Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Scienze Statistiche

Ciclo **XXVIII**

Settore Concorsuale di afferenza: 13/D1

Settore Scientifico disciplinare: SECS/S-01

TITOLO TESI

Large covariance matrix estimation by composite minimization

Presentata da: **Matteo Farnè**

Coordinatore Dottorato **Relatore**

Prof.ssa Alessandra Luati **Prof.ssa Angela Montanari**

Esame finale anno 2016

Large covariance matrix estimation by composite minimization

Matteo Farnè

Dipartimento di Scienze Statistiche
Università di Bologna
email: *matteo.farne2@unibo.it*

"*Illi autem octo cursus, in quibus eadem vis est duorum, septem efficiunt distinctos intervallis sonos, qui numerus rerum omnium fere nodus est; quod docti homines nervis imitati atque cantibus aperuerunt sibi reditum in hunc locum, sicut alii, qui praestantibus ingeniis in vita humana divina studia coluerunt.*"

Marcus Tullius Cicero
Somnium Scipionis, 6.18

Acknowledgments

At the end of a particularly exciting and laborious work, I feel the need to address my thanks to the people who accompanied and helped me to complete such a huge effort.

First of all, I desire to express my gratitude to Professor Angela Montanari, my supervisor and mentor, for the continuous support and encouragement, for the valuable advices and suggestions, for her passion and for the unique chances of growth and maturation she offered me across all my study path.

I want to thank Professor Trevor Hastie, for the brief but decisive research period spent at Stanford, which gave me the necessary insights to develop my own theory and to effectively provide original solutions to my research problem.

I have to thank Dr Yuan Liao for helpful discussions and precious references, Professor Roberto Rocci for his punctual annotations, Professor Enrico Bernardi for his useful ideas on the mathematical part and Professor Nicola Guglielmi for his strong directions on the algorithmic part.

A particularly warm acknowledgment goes to the Supervisory Statistics Division of the European Central Bank, where I spent a semester as PhD trainee: they welcomed me and allowed me to complete my PhD studies developing crucial skills for real data analysis in large dimensions. A specific thanksgiving goes to Angelos Vouldis, wonderful research supervisor, to Micheal Fedesin, for his trust and enthusiasm, and to Patrick Hogan for his valuable managing.

Finally, I want to thank my family, for the long and patient guide, aid and support, my friends, for they often alleviated this labour, and Silvia, for her continuous, passionate and enthusiastic listening and incitement.

Abstract

The present thesis concerns large covariance matrix estimation via composite minimization under the assumption of low rank plus sparse structure. Existing methods like POET (Principal Orthogonal complEment Thresholding) perform estimation by extracting principal components and then applying a soft thresholding algorithm. In contrast, our method recovers the low rank plus sparse decomposition of the covariance matrix by least squares minimization under nuclear norm plus l_1 norm penalization. This non-smooth convex minimization procedure is based on semidefinite programming and subdifferential methods, resulting in two separable problems solved by a singular value thresholding plus soft thresholding algorithm.

The most recent estimator in literature is called LOREC (Low Rank and sparsE Covariance estimator) and provides non-asymptotic error rates as well as identifiability conditions in the context of algebraic geometry. Our work shows that the unshrinkage of the estimated eigenvalues of the low rank component improves the performance of LOREC considerably. The same method also recovers covariance structures with very spiked latent eigenvalues like in the POET setting, thus overcoming the necessary condition $p \leq n$. In addition, it is proved that our method recovers structures with intermediate degrees of spikiness, obtaining a loss which is bounded accordingly.

Then, an ad hoc model selection criterion which detects the optimal point in terms of composite penalty is proposed. Empirical results coming from a wide original simulation study where various low rank plus sparse settings are simulated according to different parameter values are described outlining in detail the improvements upon existing methods. Two real datasets are finally explored highlighting the usefulness of our method in practical applications.

Keywords: covariance matrix, nuclear norm, thresholding, low rank plus sparse decomposition, unshrinkage.

Contents

Chapter 1

Introduction

The present thesis concerns large dimensional covariance matrix estimation. 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 ([2]). However, it 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.

In recent years, datasets that do not fit into this framework have become very common, the data are very high-dimensional and sample sizes can be very small relative to dimension. The most traditional covariance estimator, the sample covariance matrix, is shown to be dramatically ill-conditioned in a large dimensional context, where the process dimension p is closer to or even larger than the sample dimension n, even in the case that the true covariance matrix is well-conditioned. Some solutions to this drawback have been proposed in the asymptotic context (for example [75] [15] [45]). An alternative recent approach is by numerical optimization, which provides in the non-asymptotic context, some solutions improving upon the mentioned ones.

As described in the existing literature, two key properties of the matrix estimation process assume a particular relevance in large dimensions:

1. well conditioning, i.e. numerical stability;
2. 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 amplify dramatically any error in the data. A large dimension may cause the impossibility to identify the unknown covariance structure and the difficulty to interpret the results.

The first property is strongly related to regularization techniques. A basic reference in this respect is Tibshirani (1996) ([108]), where the LASSO estimation algorithm in the context of regression models was first proposed. The second property can be ensured by dimensionality reduction methods, which can be used to reduce the parameter space dimensionality.

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 ([75]), others consider tapering the sample covariance matrix, that is, gradually shrinking the off-diagonal elements toward zero ([54]). At the same time, a common approach is to encourage sparsity, either by a penalized likelihood approach ([53]) or by thresholding the sample covariance matrix ([100]).

For this reason, our research studies a specific regularization problem under the assumption of low rank plus sparse decomposition for the covariance matrix. Such a problem is solved exploiting non-smooth convex optimization methods. This approach allows to properly address both reconditioning and dimensionality reduction issues and is proved to be effective even in a large dimensional context.

Our dissertation moves from a detailed outline of asymptotic approaches. In Chapter 2, we provide a thorough description of the motivation to our work and a review of some relevant asymptotic methods for covariance estimation. Maximum likelihood estimators and unbiased finite estimators are described ([2]). Specific treatment to the conditioning problem for covariance matrix estimates is given. The covariance shrinkage estimator derived by Ledoit and Wolf in the general asymptotic framework is described ([75]). Sparse covariance estimators are shown together with the underlying assumptions and the estimation error rates, with particular reference to the thresholding estimator of [15]. POET (Principal Orthogonal complEment Thresholding) estimator ([45]), which combines Principal Component Analysis and thresholding algorithms, is analyzed in detail.

In Chapter 3, we define the regularization problem above mentioned. It is a nuclear norm plus l_1 norm approximation problem, and works under the assumption of low rank plus sparse structure for the covariance matrix. It is composed by a least squares loss and a composite non-smooth penalty, which is the sum of the nuclear norm of the low rank component and the l_1 norm of the sparse component.

The numerical rationale behind the problem formulation is provided. It is shown how this problem can be recast from the point of view of numerical

analysis as a semi-definite program (SDP). Non standard optimization tools, as subgradient minimization methods, are needed to solve it. We describe the most recent solution algorithm and point out its effectiveness.

