC estimator in [27]. Therein, the reference matrix class imposed to  $\Sigma^*$  is

$$\Sigma^*(\epsilon_0) = \left\{ \Sigma^* \in \mathbb{R}^{p \times p} : 0 < \epsilon_0 \leq \lambda_i(\Sigma^*) \leq \epsilon_0^{-1}, i \in \{1, \dots, p\} \right\}$$

which is the class of positive definite matrices having uniformly bounded eigenvalues. In the context so far described, Luo (cf. [27]) proves that  $\mathbf{L}$  and  $\mathbf{S}$  can be identified and recovered with bounded error, and the rank of  $\mathbf{L}$  as well as the sparsity pattern of  $\mathbf{S}$  are exactly recovered.

The proof reproduces a similar proof in [5], but neglects a fundamental assumption on which that paper relies, i.e. that  $\max_{i \leq p} \sum_{j \leq p} |\Sigma_{ij}^*|^q = o(p)$  for some  $q \in [0, 1[$ . As stated in the supplementary material (Section 1), this can happen only if  $\mathbf{L}^*$  is sparse, which contradicts rank-sparsity incoherence, thus making the model in [27] not identifiable.

#### 2.4. Contribution of the paper

In this article, we propose an estimation algorithm for  $\Sigma^*$  under the assumption in (1) based on a nuclear norm plus  $l_1$  norm penalization, as in [27]; however, contrary to [27], we derive the properties of the estimator under the sparsity assumption on  $\mathbf{S}^*$  (see Assumption 4) and not on  $\Sigma^*$ . This allows to avoid the non-identifiability trap and to enormously enlarge the set of recoverable pairs of matrices. We explicitly control the magnitude of  $\xi(T(\mathbf{L}^*))$  and  $\mu(\Omega(\mathbf{S}^*))$  with respect to  $p$ . More importantly, we allow for the generalized spikiness of the eigenvalues of  $\mathbf{L}^*$  (cf. Yu and Samworth, [15], p. 656), thus modelling a large variety of real situations. In addition, we overcome the strict assumption  $p \leq n$  by linking  $n$  to the degree of sparsity of  $\mathbf{S}^*$ . We call the resulting estimator ALCE (ALgebraic Covariance Estimator). In the end, since the singular value thresholding procedure has a significantly strong impact on sample eigenvalues when  $p$  is large and the latent eigenvalues are spiked, we apply an un-shrinkage step to the estimates of the latent eigenvalues. We name the resulting estimator UNALCE (UNshrunk ALCE). We prove that UNALCE is both algebraically and parametrically consistent. Within the class of algebraically consistent estimates, it minimizes the overall loss in Frobenius norm, given the finite sample and the threshold pair  $(\psi, \rho)$  in (4).

### 3. The ALgebraic Covariance Estimator (ALCE)

#### 3.1. Component estimates and consistency

Let us suppose that the eigenvalues of  $\Sigma^*$  are intermediately spiked. This amounts to assume the generalized spikiness of latent eigenvalues in the sense of Yu and Samworth ([15], p. 656):

**Assumption 1.** All the eigenvalues of the  $r \times r$  matrix  $p^{-\alpha} \mathbf{B}^T \mathbf{B}$  are bounded away from 0 for all  $p$  and  $\alpha \in [0, 1]$ .

If  $p$  is finite, Assumption 1 is equivalent to state that

$$\begin{aligned} \lambda_{1, \dots, r}(\Sigma^*) &> \delta_\alpha p^\alpha, \\ \lambda_{r+1, \dots, p}(\Sigma^*) &< \delta_\alpha p^\alpha, \end{aligned} \tag{9}$$

for some  $\delta_\alpha > 0$ . We aim to study the properties of the covariance estimates obtained by heuristics (4) under the generalized spikiness assumption in a non-asymptotic context.

In order to reach this goal, we need to impose the following assumptions in our finite sample context.

**Assumption 2.** There exist  $k_L, k_S > 0$ ,  $\delta \in [0, 0.5]$ , such that  $\xi(T(\mathbf{L}^*)) = \sqrt{r/(k_L^2 p^{2\delta})}$ ,  $\mu(\Omega(\mathbf{S}^*)) = k_S p^\delta$ ,  $k_S/k_L \leq 1/54$  with  $\delta < \alpha$ .

**Assumption 3.** There exist  $r_1, r_2 > 0$  and  $b_1, b_2 > 0$  such that, for any  $t > 0$ ,  $k \leq n$ ,  $i \leq r$ ,  $j \leq p$ :

$$\Pr(|\mathbf{f}_{ik}| > s) \leq \exp(-b_1/t), \quad \Pr(|\epsilon_{jk}| > s) \leq \exp(-b_2/t).$$

**Assumption 4.** There exist  $c_1, c_2, c_3, \delta_2, \delta_2' > 0$ ,  $\delta' \in [0, \delta + 0.5]$  such that  $\lambda(\mathbf{S}^*)_{\min} > c_1$ ,  $\min_{i, j \leq p} \text{var}(\epsilon_{ik} \epsilon_{jk}) > c_2$  for any  $k \leq n$ ,  $i, j \leq p$ ,  $s_{ii}^* \leq c_3$  for any  $i \leq p$ ,  $s' = \max_{i \leq p} \sum_{j \leq p} \mathbb{1}(\mathbf{S}_{ij}^* = 0) \leq \delta_2 p^\delta$  with  $\delta_2 \geq k_S$  and  $\|\mathbf{S}\|_{1,v} = \max_{i \leq p} \sum_{j \leq p} |\mathbf{S}_{ij}^*| \leq \delta_2' p^{\delta'}$ .

**Assumption 5.** There exist  $\delta_3, \delta_4 > 0$  such that  $r = \delta_3 \ln p$  and  $n \geq \delta_4 p^{1.5\delta}$ .

Under those assumptions, we prove the following Theorem 1 which provides a non-asymptotic consistency result, particularly useful when  $p$  is not that large and  $\alpha < 1$ . In fact, in that case, principal components are far from convergence, and therefore, POET approach becomes suboptimal.



**Theorem 1.** Let  $T = T(\mathbf{L}^*)$  and  $\Omega = \Omega(\mathbf{S}^*)$  be the tangent spaces to (5) and (6), respectively. Suppose that Assumptions 1-5 hold. Define  $\psi = (1/\xi(T))(p^\alpha/\sqrt{n})$  and  $\rho = \gamma\psi$ , where  $\xi(T)$  has been defined in (7),  $\alpha \in [0, 1]$ ,  $\gamma \in [9\xi(T), 1/(6\mu(\Omega))]$ , and  $\mu(\Omega)$  has been defined in (8). In addition, suppose that the minimum eigenvalue of  $\mathbf{L}^*$  ( $\lambda_r(\mathbf{L}^*)$ ) is greater than  $C_2\psi/\xi^2(T)$ . Then, with probability  $1 - O(1/\min(p, n)^2)$ , the pair  $(\hat{\mathbf{L}}, \hat{\mathbf{S}})$  minimizing (4) recovers the rank of  $\mathbf{L}^*$  ( $\text{rank}(\hat{\mathbf{L}}) = \text{rank}(\mathbf{L}^*)$ ). Moreover, the matrix losses for each component are bounded as follows:

$$\|\hat{\mathbf{L}} - \mathbf{L}^*\|_2 \leq C\psi, \quad \|\hat{\mathbf{S}} - \mathbf{S}^*\|_\infty \leq C\rho.$$

We call ALCE (ALgebraic Covariance Estimator) the estimator of  $\Sigma^*$  in (2) obtained by estimating  $\mathbf{L}^*$  by  $\hat{\mathbf{L}}$  and  $\mathbf{S}^*$  by  $\hat{\mathbf{S}}$ :

$$\hat{\Sigma}_{ALCE} = \hat{\mathbf{L}}_{ALCE} + \hat{\mathbf{S}}_{ALCE}. \quad (11)$$

Theorem 1 states that, under all the prescribed assumptions, the losses of the pair  $(\hat{\mathbf{L}}_{ALCE}, \hat{\mathbf{S}}_{ALCE})$  obtained by minimizing (4) with respect to the true  $(\mathbf{L}^*, \mathbf{S}^*)$  are bounded, and the rank of  $\mathbf{L}^*$  is exactly recovered, provided that the minimum latent eigenvalue is large enough, as well as the underlying matrix varieties  $T$  and  $\Omega$  are transverse enough. Exploiting the consistency norm of [10], i.e.

$$g_\gamma = \max\left(\frac{\|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_\infty}{\gamma}, \|\hat{\mathbf{L}}_{ALCE} - \mathbf{L}^*\|_2\right),$$

it follows from Theorem 1 that

$$g_\gamma(\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*, \hat{\mathbf{L}}_{ALCE} - \mathbf{L}^*) \leq C \frac{1}{\xi(T)} \frac{p^\alpha}{\sqrt{n}}$$

with probability  $1 - O(1/\min(p, n)^2)$ .

In the proof of Theorem 1, Assumption 2 is needed in order to ensure consistency and rank recovery. In fact, an identifiability condition for problem (4), as shown in Theorem 1, is  $\xi(T(\mathbf{L}^*))\mu(\Omega(\mathbf{S}^*)) \leq 1/54$ . According to [11],  $\sqrt{r/p} \leq \xi(T(\mathbf{L}^*)) \leq 1$  and  $\min_{i \leq p} \sum_{j \leq p} \mathbb{1}(\mathbf{S}_{ij}^* \neq 0) \leq \mu(\Omega(\mathbf{S}^*)) \leq \max_{i \leq p} \sum_{j \leq p} \mathbb{1}(\mathbf{S}_{ij}^* \neq 0)$ . It follows that  $\xi(T(\mathbf{L}^*)) = 1$  with  $\delta = 0$  in the worst case scenario and  $\xi(T(\mathbf{L}^*)) = \sqrt{r/p}$  with  $\delta = 0.5$  in the best case scenario, under the condition  $k_S/k_L \leq 1/54$ . Such assumption is essential to ensure the parametric consistency of the estimated pair in terms of matrix norms and the recovery of the underlying algebraic matrix varieties under model (2) (cf. [10]). The assumption  $\delta < \alpha$  is required in order to ensure that conditions (9) and (10) hold under the condition  $\lambda_r(\mathbf{L}^*) > C_2\psi/\xi^2(T)$  of Theorem 1 and to rule out the degenerate case  $\delta = \alpha$ .

Assumption 3 is necessary to ensure that large deviation theory can be applied to  $\mathbf{f}_{ik}$ ,  $\epsilon_{jk}$ , and  $\mathbf{f}_{ik}\epsilon_{jk}$  for all  $i \leq r$ ,  $j \leq p$ , and  $k \leq n$  (cf. [15]). Assumption 4 is necessary in order to apply the results of [5] on the sparse component which prescribe that  $\mathbf{S}^*$  must be well conditioned with uniformly bounded diagonal elements. We stress that the maximum degree  $s'$  must be bounded to ensure parametric and algebraic consistency, because Assumption 2 ensures  $\mu(\Omega(\mathbf{S}^*)) = k_S p^\delta$  with  $\delta \leq 0.5$ . This condition is different from the corresponding one in [15], which prescribes  $\max_{i \leq p} \sum_{j \leq p} |\mathbf{S}_{ij}^*|^q < c_4$ ,  $q \in [0, 1]$ ,  $c_4 > 0$ .

In general, we can allow for  $\|\mathbf{S}^*\|_2 \leq \|\mathbf{S}^*\|_{1,v} \leq \delta'_2 p^{\delta'}$ ,  $\|\mathbf{S}^*\|_1 \leq p \|\mathbf{S}^*\|_{1,v} \leq \delta'_2 p^{1+\delta'}$ , and  $\|\mathbf{S}^*\|_0 = p + s \leq p s' \leq \delta_2 p^{1+\delta}$ . In addition, we can also write  $\|\mathbf{S}^*\|_2 \leq \|\mathbf{S}^*\|_{0,v} = \delta_2 p^\delta$  (as shown in [5]). The assumption  $\delta' \leq \delta + 0.5$  is needed to respect the inequality  $\|\mathbf{S}^*\|_{1,v} \leq \sqrt{p} \|\mathbf{S}^*\|_2$ . We stress that the assumption  $\delta < \alpha$  is enough to ensure  $\|\mathbf{S}^*\|_2 = o(p)$  as  $p$  diverges, thus also guaranteeing POET consistency for each  $\alpha$  given the true rank (see Yu and Samworth, [15], p. 656).

Assumption 5 prescribes that the latent rank is infinitesimal with respect to  $p$  and the sample size  $n$  is possibly smaller than  $p$ , but not smaller than  $\delta_4 p^{1.5\delta}$ . It leads to overcoming the restrictive condition  $p \leq n$ , since  $\delta \leq 0.5$ . The need for it arises in order to ensure coherence with Assumptions 1 and 4.

In Corollary 1, we prove the asymptotic consistency of ALCE estimates.

**Corollary 1.** Suppose that Assumptions 1-5 hold. If the limit  $\lim_{v \rightarrow \infty} \min_v(p_v^{2\alpha+2\delta}, n_v) = \infty$  with the path-wise restriction  $\lim_{v \rightarrow \infty} p_v^{2\alpha+2\delta}/n_v = 0$  holds, then  $\lim_{v \rightarrow \infty} \psi_v = 0$  for  $\psi_v = p_v^\alpha/(\xi(T) \sqrt{n_v})$ .

Corollary 1 shows how  $p$  and  $n$  may cause the probabilistic error to annihilate in the limit. For the terminology about limit sequences, see [3]. Moreover,  $\psi_v/p_v^{\alpha+\delta} \rightarrow 0$  as  $\lim_{v \rightarrow \infty} \min_v(p_v^{2\alpha+2\delta}, n_v) = \infty$ , thus establishing the asymptotic consistency in relative terms, resembling the ‘blessing of dimensionality’ described in [15].

245 In order to prove the recovery of the residual sparsity pattern, we add to the previous ones the following assumption.

**Assumption 6.**  $2\delta \leq \alpha \leq 2\delta + \delta'$  and  $0 < (C_3\delta)/(k_L\delta_4) < \delta'$ .

We can then prove Theorem 2.

250 **Theorem 2.** *Suppose that all the assumptions of Theorem 1 hold. If the minimum absolute value of the non-zero off-diagonal entries of  $\mathbf{S}^*$ ,  $S_{\min,off}$ , is greater than  $(C_3\psi)/\mu(\Omega)$  and Assumption 6 holds, then the matrix  $\hat{\mathbf{S}}$  minimizing (4) exactly recovers the sparsity pattern of  $\mathbf{S}^*$  with probability  $1 - O(1/\min(p, n)^2)$  ( $\text{sign}(\hat{\mathbf{S}}) = \text{sign}(\mathbf{S}^*)$ ).*

255 Theorem 2 states that the sparsity pattern of  $\mathbf{S}^*$  is also recovered if the minimum absolute non-zero off-diagonal entry of  $\mathbf{S}^*$  is large enough and Assumption 6 holds. Consequently, we can state that the condition on the minimum latent eigenvalue and the assumption  $\delta < \alpha$  are more important than the condition on the minimum absolute non-zero off-diagonal entry. In fact, the former are strictly necessary both for proving parametric consistency and rank recovery. The latter is necessary only for proving sparsity pattern recovery, as an additional result, given that the former hold. The only consequence of its violation is that some non-zero elements of  $\mathbf{S}^*$  are not recovered.

260 Assumption 6 is necessary for the following reason. Since the product between the minimum absolute non-zero off-diagonal entry of  $\mathbf{S}^*$ ,  $S_{\min,off}$ , and the maximum degree of  $\mathbf{S}^*$ ,  $s'$ , cannot overcome the  $L_1$  norm of  $\mathbf{S}^*$ ,  $\max_{i \leq p} \sum_{j \leq p} |\mathbf{S}_{ij}^*|$ , it follows from the condition  $S_{\min,off} > (C_3\psi)/\mu(\Omega)$  of Theorem 2 and Assumptions 4 that

$$0 < \frac{C_3\psi}{\mu(\Omega)} s' < S_{\min,off} s' < \max_{i \leq p} \sum_{j \leq p} |\mathbf{S}_{ij}^*| \leq \delta'_2 p^{\delta'}. \quad (12)$$

Inequality (12), under Assumptions 2, 4, and 5, boils down to  $(C_3\delta_2 p^{\alpha-2\delta})/(k_L\delta_4) < \delta'_2 p^{\delta'}$  and  $0 < (C_3\delta_2 p^{\alpha-2\delta})/(k_L\delta_4)$ , which hold if and only if Assumption 6 is satisfied.

265 We stress that the conditions  $\lambda_r(\mathbf{L}^*) > (C_2\psi)/\xi^2(T)$  and  $S_{\min,off} > (C_3\psi)/\mu(\Omega)$  under Assumptions 2 and 5 become  $\lambda_r(\mathbf{L}^*) > C_2 p^\alpha$  and  $S_{\min,off} > C_3 p^{\alpha-2\delta}$ , respectively. The latter in turn leads to (12), which holds under Assumption 6. Therefore, the resultant model setting is fully consistent with Assumptions 1 and 4.

270 A representative selection of the latent eigenvalue and sparsity patterns admitted under the described conditions is reported in the supplementary material, Section 2. We emphasize that, e.g. the algebraic consistency no longer forces the latent eigenvalues to scale to  $p$ , provided that the maximum degree of the residual component is scaled accordingly. In general, it is necessary that the minimum latent eigenvalue and absolute non-zero residual entry should be large enough to ensure algebraic consistency, but they can both depend on  $p^\alpha$ , with  $\alpha$  potentially smaller than 1. In particular, if we increase  $\alpha$ , both  $\lambda_r(\mathbf{L}^*)$  and  $S_{\min,off}$  must be larger to ensure identifiability. The same happens if  $p$  increases. On the contrary, if  $r$  increases, then  $\mathbf{L}^*$  can have less spiked eigenvalues, while if  $\delta$  increases, then  $S_{\min,off}$  is allowed to be smaller.

