A finite sample estimator for large covariance matrices

Matteo Farné [1]

Department of Statistical Sciences,
University of Bologna, Italy

Angela Montanari

Department of Statistical Sciences,
University of Bologna, Italy

February 22, 2018

[1] Electronic address: `matteo.farne2@unibo.it`; Corresponding author

Abstract

The present 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. This paper proposes a new estimator of that family based on an additional least-squares re-optimization step aimed at un-shrinking the eigenvalues of the low rank component estimated at the first step. We prove that such un-shrinkage causes the final estimate to approach the target as closely as possible while recovering exactly the underlying low rank and sparse matrix varieties. In addition, consistency is guaranteed until $p \log(p) \gg n$, where p is the dimension and n is the sample size, and recovery is ensured if the latent eigenvalues scale to p^{α}, $\alpha \in [0,1]$. The resulting estimator is called UNALCE (UNshrunk ALgebraic Covariance Estimator) and is shown to outperform both LOREC and POET estimators, especially for what concerns fitting properties and sparsity pattern detection.

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 different goal the interest is sometimes in inferring the eigenstructure of the covariance matrix (as in PCA) and sometimes in estimating its inverse (as in discriminant analysis or in graphical models). Examples of application areas where these problems arise include gene arrays, fMRI, 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, Anderson (1984)). However, it became soon 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 very high-dimensional and sample sizes can be very small relative to dimension.

The most 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 assume a particular relevance in large dimensions: well conditioning (i.e. numerical stability) and identifiability. Both properties are crucial for the theoretical recovery 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 cause the impossibility to identify the unknown covariance structure thus masking the interpretation of the results.

Regularization approaches to large covariance matrices estimation have therefore started to be presented in the literature, both from theoretical and practical points of view. Some authors propose shrinkage towards the identity matrix (Ledoit and Wolf, 2004), others consider tapering the sample covariance matrix, that is, gradually shrinking the off-diagonal elements toward zero (Furrer and Bengtsson (2007), Cai et al. (2010)). At the same time, a common approach is to encourage sparsity, either by a penalized likelihood approach (Friedman et al., 2008) or by banding (Bickel and Levina, 2008b) or thresholding the sample covariance matrix (Bickel and Levina (2008a), Rothman et al. (2009), Cai and Liu (2011)). A consistent bandwidth selection method for all these approaches is described in Qiu and Chen (2015).

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

$$\hat{\varSigma} = L^* + S^* + W = \varSigma^* + W, \tag{1}$$

where $\hat{\varSigma}$ is a generic covariance estimator, W is an error term, S^* is **sparse** having at most s nonzero elements and L^* is **low rank** with rank $r < p$ and Σ^* is a positive definite matrix. The combination of regularization techniques and dimensionality reduction methods allows to lower the condition number and the parameter space dimensionality simultaneously. In Fan et al. (2013), a large covariance matrix estimator of this family, called POET (Principal Orthogonal complEment Thresholding), is derived. 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, thus allowing to reasonably use the truncated PCA of the sample covariance matrix. Furthermore, at the same time, sparsity in the sense of Bickel and Levina (2008a) is imposed to the residual matrix. The latent rank r is chosen by the IC criteria of Bai and Ng (2002).

Indeed, rank selection represents a relevant issue: if p is large, setting a large rank would cause the estimate to be non-positive definite, while setting a small rank would cause a too relevant variance loss. In the discussion of Fan et al. (2013), Yu and Samworth point out that the probability to underestimate the latent rank 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 $\varSigma_n$ are more spiky than the true ones (as proved in Ledoit and Wolf (2004)).

POET consistency holds given that a number of assumptions is satisfied. The key assumption is the pervasiveness of latent factors, which causes the PCA of $\hat{\Sigma}_n$ to asymptotically identify the eigenvalues and the eigenvectors of Σ^*. The results of Fan et al. (2013) provide the convergence rates of the relative norm of $\hat{\varSigma} - \varSigma^*$ (defined as $||\hat{\varSigma} - \varSigma^*||_{\varSigma} = p^{-1/2}||\varSigma^{*-\frac{1}{2}}\hat{\varSigma}\varSigma^{*-\frac{1}{2}} - I_p||_{Fro}$), the maximum norm of $\hat{\varSigma} - \varSigma^*$ and the spectral norm of $\hat{S}_{\hat{r}}^{\mathbf{T}} - S^*$. Under more strict conditions, $\hat{S}_{\hat{r}}^{\mathbf{T}}$ and $\hat{\varSigma}_{POET,\hat{r}}$ are proved to be non-singular with probability approaching 1.

At the same time, a number of non-asymptotic methods has been presented. In Chandrasekaran et al. (2011) the exact recovery of the covariance matrix in the noiseless context is first proved. The result is achieved 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 Chandrasekaran et al. (2012), which is an extension of Chandrasekaran et al. (2011), the exact recovery of the inverse covariance matrix by the same numerical problem 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 $n = O\left(\frac{p^3}{r^2}\right)$ (where r is the rank), 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 Agarwal et al. (2012). 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 Chandrasekaran et al. (2012), is then shown in Luo (2011b). 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 Chandrasekaran et al. (2012).

In fact, in Chandrasekaran et al. (2012) algebraic consistency is defined as follows

Definition 1.1 *A pair of symmetric matrices (S, L) with $S, L \in \mathbb{R}^{p \times p}$ is an algebraically consistent estimate of the low rank plus sparse model (1) for the covariance matrix $\varSigma^*$ if the following conditions hold:*

1. *The sign pattern of S is the same of S^*: $sign(S_{ij}) = sign((S^*)_{i,j})$, $\forall i, j$. Here we assume that $sign(0) = 0$.*

2. *The rank of L is the same as the rank of L^*.*

3. *Matrices $L+S$, S and L are such that: $L+S \succ 0$, $S \succ 0$, $L \succeq 0$.*

Parametric consistency holds if the estimates of (S, L) are close to (S^*, L^*) in some norm with high probability. In Chandrasekaran et al. (2012) such norm is $g_\gamma = \max\left(\frac{||\hat{S}-S^*||_\infty}{\gamma}, ||\hat{L}-L^*||_2\right)$.

LOREC shows several advantages respect to POET. The most important is that the estimates are both algebraically and parametrically consistent, while POET provides only parametric consistency. In spite of that, LOREC suffers from some drawbacks, especially concerning fitting properties. What is more, the strict condition $p \leq n$ is required, while POET allows for $p\log(p) \gg n$.

For these reasons, we propose a new estimator, UNALCE (UNshrunk ALgebraic Covariance Estimator), based on the unshrinkage 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 L^* and Σ^* are proportional to p^α, $\alpha \in [0,1]$ (the so called generalized spikiness context), and we drop the LOREC sparsity assumption on Σ^*. We prove that under POET conditions, our estimator allows for $p\log(p) \gg n$, and that under the generalized spikiness context it allows for $p^\alpha \log(p) \gg n$. We derive absolute bounds for the low rank, the sparse component, and the overall estimate, as well as the conditions for positive definiteness and invertibility. 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 first define ALCE (ALgebraic Covariance Estimator) with the necessary assumptions for algebraic and parametric consistency, and then define UNALCE, 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 3 we propose a new model selection criterion specifically tailored to our model setting. In Section 4 we describe and comment some simulation studies, showing that our estimator outperforms POET in different situations. In Section 5 we provide a real Euro Area banking data example which clarifies the effectiveness of our approach. Finally, in Section 6 we draw the conclusions and discuss the most relevant findings.

2 Numerical estimation and spiked eigenvalues: the ALCE approach

2.1 The data

Let us suppose 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 $\mathit{\Sigma}^*$ satisfies the following relationship:

$$\mathit{\Sigma}^* = L^* + S^*, \tag{2}$$

where:

1. L^* is a positive semidefinite symmetric $p \times p$ matrix with at most rank $r \ll p$;
2. S^* is a positive definite $p \times p$ sparse matrix with at most $s \ll p(p-1)/2$ nonzero elements.

According to the spectral theorem, we can write $L^* = U_L D U_L' = BB'$, where $B = U_L D^{1/2}$, U_L is a $p \times r$ orthogonal matrix, D is a $r \times r$ diagonal matrix, with $d_{jj} > 0$, $\forall j = 1, \ldots, r$. Suppose that the $p \times 1$ random vector $\mathbf{x}$ is generated according to the following model:

$$\mathbf{x} = B\mathbf{f} + \epsilon, \tag{3}$$

with

$$\mathbf{f} = N_r(\mathit{0}, I_r); \tag{4}$$

$$\epsilon = N_p(\mathit{0}, S^*), \tag{5}$$

where $\mathbf{f}$ is a $r \times 1$ random vector, and ϵ is $p \times 1$ random vector. $\mathbf{x}$ is assumed to be a zero mean random vector, without loss of generality. Given a sample $\mathbf{x_i}$, $i = 1, \ldots, n$ $\Sigma_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_i}\mathbf{x_i}'$

is the $p \times p$ sample covariance matrix.

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

$$\begin{aligned} E(\mathbf{x}\mathbf{x}') = \mathbf{E}((B\mathbf{f}+\epsilon)(B\mathbf{f}+\epsilon)') = \\ = E(B'\mathbf{f}'\mathbf{f}B) + E(B\mathbf{f}\epsilon') + \mathbf{E}(\epsilon B'\mathbf{f}') + \mathbf{E}(\epsilon\epsilon') = \\ = BB' + S^* = \mathit{\Sigma}^* \end{aligned} \tag{6}$$

under the usual assumption $\mathbf{f} \perp \epsilon$, i.e. $cov(\mathbf{f}, \epsilon) = \mathbf{E}(\mathbf{f}\epsilon') = \mathbf{E}(\epsilon\mathbf{f}') = \mathit{0}$ ($r \times p$ null matrix). If we assume a normal distribution for $\mathbf{f}$ and ϵ, we know that the matrix $W := \Sigma_n - (BB' + S^*)$ is distributed as a re-centered Wishart noise. In any case, the normality assumption is not essential in the finite sample context.

2.2 LOREC approach

The approach of Luo (2011b) is based on numerical analysis, exploiting the theory of non-smooth convex optimization provided by Rockafellar (2015) and Clarke (1990). Under model (1), this method allows at the same time to consistently estimate the covariance matrix $\mathit{\Sigma}^*$ and to catch the sparsity pattern of S^* and the spikiness pattern of the eigenvalues of L^* simultaneously.

The estimation problem is stated as

$$\min_{L,S} \frac{1}{2}||(L+S) - \mathit{\Sigma}_n||^{\mathit{2}}_{\mathit{Fro}} + \lambda rank(L) + \rho||S||_{\mathit{0}}, \tag{7}$$

where $||S||_{\mathit{0}}$ is the number of nonzero elements. This is a combinatorial problem, which is known to be NP-hard, since both $rank(L)$ and $||S||_{\mathit{0}}$ are not convex. A very well known convex relaxation of problem (7) is

$$\min_{L,S} \frac{1}{2}||(L+S) - \mathit{\Sigma}_n||^{\mathit{2}}_{\mathit{Fro}} + \lambda||L||_* + \rho||S||_{\mathit{1}}, \tag{8}$$

where λ and ρ are non-negative **threshold** parameters, $||S||_{\mathit{1}} = \sum_{i=\mathit{1}}^{n}\sum_{j=\mathit{1}}^{n}|s_{ij}|$ is the l_1

norm of S, $||L||_* = \sum_{i=1}^{r} |d_i| = \sum_{i=1}^{r} d_i = ||diag(D)||_1$ is the nuclear norm of L^*, where D is the diagonal matrix of the eigenvalues of L. Basic references are Tibshirani (1996) for the former and Fazel et al. (2001) for the latter.

From an algebraic point of view, (8) is an algebraic matrix variety recovery problem. In the noisy covariance matrix setting described in equation (1), matrices L^* and S^* are assumed to come from the following sets of matrices:

$$B(r) = \{L \in \mathbb{R}^{p\times p} \mid L = UDU', U \in \mathbb{R}^{p\times r}\text{orthogonal}, D \in \mathbb{R}^{r\times r}\text{diagonal}\} \tag{9}$$

$$\mathcal{A}(s) = \{S \in \mathbb{R}^{p\times p} \mid |support(S)| \leq s\}. \tag{10}$$

$\mathcal{B}(r)$ is the variety of matrices with **at most** rank r. $\mathcal{A}(s)$ is the variety of (entrywise) sparse matrices with **at most** s nonzero elements, where $support(S)$ is the orthogonal complement of $ker(S)$.