In Chapter 4, we provide a wide review of existing non-asymptotic methods. The evolution path of the most recent works is figured out. The most recent developments of the numerical approach under the assumption of low rank plus sparse structure for the covariance matrix are described, starting from the basic contribution by Chandraskeran et al. ([30]) which first proves the exact recovery of the covariance matrix in the noiseless context. This 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.

Then, the first approximate solution to recovery and identifiability in the noisy context, coming from [1], is described. In the following, the extension of [30] providing the first exact solution of the numerical problem in the noisy graphical model setting ([31]) is shown in detail. In that context, the objective is a least square loss penalized by the above mentioned composite penalty, and its optimization allows to recover the inverse covariance matrix. In conclusion, the extension of this framework to the covariance matrix estimation context, coming from [77], is explained. The resulting estimator is called LOREC (LOw Rank and sparsE Covariance estimator).

In the last chapter (Chapter 5), an improvement over the solution described in [77] is proposed, based on the unshrinkage of the estimated eigenvalues of the low rank component. Luo's approach is completed by deriving the rates of the sparse component estimate, and the conditions for its positive definiteness and invertibility. In addition, the rates of LOREC under the conditions of POET, and, more importantly, in a context where the eigenvalues of the low rank component are allowed to grow with $p^\alpha, \alpha \in [0, 1]$ (generalized spikiness context) are provided.

In the following, we show the results of our procedure on both simulated and real data sets. We illustrate a new model selection criterion which is proved to be effective in our context. An original simulation study is presented where extensive simulation results are pointed out, as well as the simulation algorithm and the estimation assessment framework.

In the end, the performance of our new proposed estimator is compared to the one of LOREC and POET under various settings. Two real examples are provided where our model is effective respect to the competitors. In particular, the second example is a banking supervisory data set which collects supervisory reporting indicators of the most relevant Euro Area banks. We explicitly thank the Supervisory Statistics Division of the European Central Bank, where the author spent a semester as a PhD trainee, for the allowance to use these data in anonymous form for research purposes.

The Conclusions (Chapter 6) sum up the main findings of our research.

Chapter 2

Covariance matrix estimation: state of the art

In this chapter, a short review of existing solutions to the problem of covariance matrix estimation is provided. Particular attention is given to the two properties displayed in the Introduction (well conditioning and identifiability) and to the performance of existing methods in the large dimensional context. An exhaustive review can be found in Pourhamadi (2013) ([95]).

This Chapter shows a path across existing estimators aimed at outlining the two mentioned features (well conditioning and identifiability) for each estimation setting, especially when p is very large compared to the sample size n or even larger. This is why, for each estimator, a detailed discussion of the asymptotic framework and the assumptions needed to ensure consistency (i.e. the convergence to the theoretical covariance matrix) is provided.

Existing approaches to the estimation problem are described in this Chapter, while non-asymptotic approaches will be the object of next chapters. The description of past approaches is intended to display the main issues encountered by existing methods, with particular reference to the large dimensional context, and the reasons why we need to develop an alternative numerical approach to the covariance estimation problem.

The first paragraph (2.1) is devoted to covariance matrix estimation under the assumption of normality for the data. The maximum likelihood estimator, i.e. the sample covariance matrix, is introduced and justified. The unbiased sample covariance matrix, under the assumption of fixed n, is then outlined. A specific remark on the asymptotic distribution of the sample covariance matrix under the assumption of independence and identical distribution for the data concludes the section.

In the second paragraph (2.2) the conditioning properties of the sample covariance matrix are explored. The reason why the sample covariance matrix is bad-conditioned when the dimension is close to the sample size is deeply explained and analyzed, as well as the reason why the inverse

covariance matrix dramatically amplifies the estimation error in case of bad-conditioning.

The third paragraph (2.3) widely describes a successful attempt to address the problem of reconditioning the sample covariance matrix when the dimension is larger than the sample size: the shrinkage estimator by Ledoit & Wolf ([75]). Their motivations, their results and their asymptotic context are properly highlighted, trying to retain the key elements of their approach.

The fourth paragraph (2.4) briefly outlines existing sparsity estimators, with particular reference to the thresholding estimator by Bickel & Levina ([15]), which is described in detail with respect to model assumptions and convergence rates. There we point out the strong link between sparsity assumptions and shrinkage thresholding. That family of estimators shows how it is possible to use sparsity to recondition the covariance estimate and to significantly reduce the number of parameters.

The fifth paragraph (2.5) describes covariance matrices estimator based on factor model assumptions. A brief overview of factor model specifications and underlying assumptions across history is provided, discussing the different asymptotic contexts. The relationship between Principal Component Analysis (PCA, [72]) and factor modelling (see [59]) is crucial in this respect. Finally, POET estimator ([45]), based on the assumption of approximate factor model with a sparse residual matrix, is widely illustrated, pointing out the crucial assumptions for consistency and identifiability.

In [45], the population covariance matrix is assumed to be the sum of a low rank and a sparse component. POET works under the assumption of sparse residual covariance matrix and pervasive eigenvalues of the low rank component (as $p \to \infty$). This structure is particularly convenient in a large dimensional context, and tackles both the issues mentioned above, as we will widely explain. For the same reasons, the factor analysis assumption is a key to approach covariance estimation in large dimensions. The asymptotic correspondence between PCA and factor estimation is there established according to the underlying assumptions and then exploited.

Before starting, we describe the basic matrix terminology. We restrict our analysis to the real case. The spectral theorem ensures that, when M is a positive semidefinite squared p - dimensional real matrix with rank r, there exists an orthogonal $p \times r$ matrix U and a diagonal $r \times r$ matrix Λ such that

$$M = U\Lambda U' = \sum_{i=1}^{r} \lambda_i u_i u_i', \tag{2.1}$$

which is the eigenvalue decomposition of M. Scalars $\lambda_1, \ldots, \lambda_r$ are called the eigenvalues of M and are strictly larger than 0. The r columns of U are the eigenvectors of M. If M is symmetric, the eigenvalues coincide with the singular values $\sigma_{1,\ldots,r}$, which are the square roots of the eigenvalues of $M'M$,

i.e. the absolute values of the eigenvalues of M. A fortiori, this happens if M is a covariance matrix, which is symmetric and positive definite.

The relevant norms we are going to use throughout the entire thesis are (see also [62]):

- $||M||_2 = \sqrt{\sigma_{max}(M'M)}$ is the spectral norm of M, which is its largest singular value.
- $||M||_\infty = \max_{i,j} |m_{ij}|$ is the infinity norm of M, which is the largest entry in magnitude.
- $||M||_F = trace(M'M) = \sqrt{\sum_i \sum_j m_{ij}^2}$ is the Frobenius norm of M, which is the square root of the sum of the entries of M.
- $||M||_* = trace(\sqrt{M'M}) = \sum_{i=1}^{p} \sigma_i$, sum of the singular values of M.$||M||_*$ is called nuclear norm. If M is a Positive SemiDefinite matrix (PSD), $||M||_* = tr(M)$, because the eigenvalues and the singular values exacly coincide.
- $||M||_1 = \sum_i \sum_j |m_{ij}|$: sum of the absolute values of the entries of M.