### 3.2. Error bounds for $\hat{\mathbf{S}}_{ALCE}$ and $\hat{\Sigma}_{ALCE}$ in spectral and Frobenius norm

275 Within the same framework, we can complete our analysis with the bounds for  $\hat{\mathbf{S}}_{ALCE}$ .

From  $\|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_2 \leq s' \|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_\infty$ , we obtain

$$\|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_2 \leq C s' \xi(T) \psi = \phi_{\mathbf{S}} \quad (13)$$

From  $\|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_F \leq \sqrt{p s'} \|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_\infty$ , we obtain

$$\|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_F \leq C \sqrt{p s'} \xi(T) \psi. \quad (14)$$

$\hat{\mathbf{S}}_{ALCE}$  is positive definite if and only if  $\lambda_p(\mathbf{S}^*) > \phi_{\mathbf{S}}$ . Bounds (13) and (14) hold with probability  $1 - O(1/\min(p, n)^2)$ . For the inverse of  $\hat{\mathbf{S}}_{ALCE}$ ,  $\hat{\mathbf{S}}_{ALCE}^{-1}$ , the same bounds hold with probability  $1 - O(1/\min(p, n)^2)$ :

$$\|\hat{\mathbf{S}}_{ALCE}^{-1} - \mathbf{S}^{*-1}\|_2 \leq C s' \xi(T) \psi = \phi_{\mathbf{S}}, \quad \|\hat{\mathbf{S}}_{ALCE}^{-1} - \mathbf{S}^{*-1}\|_F \leq C \sqrt{p s'} \xi(T) \psi.$$

280 if and only if  $\lambda_p(\mathbf{S}^*) \geq 2\phi_{\mathbf{S}}$ .

From Theorem 1, we can derive with probability  $1 - O(1/\min(p, n)^2)$  the following bounds for  $\hat{\Sigma}_{ALCE}$ :

$$\|\hat{\Sigma}_{ALCE} - \Sigma^*\|_2 \leq C(s'\xi(T) + 1)\psi = \phi, \quad \|\hat{\Sigma}_{ALCE} - \Sigma^*\|_F \leq C(\sqrt{ps'}\xi(T) + \sqrt{r})\psi.$$

$\hat{\Sigma}_{ALCE}$  is positive definite if and only if  $\lambda_p(\Sigma^*) > \phi$ . The same bounds hold for the inverse covariance estimate  $\hat{\Sigma}_{ALCE}^{-1}$  with probability  $1 - O(1/\min(p, n)^2)$ :

$$\|\hat{\Sigma}_{ALCE}^{-1} - \Sigma^{*-1}\|_2 \leq C(s'\xi(T) + 1)\psi = \phi, \quad \|\hat{\Sigma}_{ALCE}^{-1} - \Sigma^{*-1}\|_F \leq C(\sqrt{ps'}\xi(T) + \sqrt{r})\psi$$

given that  $\lambda_p(\Sigma^*) \geq 2\phi$ .

Overall, ALCE estimator allows to recover a relaxed spiked eigen-structure, thus overcoming the condition  $p \leq n$ , even using the sample covariance matrix as estimation input (the ratio  $p/n$  directly impacts the error bound). Our bounds are in absolute norms and reflect the underlying degree of spikiness  $\alpha$ . Our theory relies on the probabilistic convergence of the sample covariance matrix under the assumption that the data follow an approximate factor model with a sparse residual. If all the assumptions of Theorems 1 and 2 and Corollary 1 hold with  $\lambda_p(\mathbf{S}^*) > \phi_S$  and  $\lambda_p(\Sigma^*) > \phi$ , then both algebraic and parametric consistency are ensured in the sense of Definitions 1 and 2, respectively.

Compared to LOREC, ALCE minimizes the same heuristics but is consistent for a much wider range of real situations, including high-dimensional settings ( $p > n$ ). However, they both share a problem related to input eigenvalues: as  $p$  increases and the latent eigenvalues are spiked, the nuclear norm heuristics may lead to eigenvalue over-shrinkage, as shown in the following Section. For this reason, we further improve ALCE by un-shrinking the estimates of latent eigenvalues.

## 4. The UNALCE estimator: A re-optimized ALCE solution

### 4.1. Motivation

As previously mentioned, when  $p$  is large and the latent eigenvalues are spiked, the singular value thresholding procedure may lead to eigenvalue over-shrinkage, because in that case, the top  $r$  eigenvalues of  $\Sigma_n$  estimate increasingly better the latent eigenvalues as  $p$  increases. Therefore, shrinking the top  $r$  sample eigenvalues leads to too small estimates of the latent eigenvalues, and this also inevitably affects the residual and overall estimate.

Let us define  $\Delta_L = \hat{\mathbf{L}}_{ALCE} - \mathbf{L}^*$ ,  $\Delta_S = \hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*$ ,  $\Delta_\Sigma = \hat{\Sigma}_{ALCE} - \Sigma^*$ . Another key aspect of Theorem 1 is that the two losses in  $\mathbf{L}^*$  and  $\mathbf{S}^*$  are bounded separately. This fact results in a negative effect on the overall performance of  $\hat{\Sigma}_{ALCE}$ , represented by the loss  $\|\Delta_\Sigma\|_2$ , since  $\|\Delta_\Sigma\|_2$  is simply derived as a function of  $\|\Delta_L\|_2$  and  $\|\Delta_S\|_2$  according to the triangle inequality  $\|\Delta_\Sigma\|_2 \leq \|\Delta_L\|_2 + \|\Delta_S\|_2$ . Therefore, the need arises to also correct for this drawback, re-shaping  $\hat{\Sigma}_{ALCE}$ , as the ALCE solution is somehow sub-optimal for the whole covariance matrix.

We approach these issues by a finite-sample analysis, which could be referred to as a re-optimized least squares method. We refer to the usual objective function (4) with  $\|\mathbf{S}\|_1 = \|\mathbf{S}\|_{1,off} = \sum_{i=1}^{p-1} \sum_{j=i+1}^p |\mathbf{S}_{ij}|$ , which is the  $l_1$  norm of  $\mathbf{S}$  excluding the diagonal entries, consistently with POET approach. We define  $\mathbf{Y}_{pre}$  and  $\mathbf{Z}_{pre}$  as the last updates in the gradient step of the minimization algorithm of (4) (see Section 2 in the supplementary material).  $\mathbf{Y}_{pre}$  and  $\mathbf{Z}_{pre}$  are the two matrices we condition upon in order to derive our finite-sample re-optimized estimates.

We note some analogies between our approach and the restricted maximum likelihood (REML) method as explained in [22, 29]. More precisely, the minimization of (4) acts as the ML estimator of fixed effects, while our re-optimized least squares step acts as the estimator of variance components.

Let us define the recovered rank  $\hat{r} = \text{rank}(\hat{\mathbf{L}}_{ALCE})$  and the recovered number of residual non-zeros  $\hat{s} = |\text{support}(\hat{\mathbf{S}}_{ALCE})|$ . In the second step, we exploit the consistency properties of the varieties  $\hat{\mathcal{B}}(\hat{r})$  and  $\hat{\mathcal{A}}(\hat{s})$  recovered in the first step, defined as

$$\hat{\mathcal{B}}(\hat{r}) = \{\mathbf{L} \in \mathbb{R}^{p \times p} \mid \mathbf{L} = \hat{\mathbf{U}}_{ALCE} \mathbf{D} \hat{\mathbf{U}}_{ALCE}^\top, \mathbf{D} \in \mathbb{R}^{\hat{r} \times \hat{r}} \text{ diagonal}\}, \quad (15)$$

$$\hat{\mathcal{A}}(\hat{s}) = \{\mathbf{S} \in \mathbb{R}^{p \times p} \mid |\text{support}(\mathbf{S})| \leq \hat{s}\}. \quad (16)$$

In particular, based on Theorems 1 and 2, we rely on the parametric guarantees offered by  $\hat{\mathcal{B}}(\hat{r})$  and  $\hat{\mathcal{A}}(\hat{s})$ , and we condition upon the recovered latent rank and residual off-diagonal sparsity pattern. In this way, conditioning on the first step, we can focus on re-optimizing our pair of estimates to improve the overall fitting as much as possible, restricting our search into  $\hat{\mathcal{B}}(\hat{r})$  and  $\hat{\mathcal{A}}(\hat{s})$ .

#### 4.2. Optimality

The recovered varieties  $\hat{\mathcal{B}}(\hat{r})$  and  $\hat{\mathcal{A}}(\hat{s})$  ensure the algebraic consistency of  $(\hat{\mathbf{S}}_{ALCE}, \hat{\mathbf{L}}_{ALCE})$  under all the assumptions of Theorems 1 and 2. One might look for the solution (say  $(\hat{\mathbf{L}}_{New}, \hat{\mathbf{S}}_{New})$ ) of the problem

$$\min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{r}), \mathbf{S} \in \hat{\mathcal{A}}(\hat{s})} \text{STL}(\mathbf{L}, \mathbf{S}) = \|\Sigma_n - (\mathbf{L} + \mathbf{S})\|_F^2, \quad (17)$$

where  $\text{STL}(\mathbf{L}, \mathbf{S})$  stands for *Sample Total Loss*. The sample covariance matrix follows the model  $\Sigma_n = \mathbf{L}^* + \mathbf{S}^* + \mathbf{W}$ , given a sample of  $p$ -dimensional data vectors  $\mathbf{x}_k$ ,  $k \in \{1, \dots, n\}$ . Our problem essentially is as follows: which pair  $\mathbf{L} \in \hat{\mathcal{B}}(\hat{r}), \mathbf{S} \in \hat{\mathcal{A}}(\hat{s})$  satisfying algebraic consistency shows the best approximation properties of  $\Sigma_n$ ?

We prove the following result.

**Theorem 3.** Define the spectral decomposition of  $\hat{\mathbf{L}}_{ALCE}$  as  $\hat{\mathbf{U}}_{ALCE} \hat{\mathbf{D}}_{ALCE} \hat{\mathbf{U}}_{ALCE}^\top$  and  $\hat{\mathbf{L}}_{New} = \hat{\mathbf{U}}_{ALCE} (\hat{\mathbf{D}}_{ALCE} + \check{\psi} \mathbf{I}_r) \hat{\mathbf{U}}_{ALCE}^\top$ , where  $\check{\psi} > 0$  is any prescribed threshold parameter. Define  $\hat{\mathbf{S}}_{New}$  such that its off-diagonal elements are the same as  $\hat{\mathbf{S}}_{ALCE}$ , and  $\hat{\Sigma}_{New}$  such that its diagonal elements are the same as  $\hat{\Sigma}_{ALCE}$ , respectively. In addition, set  $\text{diag}(\hat{\mathbf{S}}_{New}) = \text{diag}(\hat{\Sigma}_{ALCE}) - \text{diag}(\hat{\mathbf{L}}_{New})$ . Then, supposing that all the assumptions of Theorems 1 and 2 hold, the minimum

$$\min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{r}), \mathbf{S} \in \hat{\mathcal{A}}(\hat{s})} \|\Sigma_n - (\mathbf{L} + \mathbf{S})\|_F^2$$

conditioning on  $\mathbf{Y}_{pre}$  and  $\mathbf{Z}_{pre}$  is achieved with probability  $1 - O(1/\min(p, n)^2)$  if and only if  $\mathbf{L} = \hat{\mathbf{L}}_{New}$  and  $\mathbf{S} = \hat{\mathbf{S}}_{New}$ .

Theorem 3 essentially states that the sample total loss (17) is minimized if we un-shrink the eigenvalues of  $\hat{\mathbf{L}}_{ALCE}$  (re-adding the threshold  $\check{\psi}$ ). We call the resulting overall estimator  $\hat{\Sigma}_{New} = \hat{\mathbf{L}}_{New} + \hat{\mathbf{S}}_{New}$  UNALCE (UNshrunk ALgebraic Covariance Estimator). We stress the importance of conditioning on  $\mathbf{Y}_{pre}$  and  $\mathbf{Z}_{pre}$ . Since  $\mathbf{Y}_{pre}$  and  $\mathbf{Z}_{pre}$  are the matrices minimizing  $0.5\|\Sigma_n - (\mathbf{L} + \mathbf{S})\|_F^2$  and  $\check{\psi}\|\mathbf{L}\|_* + \check{\rho}\|\mathbf{S}\|_1$  jointly considered (see (4)), our finite-sample re-optimization step aims to re-compute  $\min \|\Sigma_n - (\mathbf{L} + \mathbf{S})\|_F^2$ , once the effect of the composite penalty  $\check{\psi}\|\mathbf{L}\|_* + \check{\rho}\|\mathbf{S}\|_1$  has been removed.

As shown in Appendix A, problem (17) can be decomposed into two problems: one involving  $\mathbf{L}$  and the other involving  $\mathbf{S}$  (see (A.8)). The problem in  $\mathbf{L}$  is solved by the covariance matrix formed by the top  $\hat{r}$  principal components of  $\mathbf{Y}_{pre}$ , which belongs by construction to  $\hat{\mathcal{B}}(\hat{r})$  and is equal to  $\hat{\mathbf{U}}_{ALCE} (\hat{\mathbf{D}}_{ALCE} + \check{\psi} \mathbf{I}_r) \hat{\mathbf{U}}_{ALCE}^\top = \hat{\mathbf{L}}_{UNALCE}$ . The problem in  $\mathbf{S}$  collapses to the problem in  $\mathbf{L}$  under the prescribed assumptions on the off-diagonal elements of  $\hat{\mathbf{S}}_{UNALCE}$  (which ensures  $\hat{\mathbf{S}}_{UNALCE} \in \hat{\mathcal{A}}(\hat{s})$ ) and on the diagonal elements of  $\hat{\Sigma}_{UNALCE}$ . The new estimate of the diagonal of  $\mathbf{S}^*$  is simply the difference between the diagonal of the original  $\hat{\Sigma}_{ALCE}$  and that of the newly computed  $\hat{\mathbf{L}}_{UNALCE}$ . Note that our re-optimization step depends entirely on  $\Sigma_n$ , as  $\mathbf{Y}_{pre}$  and  $\mathbf{Z}_{pre}$  are functions of  $\Sigma_n$ .

Fig. 1 reports the proportion of latent variance  $\hat{\theta} = (\sum_{i=1}^p \hat{\mathbf{L}}_{ii}) / (\sum_{i=1}^p \hat{\Sigma}_{ii})$  estimated by UNALCE and ALCE for three selected latent eigenvalue thresholds  $\psi$  over twenty sparsity thresholds  $\rho$ . We note that  $\hat{\theta}$  gets systematically closer to the true  $\theta = 0.7$  for  $\hat{\Sigma}_{UNALCE}$  with respect to  $\hat{\Sigma}_{ALCE}$  for all threshold pairs, and the performance difference is proportional to  $\psi$ . The sample used is drawn from our Setting 1 (see the supplementary material, Section 2 for more details).

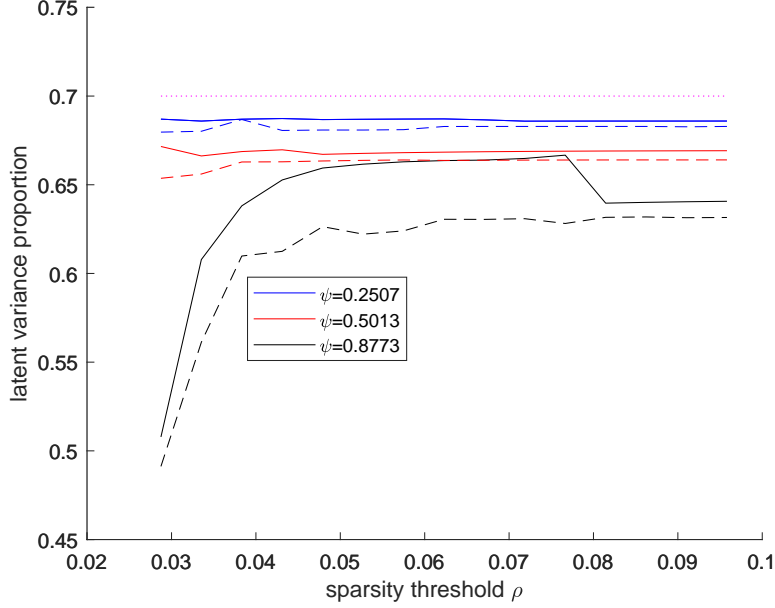
#### 4.3. Consequences

Four consequences of Theorem 3 are reported in Corollary 2.

**Corollary 2.** Under the assumptions of Theorem 3, the differences between the total losses from the target in the spectral norm of  $\hat{\mathbf{L}}_{ALCE}$  and  $\hat{\mathbf{L}}_{UNALCE}$  and of  $\hat{\mathbf{S}}_{ALCE}$  and  $\hat{\mathbf{S}}_{UNALCE}$  are strictly positive and bounded with probability  $1 - O(1/\min(p, n)^2)$  by  $\check{\psi}$ . The differences between the total losses from the target in the Frobenius norm of  $\hat{\mathbf{L}}_{ALCE}$  and  $\hat{\mathbf{L}}_{UNALCE}$  and of  $\hat{\mathbf{S}}_{ALCE}$  and  $\hat{\mathbf{S}}_{UNALCE}$  are strictly positive and bounded with probability  $1 - O(1/\min(p, n)^2)$  by  $\sqrt{r}\check{\psi}$ .

Two further relevant consequences of Theorem 3 are reported in Corollary 3.

**Corollary 3.** Under the assumptions of Theorem 3, the difference between the sample total losses in the spectral norm of  $\hat{\Sigma}_{ALCE}$  and  $\hat{\Sigma}_{UNALCE}$  is strictly positive and bounded with probability  $1 - O(1/\min(p, n)^2)$  by  $\check{\psi}$ . The difference between the sample total losses in the Frobenius norm of  $\hat{\Sigma}_{ALCE}$  and  $\hat{\Sigma}_{UNALCE}$  is strictly positive and bounded with probability  $1 - O(1/\min(p, n)^2)$  by  $\sqrt{r}\check{\psi}$ .



**Fig. 1** This figure shows the proportion of latent variance  $\theta$  estimated by UNALCE (solid line) and ALCE (dashed line) in correspondence to three selected values of latent eigenvalue thresholds  $\psi$  across twenty values of sparsity thresholds  $\rho$ . The reference value  $\theta = 0.7$  is represented as a dotted line. The used sample is drawn from Setting 1 (see Section 2 in the supplementary material).

The following result compares the losses of  $\hat{\Sigma}_{UNALCE}$  and  $\hat{\Sigma}_{ALCE}$  from the target  $\Sigma^*$ .