In Chandrasekaran et al. (2011) the notion of rank-sparsity incoherence is developed, which is defined as the uncertainty principle between the sparsity pattern of a matrix and its row/column space. Denoting by $T(L)$ and $\Omega(S)$ the tangent spaces to $\mathcal{B}(r)$ and $\mathcal{A}(s)$ respectively, the following rank-sparsity incoherence measures between $\Omega(S^*)$ and $T(L^*)$ are defined:

$$\xi(T(L^*)) = \max_{N\in T(L^*), ||N||_2 \leq 1} ||N||_\infty, \tag{11}$$

$$\mu(\Omega(S^*)) = \max_{N\in \Omega(S^*), ||N||_\infty \leq 1} ||N||_2. \tag{12}$$

In order to identify $T(L^*)$ and $\Omega(S^*)$, we need that quantities $\xi(T(L^*))$ and $\mu(\Omega(S^*))$ to be small as possible, because the smaller they are, the better is the decomposition. The product $\mu(\Omega(S^*))\xi(T(L^*))$ is the rank-sparsity incoherence measure and bounding it controls both for identification and recovery.

The reference matrix class for the covariance matrix in Luo (2011b) is

$$\Sigma^*(\epsilon_0) = \{M \in \mathbb{R}^{p\times p} : 0 < \epsilon_0 \leq \Lambda_i(M) \leq \epsilon_0^{-1}, \forall i = 1, \dots, p\} \tag{13}$$

which is the class of positive definite matrices having uniformly bounded eigenvalues ($\mathit{\Lambda}_i(M)$, $i = \mathit{1}, \ldots, p$, are the eigenvalues of M). In the context so far described, Luo proves that L and S can be identified and recovered with bounded error, and the rank of L as well as the sparsity pattern of S are exactly recovered (see (Luo, 2011b) for the details).

The key model-based results for deriving consistency bounds are a lemma by Bickel and Levina (2008a) for the sample loss in infinity (element-wise) norm:

$$||\Sigma_n - \Sigma^*||_\infty \leq O\left(\sqrt{\frac{\log p}{n}}\right), \tag{14}$$

and a lemma by Davidson and Szarek (2001) for the sample loss in spectral norm:

$$||\Sigma_n - \Sigma^*||_2 \leq O\left(\sqrt{\frac{p}{n}}\right), \tag{15}$$

where $\Sigma_n = \hat{\Sigma}_{n-1}$ is the $p \times p$ unbiased sample covariance matrix. We note that if the input is $\hat{\Sigma}_{n-1}$, the condition $p \leq n$ is unavoidable.

From a theoretical point of view, LOREC approach presents some deficiencies and incongruities. Differently from POET approach, where the sparsity assumption is imposed to the sparse component S^*, LOREC approach imposes it directly to the covariance matrix Σ^*. As a consequence, the assumption $\Sigma^* \in \Sigma^*(\epsilon_0)$ (see (13)) is necessary and causes, jointly with the identifiability assumptions, uncertainty on the underlying structure of Σ^*.

In fact, assuming uniformly bounded eigenvalues may conflict with the main necessary identifiability condition: the transversality between Ω and T. Since the eigenvalue structures of Σ^* and S^* are somehow linked, allowing class (13) for Σ^* may cause S^* to be not enough sparse, and simultaneously the row/column space of L^* to have high values of incoherence, because we have no spiked eigenvalues. This may result in possible non-identifiability issues.

2.3 ALCE estimator

Let us suppose that the eigenvalues of L^* are spiked in the sense of Proposition 1 in Fan et al. (2013), and that all propositions and assumptions of POET approach hold in our finite sample context.

In other words, we suppose that

$$\lambda_{1,\dots,r}(\Sigma^*) \gg \delta p,$$
$$\lambda_{r+1,\dots,p}(\Sigma^*) \ll \delta p,$$

$\delta \neq 0$, that is, the eigenvalues of $p^{-1}B'B$ are bounded away from 0 and ∞.

Suppose that $p = o(n^2)$ and that all the other assumptions of Theorem 2 in Fan et al. (2013) hold. In particular, suppose that the following bound can be proved

$$||\Sigma_n - \Sigma^*|| = O\left(\frac{p}{\sqrt{n}}\right) \tag{16}$$

which is equivalent to state that

$$P\left(||\Sigma_n - \Sigma^*|| \geq C_1 \frac{p}{\sqrt{n}}\right) \leq 1 - C_2 e^{-C_3 p^2}, \tag{17}$$

where C_1, C_2, C_3 are positive constants.

In addition, suppose that all the conditions for proving LOREC consistency hold. We only drop assumption (13), which is possibly dangerous for model identifiability as explained in Section 2.2. As a consequence, we can simply write, using the standard norm property $||.||_\infty \leq ||.||_2$ (Chandrasekaran et al., 2012),

$$P\left(||\Sigma_n - \Sigma^*||_\infty \geq C_1 \xi(T) \frac{p}{\xi(T)\sqrt{n}}\right) \leq 1 - C_2 e^{-C_3 p^2}. \tag{18}$$

Replacing bound (15) by bound (16) in the proof of Theorem 1 of Luo (2011b) and recalling that the consistency norm in Chandrasekaran et al. (2012) is

$$g_\gamma = \max\left(\frac{||\hat{S} - S^*||_\infty}{\gamma}, ||\hat{L} - L^*||_2\right), \tag{19}$$

we can conclude, exploiting the results therein described, that

$$g_\gamma(\hat{S}-S^*,\hat{L}-L^*) \preceq \frac{1}{\xi(T)}\frac{p}{\sqrt{n}}. \tag{20}$$

In this way it is possible to generalize the numerical approach based on (8) to the POET pervasiveness context. It is straightforward that the success of this approach depends on the coherence between the assumptions of spiked eigen-structure of Fan et al. (2013) and the identifiability assumptions for the recovery of underlying matrix varieties.

In order to relax the strong assumption of pervasiveness of latent factors, we define the generalized pervasiveness context for $\alpha \in (0,1]$ as follows (Fan et al. (2013), Yu and Samworth, p. 656):

Definition 2.1 *The eigenvalues of Σ^* follow a α-generalized spikiness structure if and only if all the eigenvalues of the $r \times r$ matrix $p^{-\alpha}B'B$ are bounded away from 0 and ∞ as $p \to \infty$.*

If $\alpha = 1$, we fall back to the POET setting.

Since this approach is non-asymptotic in nature, we need to study the behaviour of the model-based quantity $P(||\Sigma_n - \Sigma^*||)$, which is the only probabilistic component. Exploiting the property $|||.||_\infty \leq ||.||_2$, $P(||\Sigma_n - \Sigma^*||_\infty)$ is bounded as a consequence. Therefore, our aim is to generalize (17) showing that

$$P\left(||\Sigma_n - \Sigma^*|| > C_1\frac{p^\alpha}{\sqrt{n}}\right) \leq 1 - C_2 e^{-C_3 p^{2\alpha}}, \tag{21}$$

$\alpha \in (0,1]$.

In order to do that, Proposition 1 and 2 of Fan et al. (2013) need to be modified as follows to fit into the generalized spikiness context.

Proposition 2.1 *All the eigenvalues of the $r \times r$ matrix $B'B$ are bounded away from 0 for all large p^α, $\alpha \in (0,1)$. Under the assumptions $cov(f_t) = I_r$ and $B'B$ diagonal we have:*

$$|\lambda_j - ||\tilde{b}_j||^2| \leq ||S^*||, \qquad j \leq r$$

$$|\lambda_j| \leq ||S^*||, \qquad j > r.$$

In addition, for $j \leq r$, $\liminf_{p\to\infty} ||\tilde{b}_j||^2/(p^{\alpha}) > 0$.

Proposition 2.2 *Under the assumptions of Proposition 1, if $||\tilde{b}_j||_{j=1}^{r}$ are distinct, then $||u_j - \tilde{b}_j/||\tilde{b}_j|||| = O(p^{-\alpha}||S^*||)$.*

The resulting theorem is now reported.

Theorem 2.1 *Let $\Omega = \Omega(S^*)$ and $T = T(L^*)$. Suppose that all the assumptions of Theorem 2 in Fan et al. (2013) hold, given that Propositions 1 and 2 of Fan et al. (2013) are replaced by Propositions (2.1) and (2.2) respectively, $m_p = o_p(p^{\alpha})$ and $r = O(\log p^{\alpha})$, $\alpha \in (0,1)$. Suppose $\mu(\Omega(S^*))\xi(T(L^*)) \leq \frac{1}{54}$,*

$$\lambda = \left(\frac{1}{\xi(T)}\frac{p^{\alpha}}{\sqrt{n}}\right),$$

with at most $p^{\alpha}\log(p) \gg n$ and $\rho = \gamma\lambda$, where $\gamma \in [9\xi(T), 1/(6\mu(\Omega))]$. In addition, suppose that the minimum singular value of L^* ($\lambda_r(L^*)$) is greater than $C_2\frac{\lambda}{\xi^2(T)}$ and the smaller absolute value of the nonzero entries of S^*, $S_{max,off}$, is greater than $C_3\frac{\lambda}{\mu(\Omega)}$. Then, with probability greater than $1 - C_4 p^{-C_5}$, the pair $(\hat{L}, \hat{S})$ minimizing (8) recovers the rank of L^* and the sparsity pattern of S^* exactly:

$$rank(\hat{L}) = rank(L^*) \; and \; sign(\hat{S}) = sign(S^*).$$

Moreover, with probability greater than $1 - C_4 p^{-C_5}$, the matrix losses for each component are bounded as follows:

$$||\hat{L} - L^*||_2 \leq C\lambda, \qquad ||\hat{S} - S^*||_{\infty} \leq C\rho.$$

The proof is reported in the Appendix. It is straightforward that the success of this approach depends on the coherence between the relaxed spikiness assumption and the identifiability conditions of the low rank plus sparse problem. In particular, if we increase α (*ceteris paribus*), both $\lambda_r(L^*)$ and $S_{max,off}$ must be larger to ensure identifiability. The same happens if p

increases, because, according to Chandrasekaran et al. (2011), both $\xi(T(L^*))$ and $\mu(\Omega(S^*))$ depend inversely on p. On the contrary, to ensure consistency, if r increases L^* can have less spiked eigenvalues, while if s increases $S_{max,off}$ can be larger. This occurs because, according to Chandrasekaran et al. (2011), $\xi(T(L^*))$ and $\mu(\Omega(S^*))$ depend directly on r and s respectively.

If all these assumptions are valid, we can write

$$\lambda = \left(\frac{1}{\xi(T)} \frac{p^\alpha}{\sqrt{n}} \right). \tag{22}$$

In this relaxed setting, the probabilistic bound is finite until $p^\alpha \log(p) \gg n$, because the bound (21) holds until $p^\alpha = o(n^2)$. In this way we overcome the problem of the restrictive condition $p \leq n$. We call the resulting covariance estimator ALCE (ALgebraic Covariance Estimator): $\hat{\Sigma}_{ALCE} = \hat{L}_{ALCE} + \hat{S}_{ALCE}$.

From Theorem 2.1, we can derive with probability larger than $1 - C_1 p^{-C_2}$ the following bounds for $\hat{\mathit{\Sigma}}_{ALCE}$:

$$||\hat{\mathit{\Sigma}}_{ALCE} - \mathit{\Sigma}^*||_2 \leq C(s'\xi(T) + 1)\lambda = \phi, \tag{23}$$

$$||\hat{\mathit{\Sigma}}_{ALCE} - \mathit{\Sigma}^*||_{Fro} \leq C(\sqrt{ps'}\xi(T) + \sqrt{r})\lambda, \tag{24}$$

which hold if and only if $\lambda_{min}(\mathit{\Sigma}^*) > \phi$. Here r is the true latent rank of L^*, while s' is defined as the maximum number of non-zero elements per column. The same bounds hold for the inverse covariance estimate $\hat{\mathit{\Sigma}}^{-1}_{ALCE}$ with the same probability:

$$||\hat{\mathit{\Sigma}}^{-1}_{ALCE} - \mathit{\Sigma}^{*-1}||_2 \leq C(s'\xi(T) + 1)\lambda = \phi \tag{25}$$

$$||\hat{\mathit{\Sigma}}^{-1}_{ALCE} - \mathit{\Sigma}^{*-1}||_{Fro} \leq C(\sqrt{ps'}\xi(T) + \sqrt{r})\lambda \tag{26}$$

given that $\lambda_{min}(\mathit{\Sigma}^*) \geq 2\phi$.