For a p-dimensional vector x, the relevant norms for our purpose are:

- $||x||_2 = \sqrt{\sum_i x_i^2}$, the Euclidean norm of x.
- $||x||_1 = \sum_{i=1}^{p} |x_i|$, the l_1 norm of x.
- $||x||_\infty = \max_i |x_i|$, the maximum norm of x.

2.1 Sample covariance matrix estimators

In this paragraph we focus on the most used estimator of the covariance matrix: the sample covariance matrix. First, we will derive it as the maximum likelihood estimator of the covariance matrix under the assumption of multivariate normality for our data (2.1.1). Maximum likelihood estimators are consistent when $n \to \infty$. This is why we then derive the unbiased covariance estimator under the assumption of n finite (2.1.2), which is a slightly modified version of the sample covariance matrix. These two estimators asymptotically converge when $n \to \infty$, under the assumption of p fixed. In the end of this paragraph, we give a flash about the behaviour of this estimator under the assumption of independence and identical distribution for our data when $n \to \infty$ (2.1.3).

Our main reference for this argument is the famous book by Anderson ([2]).

2.1.1 The Maximum Likelihood covariance estimator

Suppose we have a sample $(x_1, \dots x_n)$, from a real-valued $p-$dimensional normal random variable $x \sim N_p(\mu^*, \Sigma^*)$, with $p \leq n$. The $p \times p$ matrix $\Sigma^* = E((x - \mu^*)(x - \ \mu^*)')$ is real positive definite and symmetric, while $\mu^* = E(x)$ is a $p \times 1$ vector.

The density of x is the following:

$$f(x|\mu^*, \Sigma^*) = (2\pi)^{-\frac{1}{2}p} |\Sigma^*|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x - \mu^*)'\Sigma^{*-1}(x - \mu^*)\right].$$

where μ^* is a $p \times 1$ vector and Σ^* is a $p \times p$ invertible (positive definite) matrix.

The likelihood function is

$$L(\mu^*, \Sigma^*) = \prod_{i=1}^{n} N(x_i|\mu^*, \Sigma^*) =$$

$$= (2\pi)^{-\frac{1}{2}pn} |\Sigma^*|^{-\frac{1}{2}n} \exp\left[-1/2 \sum_{i=1}^{n}(x_i - \mu^*)'\Sigma^{*-1}(x_i - \mu^*)\right].$$

The log-likelihood is then

$$\log L(\mu^*, \Sigma^*) = -\frac{1}{2}pn \log 2\pi - \frac{1}{2}n \log |\Sigma^*| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu^*)'\Sigma^{*-1}(x_i - \mu^*).$$

We denote by $\hat{\mu}_{ML}$ and $\hat{\Sigma}_{ML}$ the vector and the positive definite matrix maximizing $\log L$. They are the maximum likelihood estimators of μ^* and Σ^*. Since $\log L$ is an increasing function of L, $\log L$ and L share the same maximum respect to our parameter estimates.

The following important theorem holds:

Theorem 2.1.1. *If $x_1, \dots x_n$ constitute a sample from $N(\mu^*, \Sigma^*)$ with $p < n$, the maximum likelihood estimators of μ^* and Σ^* are $\hat{\mu}_{ML} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\hat{\Sigma}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$ respectively.*

The proof can be found in Anderson (1958), page 67 and following. It exploits the properties of the arithmetic mean and of positive definite matrices. The key argument is that $\log L$ can be rewritten in the following way:

$$-\frac{1}{2}pn \log 2\pi - \frac{1}{2}\log |\Sigma^*| - \frac{1}{2}tr\Sigma^{*-1}D - \frac{1}{2}n(x_i - \mu^*)\Sigma^{*-1}(x_i - \mu^*)',$$

where $D = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$.

In order to perform maximization, the necessary assumption is that Σ^* is a positive definite matrix. This condition is necessary to ensure that the

term $n(x_i - \mu^*)\Sigma^{*-1}(x_i - \mu^*)'$ achieves a maximum for $\mu^* = \bar{x}$ and the term $\log|\Sigma^*| - tr(\Sigma^{*-1}D)$ achieves a maximum for $\Sigma^* = \frac{1}{n}D$.

ML estimators show a number of interesting optimality properties. In particular, they are consistent and asymptotically efficient ([34]). A theorem by Cramer ensures that $\hat{\mu}_{ML}$ and $\hat{\Sigma}_{ML}$ are minimum variance (asymptotically) unbiased estimators. These properties hold if and only if $n \to \infty$.

Note that also the condition $p < n$ is necessary in order to perform maximization. In order to see this point, we need to recall a basic theorem ([2], p.77):

Theorem 2.1.2. *The maximum likelihood estimator $\hat{\mu}_{ML} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, from $N(\mu^*, \Sigma^*)$, is distributed according to $N(\mu^*, \frac{1}{n}\Sigma^*)$ and independently of $\hat{\Sigma}_{ML} = \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$. $n\hat{\Sigma}$ is distributed according to $\sum_{i=1}^{n-1} z_i z_i'$, where $z_i \sim N(0, \Sigma^*)$, and $z_1, \ldots, z_{n-1}$ are independent.*

This theorem states that under the multivariate normality assumption for the data, $n\hat{\Sigma}$ is the sum of $n-1$ squared p dimensional matrices having rank 1. If $p \geq n$, $n\hat{\Sigma}$ will never have full rank p.

In addition, it has been shown by Wishart ([113]) that $D = n\hat{\Sigma}$ is a matrix-valued stochastic process having the following distribution:

$$f(D|\Sigma^*) = \frac{|D|^{\frac{1}{2}(n-p-1)} \exp\left(-\frac{1}{2}tr(\Sigma^{*-1}D)\right)}{2^{\frac{1}{2}np}\pi^{\frac{p(p-1)}{4}}|\Sigma^*|^{\frac{1}{2}n}\prod_{i=1}^{p}\Gamma[\frac{1}{2}(n+1-i)]}$$

which is a Wishart distribution with $\nu = n-1$ degrees of freedom, where $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$ is the usual Gamma function. The proof is reported in [2] (p.252 and following). It exploits massively the linear transforms of random variables, and is based on the properties of Gram-Schmidt orthogonalization algorithm.

This results was first derived for a bi-variate distribution by Fisher ([51]) where the distribution of the correlation coefficient (first defined by Karl Pearson in [91]) was also derived.

We can now understand why $p < n$ is a necessary condition. If $n \leq p$, $f(D|\Sigma^*)$ is no longer a density, such that it is no longer possible to derive the asymptotic distribution for $\hat{\Sigma}$ (i.e., all the usual optimality properties of ML estimators are lost). In fact, $|D|$ would be zero, and the distribution would thus be degenerate, having null measure in $R^{p\times p}$ everywhere. Note also that if $n = p+1$ $f(D|\Sigma^*)$ has not a mode, analogously to the χ^2 distribution with two degrees of freedom.