**Theorem 4.** *Under the assumptions of Theorem 3, the difference between the total losses from the target  $\Sigma^*$  in the spectral norm of  $\hat{\Sigma}_{ALCE}$  and  $\hat{\Sigma}_{UNALCE}$  is strictly positive and bounded with probability  $1 - O(1/\min(p, n)^2)$  by  $\check{\psi}$ . The difference between the total losses from the target in the Frobenius norm of  $\hat{\Sigma}_{ALCE}$  and  $\hat{\Sigma}_{UNALCE}$  is strictly positive and bounded with probability  $1 - O(1/\min(p, n)^2)$  by  $\sqrt{r}\check{\psi}$ .*

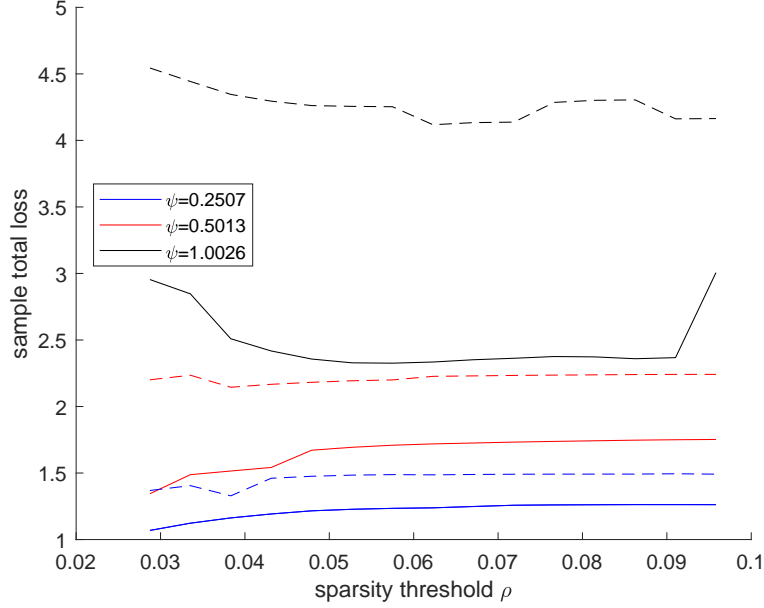
The rationale of the reported claims is as follows. We accept to pay the price of a non-optimal solution in terms of nuclear norm (we allow to increase  $\|\hat{\mathbf{L}}\|_*$  by  $r\check{\psi}$ ), but we have a better fitting performance for the whole covariance matrix, decrementing the squared Frobenius loss of  $\hat{\Sigma}$  by a quantity bounded by  $r\check{\psi}^2$ . The  $l_1$  norm of  $\mathbf{S}$  excluding the diagonal,  $\|\hat{\mathbf{S}}\|_{1,off}$ , is unvaried, while the norm  $\|\hat{\mathbf{S}}\|_1$  (including the diagonal) is decreased by a quantity bounded by  $\sqrt{r}\check{\psi}$ .

In Fig. 2 and 3, we report the differences between the sample total losses and the total losses of ALCE and UNALCE computed over the same sample of Fig. 1 for three selected latent eigenvalue thresholds  $\psi$  over twenty sparsity thresholds  $\rho$ . We note that the gain is relevant for UNALCE over all threshold pairs, is proportional to  $\psi$ , and never overcomes its theoretical maximum  $\sqrt{r}\psi$  (in Fig. 2 and 3  $r = 4$ ). We stress that the gain is positive for each prescribed threshold pair  $(\check{\psi}, \check{\rho})$ , satisfying the conditions of Theorem 1, while the overall performance also depends on the threshold selection criterion (see Section 5).

A consequence of Corollaries 2 and 3 and Theorem 4 is that we can reduce the numerical instability of our estimates as much as possible in terms of the expected variance of estimated eigenvalues. In fact, defining  $\mu_{\mathbf{L}} = E(\text{tr}(\mathbf{L})/p)$ ,  $\mu_{\mathbf{S}} = E(\text{tr}(\mathbf{S})/p)$ , and  $\mu_{\Sigma} = E(\text{tr}(\Sigma)/p)$  and recalling the following equalities according to [24]

$$\frac{1}{p}E \left\{ \sum_{i=1}^p (\hat{\lambda}_{L,i} - \mu_L)^2 \mid \Sigma_n \right\} = \frac{1}{p} \sum_{i=1}^p (\lambda_{L,i} - \mu_L)^2 + E(\|\hat{\mathbf{L}} - \mathbf{L}^*\|^2 \mid \Sigma_n),$$

$$\frac{1}{p}E \left\{ \sum_{i=1}^p (\hat{\lambda}_{S,i} - \mu_S)^2 \mid \Sigma_n \right\} = \frac{1}{p} \sum_{i=1}^p (\lambda_{S,i} - \mu_S)^2 + E(\|\hat{\mathbf{S}} - \mathbf{S}^*\|^2 \mid \Sigma_n),$$



**Fig. 2** This figure shows the sample total losses of  $\hat{\Sigma}_{UNALCE}$  (solid line) and  $\hat{\Sigma}_{ALCE}$  (dashed line) in one sample drawn from Setting 1 (see Section 2 in the supplementary material) in correspondence to three selected values of latent eigenvalue thresholds  $\psi$  across twenty values of sparsity thresholds  $\rho$ .

$$\frac{1}{p} \mathbb{E} \left\{ \sum_{i=1}^p (\hat{\lambda}_{\Sigma, i} - \mu_{\Sigma})^2 \mid \Sigma_n \right\} = \frac{1}{p} \sum_{i=1}^p (\lambda_{\Sigma, i} - \mu_L)^2 + \mathbb{E}(\|\hat{\Sigma} - \Sigma^*\|^2 \mid \Sigma_n),$$

we note that, under the assumptions of Theorem 3 the UNALCE estimated eigenvalues are maximally concentrated with probability  $1 - O(1/\min(p, n)^2)$ , because  $\hat{\mathbf{L}}_{UNALCE} = \min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{\rho})} (\|\mathbf{L} - \mathbf{L}^*\|^2 \mid \Sigma_n)$ ,  $\hat{\mathbf{S}}_{UNALCE} = \min_{\mathbf{S} \in \hat{\mathcal{A}}(\hat{\delta})} (\|\mathbf{S} - \mathbf{S}^*\|^2 \mid \Sigma_n)$ ,  $\hat{\Sigma}_{UNALCE} = \min_{\Sigma = \mathbf{L} + \mathbf{S}, \mathbf{L} \in \hat{\mathcal{B}}(\hat{\rho}), \mathbf{S} \in \hat{\mathcal{A}}(\hat{\delta})} (\|\Sigma - \Sigma^*\|^2 \mid \Sigma_n)$ , given the finite sample and a threshold pair  $(\check{\psi}, \check{\rho})$  satisfying the conditions of Theorem 1.

The following Corollary extends our framework to the performance of the inverse covariance estimate  $\hat{\Sigma}_{UNALCE}^{-1}$ .

**Corollary 4.** *Under the assumptions of Theorem 3, the difference between the total losses from the target in the spectral norm of  $\hat{\Sigma}_{ALCE}^{-1}$  and  $\hat{\Sigma}_{UNALCE}^{-1}$  is strictly positive and bounded with probability  $1 - O(1/\min(p, n)^2)$  by  $\check{\psi}$ . The difference between the total losses from the target in the Frobenius norm of  $\hat{\Sigma}_{ALCE}^{-1}$  and  $\hat{\Sigma}_{UNALCE}^{-1}$  is strictly positive and bounded with probability  $1 - O(1/\min(p, n)^2)$  by  $\sqrt{r}\check{\psi}$ .*

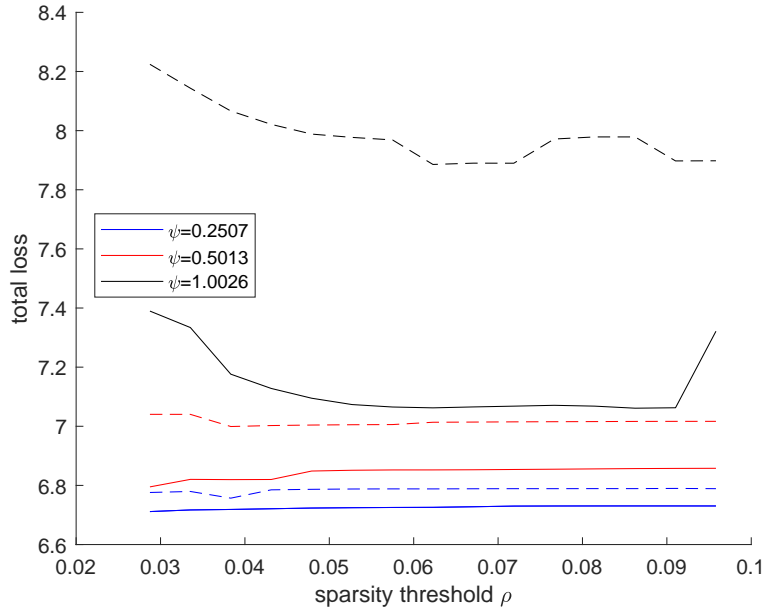
Finally, we study how the necessary conditions to ensure the positive definiteness of UNALCE estimates evolve with respect to the ALCE ones. The following Corollary holds.

**Corollary 5.**  *$\hat{\mathbf{L}}_{UNALCE}$  is positive semi-definite if  $\lambda_r(\mathbf{L}^*) > C_2 p^\alpha - \check{\psi}$ .  $\hat{\mathbf{S}}_{UNALCE}$  is positive definite if  $\lambda_p(\mathbf{S}^*) > \phi_S + r\check{\psi}/p$ .  $\hat{\Sigma}_{UNALCE}$  is positive definite if  $\lambda_p(\Sigma^*) > \phi + r\check{\psi}/p$ .*

Theorems 1, 2, and 3 and Corollaries 1 and 5 ensure the algebraic and parametric consistency of the pair  $(\hat{\mathbf{L}}_{UNALCE}, \hat{\mathbf{S}}_{UNALCE})$  in the sense of Definitions 1 and 2.

## 5. A new model selection criterion: MC

In empirical applications, the selection of thresholds  $\psi$  and  $\rho$  in equation (4) requires a model selection criterion consistent with the described estimation method and the consistency norm  $g_\gamma$  (recall that



**Fig. 3** This figure shows the total losses of  $\hat{\Sigma}_{UNALCE}$  (solid line) and  $\hat{\Sigma}_{ALCE}$  (dashed line) drawn from Setting 1 (see Section 2 in the supplementary material) in correspondence to three selected values of latent eigenvalue thresholds  $\psi$  across twenty values of sparsity thresholds  $\rho$ .

$g_\gamma = \max(\|\mathbf{S} - \mathbf{S}^*\|_\infty / \gamma, \|\mathbf{L} - \mathbf{L}^*\|_2)$ . Our aim is to detect the optimal threshold pair  $(\psi, \rho)$  with respect to the spikiness/sparsity trade-off. In order to exploit  $g_\gamma$  with model selection purposes, we need to make the two terms comparable, i.e. the need of rescaling both arguments of  $g_\gamma$  arises.

400 First, we note that if all the estimated latent eigenvalues are equal, then we have  $\|\hat{\mathbf{L}}\|_* = \hat{r}\|\hat{\mathbf{L}}\|_2$ . As the condition number of  $\hat{\mathbf{L}}$  increases, we have  $\hat{r}\|\hat{\mathbf{L}}\|_2 > \|\hat{\mathbf{L}}\|_*$ . Consequently, the quantity  $\hat{r}\|\hat{\mathbf{L}}\|_2$  acts as a penalization term against the presence of too small eigenvalues. Analogously, if  $\hat{\mathbf{S}}$  is diagonal, it holds  $\|\hat{\mathbf{S}}\|_\infty = \|\hat{\mathbf{S}}\|_{1,v}$ . As the number of non-zeros increases, it holds  $\|\hat{\mathbf{S}}\|_{1,v} > \|\hat{\mathbf{S}}\|_\infty$ . Therefore, the quantity  $\|\hat{\mathbf{S}}\|_{1,v}$  acts as a penalization term against the presence of too many non-zeros.

405 In order to compare the magnitude of the two quantities, we divide the former by the trace of  $\hat{\mathbf{L}}$ , estimated by  $\hat{\theta}\text{trace}(\mathbf{\Sigma}_n)$ , and the latter by the trace of  $\hat{\mathbf{S}}$ , estimated by  $(1 - \hat{\theta})\text{trace}(\mathbf{\Sigma}_n)$ . Our maximum criterion  $MC$  can be therefore defined as follows:

$$MC(\psi, \rho) = \max \left\{ \frac{\hat{r}\|\hat{\mathbf{L}}\|_2}{\hat{\theta}}, \frac{\|\hat{\mathbf{S}}\|_{1,v}}{\gamma(1 - \hat{\theta})} \right\}, \quad (18)$$

where  $\gamma = \rho/\psi$  is the ratio between the sparsity and the latent eigenvalue threshold.

410  $MC$  criterion is by definition mainly intended to catch the proportion of variance explained by the factors. For this reason, it tends to choose quite sparse solutions with a small number of non-zeros and a small proportion of absolute residual covariance, unless the non-zero entries of  $\hat{\mathbf{S}}$  are prominent, as Theorem 2 prescribes. The  $MC$  method performs considerably better than the usual cross-validation using  $H$ -fold Frobenius loss (cf. [27]). In fact, minimizing a loss based on a sample approximation such as the Frobenius one causes the parameter  $\hat{\theta}$  to be significantly shrunk. The threshold setting which shows a minimum for  $MC$  criterion (given that the estimate  $\hat{\Sigma}$  is positive definite) is the best in terms of composite penalty, considering the latent low rank and sparse structure simultaneously.

415 Since we apply  $MC$  criterion to choose thresholds both for UNALCE and ALCE, we observe that the overall performance of the two methods is very similar, even if a little margin in favour of UNALCE is always present (see Section 2.2 in the supplementary material for more details).

**Table 1**

Supervisory data: this table reports the main results of  $\hat{\Sigma}_{UNALCE}$  and  $\hat{\Sigma}_{POET}$  estimated on a selection of 382 supervisory indicators referred to 365 Euro Area banks with reference date Q4,2014. In particular,  $\hat{r}$  is the latent rank,  $\hat{s}$  is the number of recovered off-diagonal non-zeros in  $\hat{\Sigma}$ ,  $\hat{\pi}_{\hat{s}}$  is the percentage of recovered non-zeros over the number of off-diagonal elements in  $\hat{\Sigma}$ ,  $\hat{\theta} = (100 \sum_{i=1}^p \hat{\mathbf{L}}_{ii}) / (\sum_{i=1}^p \hat{\Sigma}_{ii})$  is the percentage of latent variance,  $\hat{\rho}_{\hat{s}} = (100 \sum_{i=1}^p \sum_{j=i+1}^p |\hat{\Sigma}_{ij}|) / (\sum_{i=1}^p \sum_{j=i+1}^p |\hat{\mathbf{L}}_{ij}|)$  is the percentage of absolute residual covariance,  $\|\hat{\Sigma} - \Sigma_n\|_F$  is the sample total loss,  $cond(\hat{\Sigma}) = \lambda_{max}(\hat{\Sigma}) / \lambda_{min}(\hat{\Sigma})$  is the condition number of the overall estimate,  $cond(\hat{\mathbf{S}}) = \lambda_{max}(\hat{\mathbf{S}}) / \lambda_{min}(\hat{\mathbf{S}})$  is the condition number of the sparse estimate, and  $cond(\hat{\mathbf{L}}) = \lambda_{max}(\hat{\mathbf{L}}) / \lambda_{min}(\hat{\mathbf{L}})$  is the condition number of the low rank estimate.

Supervisory data	UNALCE	POET
$\hat{r}$	6	6
$\hat{s}$	328	404
$\hat{\pi}_{\hat{s}}$	0.45	0.56
$\hat{\theta}$	32.47	61.23
$\hat{\rho}_{\hat{s}}$	16.87	1.61
$\ \hat{\Sigma} - \Sigma_n\ _F$	0.0337	0.0645
$cond(\hat{\Sigma})$	$6.35e + 15$	$6.68e + 15$
$cond(\hat{\mathbf{S}})$	$2.78e + 15$	$1.11e + 15$
$cond(\hat{\mathbf{L}})$	3.1335	2.5625

## 6. A Euro Area banking data example

This Section provides a real example on the performance of POET and UNALCE based on a selection of Euro Area banking data. We acknowledge the assistance of the European Central Bank, where one of the authors spent a semester as a PhD trainee, in providing access to high-level banking data. Here, we use the covariance matrix computed on a selection of balance sheet indicators relative to the last quarter of 2014 for some of the most relevant Euro Area banks. The overall number of banks (our sample size) is  $n = 365$ . These indicators are the ones needed for supervisory reporting, and they include capital and financial variables.

The chosen raw variables were rescaled to the total asset of each bank. Then, a screening based on the importance of each variable, intended as the absolute amount of correlation with all the other variables, was performed in order to remove identities. The resulting very sparse data matrix contains  $p = 382$  variables; here, we are in the typical  $p > n$  case, where the sample covariance matrix is completely ineffective. We plot sample eigenvalues in the left panel of Fig. 4.

UNALCE estimation method selects a solution with a latent rank equal to  $\hat{r} = 6$ . The number of surviving non-zeros in the sparse component is  $\hat{s} = 328$ , which corresponds to a percentage  $\hat{\pi}_{\hat{s}} = 0.45\%$  of 72771 off-diagonal elements. Conditioning properties are inevitably very bad. In order to obtain a POET estimate, we exploit the algebraic consistency of  $\hat{\Sigma}_{UNALCE}$ , setting the rank to  $\hat{r} = 6$ , and we perform cross-validation for threshold selection. The number of non-zeros estimated by POET is  $\hat{s} = 404$  ( $\hat{\pi}_{\hat{s}} = 0.56\%$ ). The results of both methods are reported in Table 1.

Apparently, one could argue that POET estimate is better; the estimated percentage of latent variance  $\hat{\theta}$  is 61.23%, and the percentage of absolute residual covariance  $\hat{\rho}_{\hat{s}}$  is 1.61%. On the contrary, UNALCE method outputs  $\hat{\theta} = 32.47\%$  and  $\hat{\rho}_{\hat{s}} = 16.87\%$ . A relevant question thus arises: how much is the true percentage of variance explained by the factors? In fact, such a large percentage of latent variance, which depends on the use of the first six principal components, causes the absolute residual covariance percentage to be very low. Therefore, POET procedure gives *a priori* preference to the low rank part. This pattern does not change even if we choose a lower value for the rank.

On the contrary, the UNALCE estimate, which depends on a double-step iterative thresholding procedure, requires a larger magnitude of the non-zero elements in the sparse component. In fact, the percentage of lost covariance during the procedure is here 29.39%. Consequently, via rank/sparsity detection, UNALCE shows better approximation properties compared to POET; its Sample Total Loss is sensibly lower than that of POET (0.337 VS 0.645).

For UNALCE, the covariance structure appears so complex that a relevant percentage of absolute residual covariance is present. This allows us to explore the importance of variables, i.e. to explore which variables have the largest systemic power (the most relevant communality) or the largest idiosyncrasy (the most relevant residual variance).