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

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

$$||\hat{S} - S^*||_2 \leq Cs'\xi(T)\lambda = \phi_S. \tag{27}$$

From $||\hat{S} - S^*||_{Fro} \leq \sqrt{ps'}||\hat{S} - S^*||_\infty$, we obtain

$$||\hat{S} - S^*||_{Fro} \leq C\sqrt{ps'}\xi(T)\lambda. \tag{28}$$

$\hat{S}$ is positive definite if and only if $\lambda_{min}(S^*) > \phi_S$. $\hat{S}^{-1}$ has the same bound of $\hat{S}$ if and only if $\phi_S^{-1} \geq 2\lambda_{min}(S^*)^{-1}$.

We note that if α decreases to 0, we obtain $p = O(1) = o(n)$. Therefore, in absence of a spikiness structure (because $r = \log(p^0) = \log(1) = 0$), the convergence rate of the sample covariance matrix simply becomes $O(\sqrt{\frac{1}{n}})$. It is easy to note that imposing $p = o(n)$ the bound (15) assumes the same shape. Therefore, we can affirm that (21) holds for $\alpha \in [0, 1]$. In this way ALCE approach encompasses both the POET case and the classic case (small and fixed data dimension).

To sum up, by ALCE estimator we offer the chance to recover consistently a relaxed spiked eigen-structure, thus overcoming the condition $p \leq n$, even using the sample covariance matrix as estimation input. In this approach, the recovery quality directly depends on the spikiness of latent eigenvalues, because the larger α, the further are identifiability and invertibility conditions from being satisfied, as well as the worse is the error bound. In addition, the ratio $\frac{p}{n}$ directly impacts on the error bound. We emphasize that our bounds are in absolute norms, and reflect the underlying degree of spikiness.

However, we remark that ALCE approach works if and only if the identifiability and consistency assumptions of the low rank plus sparse variety identification problem are satisfied. In particular, the more spiky the low rank component is, the sparsest must be the sparse component, in order to ensure a degree of transversality sufficiently low.

Finally, we remark that the proposed theory is able to address the large dimensional context, where $p > n$. Sparse factor model assumptions together with the numerical approach are the keys to provide recovery in such settings. This result is obtained by a combined use of random matrix theory and the probabilistic convergence theory of the sample covariance matrix under the sparse factor model assumption.

2.4 UNALCE estimator: a re-optimized ALCE solution

Let us define $\Delta_L = \hat{L}_{ALCE} - L^*$,$\Delta_S = \hat{S}_{ALCE} - S^*$,$\Delta_\Sigma = \hat{\Sigma}_{ALCE} - \Sigma^*$. A key aspect of Theorem 2.1 is that the two losses in L^* and 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 rises to correct for this drawback, re-shaping $\hat{\Sigma}_{ALCE}$, as ALCE approach is somehow sub-optimal for the whole covariance matrix.

We approach this problem by a finite-sample analysis, which could be referred to as a re-optimization least squares method. We refer to the usual objective function (8) with $||S||_1 = ||S||_{1,off} = \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} |s_{ij}|$, i.e. the l_1 norm of S excluding the diagonal entries, consistently with POET approach.

Suppose that $\hat{\mathcal{B}}(\hat{r})$ and $\hat{\mathcal{A}}(\hat{s})$ are the varieties ensuring the algebraic consistency of (8). One might look for the solution (say $(\hat{L}_{New}, \hat{S}_{New})$) of the problem

$$\min_{L \in \hat{\mathcal{B}}(\hat{r}), S \in \hat{\mathcal{A}}(\hat{s})} TL(L,S) = ||(\Sigma_n - (L+S)||^2_{Fro}, \tag{29}$$

where $TL(L,S)$ stands for *Total Loss*. The sample covariance matrix follows the model $\Sigma_n = L^* + S^* + W$, where we might assume $W \sim Wishart(\mathbf{0}_{p \times p}, n-1)$, given a sample of $p-$dimensional data vectors $\mathbf{x_i}$, $i = 1, \ldots, n$. Our problem essentially is: which pair $L \in \hat{\mathcal{B}}(\hat{r}), S \in \hat{\mathcal{A}}(\hat{s})$ satisfying algebraic consistency shows the best approximation properties of Σ_n?

We prove the following original result.

Theorem 2.2 *Suppose that $\hat{L}_{ALCE}$ and $\hat{S}_{ALCE}$ are the ALCE solutions, with*

$$\hat{\Sigma}_{ALCE} = \hat{L}_{ALCE} + \hat{S}_{ALCE}.$$

Suppose that $\hat{\mathcal{B}}(r)$, and $\hat{\mathcal{A}}(s)$ are the recovered matrix varieties, and that $\hat{L} = \hat{U}\hat{D}\hat{U}'$ is the eigenvalue decomposition of $\hat{L}_{ALCE}$. Define $\hat{S}_{New}$ and $\hat{\varSigma}_{New}$ such that their off-diagonal elements are the same as $\hat{S}_{ALCE}$ and $\hat{\varSigma}_{ALCE}$ respectively. Then, the minimum

$$\min_{L \in \hat{\mathcal{B}}(\hat{r}), S \in \hat{A}(\hat{s})} \|\varSigma_n - (L+S)\|_{Fro}^2$$

is achieved if and only if

$$\hat{L}_{new} = \hat{U}(\hat{D} + \lambda I_r)\hat{U}' \quad \text{and if} \quad diag(\hat{S}_{New}) = diag(\hat{\varSigma}_{ALCE}) - diag(\hat{L}_{new})$$

where λ is the threshold parameter in (8). We call the resulting overall estimate $\hat{\varSigma}_{new} = \hat{L}_{new} + \hat{S}_{new}$ UNALCE (UNshrunk ALgebraic Covariance Estimator).

Theorem (2.2) essentially states that the sample total loss (29) is minimized if we un-shrink the eigenvalues of $\hat{L}_{ALCE}$ (re-adding the threshold λ). The re-shaped ALCE low rank solution $\hat{L}_{UNALCE}$ belongs to the same low rank matrix variety as $\hat{L}_{ALCE}$ (the eigenvectors do not change). Assuming the invariance of the off-diagonal elements of $\hat{S}_{ALCE}$ and of the diagonal elements of $\hat{\varSigma}$, the reshaped sparse ALCE solution minimizing (29) has a new diagonal, which is simply the difference between the diagonal of the original $\hat{\varSigma}_{ALCE}$ and the diagonal of the newly computed $\hat{L}_{UNALCE}$. The resulting $\hat{S}_{UNALCE}$ belongs to the same sparse matrix variety as S_{ALCE}, because the sparsity pattern is exactly the same.

Four consequences of Theorem (2.2) are reported in Corollary 2.1.

Corollary 2.1 *The gains in terms of spectral loss for $\hat{L}_{UNALCE}$, $\hat{S}_{UNALCE}$ in comparison to $\hat{L}_{ALCE}$, $\hat{S}_{ALCE}$ respectively are all strictly positive and bounded by λ:*

$$0 < ||\hat{L}_{ALCE} - L^*||_2 - ||\hat{L}_{UNALCE} - L^*||_2 \leq \lambda, \tag{30}$$

$$0 < ||\hat{S}_{ALCE} - S^*||_2 - ||\hat{S}_{UNALCE} - S^*||_2 \leq \lambda. \tag{31}$$

The gains in terms of Frobenius norm are all strictly positive and bounded as follows:

$$0 < ||\hat{L}_{ALCE} - L^*||_{Fro} - ||\hat{L}_{UNALCE} - L^*||_{Fro} \leq \sqrt{2r}\lambda, \tag{32}$$

$$0 < ||\hat{S}_{ALCE} - S^*||_{Fro} - ||\hat{S}_{UNALCE} - S^*||_{Fro} \leq \sqrt{r}\lambda. \tag{33}$$

Two further relevant consequences of Theorem (2.2) are reported in Corollary 2.2.

Corollary 2.2 *The gain in terms of spectral sample total loss for $\hat{\Sigma}_{UNALCE}$ respect to $\hat{\Sigma}_{ALCE}$ is strictly positive and bounded by λ:*

$$0 < ||\Sigma_n - \hat{\Sigma}_{ALCE}||_2 - ||\Sigma_n - \hat{\Sigma}_{UNALCE}||_2 \leq \lambda. \tag{34}$$

The gain in terms of Frobenius sample total loss for $\hat{\Sigma}_{UNALCE}$ respect to $\hat{\Sigma}_{ALCE}$ is strictly positive and bounded by $\sqrt{r}\lambda$:

$$0 < ||\Sigma_n - \hat{\Sigma}_{ALCE}||_{Fro} - ||\Sigma_n - \hat{\Sigma}_{UNALCE}||_{Fro} \leq \sqrt{r}\lambda. \tag{35}$$

The proofs of Corollaries 2.1 and 2.2 are reported in Appendix A.

The rationale of the two Corollaries is the following. We accept to pay the price of a non-optimal solution in terms of nuclear norm (we allow to increment $||\hat{L}||_*$ by $r\lambda$) but we have a best fitting performance for the whole covariance matrix, decrementing the squared Frobenius loss of $\hat{\Sigma}$ by a quantity bounded by $r\lambda^2$. The l_1 norm of S excluding the diagonal, $||\hat{S}||_{off}$, is unvaried, while the norm $||S||_1$ (included the diagonal) is decreased by a quantity bounded by $\sqrt{r}\lambda$.

The following result describes the losses for $\hat{\Sigma}_{UNALCE}$ from the target Σ^* respect to $\hat{\Sigma}_{ALCE}$.

Theorem 2.3 *The gains in terms of spectral loss and Frobenius loss for $\hat{\Sigma}_{UNALCE}$ respect* *to $\hat{\Sigma}_{ALCE}$ are strictly positive and bounded as follows:*

$$0 < ||\hat{\Sigma}_{ALCE} - \Sigma^*||_2 - ||\hat{\Sigma}_{UNALCE} - \Sigma^*|||_2 \leq \lambda, \tag{36}$$

$$0 < ||\hat{\Sigma}_{ALCE} - \Sigma^*||_{Fro} - ||\hat{\Sigma}_{UNALCE} - \Sigma^*|||_{Fro} \leq \sqrt{r}\lambda. \tag{37}$$

We argue that the gain is strictly positive and bounded by λ for the spectral loss, by $\sqrt{r}\lambda$ for the Frobenius loss. The following Corollary extends our framework to the performance of $(\hat{\Sigma}_{UNALCE})^{-1}$.

Corollary 2.3 *The gains in terms of spectral loss and Frobenius loss for $\hat{\Sigma}^{-1}_{UNALCE}$ respect to $\hat{\Sigma}^{-1}_{ALCE}$ are strictly positive and bounded as follows:*

$$0 < ||\hat{\Sigma}^{-1}_{ALCE} - \Sigma^{*-1}||_2 - ||(\hat{\Sigma}^{-1}_{UNALCE} - \Sigma^{*-1}||_2 \leq \lambda. \tag{38}$$

$$0 < ||\hat{\Sigma}^{-1}_{ALCE} - \Sigma^{*-1}||^2_{Fro} - ||\hat{\Sigma}^{-1}_{UNALCE} - \Sigma^{*-1}||^2_{Fro} \leq r\lambda^2. \tag{39}$$

The proof of Theorem 2.3 and Corollary 2.3 are reported in Appendix A.

The outlined results allow us to improve the estimation performance given the finite sample. However, the non-asymptotic bounds for $\hat{L}_{UNALCE}$, $\hat{S}_{UNALCE}$ and $\hat{\Sigma}_{UNALCE}$ are exactly the ones of $\hat{L}_{ALCE}, \hat{S}_{ALCE}$ and $\hat{\Sigma}_{ALCE}$. UNALCE improves systematically the fitting performance of ALCE, inheriting all its algebraic and parametric consistency properties.