In the same way, denoting by T the quantity $T = (\bar{x} - \mu^*)'W^{-1}(\bar{x} - \mu^*)$, where $W = \frac{D}{n-1}$, it has been shown by Hotelling ([64]) that

$$\frac{\nu - p - 1}{vp}T^2 \sim F_{p,\nu-p+1},$$

where F is Fisher's distribution with p and $\nu - p + 1$ degrees of freedom ($\nu = n-1$). T^2 is called Hotelling's T-squared distribution. It is non-singular if and only if both $\hat{\mu}$ and $\hat{\Sigma}$ are non-singular, i.e. if Σ^* is positive definite and $\nu - p + 1 > 0$ (equivalent to $n > p$).

So, both the sample mean and the sample covariance matrix are ML estimators of the true mean and the true covariance matrix if and only if the true covariance matrix is positive definite and the dimension p is strictly smaller than the sample size n. In particular, the distribution of the sample covariance matrix is $\frac{n}{n-1}Wishart(\Sigma^*, n-1)$. This means that $\hat{\Sigma}$ is biased if n is finite. Note that this distribution does not change even when the true mean μ^* is known, unless $\bar{x}$ is replaced by the true μ^*. In that case, the degrees of freedom are n and the resulting estimator $(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu^*)(x_i - \mu^*)')$ is unbiased.

2.1.2 The unbiased covariance estimator: fixed n context

In order to derive the **finite sample** unbiased estimator of the covariance matrix, the key result is Theorem 2.1.2 about the distribution of $D = n\hat{\Sigma} = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$ shown above.

A corollary of that theorem states:

Corollary 2.1.1. *Let $x_1, \ldots, x_n (n > p)$ be independently distributed, each according to $N(\mu^*, \Sigma^*)$. The distribution of $\hat{\Sigma}_\nu = \frac{1}{\nu}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$ is $Wishart(\Sigma^*, \nu)$, where $\nu = n - 1$.*

This result means that $\hat{\Sigma}_{n-1} = (\frac{1}{n-1})\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$ is the unbiased estimator of the covariance matrix when the dimension n is finite. This estimator will be the input of our new estimation procedure in Chapter 4. Clearly, $\hat{\Sigma}_{n-1}$ and $\hat{\Sigma}_n$ converge asymptotically to the same estimator.

We are now going to derive the asymptotic (normal) distribution of the sample covariance matrix in the more general case of IID data.

2.1.3 Covariance matrix estimation: the IID data context

Let us suppose $x_i \sim IID(\mu^*, \Sigma^*)$, $i = 1 \ldots, n$. We want to derive the asymptotic distribution of $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})'(x_i - \bar{x})$. Under the IID hypothesis, we have:

$$E(x_i x_i') = E(x_i)E(x_i') = \Sigma^* + \mu^* \mu^{*'},$$

$$V(x_i x_i') = V(x_i) + V(x_i) = \Sigma^* + \Sigma^* = 2\Sigma^*.$$

Our target can be rewritten as the sum of three components:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})' = \sum_{i=1}^{n}\frac{x_i x_i'}{n} - 2\sum_{i=1}^{n}\bar{x}\frac{x_i'}{n} + \sum_{i=1}^{n}\frac{\bar{x}\bar{x}'}{n}$$

Since $\sum_{i=1}^{n} \frac{x_i}{n} \overset{prob}{\to} \mu^*$, we have that

$$-2\bar{x}\sum_{i=1}^{n}\frac{x_i}{n} + \sum_{i=1}^{n}\frac{\bar{x}\bar{x}'}{n} = -2\bar{x}\bar{x}' + \bar{x}\bar{x}' = -\bar{x}\bar{x}'.$$

converges in probability as follows:

$$-\bar{x}\bar{x}' \overset{prob}{\to} -\mu^*\mu^{*'} \tag{2.2}$$

Now, the first component $\sum_{i=1}^{n} \frac{x_i x_i'}{n}$ can be rewritten as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{(x_i x_i')}{\sqrt{n}}$$

So, for the Central Limit theorem, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{x_i x_i' - (\Sigma^* + \mu^*\mu^{*'})}{\sqrt{n}} \overset{CLT}{\to} \frac{1}{n}N(\mu^*\mu^{*'} + \Sigma^*, 2\Sigma^*).$$

Recalling (2.2), we have that

$$\hat{\Sigma}_n \overset{distrib}{\to} \frac{1}{\sqrt{n}}N(\Sigma^*, 2\Sigma^*). \tag{2.3}$$

These results find confirmation in [58].

2.2 The sample covariance matrix: conditioning properties

We are now going to briefly talk about matrix conditioning. Let us suppose p and n are fixed. If $n > p$, the expected value of $\Sigma_{\nu=n-1}$ is Σ^*, and the entries of its covariance matrix are $V(\hat{\sigma}_{n,ij}) = \frac{(\sigma_{ij}^{*2} + \sigma_{ii}^*\sigma_{jj}^*)}{(n-1)}$. This highlights why the variance of $\hat{\Sigma}_n$ increases as the true condition number of Σ^* increases. If the condition number $c = \sigma_{max}/\sigma_{min}$ increases, the correlation between the components x_i and x_j increases, because Σ^* is closer to collinearity. Consequently, $V(\hat{\sigma}_{n,ij})$ increases, because σ_{ij}^{*2} is closer to its maximum, which is $\sigma_{ii}^*\sigma_{jj}^*$ (for the Cauchy-Schwartz inequality).

Coming back to the main point, it is crucial to study the behaviour of the sample eigenvalues. In the matrix estimation context there is a relevant issue about numerical conditioning, i.e. the behaviour of sample maximum and minimum singular values, of a $p \times n$ data matrix X.

Theorem 2.2.1 (Theorem ([39])). *Given natural numbers n, p with $p < n+1$ let X be a $p \times n$ matrix with i.i.d. Gaussian entries that have zero-mean*

and variance $\frac{1}{n}$. *Then the largest and smallest singular values* $\sigma_{min}(X)$ *and* $\sigma_{max}(X)$ *are such that*

$$max\left\{Pr\left[\lambda_{max} \geq 1 + \sqrt{\frac{p}{n}} + t\right], Pr\left[\lambda_{min} \leq 1 - \sqrt{\frac{p}{n}} - t\right]\right\} \leq \exp\left\{\frac{-nt^2}{2}\right\},$$

for any $t > 0$.

This theorem was proved by using arguments from random matrix theory and the geometry of Banach spaces. It is an essential result to provide a probabilistic bound for the error distance $||\hat{\Sigma}_n - \Sigma^*||_2$, where $\hat{\Sigma}_n = \frac{1}{n}X'X = \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$.