In the right panel of Fig. 4, we plot in descending order the degree of each variable with respect to the estimated residual component  $\hat{\Sigma}_{UNALCE}$ . The degree of the variable  $i$  with respect to a  $p$ -dimensional covariance matrix  $\mathbf{M}$  is



**Table 2**

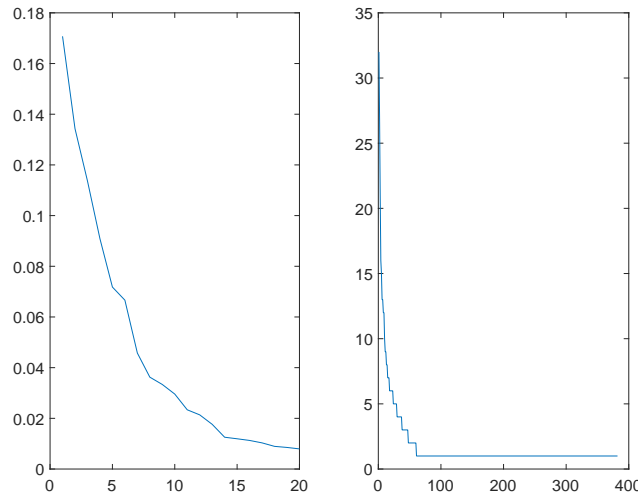
Supervisory data: this table reports the top six variables by estimated degree with respect to  $\hat{\mathbf{S}}_{UNALCE}$ , defined for variable  $i$  as  $deg_{\hat{\mathbf{S}}_{UNALCE},i} = \sum_{j=1}^p \mathbb{1}(\hat{\mathbf{S}}_{UNALCE,ij} \neq 0)$ ,  $i \in \{1, \dots, 382\}$ . This measure counts how many variables are related to variable  $i$  that their estimated residual covariance is not null. Therefore, the reported variables are the most connected with all the others.

Supervisory indicator	Estimated degree
Financial assets designated at fair value through profit or loss	34
Central banks, Impaired assets [gross carrying amount]	25
Credit institutions, Collective allowances for incurred but not reported losses	20
Other financial corporations, Collective allowances for incurred but not reported losses	19
Cash, Cash balances at central banks and other demand deposits	16
Other financial corporations, Specific allowances for financial assets, collectively estimated	16

defined as

$$deg_{\mathbf{M},i} = \sum_{j=1}^p \mathbb{1}(\mathbf{M}_{ij} \neq 0). \quad (19)$$

We observe that only 62 out of 382 variables have at least one non-zero residual covariance with other variables.



**Fig. 4** Supervisory data: the left panel of this figure shows the twenty largest eigenvalues of the sample covariance matrix computed on a selection of 382 supervisory indicators referred to 365 Euro Area banks with reference date Q4,2014. The right panel of this figure plots in descending order the estimated degree of each supervisory indicator with respect to  $\hat{\mathbf{S}}_{UNALCE}$ , defined for variable  $i$  as  $deg_{\hat{\mathbf{S}}_{UNALCE},i} = \sum_{j=1}^p \mathbb{1}(\hat{\mathbf{S}}_{UNALCE,ij} \neq 0)$ ,  $i \in \{1, \dots, 382\}$ .

In Table 2, we report the top six variables by estimated degree. These variables are related to the largest number of other variables. They are mainly credit-based indicators: financial assets through profit and loss, impaired assets, allowances to credit institutions and non-financial corporations, and cash.

In Table 3, we report the top five variables by estimated communality, defined for variable  $i$  as

$$comm_i = \frac{\hat{\mathbf{L}}_{UNALCE,ii}}{\hat{\mathbf{\Sigma}}_{UNALCE,ii}}, \quad i \in \{1, \dots, 382\}. \quad (20)$$

The results are very meaningful; the most systemic variables are debt securities, loans and advances to households, specific allowances for financial assets, and advances which are not loans to central banks. All these are fundamental indicators for banking supervision because they represent key indicators for the assessment of bank performance.

**Table 3**

Supervisory data: this table reports the top six variables by estimated communality via UNALCE, defined for variable  $i$  as  $comm_i = \hat{\mathbf{L}}_{UNALCE,ii} / \hat{\mathbf{\Sigma}}_{UNALCE,ii}$ . Therefore, the reported variables have a strong explanation power for banking supervision.

Supervisory indicator	Estimated communality
Debt securities	0.8414
Households, Carrying amount	0.8210
Non-financial corporations, Specific allowances for financial assets	0.8110
Loans and advances, Specific allowances for financial assets, collectively estimated	0.7592
Advances that are not loans, Central banks	0.7439

**Table 4**

Supervisory data: this table reports the top six variables by estimated idiosyncrasy via UNALCE, defined for variable  $i$  as  $idio_i = \hat{\mathbf{S}}_{UNALCE,ii} / \hat{\mathbf{\Sigma}}_{UNALCE,ii}$ . Therefore, the reported variables have a marginal explanation power for banking supervision.

Supervisory indicator	Estimated idiosyncrasy
Credit card debt, Central banks	0.9995
Other collateralized loans, Other financial corporations	0.9986
Equity instruments, Central banks, Carrying amount	0.9971
Equity instruments, Other financial corporations, Carrying amount	0.9970
General governments, Carrying amount of unimpaired assets	0.9970

In Table 4, we report the top five variables by estimated idiosyncrasy, defined for variable  $i$  as

$$idio_i = \frac{\hat{\mathbf{S}}_{UNALCE,ii}}{\hat{\mathbf{\Sigma}}_{UNALCE,ii}}, \quad i \in \{1, \dots, 382\}. \quad (21)$$

We note that those indicators have a marginal power in the explanation of the common covariance structure and are much less relevant for supervisory analysis than the previous five.

In conclusion, our UNALCE procedure offers a more realistic view of the underlying covariance structure of a set of variables, allowing a larger part of covariance to be explained by the residual sparse component compared to POET.

## 7. Conclusions

In this work, we propose an estimator for large covariance matrices which are assumed to be the sum of a low rank and a sparse component. Estimation is performed by solving a regularization problem where the objective function is composed of a smooth Frobenius loss and a non-smooth composite penalty, which is the sum of the nuclear norm of the low rank component and the  $l_1$  norm of the sparse component. Our estimator is called UNALCE (UNshrunk ALgebraic Covariance Estimator). UNALCE provides consistent recovery of the low rank and the sparse component, as well as of the overall covariance matrix, under a generalized assumption of spikiness of latent eigenvalues and sparsity of the residual component. Thanks to the addition of an un-shrinkage step of the estimated latent eigenvalues, we can also improve numerical properties and minimize the overall loss given the finite sample and the threshold pair, while preserving algebraic consistency. In addition, we can overcome the restrictive condition  $p \leq n$ .

Moreover, in this paper, we also compare UNALCE and POET (Principal Orthogonal complEment Thresholding, see [15]), an asymptotic estimator which performs principal component analysis in order to recover the low rank component and uses a thresholding algorithm to recover the sparse component. Both estimators provide the usual parametric consistency, while UNALCE also provides the algebraic consistency of the estimate, i.e. the rank and position of residual non-zeros are simultaneously recovered by the solution algorithm. This automatic recovery is a crucial advantage compared to POET; the latent rank, in fact, is automatically selected and the sparsity pattern of the residual component is recovered considerably better.

In particular, we prove that UNALCE can effectively recover the covariance matrix even in the presence of spiked eigenvalues with rate  $O(p)$ , exactly as POET estimator does, allowing  $n$  to be as small as  $O(p^{1.5\delta})$ , where the maximum number of non-zeros per row in the sparse component is proportional to  $O(p^\delta)$ . Moreover, we prove that the recovery

is actually effective even if the eigenvalues show an intermediate degree of spikiness  $p^\alpha$ ,  $\alpha \in [0, 1]$ . The resulting loss is bounded accordingly to  $p^\alpha$ , and all latent eigenvalues are recovered under the assumption  $\delta < \alpha$ . In this way, we obtain a generalised estimator of large covariance matrices by low rank plus sparse decomposition.

490 A real example of a set of Euro Area banking data shows that our tool is particularly useful for mapping the covariance structure among variables even in a large dimensional context. The variables with the largest systemic power, i.e. the ones mostly affecting the common covariance structure, can be identified, as well as the variables having the largest idiosyncratic power, that is, the ones characterized by the largest residual variance. In addition, the variables showing the largest idiosyncratic covariances can be identified. Particular forms of the residual covariance  
495 pattern can thus be detected, if present.

Our research may provide a basis for possible future developments in many directions. In the time series context, this procedure can be potentially extended to covariance matrix estimation under dynamic factor models. Another fruitful extension of our procedure is related to the spectral matrix estimation context. Finally, this tool can be potentially used in the Big data context, where both the dimension and the sample size are very large. This poses  
500 new computational and theoretical challenges, the solution of which is crucial in order to further extend the power of statistical modelling and its effectiveness in detecting patterns and underlying drivers of real phenomena.

## Appendix A. Proofs

### Proof of Theorem 1

First, we note that the deterministic analysis needed to ensure the identifiability of the matrix varieties  $\mathcal{B}(r)$  and  
505  $\mathcal{A}(s)$  is directly inherited by [26]. In fact, Propositions 12, 13, and 14 in [26] may be directly applied to our setting, provided that the assumption  $\xi(T(\mathbf{L}^*))\mu(\Omega(\mathbf{S}^*)) \leq 1/54$  and the conditions  $\lambda_r(\mathbf{L}^*) > C_2\psi/\xi^2(T)$  and  $\rho = \gamma\psi$  hold with  $\gamma \in [9\xi(T), 1/(6\mu(\Omega))]$ . In that case, it descends from the mentioned Propositions that  $g_\gamma(\hat{\mathbf{S}} - \mathbf{S}^*, \hat{\mathbf{L}} - \mathbf{L}^*)$  is bounded,  $\hat{\mathbf{L}} \in T(\mathbf{L}^*)$ ,  $\hat{\mathbf{S}} \in \Omega(\mathbf{S}^*)$ , and  $\text{rank}(\hat{\mathbf{L}}) = \text{rank}(\mathbf{L}^*)$ .

We stress that the remaining assumptions of Theorem 1 are not needed for this purpose. We also remark that  
510 parametric and rank consistency are not affected even if Assumption 6 and the condition  $S_{\min,off} > (C_3\psi)/\mu(\Omega)$  do not hold. The only consequence of that is that some residual non-zeros are not recovered (cf. [10], Corollary D.4 and D.6, and Proposition D.5 for more details).

Hence, we now focus on probabilistic analysis. Recalling that  $\Sigma_n = (n-1)^{-1} \sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^\top$  and  $\mathbf{x}_k = \mathbf{B}\mathbf{f}_k + \epsilon_k$ , where  $\mathbf{f}_k$  and  $\epsilon_k$ ,  $k \in \{1, \dots, n\}$ , are respectively the vectors of factor scores and residuals for each observation, we can decompose the error matrix  $\mathbf{E}_n = \Sigma_n - \Sigma^*$  in four components as follows (cf. [15]):

$$\mathbf{E}_n = \Sigma_n - \Sigma^* = \hat{\mathbf{D}}_1 + \hat{\mathbf{D}}_2 + \hat{\mathbf{D}}_3 + \hat{\mathbf{D}}_4,$$

where:

$$\hat{\mathbf{D}}_1 = \left( n^{-1} \mathbf{B} \sum_{k=1}^n \mathbf{f}_k \mathbf{f}_k^\top - I_r \right) \mathbf{B}^\top, \hat{\mathbf{D}}_2 = n^{-1} \sum_{k=1}^n (\epsilon_k \epsilon_k^\top - \mathbf{S}^*), \hat{\mathbf{D}}_3 = n^{-1} \mathbf{B} \sum_{k=1}^n \mathbf{f}_k \epsilon_k^\top, \hat{\mathbf{D}}_4 = \hat{\mathbf{D}}_3^\top.$$

Following [15], we note that

$$\|\hat{\mathbf{D}}_1\|_2 \leq \left\| \frac{1}{n} \sum_{k=1}^n \mathbf{f}_k \mathbf{f}_k^\top - \mathbb{E}(\mathbf{f}_k \mathbf{f}_k^\top) \right\|_2 \|\mathbf{B}\mathbf{B}^\top\|_2 \leq r \left\| \frac{1}{n} \sum_{k=1}^n \mathbf{f}_k \mathbf{f}_k^\top - \mathbb{E}(\mathbf{f}_k \mathbf{f}_k^\top) \right\|_\infty p^\alpha$$

where the second inequality depends on standard matrix norm properties and Assumption 1.

515 Under Assumption 3, we can apply Lemma 4 in [15], which claims

$$\max_{i,j \leq r} \left| \frac{1}{n} \sum_{k=1}^n \mathbf{f}_{ik} \mathbf{f}_{jk} - \mathbb{E}(\mathbf{f}_{ik} \mathbf{f}_{jk}) \right| \leq C \frac{1}{\sqrt{n}}, \quad (\text{A.1})$$

with probability  $1 - O(1/n^2)$ . Consequently, we obtain

$$\|\hat{\mathbf{D}}_1\|_2 \leq Cr \sqrt{\frac{1}{n}} p^\alpha \leq Cp^\alpha \sqrt{\frac{1}{n}} \quad (\text{A.2})$$

because Assumption 5 prescribes that  $r = \delta_3 \ln p$  and  $\ln p = o(n)$ .

Consider now the uniformity class of sparse matrices in [5]:

$$\left\{ \mathbf{S}^* : \mathbf{S}_{ii}^* \leq c_3, \quad \max_{i \leq p} \sum_{j \leq p} \mathbb{1}(\mathbf{S}_{ij}^* \neq 0) \leq c_0(p), \quad \forall i \right\}.$$

Therein, our Assumption 4 holds with  $\delta = 1$  (assuming  $q = 0$ ). Therefore, it is possible to write

$$\lambda_{\max}(\mathbf{S}^*) \leq \max_{i \leq p} \sum_{j \leq p} \mathbb{1}(\mathbf{S}_{ij}^* \neq 0) \leq c_3 c_0(p),$$

520 since the quantity  $c_0(p)$  is constant with respect to  $p$ . Consequently, Lemma A.3 on p. 220 in [6] can be applied, which leads to the claim

$$\max_{i, j \leq p} \left| \frac{1}{n} \sum_{k=1}^n \epsilon_{ik} \epsilon_{jk} - \mathbb{E}(\epsilon_{ik} \epsilon_{jk}) \right| \leq C \sqrt{\frac{\ln p}{n}}, \quad (\text{A.3})$$

that holds with probability  $1 - O(1/p^2)$ .

Under Assumptions 2 and 4, however, the quantity  $c_0(p)$  must be replaced by  $c_0(p^\delta)$ ,  $\delta \leq 0.5$ . Consequently, with respect to  $p$ , the rate in (A.3) is now too strong. Therefore, applying the recalled Lemma A.3 in [6], the following claim holds with probability  $1 - O(1/p^2)$ :

$$\|\hat{\mathbf{D}}_2\|_\infty = \max_{i, j \leq p} \left| \frac{1}{n} \sum_{k=1}^n \epsilon_{ik} \epsilon_{jk} - \mathbb{E}(\epsilon_{ik} \epsilon_{jk}) \right| \leq C p^{\delta-1} \sqrt{\frac{\ln p}{n}}. \quad (\text{A.4})$$

525 Consequently, by (A.4), we can derive

$$\|\hat{\mathbf{D}}_2\|_2 \leq C p \|\hat{\mathbf{D}}_2\|_\infty = C p^\delta \sqrt{\frac{\ln p}{n}} = C p^\delta \sqrt{\frac{1}{n}}, \quad (\text{A.5})$$

because  $\ln p \ll n$  by Assumption 5.

To conclude, we study the random term  $\max_{i \leq r, j \leq p} \left| n^{-1} \sum_{k=1}^n \mathbf{f}_{ik} \epsilon_{jk} \right|$ . We know from Lemma 3 in [15] that this term has exponential-type tails, due to Assumption 3. Thus, we only need to study how its standard deviation evolves in our context. We consider the following Cauchy-Schwarz inequality:

$$\max_{i \leq r, j \leq p} \left| \frac{1}{n} \sum_{k=1}^n \mathbf{f}_{ik} \epsilon_{jk} \right| \leq C \max_i \sqrt{\hat{V}(\mathbf{f}_i)} \max_j \sqrt{\hat{V}(\epsilon_j)}.$$

From (A.1), we know that  $\max_i \sqrt{\hat{V}(\mathbf{f}_i)} \leq C/\sqrt{n}$  with probability  $1 - O(1/n^2)$ . From (A.4), we know that  $\max_j \sqrt{\hat{V}(\epsilon_j)} \leq C p^{(\delta-1)/2} \sqrt{(\ln p)/n}$  with probability  $1 - O(1/p^2)$ . It follows that with probability  $1 - O(1/\min(p, n)^2)$ , it holds

$$\left\| n^{-1} \sum_{k=1}^n \mathbf{f}_k \epsilon_k^\top \right\|_2 \leq \sqrt{pr} \left\| n^{-1} \sum_{k=1}^n \mathbf{f}_k \epsilon_k^\top \right\|_\infty = C \sqrt{pr} \sqrt{\frac{1}{n}} p^{(\delta-1)/2} \sqrt{\frac{\ln p}{n}}.$$

Exploiting Assumption 5, the bound then becomes  $C p^{\delta/2} \sqrt{n^{-1}}$ , since  $r = \delta_3 \ln p$  and  $\ln p \ll n$ .

Consequently, we obtain with probability  $1 - O(1/p^2)$  the following claim

$$\|\hat{\mathbf{D}}_3\|_2 \leq \left\| \frac{1}{n} \sum_{k=1}^n \mathbf{f}_k \epsilon_k^\top \right\| \times \|\mathbf{B}\| \leq C \left( p^{\frac{\delta}{2}} \sqrt{\frac{1}{n}} \right) \left( p^{\frac{\alpha}{2}} \right) = C p^{\frac{\alpha}{2} + \frac{\delta}{2}} \sqrt{\frac{1}{n}}, \quad (\text{A.6})$$

because  $\|\mathbf{B}\| = O(p^{\alpha/2})$  by Assumption 1.

Putting (A.2), (A.5), and (A.6) together, the following bound is proved with probability  $1 - O(1/\min(p, n)^2)$

$$\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*\|_2 \leq C \frac{p^\alpha}{\sqrt{n}}, \quad (\text{A.7})$$

because  $\delta < \alpha$  from Assumption 2. In fact, if  $\delta > \alpha$ , the condition of Theorem 1  $\lambda_r(\mathbf{L}^*) > (C_2\psi)/\xi^2(T)$  would result in  $\lambda_r(\mathbf{L}^*) > C_2p^\delta$ , thus violating Assumption 1 under Assumption 5.

In other words, the bound (A.7) means  $\|\mathbf{E}_n\|_2 \rightarrow 0 \Leftrightarrow p^\alpha/\sqrt{n} \rightarrow 0$ . Exploiting the basic property  $\|\cdot\|_\infty \leq \|\cdot\|_2$  and the minimum for  $\gamma$  in the range of Theorem 1, we can also write  $\|\mathbf{E}_n\|_\infty \rightarrow 0 \Leftrightarrow \xi(T)p^\alpha/\sqrt{n} \rightarrow 0$ .