In order to conclude, we remark that the un-shrinkage of the estimated eigenvalues of $\hat{L}$ makes the necessary condition for positive definiteness and invertibility of $\hat{S}$ and $\hat{\Sigma}$ milder. In fact, in empirical analysis, one can consider that parameters $\phi = C(s'\xi(T) + 1)\lambda$ and $\phi_S = Cs'\xi(T)\lambda$ can be decreased by a quantity bounded by λ (s' is the maximum number of non zero elements per column).

3 A new model selection criterion: *MC*

In empirical applications, the selection of thresholds λ and ρ in equation (8) requires a model selection criterion consistent with the described estimation method. The motivation rises from the consistency norm g_γ used in Luo (2011b) (see (19)). Our aim is to detect the optimal threshold pair (λ, ρ) in respect to the spikiness/sparsity trade-off. In order to exploit (19) with model selection purposes, we need to make the two terms comparable, i.e., the need of rescaling both arguments of g_γ rises.

Considered that $\frac{||\hat{S}||_\infty}{\gamma}$ contains a maximum norm, we can re-scale it to the trace of $\hat{S}$, which is estimated by $(1-\hat{\theta})trace(\varSigma_n)$ ($\hat{\theta}$ is the estimated proportion of latent variance). Similarly, in order to compare the magnitude of the two quantities, we multiply $||\hat{L}||_2$ by r, since $||\hat{L}||_* \leq r||\hat{L}||_2$, and then divide it by the trace of $\hat{L}$, estimated by $\hat{\theta}trace(\varSigma_n)$. Our maximum criterion MC can be therefore defined as follows:

$$MC = \max\left\{\frac{\hat{r}||\hat{L}||_2}{\hat{\theta}trace(\varSigma_n)}, \frac{||\hat{S}||_\infty}{\gamma(1-\hat{\theta})trace(\varSigma_n)}\right\}, \tag{40}$$

where $\gamma = \frac{\rho}{\lambda}$ is the ratio between the sparsity and the spikiness threshold.

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 residual covariance, unless the non-zero entries of S are prominent, as Theorem 2.1 prescribes. The MC method performs considerably better than the usual cross-validation using H-fold Frobenius loss (used in Luo (2011b)). In fact, minimizing a loss based on sample approximation like the Frobenius one causes the parameter $\hat{\theta}$ to be shrunk too much. The threshold setting which shows a minimum for MC criterion (given that the estimate $\hat{\varSigma}$ is positive definite) is the best in terms of composite penalty, taking into account the latent low rank and sparse structure simultaneously.

4 A simulation study

4.1 Simulation settings

In order to compare the performance of UNALCE, LOREC and POET, we take into consideration three simulated low rank plus sparse settings reported in Table 1. From now, for easiness of notation, we denote $L^*, S^*, \varSigma^*$ by $L, S, \varSigma$ respectively. The key simulation parameters are:

Table 1: Simulated settings

	p	n	r	τ	θ	c	s	ρ_{corr}
Setting 1	100	1000	4	1	0.7	2	118	0.0045
Setting 2	150	150	5	1	0.8	2	378	0.0033
Setting 3	200	100	6	1	0.8	2	631	0.0036

- the dimension p, the sample size n;
- the rank r and the condition number c of the low rank component L;
- the trace of L, $\tau\theta p$, where τ is a magnitude parameter and θ is the percentage of variance explained by L;
- the (half) number non-zeros s in the sparse component S;
- the percentage of (absolute) residual covariance ρ_{corr} .

Setting 1 has $\frac{p}{n} = 0.1$, while **Setting 2** has $\frac{p}{n} = 1$ and **Setting 3** $\frac{p}{n} = 2$.

Lots of quantities are computed in order to describe comparatively the performance of the three methods on the same data. The computation algorithm (a singular value thresholding plus soft thresholding one) is described in Luo (2011a), and is applied to the generated unbiased sample covariance matrix Σ_n. We call the low rank estimate $\hat{L}$, the sparse estimate $\hat{S}$, and the covariance matrix estimate $\hat{\Sigma} = \hat{S} + \hat{L}$.

The error norms used are:

1. $Loss = ||\hat{S} - S||_{Fro} + ||\hat{L} - L||_{Fro}$,
2. $Total\ Loss = ||\hat{\Sigma} - \Sigma||_{Fro}$,
3. $Sample\ Total\ Loss = ||\hat{\Sigma} - \Sigma_n||_{Fro}$.

The estimated proportion of total variance $\hat{\theta}$ and the residual covariance proportion $\hat{\rho}_{corr}$ are computed. The performance of $\hat{S}$ is assessed by the following measures. Let us denote by nz the number of nonzeros in $\hat{S}$ (recall that s is the number of nonzeros in S), by fp the false non-zeros, by fn the false zeros, by $fpos$ the false positive and by $fneg$ the false negative elements. We define:

1. the estimated proportion of non-zeros $perc_{nz} = nz/numvar$, where $numvar = p(p-1)/2$ is the number of off-diagonal elements,

2. the *error* measure: $err = \frac{fp+fn}{numvar}$

3. $errplus = \frac{fpos+fneg}{s}$, which is the same as err but computed for non-zeros only, distinguishing between positive and negative in the usual way.

4. the overall error rate $errtot$ using the number of false zeros, false positive, and false negative elements: $errtot = \frac{fpos+fneg+fn}{numvar}$.

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

4.2 Simulation results

We start analyzing the performance of $\hat{\varSigma}_{UNALCE}$ in comparison to the one of $\hat{\varSigma}_{LOREC}$ on our reference setting (**Setting 1**). In Figure 1 and 2 we report the differences between the Sample Total Losses and the Total Losses of LOREC and UNALCE for a grid of $20 \times 20 = 400$ threshold pairs. We note that the gain is positive everywhere, with the exception of the threshold pairs which do not return the exact rank (because they do not satisfy the range of Theorem 2.1). This pattern is more remarkable for *Sample Total Loss* than for *Total Loss.* For both losses and each λ, we note that, as explained, the gain across ρ never overcomes its maximum $\sqrt{r}\lambda$ (plotted for each λ).

In Figure 3 we report the plot of the estimated proportion of latent variance θ across thresholds for $\hat{\varSigma}_{UNALCE}$ (in black the true $\theta = 0.7$). In Figure 4 the same plot is reported for $\hat{\varSigma}_{LOREC}$. The shape is exactly the same as for $\hat{\varSigma}_{UNALCE}$, the only difference is that all patterns are negatively shifted. In particular, $\hat{\theta}$ gets closer to θ for $\hat{\varSigma}_{UNALCE}$ respect to $\hat{\varSigma}_{LOREC}$ in correspondence to all threshold combinations.

The most relevant results for **Setting 1** are reported in Tables 2, 3 and 4. Table 2 shows that UNALCE outperforms POET concerning all losses, and shows the superior performance

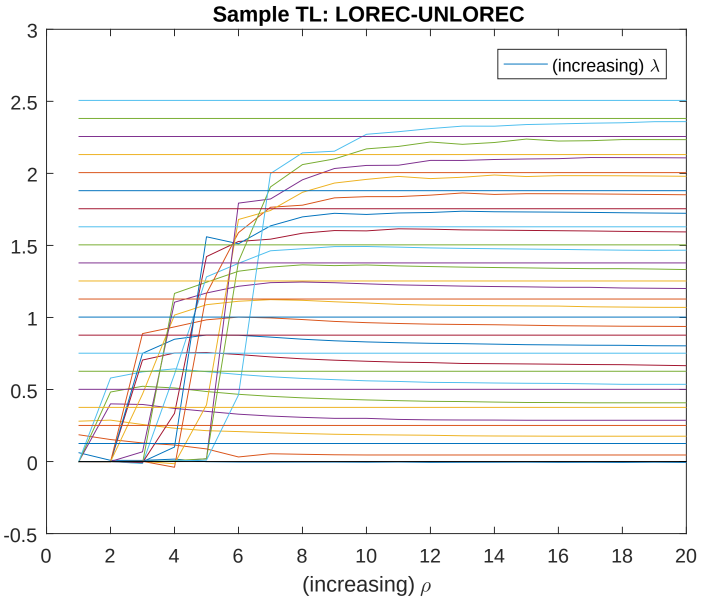


Figure 1: *Sample Total Loss* difference - $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{UNALCE}$ - **Setting 1**

of UNALCE concerning the proportion of latent variance, of residual covariance, of detected non-zeros. Table 3 enlightens the capability to catch the sparsity pattern of UNALCE. In particular, UNALCE shows a lower classification error rate err and total error rate $errtot$ as well as a better correct classification rate of positive and negative elements. Table 4 gives information on the conditioning and eigen-structure properties of UNALCE and POET. We note that for $\frac{p}{n} = 0.1$ POET detects slightly better the vector of eigenvalues of Σ, while it gets much further than UNALCE from the true eigenvalues of L and S. Similarly, POET detects slightly better the maximum eigenvalue of Σ and L, while it completely misses the spectral norm of S. We note that $\hat{\Sigma}_{UNALCE}$ has a slightly larger condition number than POET. The same stands for $\hat{S}$. About $cond(\hat{L})$, instead, we note that POET goes slightly closer than UNALCE to the true one.

Concerning **Setting 2**, simulations show that UNALCE is still better for all losses. POET in this case goes closer to the true number of non-zeros. However, both the proportion of latent variance and of residual covariance are recovered much better by UNALCE. The first fact depends on the natural upper bias of sample eigenvalues (see Ledoit and Wolf (2004) for that). The second depends on the non-algebraic recovery of POET. As for **Setting 1**, both

Table 2: **Setting 1**: all losses

Setting 1	UNALCE	POET	Target
Sample TL	0.7631	2.7323	
Total Loss	6.6899	7.0287	
Loss	7.2170	8.9130	
$\hat{r}$	4	4	4
nz	99	432	118
$perc_{nz}$	0.0200	0.0870	0.0230
$\hat{\theta}$	0.6973	0.7314	0.7000
$\hat{\rho}_{corr}$	0.0025	0.0004	0.0045

Table 3: **Setting 1**: sparsity pattern

Setting 1	UNALCE	POET
senspos	0.5094	0.0640
specpos	0.7231	0
spec	0.995	0.9389
err	0.0135	0.0238
errplus	0.0085	0
errtot	0.0137	0.0238

Table 4: **Setting 1**: eigen-structure properties

Setting 1	UNALCE	POET	Target
$cond(\hat{\Sigma})$	6.42E+04	3.50E+04	9.49E+07
$cond(\hat{S})$	2.75E+04	3.26E+03	2.26E+07
$cond(\hat{L})$	1.2904	1.3083	2
$\|\|eig(\hat{\Sigma}) - eig(\Sigma)\|\|$	5.5113	5.4882	
$\|\|eig(\hat{S}) - eig(S)\|\|$	0.1681	6.3552	
$\|\|eig(\hat{L}) - eig(L)\|\|$	5.4970	8.0727	
$\|\|\hat{\Sigma}\|\|_2$	21.0400	21.2107	24.4886
$\|\|\hat{L}\|\|_2$	20.0038	21.1821	23.3333
$\|\|\hat{S}\|\|_2$	3.6894	0.1340	3.7756

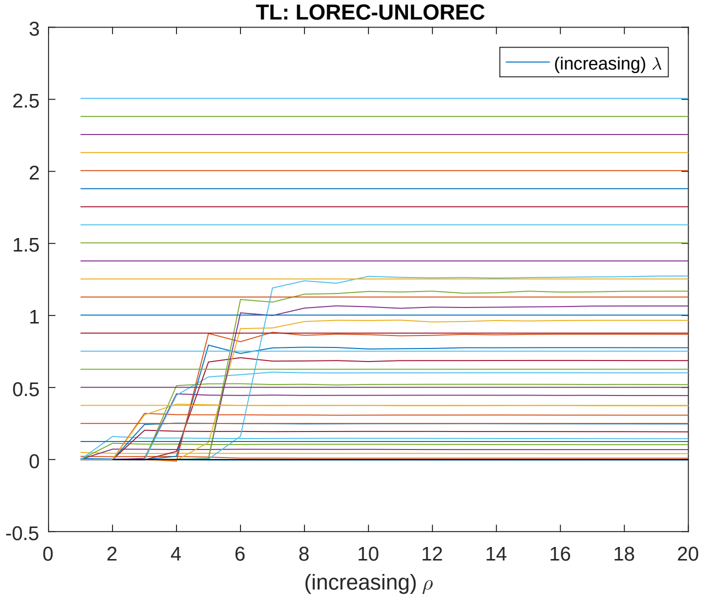


Figure 2: *Total Loss* difference - $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{UNALCE}$ - **Setting 1**

measures *err* and *errtot* are better for UNALCE.