In fact, the following Lemma holds:

Lemma 2.2.1. *Let* $\psi = ||\Sigma^*||_2$. *Given any* $\delta > 0$ *and* $\phi > 0$ *with* $\psi \leq 8\phi$, *let the number of samples* n *be such that* $n \geq \frac{64p\phi^2}{\delta^2}$. *Then we have that*

$$Pr[||\Sigma_n - \Sigma^*||_2 \geq \delta] \leq 2\exp\left(-\frac{n\delta^2}{128\psi^2}\right).$$

This Theorem is based on a specific assumption on ψ, the largest eigenvalue of Σ^*. By appropriately setting the parameter ψ, we can obtain the probabilistic bound accordingly.

This Lemma relies on the fact that the spectral norm is unitarily invariant, such that it is possible to assume a diagonal structure for $\hat{\Sigma}$ without loss of generality and then apply the previous theorem 2.2.1.

It is remarkable that without further assumptions, $\hat{\Sigma}_n$ is not invertible if $p > n$ (since it is perfectly collinear, having clearly at most rank n, and for the rest null eigenvalues). Even if $p \leq n$, in the case the ratio p/n is less than 1 but **not negligible**, the estimated (maximum and minimum) eigenvalues are numerically unstable, since the probabilistic bound is too large. This may result in bad conditioning (i.e. too large condition number) for $\hat{\Sigma}_n$. This is why in the Big Data context, when p is very large, it is frequent to have an ill-conditioned sample covariance matrix, since it is difficult to have enough observation to keep the ratio p/n negligible ([75]).

The example in figure (2.1) clearly outlines the described drawback. The eigenvalues of the covariance matrix of a simulated $n \times p$ process $\epsilon_i = N_p\ (0, \frac{1}{n}I)$, $p = 100$, $n = [10, 50, 100, 500, 1000, 10000]$ are plotted. The figure displays how the dispersion of the eigenvalues decreases as p/n decreases. All distributions tend to the Marcenko-Pastur distribution, which is proved to be the limiting distribution of the eigenvalues of IID random variables (in the Kolmogorov asymptotic framework, see [79]). The rank is always equal to $\min(p, n-1)$. If $p = n$, the matrix is thus singular.

We have provided this simple example to state that without further assumption on the eigen-structure (values and vectors) of Σ^*, the condition

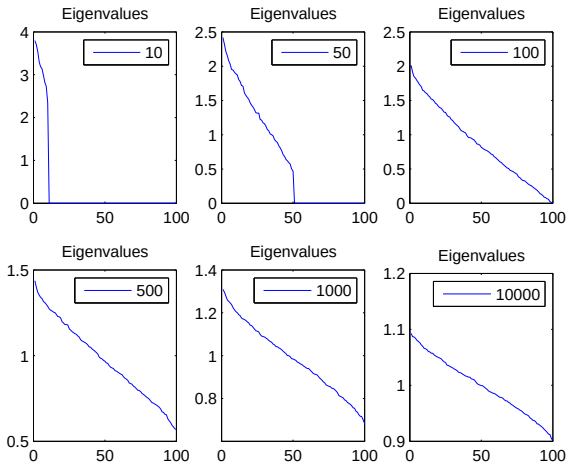

Figure 2.1: Eigenvalues of the sample covariance matrix of $\epsilon_i = N_p(0, \frac{1}{n}I)$, $p = 100$, n varying

$p \leq n$ is unavoidable in order to guarantee the positive definiteness (and thus the invertibility) of our covariance estimate. Anyway, the recovery of the eigen-structure of a covariance matrix is strongly related to the underlying assumptions and to the asymptotic context.

We now enumerate three parameter settings relevant for our dissertation:

1. p and n fixed: this is the case of $\hat{\Sigma}_{n-1}$, and all numerical estimators we will analyze in next chapters ([31], [1], [77],[15])

2. p fixed, $n \to \infty$: this is the case of $\hat{\Sigma}_{ML}$, or of the approximate factor model ([29])

3. $\frac{p_n}{n} \to c$ when $n \to \infty$: here we find the General asymptotic framework, used by Ledoit and Wolf to ensure the consistency of their estimator ([75]), and the Kolmogorov asymptotic framework (where also $p \to \infty$). Also consistency properties of the thresholding estimator ([15]) and of POET estimator ([45]) are derived under a similar framework, where a function of p and n tends to 0 while $n \to \infty$. See for more explanations sections (2.4) and (2.5).

In the second context, with fixed p and n, the outlined results concerning numerical conditioning for the sample covariance matrix hold, and the condition $p \leq n$ is unavoidable without further assumptions to derive finite sample bounds. This is why one of the aims of the present work is trying to exploit results from the third asymptotic framework (in terms of model assumptions) to establish bounds under the finite sample context dropping the condition $p \leq n$.

2.2.1 Matrix conditioning as an ill-posed inverse problem

We are now explaining in detail why a bad-conditioned sample matrix is a fatal drawback for us. The reason stands in the consequences deriving from the inversion of a bad-conditioned matrix.

Let us now consider the standard linear system $Ax = b$, where A is $p \times p$, and x, b are $p \times 1$. If our aim is to derive b (the output), we are solving the direct problem. If our aim is to derive x (the input), we are solving the inverse problem. If A is full rank, Cramer's theorem is ensuring that the inverse problem has exact solution $x^* = A^{-1}b$. Otherwise, if A has rank $r < p$, we need to solve the least squares problem

$$\min_{x \in \mathbb{R}^p} ||Ax - b||_2,$$

and we have

$$x^* = \sum_{i=1}^{r} \frac{|u_i' b|}{\lambda_i} u_i \tag{2.4}$$

$$||Ax^* - b||^2 = \sum_{i=r+1}^{p} ||u_i' b||^2.$$

This fundamental result was proved in [40].

How much is solution the x^* reliable? Hadamard([57]) outlined the three characteristics of a well-posed problem:

- existence: the problem admits one solution
- uniqueness: the problem has at most one solution
- stability: the problem is not sensitive to data perturbation.

In our context, if A is full rank, the inverse problem may be ill-posed since it violates the stability condition. If A is not full rank, the inverse problem is ill-posed since it violates the existence and the uniqueness condition (there are only approximate solutions, no exact ones). The least squares system serves for identifying in any case a solution even if there would be none.

Anyway, (2.1) and (2.4) enable us to understand why the inverse of bad-conditioned matrices are numerically unstable. The solution of the direct problem is $Ax = U\Lambda U'x = \sum_{i=1}^{p} \lambda_i (u_i' x) u_i$, which dampens the components corresponding to the smallest eigenvalues of A. On the contrary, (2.4) shows us that the solution of the inverse problem amplifies the effects of the same components. If we assume that b is perturbed, i.e. $b_\epsilon = b + \epsilon$, we note that $x_\epsilon = x^* + \sum_{i=1}^{r} \frac{|u_i' \epsilon|}{\lambda_i} u_i$. So, if A is bad conditioned (i.e. we have very small eigenvalues), the effect of data perturbation is amplified, and the solution may not be effectively usable in applications.