535 In order to prove Theorem 1, we observe from [26] that the only probabilistic component of the error norm  $g_\gamma(\hat{\mathbf{S}} - \mathbf{S}^*, \hat{\mathbf{L}} - \mathbf{L}^*)$  is  $g_\gamma(\mathbf{E}_n) = \max(\|\mathbf{E}_n\|_\infty/\gamma, \|\mathbf{E}_n\|_2)$ , which in turn depends on  $\|\mathbf{E}_n\|_2$  and  $\|\mathbf{E}_n\|_\infty$ . Therefore, setting  $\psi = (1/\xi(T))(p^\alpha/\sqrt{n})$ , it follows that the claims

$$g_\gamma(\hat{\mathbf{S}} - \mathbf{S}^*, \hat{\mathbf{L}} - \mathbf{L}^*) \leq C \frac{1}{\xi(T)} \frac{p^\alpha}{\sqrt{n}}, \text{rank}(\hat{\mathbf{L}}) = \text{rank}(\mathbf{L}^*)$$

hold with probability  $1 - O(1/\min(p, n)^2)$  under all the assumptions of Theorem 1. Parametric and rank consistency are thus guaranteed.

#### 540 **Proof of Corollary 1**

We observe that, under Assumption 2, the bound  $\psi = (1/\xi(T))(p^\alpha/\sqrt{n})$  tends to 0 if and only if  $p^{2\alpha+2\delta}/n = o(1)$  as  $\lim_{\gamma \rightarrow \infty} \min_\nu(p_\nu^{2\alpha+2\delta}, n_\nu) = \infty$ . As expected, the absolute bound vanishes only in the small dimensional case ( $n \gg p^{\alpha+\delta} \log(p)$ ).

#### **Proof of Theorem 2**

545 If, in addition to all the assumptions and conditions of Theorem 1, Assumption 6 and the condition  $S_{\min, off} > (C_3\psi)/\mu(\Omega)$  hold, then we can fully apply Corollary D.4, D.6, Proposition D.5, and Lemma D.7 in [10] and conclude that the recovered sparsity pattern is also consistent:  $\text{sign}(\hat{\mathbf{S}}_{ALCE}) = \text{sign}(\mathbf{S}^*)$ .

#### **Proof of Theorem 3**

Conditioning on  $\mathbf{Y}_{pre}$ ,  $\mathbf{Z}_{pre}$ , and  $\boldsymbol{\Sigma}_{pre} = \mathbf{Y}_{pre} + \mathbf{Z}_{pre}$ , we aim to solve

$$\min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{r}), \mathbf{S} \in \hat{\mathcal{A}}(\hat{\delta}), \boldsymbol{\Sigma} = \mathbf{L} + \mathbf{S}} \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_n\|_F^2 = \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{pre} + \boldsymbol{\Sigma}_{pre} - \boldsymbol{\Sigma}_n\|_F^2.$$

By Cauchy-Schwarz inequality, it can be shown that

$$\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{pre} + \boldsymbol{\Sigma}_{pre} - \boldsymbol{\Sigma}_n\|_F^2 \leq \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{pre}\|_F^2 + \|\boldsymbol{\Sigma}_{pre} - \boldsymbol{\Sigma}_n\|_F^2.$$

550  $\boldsymbol{\Sigma}_{pre}$  solves the problem

$$\min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{r}), \mathbf{S} \in \hat{\mathcal{A}}(\hat{\delta}), \boldsymbol{\Sigma} = \mathbf{L} + \mathbf{S}} \|\boldsymbol{\Sigma}_{pre} - \boldsymbol{\Sigma}_n\|_F^2$$

conditioning on the fact that  $\check{\psi}\|\mathbf{L}\|_* + \check{\rho}\|\mathbf{S}\|_1$  is minimum over the same set.

Then, we can write

$$\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{pre}\|_F^2 = \|\mathbf{L} + \mathbf{S} - \mathbf{Y}_{pre} - \mathbf{Z}_{pre}\|_F^2.$$

By Cauchy-Schwarz inequality, it can be shown that

$$\|\mathbf{L} + \mathbf{S} - \mathbf{Y}_{pre} + \mathbf{Z}_{pre}\|_F^2 \leq \|\mathbf{L} - \mathbf{Y}_{pre}\|_F^2 + \|\mathbf{S} - \mathbf{Z}_{pre}\|_F^2. \quad (\text{A.8})$$

Hence,

$$\min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{r}), \mathbf{S} \in \hat{\mathcal{A}}(\hat{\delta}), \boldsymbol{\Sigma} = \mathbf{L} + \mathbf{S}} \|\mathbf{L} + \mathbf{S} - \mathbf{Y}_{pre} + \mathbf{Z}_{pre}\|_F^2 \leq \min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{r})} \|\mathbf{L} - \mathbf{Y}_{pre}\|_F^2 + \min_{\mathbf{S} \in \hat{\mathcal{A}}(\hat{\delta})} \|\mathbf{S} - \mathbf{Z}_{pre}\|_F^2.$$

The problem in  $\mathbf{L}$  is solved by taking out the first  $\hat{r}$  principal components of  $\mathbf{Y}_{pre}$ . By construction, the solution is  $\hat{\mathbf{U}}_{ALCE}(\hat{\mathbf{D}}_{ALCE} + \check{\psi}\mathbf{I}_r)\hat{\mathbf{U}}_{ALCE}^\top = \hat{\mathbf{L}}_{UNALCE}$ . The problem in  $\mathbf{S}$ , assuming that the diagonal of  $\hat{\mathbf{\Sigma}}_{ALCE}$  is given and the off-diagonal elements of  $\hat{\mathbf{S}}$  are invariant, leads to:

$$\begin{aligned} \min_{\mathbf{S} \in \hat{\mathcal{A}}(\hat{\mathcal{S}})} \|\mathbf{S} - \mathbf{Z}_{pre}\|_F^2 &= \min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{r})} \|(\hat{\mathbf{\Sigma}} - \mathbf{L}) - (\mathbf{\Sigma}_{pre} - \mathbf{Y}_{pre})\|_F^2 = \min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{r})} \|(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{pre}) - (\mathbf{L} - \mathbf{Y}_{pre})\|_F^2 \leq \\ &\leq \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{pre}\|_F^2 + \|\mathbf{L} - \mathbf{Y}_{pre}\|_F^2 = \mathbf{B}' + \mathbf{B}''. \end{aligned}$$

The following question now arises: which diagonal elements of  $\mathbf{L}$  ensure the minimum of  $\mathbf{B}' + \mathbf{B}''$ ? Term  $\mathbf{B}'$  is fixed with respect to  $\mathbf{L}$  because we are assuming the invariance of diagonal elements in  $\hat{\mathbf{\Sigma}}$  ( $\text{diag}(\hat{\mathbf{\Sigma}}_{UNALCE}) = \text{diag}(\hat{\mathbf{\Sigma}}_{ALCE})$ ). The minimization of term  $\mathbf{B}''$ , given that  $\text{rank}(\mathbf{L}) = \hat{r}$ , falls back into the previous case, i.e.  $\mathbf{B}''$  is minimum if and only if  $\hat{\mathbf{L}} = \hat{\mathbf{L}}_{UNALCE} = \hat{\mathbf{U}}_{UNALCE}(\hat{\mathbf{D}}_{UNALCE} + \check{\psi}\mathbf{I}_r)\hat{\mathbf{U}}_{UNALCE}^\top$ .

Optimality holds over the Cartesian product of the set of all symmetric positive semi-definite matrices with a rank smaller or equal to  $r$ ,  $\hat{\mathcal{B}}(\hat{r})$ , and the set of all symmetric sparse matrices with the same sparsity pattern as  $\hat{\mathbf{S}}_{ALCE}$  such that  $\text{diag}(\mathbf{S}) = \text{diag}(\hat{\mathbf{\Sigma}}_{ALCE} - \mathbf{L})$ ,  $\mathbf{L} \in \hat{\mathcal{B}}(\hat{r})$  (we call this set  $\hat{\mathcal{A}}_{diag}(\hat{\mathcal{S}})$ ).

Consequently, we can write:

$$\hat{\mathbf{S}}_{UNALCE,ii} = \hat{\mathbf{\Sigma}}_{ALCE,ii} - \hat{\mathbf{L}}_{UNALCE,ii}, \hat{\mathbf{S}}_{UNALCE,ij} = \hat{\mathbf{S}}_{ALCE,ij}, i \neq j.$$

### 565 **Proof of Corollary 2**

We know that  $\|\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}\|_2 = \check{\psi}$ . We can prove that  $\hat{\mathbf{L}}_{UNALCE} = \min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{r})} \|\mathbf{L} - \mathbf{L}^*\|_F^2$ , conditioning on the event  $\min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{r}), \mathbf{S} \in \hat{\mathcal{A}}(\hat{\mathcal{S}}), \mathbf{\Sigma} = \mathbf{L} + \mathbf{S}} \|\mathbf{\Sigma} - \mathbf{\Sigma}_n\|_F^2$  under prescribed assumptions (see Theorem 3). In fact, we can write

$$\min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{r})} \|\mathbf{L} - \mathbf{L}^*\|_F^2 \leq \min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{r})} \|\mathbf{L} - \mathbf{Y}_{pre}\|_F^2 + \|\mathbf{Y}_{pre} - \mathbf{L}^*\|_F^2,$$

because  $\mathbf{Y}_{pre}$  is uniquely determined by the conditioning event. The same inequality holds in the spectral norm.

Since

$$\|\hat{\mathbf{L}}_{ALCE} - \mathbf{L}^*\|_2 \leq \|\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}\|_2 + \|\hat{\mathbf{L}}_{UNALCE} - \mathbf{L}^*\|_2,$$

it follows

$$0 < \|\hat{\mathbf{L}}_{ALCE} - \mathbf{L}^*\|_2 - \|\hat{\mathbf{L}}_{UNALCE} - \mathbf{L}^*\|_2 \leq \check{\psi}$$

given the conditioning event. Consequently, since  $\|\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}\|_F = \text{tr}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})^2 = r\check{\psi}^2$ , we obtain

$$0 < \|\hat{\mathbf{L}}_{ALCE} - \mathbf{L}^*\|_F - \|\hat{\mathbf{L}}_{UNALCE} - \mathbf{L}^*\|_F \leq \sqrt{r}\check{\psi}.$$

The analogous triangular inequality for the sparse component is

$$\|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_F^2 \leq \|\hat{\mathbf{S}}_{UNALCE} - \hat{\mathbf{S}}_{ALCE}\|_F^2 + \|\hat{\mathbf{S}}_{UNALCE} - \mathbf{S}^*\|_F^2.$$

In order to quantify  $\|\hat{\mathbf{S}}_{UNALCE} - \hat{\mathbf{S}}_{ALCE}\|_F^2$ , we need to study the behaviour of the term  $\sum_{i=1}^p (\hat{\mathbf{L}}_{UNALCE,ii} - \hat{\mathbf{L}}_{ALCE,ii})^2$ , which is less or equal to  $r\check{\psi}^2$ , because it is less or equal to  $\text{tr}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})^2$ .

Consequently, we have  $\|\hat{\mathbf{S}}_{UNALCE} - \hat{\mathbf{S}}_{ALCE}\|_F \leq \sqrt{r}\check{\psi}$ . Analogously to  $\hat{\mathbf{L}}_{UNALCE}$ , we can prove that

$$\hat{\mathbf{S}}_{UNALCE} = \min_{\mathbf{S} \in \hat{\mathcal{A}}(\hat{\mathcal{S}})} \|\mathbf{S} - \mathbf{S}^*\|_F^2,$$

conditioning on the event

$$\min_{\mathbf{L} \in \hat{\mathcal{B}}(\hat{r}), \mathbf{S} \in \hat{\mathcal{A}}(\hat{\mathcal{S}}), \mathbf{\Sigma} = \mathbf{L} + \mathbf{S}} \|\mathbf{\Sigma} - \mathbf{\Sigma}_n\|_F^2$$

under prescribed assumptions (see Theorem 3). In fact, we can write

$$\min_{\mathbf{S} \in \hat{\mathcal{A}}_{diag}(\hat{\mathcal{S}})} \|\mathbf{S} - \mathbf{S}^*\|_F^2 \leq \min_{\mathbf{S} \in \hat{\mathcal{A}}_{diag}(\hat{\mathcal{S}})} \|\mathbf{S} - \mathbf{Z}_{pre}\|_F^2 + \|\mathbf{Z}_{pre} - \mathbf{S}^*\|_F^2,$$

because  $\mathbf{Z}_{pre}$  is uniquely determined by the conditioning event.

Therefore, we can write

$$0 < \|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_F - \|\hat{\mathbf{S}}_{UNALCE} - \mathbf{S}^*\|_F \leq \sqrt{r}\check{\psi}.$$

The claim on  $\|\hat{\mathbf{S}}_{UNALCE} - \mathbf{S}^*\|_2$  is less immediate. We recall that  $\|\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}\|_2 = \|\hat{\mathbf{U}}\check{\psi}\mathbf{I}_r\hat{\mathbf{U}}^\top\|_2 = \check{\psi}$ .  $\hat{\mathbf{U}}\check{\psi}\mathbf{I}_r\hat{\mathbf{U}}^\top$  can be divided in the contribution coming from diagonal elements and the rest:  $\|\text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}) + \text{off} - \text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2$ . Both contributions are part of  $\hat{\mathbf{U}}\check{\psi}\mathbf{I}_r\hat{\mathbf{U}}^\top$ .

Given the matrix of eigenvectors  $\hat{\mathbf{U}}$ , we can write  $\text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}) = \sum_{i=1}^p \|\hat{\mathbf{u}}_i^\top\|^2 \mathbf{K}_{ii}$ , where  $\mathbf{K}_{ii}$  is a null matrix, except for the  $i$ -th diagonal element equal to  $\check{\psi}$ , and  $\hat{\mathbf{u}}_i^\top$  is the  $i$ -th row of  $\hat{\mathbf{U}}$ . Similarly, we can write  $\text{off} - \text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}) = \sum_{i=1}^p \sum_{j \neq i} \hat{\mathbf{u}}_i^\top \hat{\mathbf{u}}_j \mathbf{K}_{ij}$  where  $\mathbf{K}_{ij}$  is a null matrix, except for the element  $ij$  equal to  $\check{\psi}$ . Note that the rows of  $\hat{\mathbf{U}}$ , differently from the columns, are not orthogonal.

Since all summands are orthogonal to each other ( $\mathbf{A} \perp \mathbf{B} \Leftrightarrow \text{tr}(\mathbf{A}\mathbf{B}^\top) = 0$ ), the triangular inequalities relative to  $\|\text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2$ ,  $\|\text{off} - \text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2$  and  $\|\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}\|_2$  become equalities. Therefore, we can write:

$$\begin{aligned} \|\text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2 &= \sum_{i=1}^p \|\hat{\mathbf{u}}_i^\top\|^2 \times \|\mathbf{K}_{ii}\| = \sum_{i=1}^p \|\hat{\mathbf{u}}_i^\top\|^2 \check{\psi}; \\ \|\text{off} - \text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2 &= \sum_{i=1}^p \sum_{j \neq i} \hat{\mathbf{u}}_i^\top \hat{\mathbf{u}}_j \|\mathbf{K}_{ij}\| = \sum_{i=1}^p \sum_{j \neq i} \hat{\mathbf{u}}_i^\top \hat{\mathbf{u}}_j \check{\psi}; \\ \|\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}\|_2 &= \sum_{i=1}^p \|\hat{\mathbf{u}}_i^\top\|^2 \times \|\mathbf{K}_{ii}\| + \sum_{i=1}^p \sum_{j \neq i} \hat{\mathbf{u}}_i^\top \hat{\mathbf{u}}_j \|\mathbf{K}_{ij}\| = \check{\psi}. \end{aligned}$$

From this consideration, it follows that

$$\|\text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2 \leq \|\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}\|_2 = \check{\psi}.$$

Since, by definition,  $\|\text{diag}(\hat{\mathbf{S}}_{UNALCE} - \hat{\mathbf{S}}_{ALCE})\|_2 = \|\text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2$  (because  $\text{diag}(\hat{\mathbf{S}}_{UNALCE} - \hat{\mathbf{S}}_{ALCE}) = -\text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})$ ), and recalling that  $\hat{\mathbf{S}}_{UNALCE}$  has the best approximation property (for Theorem 3) given the conditioning event, we can conclude

$$0 < \|\hat{\mathbf{S}}_{ALCE} - \mathbf{S}^*\|_2 - \|\hat{\mathbf{S}}_{UNALCE} - \mathbf{S}^*\|_2 \leq \check{\psi}.$$

### **Proof of Corollary 3**

The relevant triangular inequality for the overall estimate is

$$\|\boldsymbol{\Sigma}_n - \hat{\boldsymbol{\Sigma}}_{ALCE}\|_2 \leq \|\hat{\boldsymbol{\Sigma}}_{UNALCE} - \hat{\boldsymbol{\Sigma}}_{ALCE}\|_2 + \|\boldsymbol{\Sigma}_n - \hat{\boldsymbol{\Sigma}}_{UNALCE}\|_2.$$

By definition,  $\|\hat{\boldsymbol{\Sigma}}_{UNALCE} - \hat{\boldsymbol{\Sigma}}_{ALCE}\|_2 = \|\text{off} - \text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2$ . For the same considerations explained before,

$$\|\text{off} - \text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_2 \leq \|\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}\|_2 = \check{\psi}.$$

Consequently, recalling that  $\hat{\boldsymbol{\Sigma}}_{UNALCE} = \min_{\boldsymbol{\Sigma}=\mathbf{L}+\mathbf{S}, \mathbf{L} \in \hat{\mathcal{B}}(\hat{r}), \mathbf{S} \in \hat{\mathcal{A}}(\hat{s})} \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_n\|_F^2$  under the described assumptions, it follows

$$0 < \|\boldsymbol{\Sigma}_n - \hat{\boldsymbol{\Sigma}}_{ALCE}\|_2 - \|\boldsymbol{\Sigma}_n - \hat{\boldsymbol{\Sigma}}_{UNALCE}\|_2 \leq \check{\psi}. \quad (\text{A.9})$$

Since  $\|\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE}\|_F^2 = \text{tr}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})^2 = r\check{\psi}^2$ , we have

$$0 < \|\text{off} - \text{diag}(\hat{\mathbf{L}}_{UNALCE} - \hat{\mathbf{L}}_{ALCE})\|_F \leq \sqrt{r}\check{\psi}.$$

We can then claim

$$0 < \|\boldsymbol{\Sigma}_n - \hat{\boldsymbol{\Sigma}}_{ALCE}\|_F - \|\boldsymbol{\Sigma}_n - \hat{\boldsymbol{\Sigma}}_{UNALCE}\|_F \leq \sqrt{r}\check{\psi}.$$

Therefore, the real gain in terms of the approximation of  $\boldsymbol{\Sigma}_n$  with respect to ALCE measured in the squared Frobenius norm is strictly positive and bounded from  $r\check{\psi}^2$ .