We also note that all hierarchies concerning conditioning properties are the same as in Table 4. We note that the unshrinkage procedure causes $cond(\hat{L})$ to be larger than necessary, and this is more evident as $\frac{p}{n}$ increases. All patterns about the error of estimated eigenvalues are the same, except that here UNALCE goes closer to the true eigenvalues of Σ than POET. On the contrary, the pattern of estimated maximum eigenvalues changes considerably: UNALCE is prevailing. This depends on the natural bias of sample eigenvalues, such that if the true eigenvalues are more spiked (since $\frac{p}{n}$ increases) the spectral norms are overestimated by POET procedure. At the same time, we observe an improvement in the capability of POET to catch the spectral norm of S respect to **Setting 1**, which depends on the increased consistency of **Setting 2** with the assumptions of POET.

Concerning **Setting 3**, we have no changes respect to the patterns previously described about all losses. The same holds for $\hat{\theta}$, $\hat{\rho}_{corr}$, nz, even if the performance of the sparsity pattern detection is in general quite worse for both estimators. This happens because non-zeros have to be very spiked when top eigenvalues are larger. The attititudes to catch positive and negative elements, as well as measures *err* and *errtot*, are still better for UNALCE than for POET,

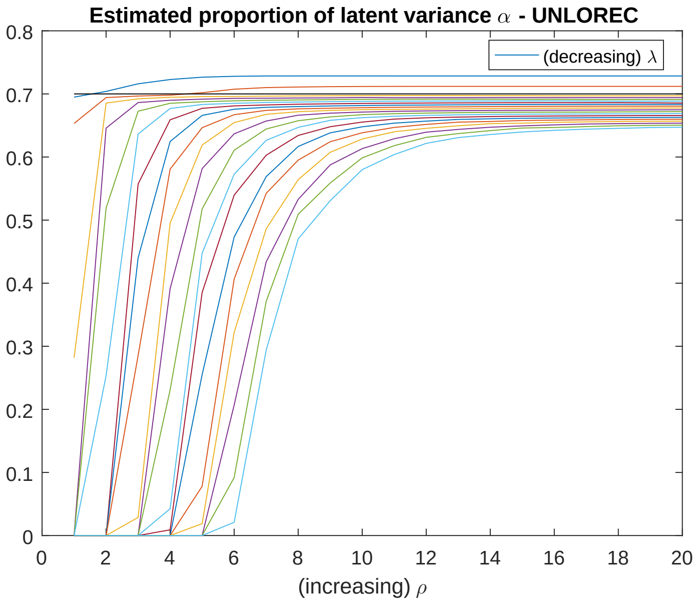


Figure 3: Estimated proportion of latent variance - $\hat{\Sigma}_{UNALCE}$ - **Setting 1**

but they are much worse than for previous settings. In fact, too spiked eigenvalues together with non-prominent non-zero residual entries may be not consistent with the assumptions of Theorem 2.1. Finally, concerning eigen-structures we find the same patterns as before, even amplified, if we exclude that POET is here able to catch the spectral norm of S better than UNALCE. We can conclude that UNALCE still performs better than POET even in the case of $\frac{p}{n} > 1$, but the performance is in general worse, as the absolute statistical bound of Theorem 2.1 prescribes. Tables for **Setting 2** and **Setting 3** are reported in the Appendix.

To sum up, our UNALCE estimator outperforms POET concerning fitting and conditioning properties, detection of sparsity pattern, and eigen-structure recovery. We note that POET does not detect positive and negative elements at all. This is because it only has parametric consistency and not also the algebraic one. In addition, in order to obtain a positive definite estimate, cross validation selects a very high threshold for POET, and this causes the sparse estimate to be almost completely diagonal if $\frac{p}{n}$ is large. On the contrary, the mathematical optimization procedure of UNALCE procedure gets closer to the target and ensures to catch the algebraic spaces behind the two components.

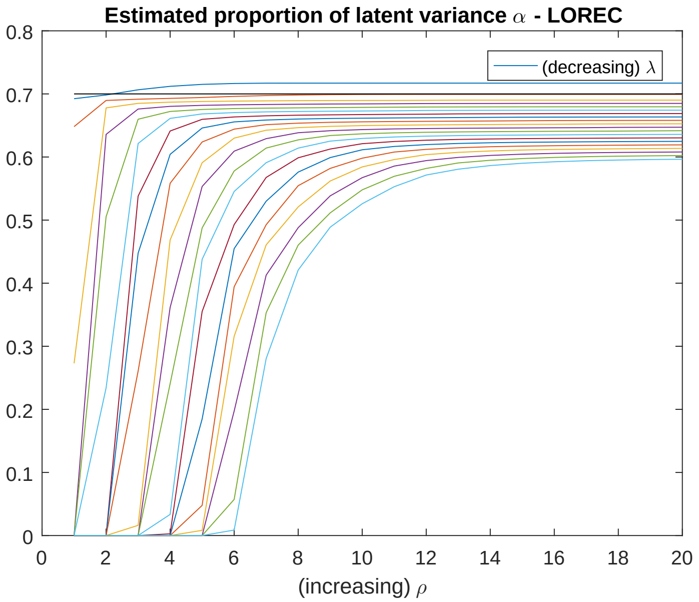


Figure 4: Estimated proportion of latent variance - $\hat{\Sigma}_{LOREC}$ - **Setting 1**

5 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 for some of the most relevant Euro Area banks by systemic power. The overall number of banks (our sample size) is $n = 365$. These indicators are the ones needed for supervisory reporting, and include capital and financial variables.

The chosen raw variables (1039) 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 Figure 5.

UNALCE estimation method selects a solution having a latent rank equal to 6. The

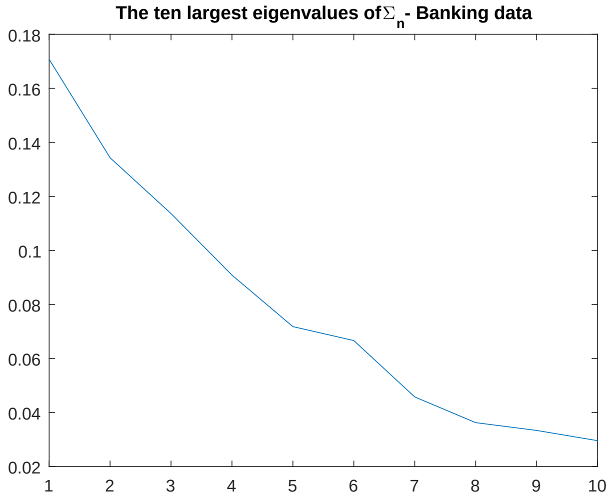


Figure 5: Supervisory data: sample eigenvalues

Supervisory data	UNALCE
$\hat{r}$	6
nz	328
$perc_{nz}$	0.0045
$\hat{\theta}$	0.3247
$\hat{\rho}_{corr}$	0.1687
Sample TL	0.0337
$cond(\hat{\Sigma})$	6.35E+15
$cond(\hat{S})$	2.78E+15
$cond(\hat{L})$	3.1335

Table 5: Supervisory data: results for $\hat{\Sigma}_{UNALCE}$

number of surviving non-zeros in the sparse component is 328, which is the 0.45% of 72772 elements. Conditioning properties are inevitably very bad. The results are reported in Table 5.

In order to to obtain a POET estimate, we exploit the algebraic consistency of $\hat{\Sigma}_{UNALCE}$ setting the rank to 6 and we perform cross-validation for threshold selection. The results are reported in Table 6, where we note that the number of estimated non-zeros is 404 (0.56%).

Apparently, one could argue that POET estimate is better: the estimated proportion of common variance is 0.6123, and the proportion of residual covariance is 0.0161. On the contrary, UNALCE method outputs $\hat{\theta} = 0.3247$ and $\hat{\rho}_{corr} = 0.1687$. A relevant question

Supervisory data	POET
$\hat{r}$	6
nz	404
$perc_{nz}$	0.0056
$\hat{\theta}$	0.6123
$\hat{\rho}_{corr}$	0.0161
Sample TL	0.0645
$cond(\hat{\Sigma})$	6.68E+15
$cond(\hat{S})$	1.11E+15
$cond(\hat{L})$	2.5625

Table 6: Supervisory data: results for $\hat{\Sigma}_{POET}$

arises: how much is the true proportion of variance explained by the factors? In fact, a so high latent proportion variance, which depends on the use of PCA with 6 components, causes the residual covariance proportion to be very low. Therefore, POET procedure gives *a priori* a preference for 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, allows for a larger magnitude of the non-zero elements in the sparse component. In fact, the proportion of lost covariance during the procedure is here 29.39%. As a consequence, via rank/sparsity detection UNALCE shows better approximation properties respect to POET: its Sample Total Loss is relevantly lower than the one of the competitor (0.337 VS 0.645).

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

In Figure 6 we plot the estimated degree (number of non-zero covariances in the residual component) sorted by variable. Only 62 out of 382 variables have at least one non-zero residual covariance.

In Figure 7 we report the top 6 variables by estimated number of non-zero residual covariances. They are mainly credit-based variables: financial assets through profit and loss, central

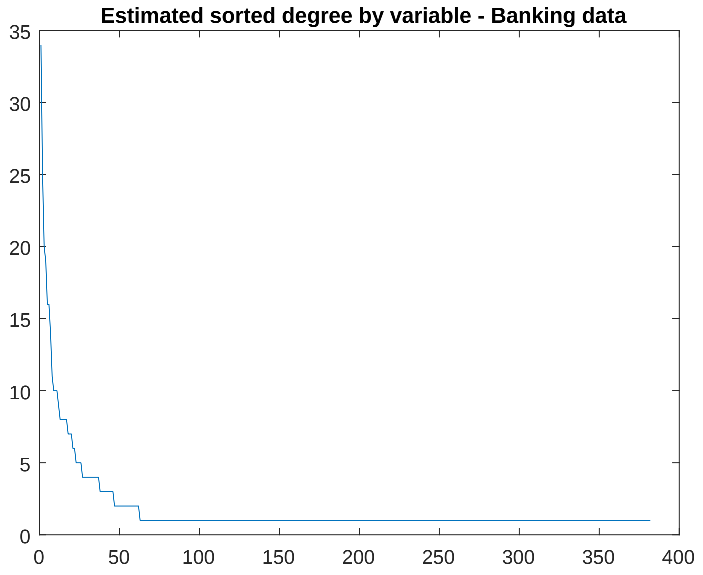
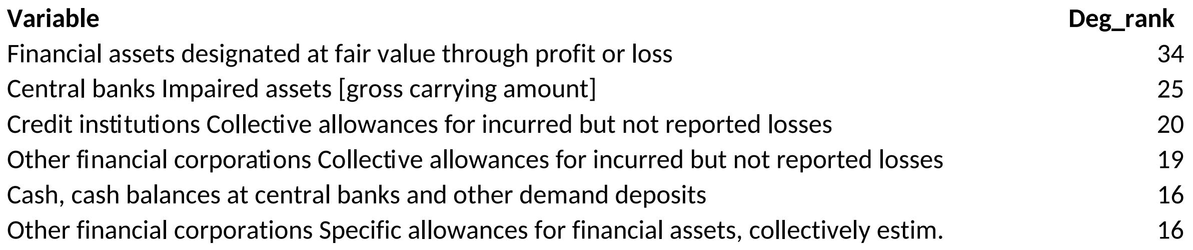


Figure 6: Banking data: sorted degree by variable

banks impaired assets, allowances to credit institutions and non-financial corporations, cash. These variables are related to the largest number of other variables.