This is why Picard ([93]) elaborated a condition under which the inverse solution is reliable. It states that $x^* = \sum_{i=1}^{r} \frac{|u_i'b|}{\lambda_i} u_i < \infty$ if and only if $|u_i'b|$ decays more rapidly than the corresponding λ_i for all i, which occurs if $\lambda_i > \tau \, \forall i$, where τ is the threshold at which the singular values are levelled by the noise.

If this condition is violated, a regularization method, like the truncated singular value decomposition (TSVD, see [55]) or Tikhonov's regression method ([109]) or other regression methods (like the ridge one), are needed. This is why the nonasymptotic approach for covariance matrix estimation essentially consists in specifying appropriate regularization problems under suitable conditions for deriving improved error rates, as we will widely describe in the following chapters.

Note that there is a huge literature dealing with the distribution of eigenvalues. We mention again Marcenko-Pastur law, which describes the behaviour of the singular values of a rectangular random matrix having Gaussian entries ([79]). Tracy and Widom ([107]) found the limiting distribution of the singular values of a large dimensional random Hermitian matrix. Johnstone ([70]) found out the limiting distribution of the largest eigenvalue in principal component analysis (for $n \leq p$, under the assumption of independent normality for the columns of the data matrix) which is proportional to a Wishart of order 1. A recent work by Chiani ([33]) derived the exact distribution of the largest eigenvalues for real Wishart matrices and Gaussian Orthogonal Ensembles.

The work in [70], in particular, outlined that for large p it can be easier to recover the top r eigenvalues if they are particularly spiked, because the distribution of the $(r+1)$-th eigenvalue is bounded by a Tracy-Widom law of lower dimensions ($n \times (p-r)$ respect to $n \times p$). Thus, the $(r+1)$-th eigenvalue of a set of p eigenvalues where r are spiked is stochastically smaller than the largest eigenvalue of a setting of $(p-r) < p$ variables non-spiked. This fact suggests that large dimensions ($p \to \infty$) can help the recovery of strong eigenvalues and somehow justifies the use of "scree-plot" to choose the number of eigenvalues.

There are also some results on the distribution of the smallest eigenvalues. We refer to [8] for a general review.

All in all, the problem of reconditioning our covariance matrix estimate is approached differently according to the related asymptotic context. In Chapter 4 we will focus on the non-asymptotic context, outlining various solutions recently provided. Now, we will focus on the description of key covariance estimators in the asymptotic context where both p and n are allowed to tend to ∞. The estimator we are about to describe belongs to the class of shrinkage estimators ([68]) which represent a widely used approach in this context as an effective regularization method. It is relevant to note that the distributional assumption of normality is no longer needed, since

the approach we are going to describe is distribution-free.

2.3 Shrinkage towards the identity: Ledoit and Wolf's approach

Ledoit and Wolf were the first to derive in [75] a consistent estimator of the covariance matrix in a new asymptotic framework, called general asymptotic framework. They proposed a way to temper the numerical instability of sample eigenvalues, explicitly reconditioning them by shrinkage. The adoption of a new asymptotic framework was needed to ensure the shrinkage intensity to be positive, avoiding it to vanish in the limit. Their estimator is also Bayesian in nature, since it is a combination of a priori and sample information. They call it Empirical Bayesian estimator.

The motivating result of their analysis it reported below.

Theorem 2.3.1. *The eigenvalues are the most dispersed diagonal elements that can be obtained by rotation of a symmetric matrix.*

The proof exploits the invariance by rotation of trace.

This causes that the largest sample eigenvalues are positively biased, while the smallest are negatively biased, and the bias increases in p/n (recall Theorem 2.2.1). The pattern of sample eigenvalues depends on the Marcenko-Pastur distribution, which holds in the Kolmogorov asymptotic framework. As described, under Kolmogorov asymptotics the ratio p/n tends to a specific constant, while both p and n tend to infinity.

Here we report the solution proposed by Ledoit and Wolf to the described problem. Their idea is to shrink the sample covariance matrix towards the identity matrix, solving the following optimization problem (thus reconditioning the eigenvalues):

$$\min_{\rho_1,\rho_2} E[||\Sigma - \Sigma^*||^2]$$
$$\text{s.t.}\Sigma = \rho_1 I_p + \rho_2 \hat{\Sigma}_n.$$

where ρ_1 and ρ_2 are nonrandom coefficients.

The theoretical solution to this problem is the **optimal linear shrinkage** estimator

$$\Sigma_{LW} = \frac{\beta^2}{\gamma^2}\mu I + \frac{\alpha^2}{\gamma^2}\hat{\Sigma}_n \tag{2.5}$$

with $E[||\Sigma_{LW} - \Sigma^*||^2] = \frac{\alpha^2\beta^2}{\gamma^2}$, where:

$$\begin{aligned}\mu &= < \Sigma, I >; \\ \alpha^2 &= ||\Sigma^* - \mu I||^2; \\ \beta^2 &= E[||\hat{\Sigma}_n - \Sigma^*||^2]; \\ \gamma^2 &= E[||\hat{\Sigma}_n - \mu I||^2].\end{aligned}$$

Their derivation exploits the natural Pythagorean relationship

$$\alpha^2 + \beta^2 = \gamma^2. \tag{2.6}$$

In this view, the ratio $\frac{\beta^2}{\gamma^2}$ is called optimal shrinkage intensity.

The most important interpretation of this approach for our purposes is the following. It is well known (Theorem 2.2.1) that the sample eigenvalues of IID data have bounded error respect to the true ones, so that, under the condition $p \leq n$ (p and n fixed), $\frac{1}{p}E(\sum_{i=1}^{p} \hat{\lambda}_i) = \frac{1}{p}\sum_{i=1}^{p} \lambda_i$, i.e. the trace of Σ^* is unbiasedly estimated.

At the same time, theorem 2.3.1 shows that sample eigenvalues have a larger dispersion around their grand mean respect to the true ones (assuming that the eigenvectors are reliable). From (2.6) we can argue that

$$\frac{1}{p}E\left[\sum_{i=1}^{p}(\hat{\lambda}_i - \mu)^2\right] = \frac{1}{p}\sum_{i=1}^{p}(\lambda_i - \mu)^2 + E[||\hat{\Sigma}_n - \Sigma||^2],$$

i.e. the excess dispersion of the sample eigenvalues is the error of the sample covariance matrix. This is why here the authors bound $[||\hat{\Sigma}_n - \Sigma||^2]$ by bounding $\frac{1}{p}E\left[\sum_{i=1}^{p}(\hat{\lambda}_i - \mu)^2\right]$, where $\mu = 1$.

So, Σ_{LW} implicitly does the reconditioning of eigenvalues, since

$$\lambda_{i,LW} = \frac{\beta^2}{\gamma^2}\mu + \frac{\alpha^2}{\gamma^2}\hat{\lambda}_i, \ \forall i = 1, \ldots, p.$$

$\frac{1}{p}E[\sum_{i=1}^{p}(\hat{\lambda}_{i,LW} - \mu)^2]$ is equal to $\frac{\alpha^2}{\gamma}$, and is even smaller than the dispersion of the true ones, for the reasons described above. Note that this method is very similar in its meaning to the $\max \log - \det$ heuristics for nuclear norm minimization (see [49]).