**Proof of Theorem 4**

Conditioning on  $\Sigma_n$ , we can easily write

$$\|\hat{\Sigma}_{UNALCE} - \Sigma^*\|_2 = \|\hat{\Sigma}_{UNALCE} - \Sigma_n + \Sigma_n - \Sigma^*\|_2 \leq \|\hat{\Sigma}_{UNALCE} - \Sigma_n\|_2 + \|\Sigma_n - \Sigma^*\|_2. \quad (\text{A.10})$$

595 The term  $\|\Sigma_n - \Sigma^*\|_2$  only depends on the estimation input  $\Sigma_n$ .  
Therefore, by (A.9) and (A.10), it is straightforward that

$$0 < \|\hat{\Sigma}_{ALCE} - \Sigma^*\|_2 - \|\hat{\Sigma}_{UNALCE} - \Sigma^*\|_2 \leq \check{\psi}.$$

Analogously, it is easy to prove that

$$0 < \|\hat{\Sigma}_{ALCE} - \Sigma^*\|_F - \|\hat{\Sigma}_{UNALCE} - \Sigma^*\|_F \leq \sqrt{r}\check{\psi}. \quad (\text{A.11})$$

**Proof of Corollary 4**

Let us recall the following expression:

$$\|(\hat{\mathbf{L}} + \hat{\mathbf{S}})^{-1} - (\Sigma^{*-1})\|_F = \|(\hat{\mathbf{L}} + \hat{\mathbf{S}})^{-1}[\hat{\mathbf{L}} + \hat{\mathbf{S}} - \Sigma^*](\Sigma^{*-1})\|_F \leq \|(\hat{\mathbf{L}} + \hat{\mathbf{S}})^{-1}\|_2 \cdot \|\hat{\mathbf{L}} + \hat{\mathbf{S}} - \Sigma^*\|_F \cdot \|\Sigma^{*-1}\|_2.$$

From (A.11), we can conclude that

$$0 < \|(\hat{\mathbf{L}}_{ALCE} + \hat{\mathbf{S}}_{ALCE})^{-1} - \Sigma^{*-1}\|_F - \|(\hat{\mathbf{L}}_{UNALCE} + \hat{\mathbf{S}}_{UNALCE})^{-1} - \Sigma^{*-1}\|_F \leq \sqrt{r}\check{\psi}.$$

Analogously, since it holds

$$\|(\hat{\mathbf{L}} + \hat{\mathbf{S}})^{-1} - (\Sigma^{*-1})\|_2 = \|(\hat{\mathbf{L}} + \hat{\mathbf{S}})^{-1}[\hat{\mathbf{L}} + \hat{\mathbf{S}} - \Sigma^*](\Sigma^{*-1})\|_2 \leq \|(\hat{\mathbf{L}} + \hat{\mathbf{S}})^{-1}\|_2 \cdot \|\hat{\mathbf{L}} + \hat{\mathbf{S}} - \Sigma^*\|_2 \cdot \|\Sigma^{*-1}\|_2,$$

600 , it is straightforward that

$$0 < \|(\hat{\mathbf{L}}_{ALCE} + \hat{\mathbf{S}}_{ALCE})^{-1} - \Sigma^{*-1}\|_2 - \|(\hat{\mathbf{L}}_{UNALCE} + \hat{\mathbf{S}}_{UNALCE})^{-1} - \Sigma^{*-1}\|_2 \leq \check{\psi}.$$

**Proof of Corollary 5**

The three claims of the corollary are proved in sequence.

1. We start to note that  $\hat{\mathbf{L}}_{UNALCE}$ ,  $\hat{\mathbf{L}}_{ALCE}$ , and  $\hat{\mathbf{U}}_{ALCE}\check{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top$  are  $r$ -ranked. Let the respective spectral decompositions be:

- 605 (a)  $\hat{\mathbf{B}}_{UNALCE}\hat{\mathbf{B}}_{UNALCE}^\top$  with  $\hat{\mathbf{B}}_{UNALCE} = \hat{\mathbf{U}}_{ALCE}\sqrt{\hat{\mathbf{D}}_{UNALCE}}$  ;
- (b)  $\hat{\mathbf{B}}_{ALCE}\hat{\mathbf{B}}_{ALCE}^\top$  with  $\hat{\mathbf{B}}_{ALCE} = \hat{\mathbf{U}}_{ALCE}\sqrt{\hat{\mathbf{D}}_{ALCE}}$  ;
- (c)  $(\hat{\mathbf{U}}_{ALCE}\sqrt{\check{\psi}})(\hat{\mathbf{U}}_{ALCE}\sqrt{\check{\psi}})^\top$ .

Consequently, we note that

$$\begin{aligned} \lambda_r(\hat{\mathbf{L}}_{UNALCE}) &= \lambda_r(\hat{\mathbf{L}}_{ALCE} + \hat{\mathbf{U}}_{ALCE}\check{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top) = \\ &= \lambda_r(\hat{\mathbf{U}}_{ALCE}\hat{\mathbf{D}}_{ALCE}\hat{\mathbf{U}}_{ALCE} + \hat{\mathbf{U}}_{ALCE}\check{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top) = \lambda_r(\hat{\mathbf{L}}_{ALCE}) + \check{\psi}, \end{aligned}$$

which proves the claim on  $\hat{\mathbf{L}}_{UNALCE}$ .

- 610 2. By Lidskii dual inequality (see [32]), we note that

$$\lambda_p(\hat{\mathbf{S}}_{UNALCE}) = \lambda_p(\hat{\mathbf{S}}_{ALCE} - \text{diag}(\hat{\mathbf{U}}_{ALCE}\check{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top)) \geq \lambda_p(\hat{\mathbf{S}}_{ALCE}) + \lambda_p(-\text{diag}(\hat{\mathbf{U}}_{ALCE}\check{\psi}\mathbf{I}_r\hat{\mathbf{U}}_{ALCE}^\top)).$$



The matrix  $-\text{diag}(\hat{\mathbf{U}}_{ALCE} \check{\Psi} \mathbf{I}_r \hat{\mathbf{U}}_{ALCE}^\top)$  is a  $p$ -dimensional squared matrix having as  $i$ -th element the quantity  $-\|\mathbf{u}_i^\top\|^2 \check{\Psi}$ , where  $\mathbf{u}_i^\top$ ,  $i \in \{1, \dots, p\}$ , is the  $i$ -th row of the matrix  $\hat{\mathbf{U}}_{ALCE}$ . Since  $\text{tr}(-\text{diag}(\hat{\mathbf{U}}_{ALCE} \check{\Psi} \mathbf{I}_r \hat{\mathbf{U}}_{ALCE}^\top)) = \text{tr}(-\hat{\mathbf{U}}_{ALCE} \check{\Psi} \mathbf{I}_r \hat{\mathbf{U}}_{ALCE}^\top) = -r\check{\Psi}$ , it follows that  $\lambda_p(\text{diag}(\hat{\mathbf{U}}_{ALCE} \check{\Psi} \mathbf{I}_r \hat{\mathbf{U}}_{ALCE}^\top)) \leq r\check{\Psi}/p$ , i.e.

$$-\frac{r}{p}\check{\Psi} \leq \lambda_p(-\text{diag}(\hat{\mathbf{U}}_{ALCE} \check{\Psi} \mathbf{I}_r \hat{\mathbf{U}}_{ALCE}^\top)) \leq 0.$$

Therefore, we obtain

$$\lambda_p(\hat{\mathbf{S}}_{UNALCE}) \geq \lambda_p(\hat{\mathbf{S}}_{ALCE}) - \frac{r}{p}\check{\Psi},$$

615 which proves the claim on  $\hat{\mathbf{S}}_{UNALCE}$ .

3. By Lidskii dual inequality, we note that

$$\begin{aligned} \lambda_p(\hat{\mathbf{\Sigma}}_{UNALCE}) &= \lambda_p(\hat{\mathbf{\Sigma}}_{ALCE} + \hat{\mathbf{U}}_{ALCE} \check{\Psi} \mathbf{I}_r \hat{\mathbf{U}}_{ALCE}^\top - \text{diag}(\hat{\mathbf{U}}_{ALCE} \check{\Psi} \mathbf{I}_r \hat{\mathbf{U}}_{ALCE}^\top)) \\ &\geq \lambda_p(\hat{\mathbf{\Sigma}}_{ALCE}) + \lambda_p(\hat{\mathbf{U}}_{ALCE} \check{\Psi} \mathbf{I}_r \hat{\mathbf{U}}_{ALCE}^\top) - \lambda_p(\text{diag}(\hat{\mathbf{U}}_{ALCE} \check{\Psi} \mathbf{I}_r \hat{\mathbf{U}}_{ALCE}^\top)). \end{aligned}$$

Recalling the argument above and noting that

$\lambda_p(\hat{\mathbf{U}}_{ALCE} \check{\Psi} \mathbf{I}_r \hat{\mathbf{U}}_{ALCE}^\top) = 0$  because  $\text{rank}(\hat{\mathbf{U}}_{ALCE} \check{\Psi} \mathbf{I}_r \hat{\mathbf{U}}_{ALCE}^\top) = \hat{r}$ , it follows

$$\lambda_p(\hat{\mathbf{\Sigma}}_{UNALCE}) \geq \lambda_p(\hat{\mathbf{\Sigma}}_{ALCE}) + 0 - \check{\Psi} = \lambda_p(\hat{\mathbf{\Sigma}}_{ALCE}) - \frac{r}{p}\check{\Psi},$$

which proves the claim on  $\hat{\mathbf{\Sigma}}_{UNALCE}$ .

## 620 Supplementary material

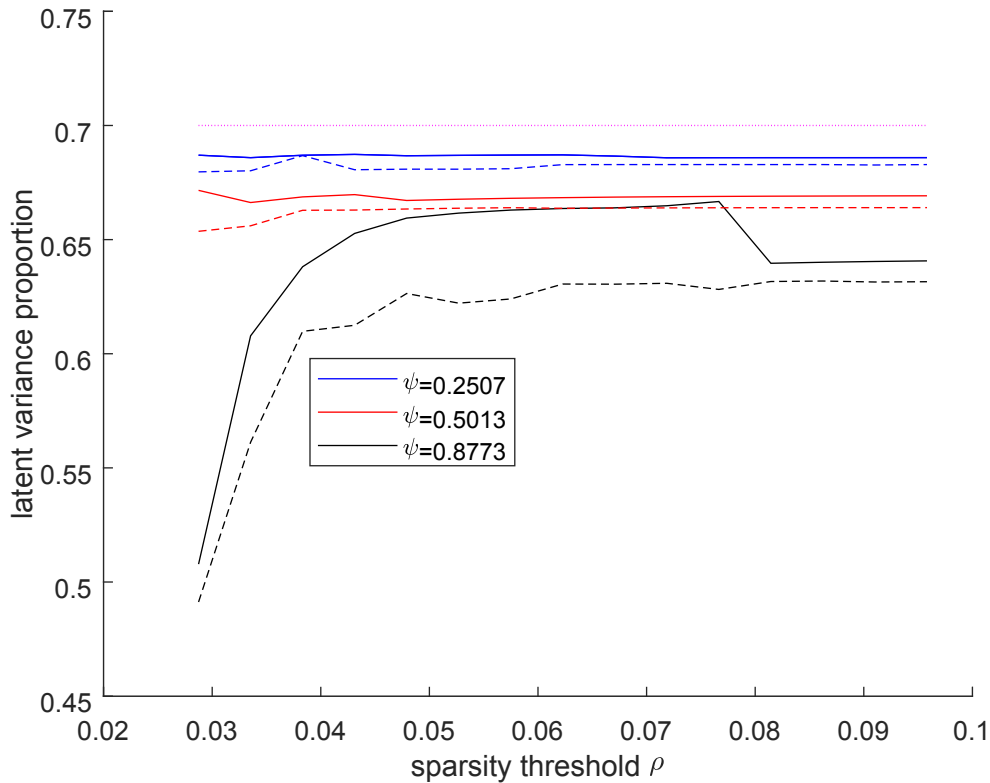
This paper is complemented by a supplement containing a discussion of LOREC assumptions and a simulation study. In addition, the MATLAB functions UNALCE.m and POET.m, performing UNALCE and POET procedures, respectively, can be downloaded at [16]. Both functions contain the detailed explanation of input and output arguments. Finally, the MATLAB dataset supervisory\_data.mat, which contains the covariance matrix,  $\mathbf{C}$ , and the relative labels of supervisory indicators, Labgood, can also be downloaded at the same link, which we refer to for the details.