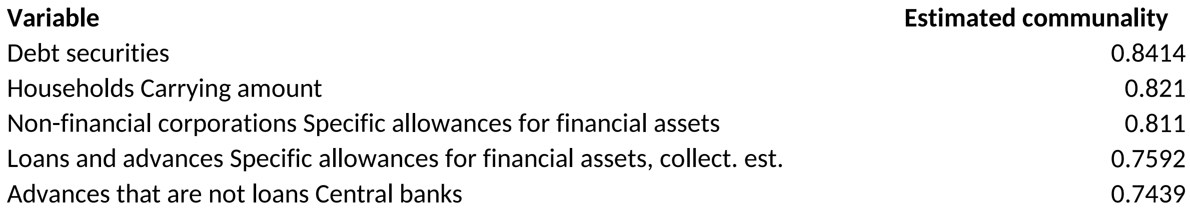
Variable	Deg_rank
Financial assets designated at fair value through profit or loss	34
Central banks Impaired assets [gross carrying amount]	25
Credit institutions Collective allowances for incurred but not reported losses	20
Other financial corporations Collective allowances for incurred but not reported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

Figure 7: Banking data: top 6 variables by degree

In Figure 8 we report the top 5 variables by estimated communality, defined as

$$\frac{\hat{l}_{UNALCE,jj}}{\hat{\sigma}_{UNALCE,jj}} \, \forall j = 1, \ldots, 382.$$

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



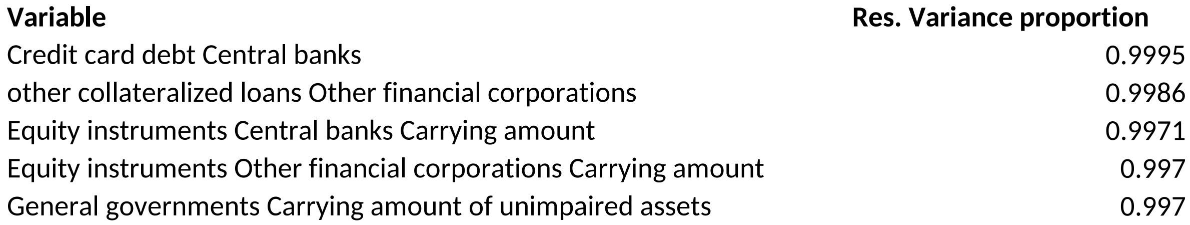
Variable	Estimated communality
Debt securities	0.8414
Households Carrying amount	0.821
Non-financial corporations Specific allowances for financial assets	0.811
Loans and advances Specific allowances for financial assets, collect. est.	0.7592
Advances that are not loans Central banks	0.7439

Figure 8: Banking data: top 5 variables by estimated communality

In Figure 9 we report the top 5 variables by estimated idiosyncratic covariance proportion

$$\frac{\hat{s}_{UNALCE,jj}}{\hat{\sigma}_{UNALCE,jj}} \forall j = 1, \dots, 382.$$

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



Variable	Res. Variance proportion
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	0.997
General governments Carrying amount of unimpaired assets	0.997

Figure 9: Banking data: top 5 variables by residual covariance proportion

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 respect to POET.

6 Conclusions

The present work describes a numerical estimator of large covariance matrices which may assumed to be the sum of a low rank and a sparse component. Estimation is performed solving a regularization problem where the objective function is composed by 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) and 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.

In this paper we compare UNALCE and POET (Principal Orthogonal complEment Thresholding, Fan et al. (2013)), an asymptotic estimator which performs PCA 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 provides also the algebraic consistency of the estimate, that is, the rank and the position of residual non zeros are simultaneously detected by the solution algorithm. This automatic recovery is a crucial advantage respect to POET: the latent rank, in fact, is automatically selected and the sparsity pattern of the residual component is recovered considerably better.

UNALCE improves over the most recent numerical estimator, LOREC (LOw Rank and sparsE Covariance estimator), for two reasons. First, it is proved that the unshrinkage of the eigenvalues of the low rank component estimated by LOREC corrects for the systematic underestimation, due to the thresholding procedure, of the variance proportion explained by the factors, which leaded to a systematic sub-optimality of the overall covariance estimate. Second, we show that it is possible to overcome the restrictive condition $p \leq n$ (p is the dimension, n is the sample length) exploiting the asymptotic theory of POET in the finite-sample context.

In particular, we prove that UNALCE can effectively recover the covariance matrix even in presence of spiked eigenvalues with rate $O(p)$, exactly as POET estimator does, allowing until $p = o(n^2)$ variables under POET assumptions. The loss from the target is bounded in absolute norm (in contrast to POET procedure). In addition, we prove that the recovery is actually effective even if we have an intermediate degree $\alpha \in [0, 1]$ of spikiness, and the loss is bounded accordingly to α with the allowance of $p = o(n^{2\alpha})$ variables. In this way we encompass both LOREC and POET theory in a generalized theory of large covariance matrix estimation by low rank plus sparse decomposition.

The performance of UNALCE is assessed comparatively to LOREC and POET in a wide empirical study which exploits a new original simulation setting particularly flexible and useful for low rank plus sparse modelling. In that context, we provide a new model selection criterion specifically thought for our minimization problem. The criterion is observed to detect the best balance between the low rank latent structure and the (residual) sparsity pattern.

Simulation results show that our method is particularly effective for recovering the proportion of latent variance, as well as the proportion of residual covariance and the number of non-zeros, both respect to LOREC (because of the unshrinkage and of the new model selection procedure) and respect to POET. Moreover, UNALCE shows better fitting properties respect to LOREC and POET under various (absolute) losses, like the composite loss of the low rank and the sparse component and the total loss. The gap is still present if we increase the ratio between dimensionality and sample size.

A real example on 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 having the largest systemic power, that is, the ones most affecting the common covariance structure, can be identified, as well as the variables having the largest idiosyncratic power, that is, the ones most characterized by the residual variance. In addition, the variables showing the most of idiosyncratic covariances with all the other ones can be identified, thus recovering the strongest related variables. Particular forms of the residual covariance pattern can thus be detected if present.

Our research may be ground 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 new computational and theoretical challenges, the solution of which is crucial to further extend the power of statistical modelling and its effectiveness in detecting patterns and underlying drivers of real phenomena.

Acknowledgments

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

A Proofs

A.1 Proof of Theorem 2.1

First of all, we report some technical results of POET approach necessary for our proof strategy. We recall that $L^* = U_L D U_L'$, where $B = U_L D^{1/2}$, and $\lambda_1(\Sigma^*), \dots, \lambda_p(\Sigma^*)$ are the eigenvalues of Σ^* in decreasing order.

Proposition A.1 (Fan et al. (2013) Proposition 1) *All the eigenvalues of the $r \times r$ matrix $B'B$ are bounded away from 0 for all large p. Under the assumptions $cov(f_t) = I_r$ and $B'B$ diagonal we have:*

$$|\lambda_j - ||\tilde{b}_j||^2| \leq ||S^*||, \qquad j \leq r$$
$$|\lambda_j| \leq ||S^*||, \qquad j > r.$$

In addition, for $j \leq r$, $\liminf_{p\to\infty} ||\tilde{b}_j||^2/p > 0$.

Proposition A.2 (Fan et al. (2013) Proposition 2) *Under the assumptions of Proposition 1, if $||\tilde{b}_j||_{j=1}^r$ are distinct, then $||u_j - \frac{\tilde{b}_j}{||\tilde{b}_j||}|| = O(p^{-1}||S^*||)$.*

Proposition A.1 and A.2 together state that U_Σ and $\tilde{B}$ are approximately the same if $||S^*|| = o(p)$. Defining the quantity $m_p = \max_{i\leq p}\sum_{j\leq p}|\sigma_{ij}|^q$, we know that the PCA of $\hat{\Sigma}_n$ asymptotically identifies the eigenvalues and the eigenvectors of Σ^* if $m_p = o(p)$, because $||S^*|| \leq O(m_p)$.

Three technical claims, proved in Fan et al. (2011) under POET assumptions, are:

Lemma A.1
$$\max_{i,j\leq r}\left|\frac{1}{T}\sum_{t=1}^{T} f_{it}f_{jt} - E(f_{it}f_{jt})\right| = O_p\left(\frac{1}{\sqrt{n}}\right)$$

Lemma A.2

$$\max_{i,j\leq r}\left|\frac{1}{T}\sum_{t=1}^{T}s_{it}s_{jt}-E(s_{it}s_{jt})\right|=O_p\left(\frac{\log(p)}{\sqrt{n}}\right)$$

Lemma A.3

$$\max_{i,j\leq r}\left|\frac{1}{T}\sum_{t=1}^{T}f_{it}s_{jt}\right|=O_p\left(\frac{\log(p)}{\sqrt{n}}\right).$$

Lemmas A.1, A.2, A.3 allow to prove a crucial Lemma for our purpose (see Fan et al. (2013)):

Lemma A.4 *Let $\hat{\lambda}_r$ be the $r-th$ largest eigenvalue of $\hat{\Sigma}_n$. Then $\hat{\lambda}_r > C_1p$ with probability approaching* 1 *for some $C_1>0$.*

In other words, Lemma A.4 states that $||\hat{\Sigma}_n\ -\ \Sigma^*||=o(p)$ with a rate proportional to $O(\frac{p}{\sqrt{n}})$, i.e. the $r-$th largest eigenvalue of $\hat{\Sigma}_n$ grows at rate $O(p)$ with probability approaching 1:

$$||\hat{\Sigma}_n-\Sigma^*||=O\left(\frac{p}{\sqrt{n}}\right).$$

The proof of Theorem 2.1 relies on Lemma A.4. Since Lemma A.4 (as it is) is the key to prove that under Fan's condition the bound (20) holds, the updated version of Lemma A.4 in the α - spiked context must be the key to prove that

$$g_\gamma(\hat{S}_n-S^*,\hat{L}_n-L^*)\preceq\frac{1}{\xi(T)}\frac{p^\alpha}{\sqrt{n}}. \tag{41}$$

This proof requires to adapt claims A.1, A.2, A.3 (coming from Fan et al. (2011)) to the generalized pervasiveness context, where the pervasiveness of latent factors has been relaxed, applying the proof technique in Fan et al. (2013), Appendix C, page 639. The bound (14) of Bickel and Levina (2008a) is also a necessary tool.

The first step of the proof consists in decomposing $E_n = \Sigma_n - \Sigma$ in its four components:

$$E_n = \Sigma_n - \Sigma = D_1 + D_2 + D_3 + D_4$$

where:

$$D_1 = (n^{-1} B \sum_{i=1}^{n} \mathbf{f}_i \mathbf{f}_i' - I_r) B'$$

$$D_2 = n^{-1} (\sum_{i=1}^{n} \epsilon_i \epsilon_i' - S)$$

$$D_3 = B n^{-1} \sum_{i=1}^{n} \mathbf{f}_i \epsilon_i'$$

$$D_4 = D_3',$$

where $\mathbf{f}_i$ and ϵ_i are respectively the vectors of factor scores and residuals for each observation. Asterisks are omitted to avoid cluttered notation.

From the inequality $||B'\Sigma^{-1}B|| \leq |cov(f)^{-1}|$ (page 194 Fan et al. (2008), Assumption (B)), Lemma A.1 follows. Lemma A.1 is unaffected the modification of Proposition A.1. Consequently, following the proof of Lemma A.4, we can argue that under the relaxed pervasiveness condition $||D_1|| \leq O(p^{\alpha}\sqrt{\frac{1}{n}})$, because now $||BB'|| = O(p^{\alpha})$. This happens also because $r \log p = o(n)$.

In order to show how Lemma A.2 changes, we need to recall some key results of Bickel and Levina (2008a). Differently from Luo's approach, in that setting (as in the POET one) the sparsity assumption is imposed to S^*, and not to Σ^*. The relevant quantity m_p in POET setting is $o(p)$, in order to have $||S|| = o(p)$, which allows to identify the low rank component via PCA. Here, since Definition 2.1 holds, we have that $m_p = o(p)$ is no longer appropriate. We impose, in order to preserve the correspondence between the bounds of the sample and theoretical eigenvalues, the assumption $m_p = o(p^{\alpha})$ (which causes $||S|| = o(p^{\alpha})$ in the POET setting).

Consider now the uniformity class of sparse matrices:

$$\left\{ S^* : s^*_{ii} \leq M, \sum_{j=1}^{p} |s^*_{ij}|^q \leq c_0(p), \; \forall i \right\}. \tag{42}$$

We have residual variances uniformly bounded by M. This assumption here is no longer valid, because M cannot longer assumed to be negligible respect to p.