2.3.1 General Asymptotics

In order to derive a feasible estimator, we now need to get into a new asymptotic framework, since the optimal shrinkage intensity β^2 vanishes as $||\hat{\Sigma}_n - \Sigma^*||^2$ vanishes when $n \to \infty$ in the standard asymptotic framework (as proved in paragraph 2.1.3, see convergence (2.3)). This fact, when p is closer to n or even larger, is inconsistent with reality. So, a new asymptotic framework, called General Asymptotics, is needed, where β^2 is not vanishing.

Consider $n = 1, 2, \ldots$ indexing a sequence of statistical models, and for every n, X_n is a $p_n \times n$ matrix of n iid observations on a system of p_n random zero mean variables with covariance matrix Σ_n.

The following assumption characterizes this context:

A1. There exists a constant K_1 independent of n such that $p_n/n \leq K_1$.

It is remarkable that in this setting p can change and even go to infinity, but it is not required. Differently from the Kolmogorov asymptotic framework (the one of Marcenko-Pastur Law), it is not even necessary this ratio tends to a finite constant.

Two further assumptions are needed to derive a consistent estimator of Σ_{LW}. If $\Sigma_n = \Gamma_n \Lambda_n \Gamma_n'$, the product $Y_n = \Gamma_n' X_n$ is a set of uncorrelated variables spanning the same space as the original variables. The following restrictions on the higher moments of Y_n are imposed:

A2. There exists a constant K_2 independent of n such that

$$\frac{1}{p_n} \sum_{i=1}^{p_n} E[(y_{i1}^n)^8] \leq K_2,$$

A3.

$$\lim_{n \to \infty} \frac{p^2}{n^2} \frac{\sum_{i,j,k,l} \in \mathbb{Q}_n Cov(y_{i1} y_{j1}, y_{k1} y_{l1})}{\text{Cardinal of} Q_n} = 0.$$

where Q_n denotes the set of all the quadruples that are made of four distinct integers between 1 and p_n.

Assumption 2 states that the eighth moment of y is bounded (on average). Assumption 3 states that products of uncorrelated random variables are themselves uncorrelated (on average, in the limit). In the case when general asymptotics degenerate into standard asymptotics ($p/n \to 0$); Assumption 3 is trivially verified as a consequence of Assumption 2.

For what previously stated, Assumption 3 is verified when random variables are normally or even elliptically distributed, since the sample covariance of (uncorrelated) normal variables is asymptotically unbiased. Anyway, **A3** is much weaker than that situation.

These assumptions are specifically needed to derive the sample counterparts of μ,γ^2, β^2.

Note that these two assumptions heavily involve the eigenstructure (eigenvalues and eigenvectors) of the true covariance matrix. Here we need to impose restrictions on eighth moments, for the particular nature of their optimal weights. Anyway, the need to control the pervasiveness of the latent structure in the covariance matrix is crucial for model recovery. We also underline how much latent factorial assumptions can impact on covariance estimation. This is why we are going to specifically discuss the relationship between factor modelling and covariance estimation in paragraph (2.5).

Under these assumptions, Ledoit and Wolf approach the study on the consistency of their estimator. In their context, the reference norm is $||A||_n =$

$\frac{1}{p_n}tr(AA')$, such that the identity matrix has always norm one, and the reference cross product is $< A_1, A_2 >_n = \frac{1}{p_n}tr(A_1A_2')$. The problem of obtaining meaningful absolute rates in high dimensions is another relevant issue. As we will see, in [45] the authors derive asymptotic rates for the relative error matrix (and not the covariance matrix itself). Instead, under the nonasymptotic setting (Chapter 4), we will obtain finite absolute rates, even under the same assumptions of [45].

We are now going to show why the sample covariance matrix is not consistent in this context, differently from the finite p context, where the covariance matrix is asymptotically consistent under the assumption of normality. The authors show that quantities $\mu_n =< \Sigma_n, I >$,$\alpha_n^2 = ||\Sigma_n - \mu_n I||^2$, $\beta_n^2 = E[||\hat{\Sigma}_n - \Sigma_n||^2]$, $\gamma_n^2 = E[||\hat{\Sigma}_n - \mu I||^2]$ are bounded in the general asymptotic framework when $n \to \infty$. Then, they prove the following important Theorem:

Theorem 2.3.2. *Define* $\theta_n^2 = Var(\frac{1}{p_n}\sum_{i=1}^{p_n} E[(y_{i1}^n)^2])$. θ_n^2 *is bounded as* $n \to \infty$, *and we have:*

$$\lim_{n\to\infty} E[||\hat{\Sigma}_n - \Sigma_n||^2] = \frac{p_n}{n}(\mu_n^2 + \theta_n^2).$$

This result states that the sample covariance matrix is not consistent under the general asymptotic framework, since its expected loss is lower bounded by $\frac{p_n}{n}(\mu_n^2)$, which does not usually vanish. (Recall that θ_n^2 vanishes asymptotically under the assumption of normality, for convergence (2.2)).

There are two interesting exceptions:

- when $\frac{p_n}{n} \to 0$, we fall into the standard asymptotic context, where the sample covariance matrix is consistent. The only difference is that more general case $p = o(n)$ is allowed, i.e. p is allowed to be unbounded and grow towards infinity;

- $\mu_n^2 \to 0$ and $\theta_n^2 \to 0$. μ_n^2 implies that most of the random variables have vanishing variances, i.e. there are $O(n)$ asymptotically degenerate variables. So, if the number of nondegenerate random variables is NOT negligible with respect to the number of observations, the sample covariance matrix is not consistent.

Inconsistency is due to the disequilibrium between the number of datapoints np_n and the number of parameters $p_n(p_n + 1)/2$. This is a key point in our analysis, which is unsolved by the approach of Ledoit and Wolf. In fact, they write there is no DIRECT consistent estimator of the covariance matrix under the general asymptotics. Their strategy is to derive a consistent estimator of their theoretical estimator, which is proved to have the minimum risk among all the linear combinations of I_p and Σ_n and is shown to be better conditioned than the sample covariance matrix.

So, shrinkage matters unless $\frac{p_n}{n}$ is negligible respect $\frac{\gamma^2}{\mu^2}$, i.e. if the dispersion of sample eigenvalues is much larger than $\frac{p_n}{n}$.

To conclude this section, we are now going to explain how Ledoit and Wolf derive a consistent estimator for Σ_{LW}.

They introduce sample counterparts of their key quantities:

$$\begin{aligned} m_n &= < \hat{\Sigma}_n, I >_n, \\ d_n^2 &= ||\hat{\Sigma}_n - m_n I||^2, \\ \bar{b}_n^2 &= \sum_{k=1}^{n} ||x_{.k}^n x_{.k}^{n'} - \hat{\Sigma}_n||, \\ b_n^2 &= \min(\bar{b}_n^2, d_n^2), \\ a_n^2 &= d_n^2 - b_n^2, \end{aligned}$$

where $x_{.k}^n$ denote the $k-th$ column of X_n.