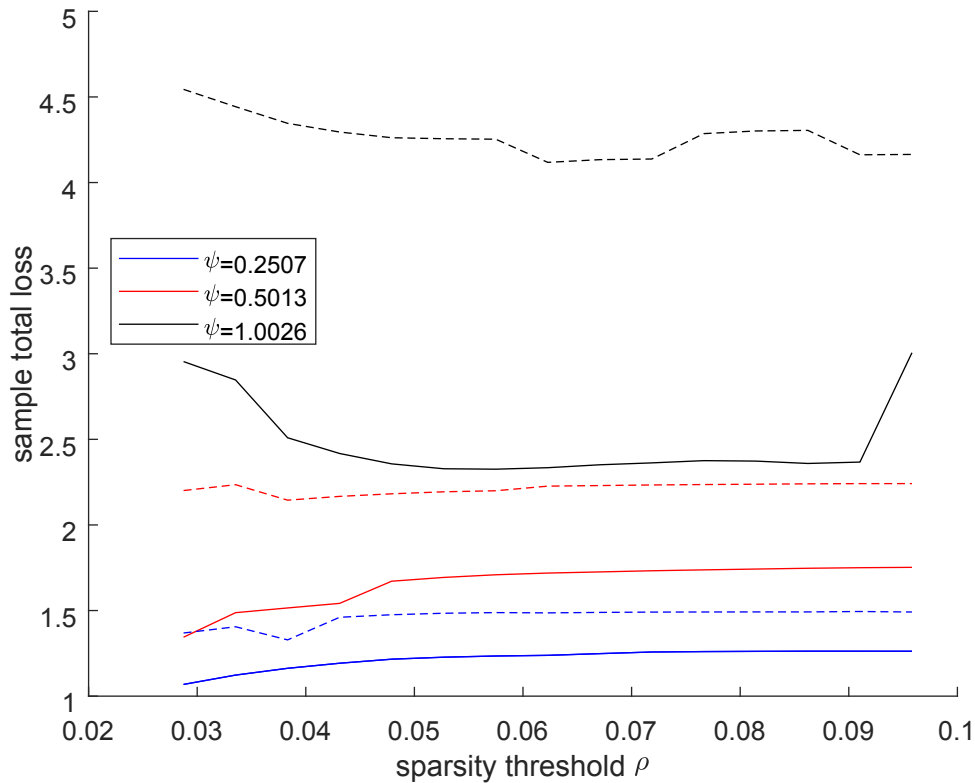
625

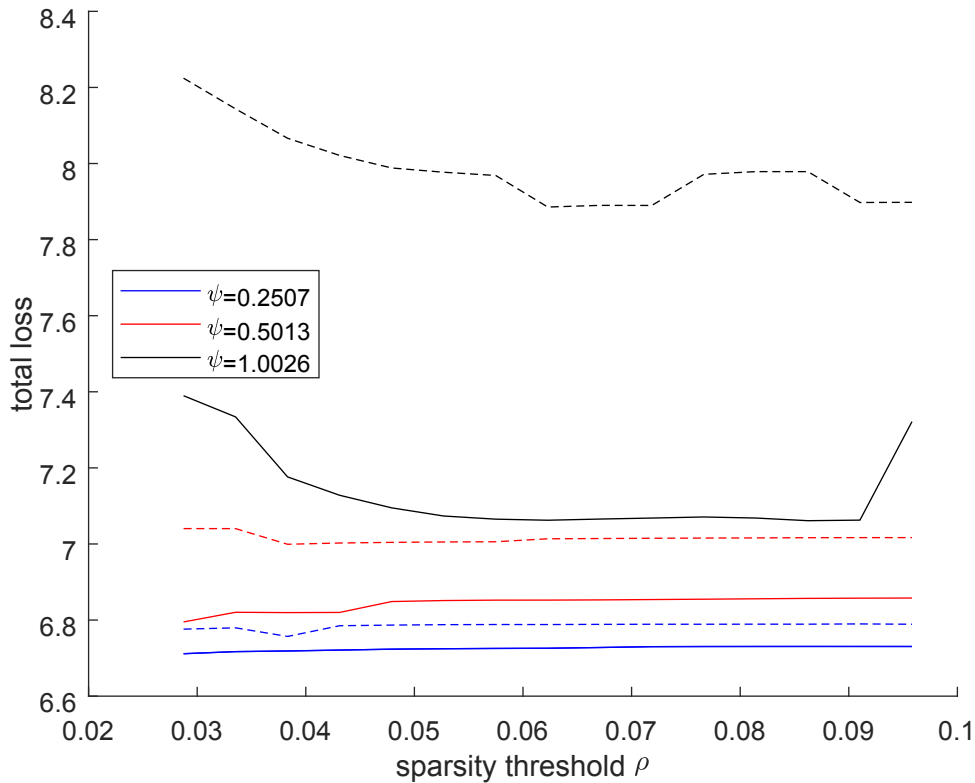
## References

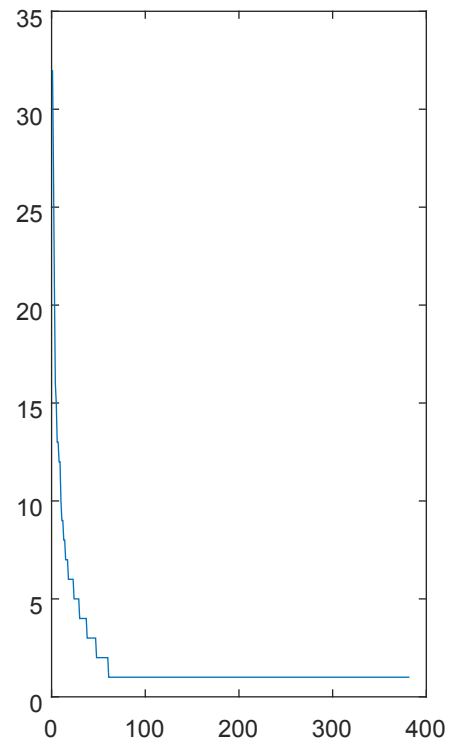
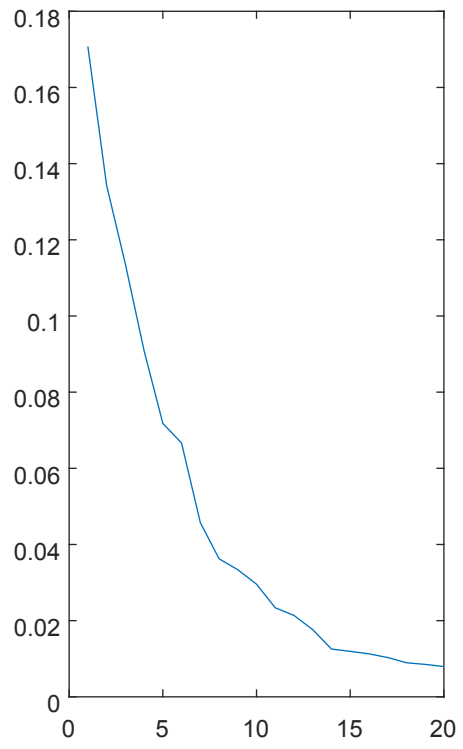
- [1] A. Agarwal, S. Negahban, M. J. Wainwright, Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions, *The Annals of Statistics* 40 (2012) 1171–1197.
- [2] T. Anderson, *Multivariate statistical analysis*, Wiley and Sons, New York, NY, 1984.
- 630 [3] J. Bai, Inferential theory for factor models of large dimensions, *Econometrica* 71 (2003) 135–171.
- [4] J. Bai, S. Ng, Determining the number of factors in approximate factor models, *Econometrica* 70 (2002) 191–221.
- [5] P. J. Bickel, E. Levina, Covariance regularization by thresholding, *The Annals of Statistics* 36 (2008) 2577–2604.
- [6] P. J. Bickel, E. Levina, Regularized estimation of large covariance matrices, *The Annals of Statistics* 36 (2008) 199–227.
- 635 [7] J.-F. Cai, E. J. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, *SIAM Journal on Optimization* 20 (2010) 1956–1982.
- [8] T. Cai, W. Liu, Adaptive thresholding for sparse covariance matrix estimation, *Journal of the American Statistical Association* 106 (2011) 672–684.
- [9] T. Cai, C.-H. Zhang, H. H. Zhou, Optimal rates of convergence for covariance matrix estimation, *The Annals of Statistics* 38 (2010) 2118–2144.
- 640 [10] V. Chandrasekaran, P. A. Parrilo, A. S. Willsky, Latent variable graphical model selection via convex optimization, *The Annals of Statistics* 40 (2012) 1935–1967.
- [11] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, A. S. Willsky, Rank-sparsity incoherence for matrix decomposition, *SIAM Journal on Optimization* 21 (2011) 572–596.
- [12] I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Communications on pure and applied mathematics* 57 (2004) 1413–1457.
- 645 [13] D. K. Dey, C. Srinivasan, Estimation of a covariance matrix under stein’s loss, *The Annals of Statistics* 13 (1985) 1581–1591.
- [14] J. Fan, Y. Liao, H. Liu, An overview of the estimation of large covariance and precision matrices, 2016.
- [15] J. Fan, Y. Liao, M. Mincheva, Large covariance estimation by thresholding principal orthogonal complements, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (2013) 603–680.

- 650 [16] M. Farné, A. Montanari, A large covariance matrix estimator under intermediate spikiness regimes, <https://data.mendeley.com/datasets/nh97vfvhkt>, 2018.
- [17] M. Fazel, Matrix rank minimization with applications, Ph.D. thesis, Stanford University, 2002.
- [18] M. Fazel, H. Hindi, S. P. Boyd, A rank minimization heuristic with application to minimum order system approximation, in: American Control Conference, 2001. Proceedings of the 2001, volume 6, IEEE, pp. 4734–4739.
- 655 [19] T. S. Ferguson, A course in large sample theory, Routledge, 2017.
- [20] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, *Biostatistics* 9 (2008) 432–441.
- [21] R. Furrer, T. Bengtsson, Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants, *Journal of Multivariate Analysis* 98 (2007) 227–255.
- [22] D. A. Harville, Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association* 72 (1977) 320–338.
- 660 [23] C. Lam, Nonparametric eigenvalue-regularized precision or covariance matrix estimator, *The Annals of Statistics* 44 (2016) 928–953.
- [24] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis* 88 (2004) 365–411.
- [25] O. Ledoit, M. Wolf, Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions, *Journal of Multivariate Analysis* 139 (2015) 360–384.
- 665 [26] X. Luo, High dimensional low rank and sparse covariance matrix estimation via convex minimization, Arxiv preprint 1111.1133v2 (2011).
- [27] X. Luo, Recovering model structures from large low rank and sparse covariance matrix estimation, Arxiv preprint 1111.1133 (2011).
- [28] Y. Nesterov, Gradient methods for minimizing composite functions, *Mathematical Programming* 140 (2013) 125–161.
- [29] H. D. Patterson, R. Thompson, Recovery of inter-block information when block sizes are unequal, *Biometrika* 58 (1971) 545–554.
- 670 [30] Y. Qiu, S. X. Chen, Bandwidth selection for high-dimensional covariance matrix estimation, *Journal of the American Statistical Association* 110 (2015) 1160–1174.
- [31] A. J. Rothman, E. Levina, J. Zhu, Generalized thresholding of large covariance matrices, *Journal of the American Statistical Association* 104 (2009) 177–186.
- [32] T. Tao, Topics in random matrix theory, volume 132, American Mathematical Society, 2012.









**Matteo Farnè:** Investigation, Methodology, Software, Data curation, Writing- Original draft preparation.

**Angela Montanari:** Conceptualization, Supervision, Writing- Reviewing and Editing, Resources, Funding acquisition

# Supplementary Material

Matteo Farné<sup>a,\*\*</sup>, Angela Montanari<sup>a</sup>

<sup>a</sup>Department of Statistical Sciences, University of Bologna, Bologna, Italy

---

---

## 1. Discussion of LOREC assumptions

5 The key model-based results for deriving the LOREC [8] consistency bounds are a lemma by [1], which details the sample loss in infinity (element-wise) norm:

$$\|\Sigma_n - \Sigma^*\|_\infty = O\left(\sqrt{\frac{\ln p}{n}}\right), \quad (1)$$

and a lemma by [5], which details the sample loss in spectral norm:

$$\|\Sigma_n - \Sigma^*\|_2 = O\left(\sqrt{\frac{p}{n}}\right). \quad (2)$$

We stress that (2) strictly requires the assumption  $p \leq n$ .

10 From a theoretical point of view, the LOREC approach is hindered by some deficiencies and incongruities. In contrast to the POET approach, where the sparsity assumption is imposed on the sparse component  $\mathbf{S}^*$ , the LOREC approach imposes it directly on the covariance matrix  $\Sigma^*$ . As a consequence, the assumption  $\Sigma^* \in \Sigma^*(\epsilon_0)$ , where

$$\Sigma^*(\epsilon_0) = \left\{ \Sigma^* \in \mathbb{R}^{p \times p} : 0 < \epsilon_0 \leq \lambda_i(\Sigma^*) \leq \epsilon_0^{-1}, i \in \{1, \dots, p\} \right\}, \quad (3)$$

is necessary and leads to some non-identifiability issues:

- to guarantee the validity of a bound (1), the assumption  $\max_{i \leq p} \sum_{j \leq p} |\Sigma_{ij}^*|^q = o(p)$  for some  $q \in [0, 1[$  must hold. Since we can write

$$\max_{i \leq p} \sum_{j \leq p} |\Sigma_{ij}^*|^q \leq \max_{i \leq p} \sum_{j \leq p} |\mathbf{L}_{ij}^*|^q + \max_{i \leq p} \sum_{j \leq p} |\mathbf{S}_{ij}^*|^q,$$

15 , the aforementioned assumption would impose the same bound on  $\mathbf{L}^*$ :  $\max_{i \leq p} \sum_{j \leq p} |\mathbf{L}_{ij}^*|^q = o(p)$ . In turn, this would mean, by definition, that  $\xi(T(\mathbf{L}^*))$  is larger. This is because, according to [3],  $\xi(T(\mathbf{L}^*))$  directly depends on the alignment of the eigenvectors of  $\mathbf{L}^*$  with the canonical basis vectors. Therefore, as  $q$  decreases, it is not clear what happens to the product  $\xi(T(\mathbf{L}^*))\mu(\Omega(\mathbf{S}^*))$ . This potentially affects identifiability.

- applying Lidskii dual inequality (see [10]) to  $\lambda_r(\Sigma^*)$ , we obtain

$$\lambda_r(\Sigma^*) = \lambda_r(\mathbf{L}^* + \mathbf{S}^*) \geq \lambda_r(\mathbf{L}^*) + \lambda_p(\mathbf{S}^*).$$

20 Therefore, the assumption  $\lambda_r(\Sigma^*) > \text{const}$  in (3) inevitably leads to  $\lambda_p(\mathbf{S}^*) > \text{const}$  and  $\lambda_r(\mathbf{L}^*) > \text{const}$ . The latter formulation conflicts with the identifiability assumption  $\lambda_r(\mathbf{L}^*) > (C_2/\xi^2(T))\sqrt{p/n}$  in [8], and affects identifiability. This is the case even when  $p = o(n)$ , because in that case  $\lambda_r(\mathbf{L}^*) > \text{const}$  implies that  $\mathbf{L}^*$  is expected to be close to the identity matrix, thus increasing the product  $\xi(T(\mathbf{L}^*))\mu(\Omega(\mathbf{S}^*))$ .

---

<sup>\*\*</sup>Declarations of interest: none

<sup>\*\*</sup>Corresponding author

Email address: [matteo.farne2@unibo.it](mailto:matteo.farne2@unibo.it) (Matteo Farné)



## 2. A simulation study

### 2.1. Simulation settings

To compare the performance of UNALCE, LOREC, and POET, we take into consideration four simulated low rank plus sparse settings, which give an exhaustive idea of the recoverable situations under our assumptions. The key simulation parameters are:

1. the dimension  $p$ , the sample size  $n$ ;
2. the rank  $r$  and the condition number  $cond(\mathbf{L}^*) = \lambda_{max}(\mathbf{L}^*)/\lambda_{min}(\mathbf{L}^*)$  of the low rank component  $\mathbf{L}^*$ ;
3. the trace of  $\mathbf{L}^*$ ,  $\tau\theta p$ , where  $\tau$  is a magnitude parameter and  $\theta$  is the percentage of variance explained by  $\mathbf{L}^*$ ;
4. the number of off-diagonal non-zeros  $s$  in the sparse component  $\mathbf{S}^*$ ;
5. the percentage of non-zeros  $\pi_s$  over the number of off-diagonal elements;
6. the percentage of the (absolute) residual covariance  $\rho_{S^*}$ ;
7.  $N = 100$  replicates for each setting.

For POET, the soft-thresholding parameter is estimated via cross-validation, cf. [6]. For ALCE and UNALCE, spikiness and sparsity thresholds are picked up via the MC criterion (see (14) in the main paper). The data generation algorithm is described in detail in [7].

In Table S.1, we summarize the features of our settings. Settings 1 and 2 vary according to the degree of spikiness and sparsity. Settings 3 and 4 are intermediately spiked and sparse and vary according to the ratio  $p/n$ . In Table S.2, we report all the norms and condition numbers relative to our settings. In addition, we report the minimum latent eigenvalue  $\lambda_r(\mathbf{L}^*)$  and the minimum residual non-zero entry in absolute value  $S_{min,off}$  (the identifiability parameters). We stress that we assume equispaced latent eigenvalues. It follows that for each setting, the eigenvalues of  $\mathbf{L}^*$  and  $\Sigma^*$  are pretty similar, while the ones of  $\mathbf{S}^*$  are much smaller.

**Table S.1**

Simulated settings: main parameters and synthetic descriptions.

Setting	$p$	$n$	$p/n$	$r$	$\tau$	$\theta$	$\pi_s$	$\rho_{S^*}$	spikiness	sparsity
1	100	1000	0.1	4	0.01	70	2.38	0.45	low	high
2	100	1000	0.1	3	0.03	80	11.72	0.72	high	low
3	150	150	1	5	0.01	80	3.20	0.33	middle	middle
4	200	100	2	6	0.01	80	3.66	0.39	middle	middle

**Table S.2**

Simulated settings: spectral norms, condition numbers, and identifiability parameters.

Setting	$\ \mathbf{L}^*\ _2$	$\lambda_r(\mathbf{L}^*)$	$cond(\mathbf{L}^*)$	$\ \mathbf{S}^*\ _2$	$S_{min,off}$	$cond(\mathbf{S}^*)$	$\ \Sigma^*\ _2$	$cond(\Sigma^*)$
1	23.33	11.67	2	3.78	0.0275	$2.26e + 07$	24.49	$9.49e + 07$
2	128	32	4	5.58	0.0226	$2.53e + 05$	130.14	$4.07e + 06$
3	32	16	2	2.56	0.0161	$2.35e + 13$	32.48	$1.58e + 10$
4	35.56	17.78	2	4.69	0.0138	$1.17e + 13$	36.39	$3.09e + 09$

Our objective function (see (4) in the main paper) is minimized according to an alternate thresholding algorithm, composed of singular value thresholding (SVT, [2]) and a soft thresholding step [4]. To speed convergence, Nesterov's acceleration scheme for composite gradient mapping minimization problems [9] is applied. Given a prescribed precision level  $\varepsilon$ , the algorithm assumes the form (cf. [8]):

1. Set  $(\mathbf{L}_0, \mathbf{S}_0) = 0.5(\text{diag}(\Sigma_n), \text{diag}(\Sigma_n))$ ,  $\eta_0 = 1$ .
2. Initialize  $\mathbf{Y}_0 = \mathbf{L}_0$  and  $\mathbf{Z}_0 = \mathbf{S}_0$ . Set  $t = 1$ .

3. Repeat: compute  $\frac{\partial \frac{1}{2} \|\mathbf{Y}_{t-1} + \mathbf{Z}_{t-1} - \boldsymbol{\Sigma}_n\|_F^2}{\partial \mathbf{Y}_{t-1}} = \frac{\partial \frac{1}{2} \|\mathbf{Y}_{t-1} + \mathbf{Z}_{t-1} - \boldsymbol{\Sigma}_n\|_F^2}{\partial \mathbf{Z}_{t-1}} = \mathbf{Y}_{t-1} + \mathbf{Z}_{t-1} - \boldsymbol{\Sigma}_n$ .

4. Apply the SVT operator  $T_\psi$  to  $\mathbf{E}_{Y,t} = \mathbf{Y}_{t-1} - 0.5(\mathbf{Y}_{t-1} + \mathbf{Z}_{t-1} - \boldsymbol{\Sigma}_n)$  and set  $\mathbf{L}_t = T_\psi(\mathbf{E}_{Y,t}) = \hat{\mathbf{U}} \hat{\mathbf{D}}_\psi \hat{\mathbf{U}}^\top$ .

5. Apply the soft-thresholding operator  $T_\rho$  to  $\mathbf{E}_{Z,t} = \mathbf{Z}_{t-1} - 0.5(\mathbf{Y}_{t-1} + \mathbf{Z}_{t-1} - \boldsymbol{\Sigma}_n)$  and set  $\mathbf{S}_t = T_\rho(\mathbf{E}_{Z,t})$ .

6. Set  $(\mathbf{Y}_t, \mathbf{Z}_t) = (\mathbf{L}_t, \mathbf{S}_t) + \{(\eta_{t-1} - 1)/\eta_t\} \{(\mathbf{L}_t, \mathbf{S}_t) - (\mathbf{L}_{t-1}, \mathbf{S}_{t-1})\}$  where  $\eta_t = 0.5 + 0.5 \sqrt{1 + 4\eta_{t-1}^2}$ .

7. Until the convergence criterion  $(\|\mathbf{L}_t - \mathbf{L}_{t-1}\|_F)/(1 + \|\mathbf{L}_{t-1}\|_F) + (\|\mathbf{S}_t - \mathbf{S}_{t-1}\|_F)/(1 + \|\mathbf{S}_{t-1}\|_F) \leq \varepsilon$  (we set  $\varepsilon = 10^{-4}$ ).

The reported scheme achieves a convergence speed proportional to  $O(t^2)$ . We define  $t^*$  as the number of steps needed for convergence. We set  $\mathbf{Y}_{pre} = \mathbf{Y}_{t^*-1} - 0.5(\mathbf{Y}_{t^*-1} + \mathbf{Z}_{t^*-1} - \boldsymbol{\Sigma}_n)$  and  $\mathbf{Z}_{pre} = \mathbf{Z}_{t^*-1} - 0.5(\mathbf{Y}_{t^*-1} + \mathbf{Z}_{t^*-1} - \boldsymbol{\Sigma}_n)$ . The computational cost of the solution algorithm is proportional to  $\frac{p^4}{\sqrt{\varepsilon}}$ , where  $\varepsilon$  is the required precision. By contrast, POET reflects the cost of a full-SVD (proportional to  $p^3$ ). For more details, see [7].

Lots of metrics are computed to comparatively describe the performance of the three methods using the same simulated data. We call the low rank estimate  $\hat{\mathbf{L}}$ , the sparse estimate  $\hat{\mathbf{S}}$ , and the covariance matrix estimate  $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{L}} + \hat{\mathbf{S}}$ . The error norms used are:

$$\text{Loss} = \|\hat{\mathbf{L}} - \mathbf{L}^*\|_F + \|\hat{\mathbf{S}} - \mathbf{S}^*\|_F, \quad (4)$$

$$\text{TotalLoss} = \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_F, \quad (5)$$

$$\text{SampleTotalLoss} = \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_n\|_F. \quad (6)$$

The estimated percentage of latent variance  $\hat{\theta}$ , residual covariance  $\hat{\rho}_s$ , and residual non-zeros  $\hat{\pi}_s$  are also computed. Their estimation performance is measured by the mean square error, defined for  $\hat{\theta}$  as

$$\text{MSE}(\hat{\theta}) = \frac{1}{N} \sum_{h=1}^N (\hat{\theta}_h - \theta)^2, \quad (7)$$

where  $\hat{\theta}_h$  is the estimate of  $\theta$  on the  $h$ -th replicate. We also compute the estimation bias for each parameter, defined for  $\hat{\theta}$  as

$$\text{bias}(\hat{\theta}) = \hat{\theta}_{mean} - \theta, \quad (8)$$

where  $\hat{\theta}_{mean}$  is the mean estimate of  $\theta$  over the  $N$  replicates.