Here we can no longer write (see Bickel and Levina (2008a), page 2580)

$$\lambda_{max}(S^*) \leq \max_i \sum_j |s^*_{ij}| \leq M^{1-q} c_0(p),$$

as Fan et al. do in their pure spikiness context.

The quantity $c_0(p)$ can still be assumed not to scale with p, because we want to have a sparse S^*, but the assumption $m_p = o_p(p^{\alpha})$ causes that M cannot longer be considered as a constant as $p \to \infty$. In order to normalize it, we need to divide by $p^{1-\alpha}$, thus obtaining that m_p grows at a rate of $O(p^{\alpha-1})$ as p increases. Imposing $M = O(p^{\alpha-1})$ in the proof which derives the rate of the sample covariance matrix under class (42) (see Bickel and Levina (2008a), page 2582), we can prove:

Lemma A.5

$$||\hat{S}_n - S||_{\infty} \leq O\left(p^{\alpha-1} \sqrt{\frac{\log p}{n}} \right).$$

Using Lemma A.5, we can apply the proof strategy of Lemma A.4 (Appendix C) to matrix D_2, obtaining

$$||D_2|| \leq p O_p(p^{\alpha-1}) O\left(\sqrt{\frac{\log(p)}{n}} \right) = O_p\left(p^{\alpha} \sqrt{\frac{\log p}{n}} \right),$$

because $||D_2| \leq p||D||_{\infty}$. Since $log(p) = o(n)$, we can write

$$||D_2|| \leq p O_p(p^{\alpha-1}) O\left(\sqrt{\frac{\log p}{n}} \right) = O_p\left(p^{\alpha} \frac{1}{\sqrt{n}} \right), \tag{43}$$

as we needed.

To conclude, we analyze Lemma A.3:

$$\max_{i\leq r, j\leq p}\left|\frac{1}{n}\sum_{k=1}^{n} f_{ik}s_{jk}\right| \leq \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\max_{i}|f_{ik}|\frac{1}{\sqrt{n}}\max_{j}\sum_{k=1}^{n}|s_{jk}| \leq \sqrt{\frac{r}{n}}pp^{\alpha-1}\sqrt{\frac{\log p}{n}},$$

Exploiting the assumption $r = O(\log(p^{\alpha}))$ (since $\log(p^{\alpha}) = o(n)$), and $n = O(p^{\alpha})$ we obtain $O(\sqrt{\frac{r}{n}}) = O(p^{-\frac{\alpha}{2}})$. This replacement is valid if and only if $n = o(p^{2\alpha})$. Therefore, the bound above becomes $O\left(p^{\frac{\alpha}{2}}\sqrt{\frac{\log p}{n}}\right)$.

Applying the proof strategy of Lemma A.4 to D_3 we obtain

$$||D_3|| \leq O\left(p^{\frac{\alpha}{2}}\sqrt{\frac{\log p}{n}}\right) O\left(p^{\frac{\alpha}{2}}\right) = O\left(p^{\alpha}\sqrt{\frac{\log p}{n}}\right),$$

because $||B|| = O(p^{\frac{\alpha}{2}})$. The condition $log(p) = o(n)$ finally leads to:

$$||D_3|| \leq O\left(\frac{p^{\alpha}}{\sqrt{n}}\right). \tag{44}$$

The bound (21) is consequently proved. Therefore we can affirm

Lemma A.6 *Let $\hat{\lambda}_r$ be the $r-th$ largest eigenvalue of $\hat{\Sigma}_n$. Then $\hat{\lambda}_r > C_1 p^{\alpha}$ with probability approaching 1 for some $C_1 > 0$.*

Lemma A.6 states that $||\Sigma_n - \Sigma^*||_2$ is $o(p)$ with rate $O(\frac{p^{\alpha}}{\sqrt{n}})$, or, in other words,

$$P\left(||E_n|| \geq C_1\frac{p^{\alpha}}{\sqrt{n}}\right) \leq 1 - C_2 e^{-C_3 p^{2\alpha}}. \tag{45}$$

The proof relies on the combined use of proof tools by Fan et al. (2013), Fan et al. (2011), Fan et al. (2008) and Bickel and Levina (2008a).

Since we have dropped assumption (13) for Σ^*, we can simply write, using the basic property $||.||_{\infty} \leq ||.||_2$ and the minimum for γ in the range of Theorem 2.1,

$$P\left(||E_n||_{\infty} \geq C_1\xi(T)\frac{p^{\alpha}}{\sqrt{n}}\right) \leq 1 - C_2 e^{-C_3 p^{2\alpha}}. \tag{46}$$

According to Chandrasekaran et al. (2012) and Luo (2011a), the only probabilistic component of the error norm $g_\gamma(\hat{S} - S^*, \hat{L} - L^*)$ is $g_\gamma(E_n)$. Therefore, following the proof of Luo (2011a) and setting $\lambda = \left(\frac{1}{\xi(T)} \frac{p^\alpha}{\sqrt{n}}\right)$, we can finally prove, under all the assumptions and conditions of Theorem 2.1, the thesis

$$g_\gamma(\hat{S} - S^*, \hat{L} - L^*) \preceq \frac{1}{\xi(T)} \frac{p^\alpha}{\sqrt{n}}. \tag{47}$$

A.2 Proof of Theorem 2.2

Given finite p and n we have

$$TL(L, S) = ||L^* + S^* + W - L - S||^{\mathit{2}}_{\mathit{Fro}} \leq$$

$$\leq ||L - L^*||^{\mathit{2}}_{\mathit{Fro}} + ||S - S^*||^{\mathit{2}}_{\mathit{Fro}} + ||W||^{\mathit{2}}_{\mathit{Fro}} = A + B + C.$$

The ALCE solution is $\hat{\Sigma}_{ALCE} = \hat{L} + \hat{S}$, $L \in \hat{\mathcal{B}}(\hat{r}), S \in \hat{\mathcal{A}}(\hat{s})$, with

$$\hat{L} = \hat{U}\hat{D}\hat{U}', \tag{48}$$

where $\hat{D} = D_\lambda$ is the diagonal eigenvalue matrix coming out from the singular value thresholding procedure, and $\hat{U}$ is the matrix of corresponding eigenvectors.

We start from a standard result: the spectral decomposition of a real square matrix M truncated to the r-th component is the r-ranked matrix best approximating M. In fact, the optimization problems $\min_{K, rank(K)=r} ||M - K|_{\mathit{2}}$ and $\min_{K, rank(K)=r} ||M - K||_{\mathit{Fro}}$ are both solved for $K = \sum_{i=\mathit{1}}^{r} \lambda_i \mathbf{u_i} \mathbf{u_i'}$, which is the spectral decomposition of M truncated to the r-th summand when r is known. This result was proved in Eckart and Young (1936), who called the identified directions "canonical components".

Aware of the best approximation property of canonical components, our question is the following: which is the matrix in the variety $\hat{\mathcal{B}}(\hat{r})$ being closer to the unknown r-ranked matrix L^*, keeping fixed $\hat{U}$?

The solution is straightforward: the matrix we are looking for has the same eigenvectors $\hat{U}$, but has the original (natural) eigenvalues. This new matrix $\hat{D}_{UNALCE}$ can be obtained simply un-shrinking the obtained eigenvalues: $\hat{D}_{UNALCE} = D_\lambda + \lambda I_r$. This is why term A is minimized as follows: $\min_{L \in \hat{\mathcal{B}}(\hat{r})} ||L - L^*||^2_{Fro} \iff \hat{L}_{UNALCE} = \hat{U}(D_\lambda + \lambda I_r)\hat{U}'$.

Suppose that $\hat{\Sigma}_{ALCE}$ is given, and assume that the **off-diagonal** elements of $\hat{S}$ are **invariant**. We can re-write term B as follows:

$$\begin{aligned}
&\min_{S \in \hat{\mathcal{A}}(\hat{s})} ||S - S^*||^2_{Fro} = \\
&= \min_{L \in \hat{B}(\hat{r})} ||(\hat{\Sigma} - L) - (\Sigma^* - L^*)||^2_{Fro} = \\
&= \min_{L \in \hat{B}(\hat{r})} ||(\hat{\Sigma} - \Sigma^*) - (L - L^*)||^2_{Fro} \leq \\
&\sum_{i=1}^{p} (\hat{\sigma}_{ii} - \sigma_{ii})^2 + \sum_{i=1}^{p} (\hat{l}_{ii} - l_{ii})^2 = \\
&= B' + B''.
\end{aligned}$$

Term B' is assumed to be fixed respect to L, i.e. we are assuming the invariance of diagonal elements in $\hat{\Sigma}_{ALCE}$ ($diag(\hat{\Sigma}_{UNALCE}) = diag(\hat{\Sigma}_{ALCE})$). The minimization of term B'', given that $rank(L) = \hat{r}$, falls back into the previous case, i.e. B'' is minimum $\iff \hat{L}_{UNALCE} = \hat{U}(D_\lambda + \lambda I_r)\hat{U}'$.

Term C depends on the quality of the estimation input Σ_n, and on the degree of correspondence with ALCE assumptions.

Consequently, we can write:

$$\hat{S}_{UNALCE,ii} = \hat{\Sigma}_{ii} - \hat{L}_{UNALCE,ii},\ \forall i.$$

$$\hat{S}_{UNALCE,ij} = \hat{S}_{ij},\ \forall i \neq j.$$

A.3 Proof of Corollary 2.1

We know that $||\hat{L}_{UNALCE} - \hat{L}_{ALCE}||_2 = \lambda$. Recalling that

$$\hat{L}_{UNALCE} = \min_{L\ \in\ \hat{B}(\hat{r})} ||L - L^*||^2_{Fro},$$

and the triangular inequality

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

we can write

$$0 < ||\hat{L}_{ALCE} - L^*||_2 - ||\hat{L}_{UNALCE} - L^*||_2 \leq \lambda. \tag{49}$$

As a consequence, $||\hat{L}_{UNALCE} - \hat{L}_{ALCE}||_{Fro} = \sqrt{2r}\lambda$ and

$$0 < ||\hat{L}_{ALCE} - L^*||_{Fro} - ||\hat{L}_{UNALCE} - L^*||_{Fro} \leq \sqrt{2r}\lambda. \tag{50}$$

The analogous triangular inequality for the sparse component is

$$||\hat{S}_{ALCE} - S^*||_2 \leq ||\hat{S}_{UNALCE} - \hat{S}_{ALCE}||_2 + ||\hat{S}_{UNALCE} - S^*||_2.$$

In order to quantify $||\hat{S}_{UNALCE} - \hat{S}_{ALCE}||_{Fro}$, we need to study the behaviour of the term $\sum_{i=1}^{p}(\hat{l}_{UNALCE,i} - l_{ii})^2$. This can be re-written as

$$\sum_{i=1}^{p}(\hat{l}_{UNALCE,ii} - \hat{l}_{ALCE,ii} + \hat{l}_{ALCE,ii} - l_{ii})^2 \leq$$

$$\leq \sum_{i=1}^{p}(\hat{l}_{UNALCE,ii} - \hat{l}_{ALCE,ii})^2 + \sum_{i=1}^{p}(\hat{l}_{ALCE,ii} - l_{ii})^2.$$

The sum $\sum_{i=1}^{p}(\hat{l}_{ALCE,ii} - l_{ii})^2$ depends on the statistical properties of $\hat{L}_{ALCE}$.
The term $\sum_{i=1}^{p}(\hat{l}_{UNALCE,ii} - \hat{l}_{ALCE,ii})^2$ is less or equal to $r\lambda^2$, because it is less or equal to $tr(\hat{L}_{UNALCE} - \hat{L}_{ALCE})^2 = r\lambda^2$.

As a consequence, we have $||\hat{S}_{UNALCE} - \hat{S}_{ALCE}||_{Fro} \leq \sqrt{r}\lambda$. Recalling that $\hat{S}_{UNALCE} = \min_{S \in \hat{\mathcal{A}}(\hat{s})} ||S - S^*||_{Fro}^2$, we can write:

$$0 < ||\hat{S}_{ALCE} - S^*||_{Fro} - ||\hat{S}_{UNALCE} - S^*||_{Fro} \leq \sqrt{r}\lambda. \tag{51}$$

The last claim is less immediate. We recall that $||\hat{L}_{UNALCE} - \hat{L}_{ALCE}||_2 = ||\hat{U}\lambda I_r \hat{U}'||_2 = \lambda$.