The performance in terms of the sparsity pattern recovery of  $\hat{\mathbf{S}}$  is assessed by the following measures. Let us denote the number of non-zeros in  $\hat{\mathbf{S}}$  by  $\hat{s}$ , the false non-zeros by  $fp$ , the false zeros by  $fn$ , the false positive elements by  $fpos$ , and the false negative elements by  $fneg$ . We define:

- the *error* measure:  $err = (fp + fn)/nv$ , where  $nv = p(p-1)/2$  is the number of off-diagonal elements;
- $errplus = (fpos + fneg)/s$ , which is the same as  $err$ , except it is computed for non-zeros only whereby positive and negative are distinguished in the usual way;
- the overall error rate  $errtot$ , using the number of false zeros, false positive, and false negative elements:  $errtot = (fpos + fneg + fn)/nv$ .

The correct classification rates of (true) non-zero and zero elements (denoted respectively by  $sens$  and  $spec$ ) are derived, as well as the correct classification rates of positive and negative elements, which are considered separately (denoted respectively by  $senspos$  and  $specpos$ ).

The performance in terms of eigenstructure recovery is measured for  $\boldsymbol{\Sigma}^*$  by  $\lambda(\hat{\boldsymbol{\Sigma}})$ , which is defined as the Euclidean distance between the estimated and true eigenvalues of  $\boldsymbol{\Sigma}^*$ :

$$\lambda(\hat{\boldsymbol{\Sigma}}) = \sqrt{\sum_{i=1}^p (\hat{\lambda}_i(\hat{\boldsymbol{\Sigma}}) - \lambda_i(\boldsymbol{\Sigma}^*))^2}. \quad (9)$$

The measure (9) is similarly defined for  $\mathbf{L}^*$  and  $\mathbf{S}^*$ .

## 2.2. Simulation results

Table S.3 shows the superior performance of UNALCE and ALCE than that of POET for the estimates of  $\theta$ ,  $\rho_{\mathcal{S}^*}$  and  $\pi_{\mathcal{S}}$  across different degrees of spikiness (Settings 1 and 2). This occurs because the percentage of residual variance detected by POET is biased upward, due to the natural bias of sample eigenvalues. The bias decreases as the dimension and degree of spikiness increase. Comparing UNALCE and ALCE, we observe a remarkable superiority on behalf of the UNALCE estimates of  $\theta$  and  $\rho_{\mathcal{S}^*}$  over Settings 1 and 2, while ALCE prevails for the estimates of  $\pi_{\mathcal{S}}$ .

Table S.4 shows the same statistics computed over Settings 3 and 4. The described pattern for the estimated latent variance percentage still holds. On the contrary, the performance of the estimated residual covariance percentage is poor, even for UNALCE and ALCE. This is because the estimation bias is significant. This occurs because Settings 3 and 4 do not fulfil the assumptions of Theorem 2, due to the small ratio of  $p/n$ . As a consequence, any estimate of  $\pi_{\mathcal{S}}$  is biased.

Table S.5 contains the fitting measures described in Section 2.1, which were computed for Settings 1 and 2. It is clear that UNALCE outperforms POET for all losses and is also generally superior to ALCE by a small margin. A similar pattern can be deduced from Table S.6, which contains the same results for Settings 3 and 4. However, it can be observed that the gap with POET progressively decreases as  $p/n$  increases because in that case, the setting becomes more consistent with the POET assumptions.

**Table S.3**

This table shows the mean square error and bias, as defined in (7) and (8), respectively, of the percentage of latent variance  $\hat{\theta}$ , residual covariance  $\hat{\rho}_{\mathcal{S}}$ , and residual non-zeros  $\hat{\pi}_{\mathcal{S}}$ , as estimated by UNALCE, ALCE, and POET over 100 runs of Settings 1 and 2.

	Setting 1			Setting 2		
	UNALCE	ALCE	POET	UNALCE	ALCE	POET
MSE( $\hat{\theta}$ )	<b>0.55</b>	0.75	10.10	<b>0.21</b>	0.24	2.45
MSE( $\hat{\rho}_{\mathcal{S}}$ )	<b>0.02</b>	0.04	0.20	<b>0.02</b>	0.03	0.23
MSE( $\hat{\pi}_{\mathcal{S}}$ )	0.89	<b>0.23</b>	5.53	11.25	<b>9.20</b>	119.74
bias( $\hat{\theta}$ )	<b>-0.48</b>	-0.69	3.14	<b>-0.14</b>	-0.17	1.51
bias( $\hat{\rho}_{\mathcal{S}}$ )	<b>-0.11</b>	-0.19	-0.45	<b>-0.36</b>	-0.38	-0.72
bias( $\hat{\pi}_{\mathcal{S}}$ )	0.61	<b>-0.06</b>	-2.35	<b>1.23</b>	1.37	10.93

**Table S.4**

This table shows the mean square error and bias, as defined in (7) and (8), respectively, of the percentage of latent variance  $\hat{\theta}$ , residual covariance  $\hat{\rho}_{\mathcal{S}}$ , and residual non-zeros  $\hat{\pi}_{\mathcal{S}}$ , as estimated by UNALCE, ALCE, and POET over 100 runs of Settings 3 and 4.

	Setting 3			Setting 4		
	UNALCE	ALCE	POET	UNALCE	ALCE	POET
MSE( $\hat{\theta}$ )	<b>1.18</b>	3.03	6.29	<b>1.90</b>	6.13	8.81
MSE( $\hat{\rho}_{\mathcal{S}}$ )	0.03	<b>0.02</b>	0.11	0.13	<b>0.12</b>	0.15
MSE( $\hat{\pi}_{\mathcal{S}}$ )	3.43	<b>2.64</b>	4.84	12.31	<b>11.88</b>	12.70
bias( $\hat{\theta}$ )	<b>-0.20</b>	-1.34	2.33	<b>-0.68</b>	-2.15	2.84
bias( $\hat{\rho}_{\mathcal{S}}$ )	-0.13	<b>-0.10</b>	-0.33	<b>-0.34</b>	-0.35	-0.39
bias( $\hat{\pi}_{\mathcal{S}}$ )	-1.57	<b>-1.38</b>	-2.05	<b>-3.51</b>	-3.61	-3.57

Tables S.7 and S.8 contain the error measures regarding the recovery of the residual sparsity pattern (see Section 2.1) for Settings 1-2 and 3-4, respectively. We note that POET, due to the lack of algebraic consistency, is completely unable to classify positive and negative elements. On the contrary, UNALCE shows a recovery rate around 70% when  $p/n$  is small. This rate is larger than for ALCE, although the recovery capability deteriorates as  $p/n$  increases.

Tables S.9 and S.10 report the Euclidean distance between the vectors of estimated and true eigenvalues (see (9)), the condition number of the estimates, and the estimated spectral norms. This table can be compared to Table S.2, which contains the true spectral norms and condition numbers across settings. All statistics suggest that UNALCE is the preferable method, with some notable exceptions due to our thresholding procedure. If  $p/n$  is small and the eigenvalues are not spiked, the spectral norm of  $\mathbf{L}^*$  and  $\mathbf{\Sigma}^*$  tend to be underestimated by UNALCE with respect to POET. On the contrary, UNALCE may overestimate the condition number of  $\mathbf{L}^*$  if  $p/n$  is large. At the same time,

**Table S.5**

This table shows the loss ( $Loss$ ), the total loss ( $TL$ ), and the sample total loss ( $SampleTL$ ) for UNALCE, ALCE, and POET estimates computed as in (4), (5) and (6) over 100 runs of Settings 1 and 2. In particular,  $Loss$ ,  $TL$  and  $SampleTL$  refer to the overall estimate,  $Loss_L$  refers to the low rank estimate and  $Loss_S$  refers to the sparse estimate. Mean values are reported with standard deviations in parentheses.

	Setting 1			Setting 2		
	UNALCE	ALCE	POET	UNALCE	ALCE	POET
$TL$	<b>6.98</b> (0.82)	6.98(0.82)	7.39(0.75)	<b>9.93</b> (2.30)	9.94(2.32)	10.47(2.21)
$SampleTL$	<b>0.72</b> (0.08)	0.89(0.09)	2.79(0.07)	<b>1.25</b> (0.15)	1.35(0.13)	3.85(0.08)
$Loss$	<b>7.63</b> (0.82)	7.64(0.82)	9.30(0.73)	<b>11.38</b> (2.29)	11.40(2.34)	13.22(2.27)
$Loss_L$	6.91(0.81)	<b>6.90</b> (0.82)	7.58(0.73)	<b>9.82</b> (2.30)	9.83(2.32)	10.74(2.27)
$Loss_S$	<b>0.72</b> (0.06)	0.74(0.05)	1.72(0.10)	<b>1.56</b> (0.18)	1.58(0.17)	2.48(0.12)

**Table S.6**

This table shows the loss ( $Loss$ ), the total loss ( $TL$ ) and the sample total loss ( $SampleTL$ ) for UNALCE, ALCE, and POET estimates computed as in (4), (5) and (6) over 100 runs of Settings 3 and 4. In particular,  $Loss$ ,  $TL$  and  $SampleTL$  refer to the overall estimate,  $Loss_L$  refers to the low rank estimate and  $Loss_S$  to the sparse estimate. Mean values are reported with standard deviations in parentheses.

	Setting 3			Setting 4		
	UNALCE	ALCE	POET	UNALCE	ALCE	POET
$TL$	<b>13.01</b> (1.98)	13.01(1.90)	13.31(2.12)	<b>20.93</b> (2.68)	20.96(2.61)	21.41(2.79)
$SampleTL$	<b>1.95</b> (0.16)	2.35(0.16)	2.90(0.12)	<b>3.91</b> (0.19)	4.66(0.29)	4.38(0.32)
$Loss$	<b>14.22</b> (2.01)	14.24(1.93)	14.89(2.18)	<b>22.51</b> (2.73)	22.57(2.63)	23.80(2.83)
$Loss_L$	12.91(2.00)	<b>12.90</b> (1.92)	13.38(2.17)	<b>20.86</b> (2.71)	20.86(2.63)	21.53(2.83)
$Loss_S$	<b>1.32</b> (0.16)	1.34(0.14)	1.47(0.12)	<b>1.65</b> (0.20)	1.71(0.22)	2.28(0.31)

if  $p/n$  is large, ALCE may provide a better estimate of the spectral norms of the targets, due to the upper bias of the largest sample eigenvalues.

## References

- 110 [1] P. J. Bickel, E. Levina, Covariance regularization by thresholding, *The Annals of Statistics* 36 (2008) 2577–2604.  
 [2] J.-F. Cai, E. J. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, *SIAM Journal on Optimization* 20 (2010) 1956–1982.  
 [3] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, A. S. Willsky, Rank-sparsity incoherence for matrix decomposition, *SIAM Journal on Optimization* 21 (2011) 572–596.  
 115 [4] I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Communications on pure and applied mathematics* 57 (2004) 1413–1457.  
 [5] K. R. Davidson, S. J. Szarek, Local operator theory, random matrices and banach spaces, *Handbook of the geometry of Banach spaces* 1 (2001) 131.  
 [6] J. Fan, Y. Liao, M. Mincheva, Large covariance estimation by thresholding principal orthogonal complements, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (2013) 603–680.  
 120 [7] M. Farné, Large Covariance Matrix Estimation by Composite Minimization, Ph.D. thesis, Alma Mater Studiorum, 2016.  
 [8] X. Luo, Recovering model structures from large low rank and sparse covariance matrix estimation, *Arxiv preprint 1111.1133* (2011).  
 [9] Y. Nesterov, Gradient methods for minimizing composite functions, *Mathematical Programming* 140 (2013) 125–161.  
 [10] T. Tao, *Topics in random matrix theory*, volume 132, American Mathematical Society, 2012.

**Table S.7**

This table shows the sparsity pattern recovery measures (defined at the end of Section 2.1) with respect to UNALCE, ALCE, and POET over 100 runs of Settings 1 and 2. Mean values are reported with standard deviations in parentheses.

	Setting 1			Setting 2		
	UNALCE	ALCE	POET	UNALCE	ALCE	POET
<i>err</i>	1.95(0.45)	<b>1.70</b> (0.17)	2.42(0.04)	<b>8.08</b> (0.95)	9.70(1.35)	9.78(1.27)
<i>errplus</i>	1.12(0.99)	1.20(0.96)	<b>0.00</b> (0.00)	<b>0.03</b> (0.10)	1.83(0.70)	2.21(0.71)
<i>errtot</i>	<b>0.70</b> (0.25)	0.91(0.23)	2.38(0.00)	<b>4.45</b> (0.95)	4.47(0.75)	11.72(0.01)
<i>senspos</i>	<b>70.19</b> (10.89)	61.64(10.42)	0.00(0.00)	<b>60.98</b> (8.55)	60.84(6.84)	0.02(0.09)
<i>specpos</i>	<b>71.05</b> (11.41)	62.14(10.76)	0.00(0.00)	<b>63.12</b> (8.00)	62.99(6.57)	0.00(0.00)
<i>spec</i>	98.69(0.56)	99.16(0.27)	<b>99.97</b> (0.04)	93.81(2.49)	93.68(2.28)	<b>99.11</b> (0.58)

**Table S.8**

This table shows the sparsity pattern recovery measures (defined at the end of Section 2.1) computed for UNALCE, ALCE, and POET over 100 runs of Settings 3 and 4. Mean values are reported with standard deviations in parentheses.

	Setting 3			Setting 4		
	UNALCE	ALCE	POET	UNALCE	ALCE	POET
<i>err</i>	<b>3.21</b> (0.30)	3.22(3.1)	4.35(0.78)	3.59(0.06)	<b>3.57</b> (0.05)	3.75(0.06)
<i>errplus</i>	1.77(1.00)	1.99(0.97)	<b>0.01</b> (0.04)	0.25(0.32)	0.41(0.32)	<b>0.00</b> (0.00)
<i>errtot</i>	2.45(3.5)	<b>2.36</b> (0.29)	3.20(0.00)	3.56(0.10)	<b>3.52</b> (0.07)	3.66(0.00)
<i>senspos</i>	23.02(11.18)	<b>25.76</b> (9.28)	0.01(0.08)	2.86(2.87)	<b>3.84</b> (2.24)	0.00(0.00)
<i>specpos</i>	23.98(11.12)	<b>26.60</b> (9.41)	0.00(0.00)	2.65(2.68)	<b>3.65</b> (2.21)	0.00(0.00)
<i>spec</i>	<b>99.16</b> (6.64)	99.05(5.8)	98.82(8.1)	<b>99.95</b> (0.05)	99.94(0.04)	99.90(0.06)

**Table S.9**

This table shows the performance in terms of eigenstructure recovery (via the measure in (9)) and the estimated spectral norms and condition numbers with respect to the overall, the low rank, and the sparse component estimates obtained by UNALCE, ALCE, and POET over 100 runs of Settings 1 and 2. Mean values are reported with standard deviations in parentheses.

	Setting 1			Setting 2		
	UNALCE	ALCE	POET	UNALCE	ALCE	POET
$\lambda(\hat{\Sigma})$	<b>5.51</b> (0.86)	5.51(0.86)	5.74(1.06)	<b>5.62</b> (2.66)	5.65(2.66)	6.07(2.60)
$\lambda(\hat{S})$	<b>0.29</b> (0.08)	0.30(0.08)	1.55(0.11)	<b>0.44</b> (0.21)	0.45(0.19)	1.86(0.17)
$\lambda(\hat{L})$	7.75(2.75)	7.73(2.65)	<b>7.16</b> (2.29)	<b>5.62</b> (2.63)	5.64(2.65)	6.16(2.84)
$cond(\hat{\Sigma})$	<b>104180</b> (38855)	58366(7785)	34048(2537)	<b>2312900</b> (1160600)	1896300(713370)	1141400(369010)
$cond(\hat{S})$	<b>21571</b> (7719)	11829(1626)	4776.3(403)	<b>130610</b> (63271)	113850(120320)	40407(12802)
$cond(\hat{L})$	1.32(0.07)	1.32(0.07)	<b>1.32</b> (0.07)	4.06(0.22)	4.07(0.23)	<b>3.97</b> (0.21)
$\ \hat{\Sigma}\ _2$	20.84(0.89)	<b>20.73</b> (0.89)	21.84(0.92)	<b>130.20</b> (5.14)	130.04(5.13)	131.58(5.14)
$\ \hat{S}\ _2$	<b>3.77</b> (0.18)	3.73(0.16)	2.75(0.12)	<b>5.68</b> (0.30)	5.59(0.36)	4.14(0.20)
$\ \hat{L}\ _2$	19.84(0.83)	19.76(0.83)	<b>21.00</b> (0.88)	128.42(5.13)	<b>128.34</b> (5.12)	130.39(5.14)

**Table S.10**

This table shows the performance in terms of eigenstructure recovery (via the measure in (9)) and the estimated spectral norms and condition numbers with respect to the overall, the low rank, and the sparse component estimates obtained by UNALCE, ALCE, and POET over 100 runs of Settings 3 and 4. Mean values are reported with standard deviations in parentheses.

	Setting 3			Setting 4		
	UNALCE	ALCE	POET	UNALCE	ALCE	POET
$\lambda(\hat{\Sigma})$	<b>6.06</b> (2.59)	6.16(2.40)	6.24(2.93)	<b>10.05</b> (3.63)	10.28(3.44)	10.57(4.02)
$\lambda(\hat{S})$	<b>0.50</b> (0.15)	0.52(0.14)	1.15(0.16)	<b>0.81</b> (0.31)	0.83(0.32)	1.92(0.38)
$\lambda(\hat{L})$	<b>6.07</b> (2.64)	6.15(2.44)	6.34(3.09)	<b>10.43</b> (3.79)	10.45(3.53)	10.57(4.21)
$cond(\hat{\Sigma})$	<b>28355</b> (10242)	11456(1804)	19279(3329)	12817(3526)	7685(1237)	<b>20171</b> (3746)
$cond(\hat{S})$	<b>2469</b> (1828)	958(148)	1132.2(181)	1424(460)	875(140)	<b>1515</b> (316)
$cond(\hat{L})$	2.41(0.32)	2.44(0.33)	<b>2.35</b> (0.30)	2.99(0.52)	3.08(0.55)	<b>2.85</b> (0.47)
$\ \hat{\Sigma}\ _2$	35.36(3.68)	<b>34.99</b> (3.69)	36.03(3.68)	42.46(4.67)	<b>41.89</b> (4.68)	43.57(4.68)
$\ \hat{S}\ _2$	<b>2.73</b> (0.30)	2.76(0.28)	1.98(0.19)	4.48(0.63)	<b>4.59</b> (0.67)	3.15(0.47)
$\ \hat{L}\ _2$	<b>34.51</b> (3.68)	34.89(3.68)	35.68(3.68)	41.98(4.66)	<b>41.37</b> (4.67)	43.17(4.67)