$\hat{U}\lambda I_r\hat{U}'$ can be divided in the contribution coming from diagonal elements and the rest: $||diag(\hat{L}_{UNALCE}-\hat{L}_{ALCE})+off-diag(\hat{L}_{UNALCE}-\hat{L}_{ALCE})||_2$. Both contributes are part of $\hat{U}\lambda I_r\hat{U}'$.

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

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

$$||diag(\hat{L}_{UNALCE}-\hat{L}_{ALCE})||=\sum_{i=1}^{p}||\hat{\mathbf{u}}_i||||K_{ii}||=\sum_{i=\mathit{1}}^{p}||\hat{\mathbf{u}}_i||\lambda \tag{52}$$

$$||off-diag(\hat{L}_{UNALCE}-\hat{L}_{ALCE})||=\sum_{i=1}^{p}\sum_{j\neq i}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_j'||K_{ij}||=\sum_{i=\mathit{1}}^{p}\sum_{j\neq i}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_j'\lambda \tag{53}$$

$$||\hat{L}_{UNALCE}-\hat{L}_{ALCE}||_2=\sum_{i=1}^{p}||\hat{\mathbf{u}}_i||||K_{ii}||+\sum_{i=\mathit{1}}^{p}\sum_{j\neq i}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_j'||K_{ij}||=\lambda. \tag{54}$$

From this consideration it follows that

$$||diag(\hat{L}_{UNALCE}-\hat{L}_{ALCE})||\leq||\hat{L}_{UNALCE}-\hat{L}_{ALCE}||_2||=\lambda.$$

Since, by definition, $||diag(\hat{S}_{UNALCE}-\hat{S}_{ALCE})||=||diag(\hat{L}_{UNALCE}-\hat{L}_{ALCE})||$ (because $diag(\hat{S}_{UNALCE}-\hat{S}_{ALCE})=-diag(\hat{L}_{UNALCE}-\hat{L}_{ALCE})$), and recalling that $\hat{S}_{UNALCE}$ has the best approximation property (for Theorem (2.2)), we can conclude

$$0<||\hat{S}_{ALCE}-S^*||_{\mathit{2}}-||\hat{S}_{UNALCE}-S^*||_{\mathit{2}}\leq\lambda. \tag{55}$$

A.4 Proof of Corollary 2.2

The relevant triangular inequality for the overall estimate is

$$||\hat{\Sigma}_{ALCE} - \Sigma^*||_2 \leq ||\hat{\Sigma}_{UNALCE} - \hat{\Sigma}_{ALCE}||_2 + ||\hat{\Sigma}_{UNALCE} - \Sigma^*||_2.$$

We know that, by definition, $||\hat{\Sigma}_{UNALCE} - \hat{\Sigma}_{ALCE}||_2 = ||off - diag(\hat{L}_{UNALCE} - \hat{L}_{ALCE})||$. For the same considerations explained before,

$$||off - diag(\hat{L}_{UNALCE} - \hat{L}_{ALCE})|| \leq ||\hat{L}_{UNALCE} - \hat{L}_{ALCE}||_2 = \lambda.$$

As a consequence, recalling that $\hat{\Sigma}_{UNALCE} = \min_{\Sigma=L+S}(TL(L,S))$ under the described assumptions, we can conclude

$$0 < ||\Sigma_n - \hat{\Sigma}_{ALCE}||_2 - ||\Sigma_n - \hat{\Sigma}_{UNALCE}||_2 \leq \lambda. \tag{56}$$

Applying the formula $||A||_{fro} \leq \sqrt{r}||A||$ (if A has rank r) we can then claim

$$0 < ||\Sigma_n - \hat{\Sigma}_{ALCE}||_{Fro} - ||\Sigma_n - \hat{\Sigma}_{UNALCE}||_{Fro} \leq \sqrt{r}\lambda. \tag{57}$$

Therefore, the real gain is terms of approximation of Σ_n respect to ALCE measured in squared Frobenius norm is strictly positive and bounded from $r\lambda^2$.

A.5 Proof of Theorem 2.3

We can easily write

$$||\hat{\Sigma}_{UNALCE} - \Sigma|| =$$

$$= ||\hat{\Sigma}_{UNALCE} - \Sigma_n + \Sigma_n - \Sigma|| \leq ||\hat{\Sigma}_{UNALCE} - \Sigma_n||^2 + ||\Sigma_n - \Sigma||^2. \tag{58}$$

The quality of the estimation input $||\Sigma_n - \Sigma||^2_{Fro}$ does not depend on the estimation method.

Therefore, by (56) and (58), it is straightforward that

$$0 < ||\hat{\Sigma}_{ALCE} - \Sigma||_2 - ||\hat{\Sigma}_{UNALCE} - \Sigma|||_2 \leq \lambda. \tag{59}$$

Analogously, it is easy to prove that

$$0 < ||\hat{\Sigma}_{ALCE} - \Sigma||_{Fro} - ||\hat{\Sigma}_{UNALCE} - \Sigma|||_{Fro} \leq \sqrt{r}\lambda. \tag{60}$$

A.6 Proof of Corollary 2.3

We recall the following expression:

$$||(\hat{L} + \hat{S})^{-1} - (\Sigma)^{-1}||_{Fro} = ||(\hat{L} + \hat{S})^{-1}[\hat{Ly} + \hat{S} - \Sigma](\Sigma)^{-1}|| \leq$$

$$\leq ||(\hat{L} + \hat{S})^{-1}|| \cdot ||[\hat{L} + \hat{S} - \Sigma]||_{Fro} \cdot ||(\Sigma)^{-1}||.$$

From (60) we can conclude that

$$0 < ||(\hat{L}_{ALCE} + \hat{S}_{ALCE})^{-1} - \Sigma^{-1}||^2_{Fro} - ||(\hat{L}_{UNALCE} + \hat{S}_{UNALCE})^{-1} - \Sigma^{-1}||^2_{Fro} \leq r\lambda^2. \tag{61}$$

Analogously, since it holds

$$||(\hat{L} + \hat{S})^{-1} - (\Sigma)^{-1}|| = ||(\hat{L} + \hat{S})^{-1}[\hat{L} + \hat{S} - \Sigma](\Sigma)^{-1}|| \leq$$

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

it is straightforward that

$$0 < ||(\hat{L}_{ALCE} + \hat{S}_{ALCE})^{-1} - \Sigma^{-1}||_2 - ||(\hat{L}_{UNALCE} + \hat{S}_{UNALCE})^{-1} - \Sigma^{-1}||_2 \leq \lambda. \tag{62}$$

B Additional tables

Table 7: **Setting 2**: all losses

Setting 2	UNALCE	POET	Target
ρ	0.0270		
λ	0.2300		
Sample TL	1.7464	2.8632	
Total Loss	11.7399	12.4781	
Loss	13.1401	14.2190	
$\hat{r}$	5	5	5
nz	210	453	378
$perc_{nz}$	0.0188	0.0405	0.0338
$\hat{\theta}$	0.8089	0.8363	0.8
$\hat{\rho}_{corr}$	0.0024	5.8249e-005	0.0033

Table 8: **Setting 3**: all losses

Setting 3	UNALCE	POET	Target
ρ	0.0700		
λ	0.5450		
Sample TL	3.5821	4.1688	
Total Loss	24.0471	24.6359	
Loss	25.6832	26.5689	
$\hat{r}$	6	6	6
nz	74	1	631
$perc_{nz}$	0.0037	5.0251e-005	0.0317
$\hat{\theta}$	0.8073	0.8318	0.8
$\hat{\rho}_{corr}$	6.1954e-004	2.5805e-008	0.0036

Table 9: **Setting 2**: sparsity pattern

Setting 2	UNALCE	POET
senspos	0.3016	0
specpos	0.2092	0
spec	0.9911	0.9580
err	0.0322	0.0744
errplus	0.0291	0
errtot	0.0332	0.0744

Table 10: **Setting 3**: sparsity pattern

Setting 3	UNALCE	POET
senspos	0.0603	0
specpos	0.0854	0
spec	0.9988	0.9999
err	0.0303	0.0318
errplus	0.0085	0
errtot	0.0079	0.0318

Table 11: **Setting 2**: eigen-structure properties

Setting 2	UNALCE	POET	Σ
$cond(\hat{\Sigma})$	2.1820e+004	2.6178e+004	6.4838e+013
$cond(\hat{S})$	1.8404e+003	1.5612e+003	3.7043e+009
$cond(\hat{L})$	2.7213	2.6545	2
$\|\|eig(\hat{\Sigma}) - eig(\Sigma)\|\|$	7.8055	8.7183	
$\|\|eig(\hat{S}) - eig(S)\|\|$	0.4372	1.2487	
$\|\|eig(\hat{L}) - eig(L)\|\|$	7.9288	8.9843	
$\|\|\hat{\Sigma}\|\|_2$	39.8818	40.6020	32.7258
$\|\|\hat{L}\|\|_2$	39.2926	40.1838	32
$\|\|\hat{S}\|\|_2$	3.1799	3.7756	2.9034

Table 12: **Setting 3**: eigen-structure properties

Setting 3	UNALCE	POET	Σ
$cond(\hat{\Sigma})$	2.1424e+004	1.8320e+004	2.0147e+014
$cond(\hat{S})$	1.6783e+003	1.0376e+003	5.3513e+009
$cond(\hat{L})$	3.4337	3.2890	2
$\|\|eig(\hat{\Sigma}) - eig(\Sigma)\|\|$	11.1913	11.9235	
$\|\|eig(\hat{S}) - eig(S)\|\|$	0.9242	1.2141	
$\|\|eig(\hat{L}) - eig(L)\|\|$	11.3453	12.1670	
$\|\|\hat{\Sigma}\|\|_2$	44.7070	45.5133	36.1000
$\|\|\hat{L}\|\|_2$	44.3976	45.2441	35.5556
$\|\|\hat{S}\|\|_2$	3.3438	2.4894	2.5701

References

Agarwal, A., S. Negahban, and M. J. Wainwright (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 1171–1197.

Anderson, T. (1984). Multivariate statistical analysis. *Wiley and Sons, New York, NY*.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

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

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

Cai, T. and W. Liu (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association 106*(494), 672–684.

Cai, T. T., C.-H. Zhang, and H. H. Zhou (2010, 08). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist. 38*(4), 2118–2144.

Chandrasekaran, V., P. A. Parrilo, and A. S. Willsky (2012, 08). Latent variable graphical model selection via convex optimization. *Ann. Statist. 40*(4), 1935–1967.

Chandrasekaran, V., S. Sanghavi, P. A. Parrilo, and A. S. Willsky (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization 21*(2), 572–596.

Clarke, F. H. (1990). *Optimization and nonsmooth analysis.* SIAM.

Davidson, K. R. and S. J. Szarek (2001). Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces 1*(317-366), 131.

Eckart, C. and G. Young (1936). The approximation of one matrix by another of lower rank. *Psychometrika 1*(3), 211–218.

Fan, J., Y. Fan, and J. Lv (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics 147*(1), 186–197.

Fan, J., Y. Liao, and M. Mincheva (2011). High dimensional covariance matrix estimation in approximate factor models. *Annals of statistics 39*(6), 3320–3356.

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(4), 603–680.

Fazel, M., H. Hindi, and S. P. Boyd (2001). A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, Volume 6, pp. 4734–4739. IEEE.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Furrer, R. and T. Bengtsson (2007). Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis 98*(2), 227–255.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis 88*(2), 365–411.

Luo, X. (2011a). High dimensional low rank and sparse covariance matrix estimation via convex minimization. *Arxiv preprint*.

Luo, X. (2011b). Recovering model structures from large low rank and sparse covariance matrix estimation. *arXiv preprint arXiv:1111.1133*.

Qiu, Y. and S. X. Chen (2015). Bandwidth selection for high-dimensional covariance matrix estimation. *Journal of the American Statistical Association 110*(511), 1160–1174.

Rockafellar, R. T. (2015). *Convex analysis.* Princeton university press.

Rothman, A. J., E. Levina, and J. Zhu (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association 104*(485), 177–186.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.