All these sample counterparts are consistent in the general asymptotic framework, i.e. they converge to μ_n^2, α_n^2, β_n^2, γ_n^2 respectively in quadratic mean.

Then, their feasible consistent estimator is

$$\hat{\Sigma}_{LW} = \frac{b_n^2}{d_n^2} m_n I_n + \frac{a_n^2}{d_n^2} \hat{\Sigma}_n \tag{2.7}$$

This estimator is consistent in the general asymptotic framework respect to Σ_{LW}, i.e. they share the same asymptotic expected loss. Thus, the expected quadratic loss $\frac{\alpha^2 \beta^2}{\gamma^2}$ can be consistently estimated in quadratic mean by $\frac{a_n^2 b_n^2}{d_n^2}$.

$\hat{\Sigma}_{LW}$ is shown to have an important optimality property: it has the same asymptotic risk as the theoretical optimal linear combination of $\hat{\Sigma}_n$ and I_n with random coefficients. In addition, its condition number is proved to be bounded in probability, which is very important for practical use.

The approach by Ledoit and Wolf is undoubtedly very elegant. However, there is still one main difficulty: their estimator is excessively better conditioned than the true covariance matrix, i.e. it is often too biased, for the presence of the identity matrix in the estimator. This is why another major point of our dissertation will deal with the need of "unshrinking" the estimated eigenvalues.

In fact, the numerical issue is not the only relevant reason for desiring a well conditioned estimate of the covariance matrix. Deep statistical reasons lie behind this need: we suppose that the true covariance matrix Σ^* is well conditioned, that is there is no multi-collinearity among our p variables. In this respect, a well conditioned estimate is crucial also for fitting purposes, i.e. to improve the statistical properties of the estimate.

We recall the previous skrinkage estimator by the same authors ([74]) in the market portfolio context. There, the authors specify for the covariance matrix a single-index model (consistently with the basic theory of asset prices, see [103]), which is essentially a one-factor latent model, and then estimate the covariance matrix deriving the optimal shrinkage intensity towards the single-index as described. This single index covariance matrix estimator is an interesting contact point between latent variable models and shrinkage methods.

Before passing to the analysis of factor-based covariance matrix estimators (paragraph (2.5)), we now briefly outline the covariance estimators based on pure sparsity assumptions, with particular reference to the use of shrinkage thresholding. In this context, sparsity means that our true covariance matrix has a prevalence of zeros.

2.4 Sparse covariance matrix estimation

In this section we list the most relevant estimators based on a pure sparsity assumption, which can be effective for reducing the number of parameters and reconditioning the estimate, removing unnecessary off-diagonal correlations. If $p/n \to c \in (0,1)$ (general asymptotic framework) the eigenvalues of $\hat{\Sigma}_n$ follow the Marcenko-Pastur law, supported on $(1-\sqrt{c})^2, (1+\sqrt{c})^2$. If p/n does not tend to a constant, we do not have any guarantee. For this reason, enforcing sparsity can be a key for obtaining a full rank estimate in high dimensions, even when $n < p+1$. However, there are lots of different types of sparsity assumptions, methods and asymptotic frameworks to prove consistency.

The natural context which gave rise to the concept of sparsity lies in a data-set showing a clear index ordering among variables. This condition arises easily for spatial data, when the variables are geographical areas for which a proximity matrix is naturally defined. Applications include spectroscopy and climate data.

For this kind of data, several methods have been developed. Banding the covariance matrix, by appropriately defining a banding parameter, is one effective solution. In that approach ([14]), the matrix reference class is $\Sigma^* \in U(\epsilon_0)$, where

$$U(\epsilon_0) = \big\{\Sigma^* \in \mathbb{R}^{p\times p} : 0 < \epsilon_0 \leq \Lambda_i(\Sigma^*) \leq \epsilon_0^{-1} < +\infty, \\ \max_j\{\sum_i |\sigma^*_{ij}| : |i-j| > k\} \leq Ck^{-\alpha}\big\}, \tag{2.8}$$

which is the class of matrices having uniformly bounded eigenvalues and banded covariance.

For any $\Sigma^* \in U(\epsilon_0)$, the natural ordering among variables is therefore

enforced imposing:

$$\left\{\Sigma^* : \sigma^*_{ii} \leq M, \max_j\{\sum_i |\sigma^*_{ij}|^q : |i-j| > k\} \leq Ck^{-\alpha}, \forall k > 0, \forall i\right\}. \quad (2.9)$$

This condition prescribes that the further two variables are, the lower their correlation is. Matrices obeying this condition are "approximately bandable" matrices.

These assumptions are made for the nature of banding operator, which is defined for any matrix M as: $B_k(M) = [m_{ij}\mathbf{1}(|i-j| \leq k)]$. It is straightforward that the banding operator would be perfectly effective if

$$|i-j| > k \rightarrow \sigma^*_{ij} = 0.$$

Choosing $k = O((\frac{\log p}{n})^{\frac{1}{2(\alpha+1)}})$ the banding operator $B_k(\hat{\Sigma}_n)$ is shown to consistently estimate Σ^* with rate $O((\frac{\log p}{n})^{\frac{\alpha}{2(\alpha+1)}})$.

This approach can be indifferently applied to the covariance matrix or to the Cholesky factor of the inverse covariance matrix. In [20], minimax properties for the rates of convergence of covariance estimators having (2.8) as matrix reference class are provided both for operator (spectral) and Frobenius norms. There the authors show that the described approach achieves sub-optimal rates. Among other possible solutions, we mention tapering, which is gradually shrinking the off-diagonal elements to zero ([54]), and alternative uses of the Cholesky factor of the precision matrix ([114][66]).

When there is no natural ordering among variables, the banding approach becomes ineffective. This situation includes the vast majority of cases, including recent relevant applications to gene expression arrays. This is why the same authors (Bickel and Levina) developed in [15] a very elegant theory to make their previous work on banding methods applicable to this case. That approach is based on the thresholding of sample covariance matrices, where the hard thresholding operator is defined as $T_s(M) = m_{ij}\mathbf{1}(|m_{ij}| \geq s)$. $T_s(M)$ preserves preserve the positive definiteness of M if and only if $\lambda_{min}(M) > s$:

$$||T_s - T_0|| \leq s \longleftrightarrow \lambda_{min}(M) > s. \quad (2.10)$$

This happens because $v'T_s(M)v \geq v'Mv - s \geq \lambda_{min} - s$.

Note that the hard thresholding operator is implicitly based on the minimization of the l_0 norm of Σ^*, which is simply the number of non-null entries. This norm is not convex, and so it is hard to establish a unique minimum. This is why alternative thresholding operators have been developed. The most used, central to our discussion in following chapters, is the soft thresholding operator: $T_s(M) = sign(m_{ij})max(|m_{ij}| - s, 0)$. Note that the thresholding parameter s can be constant or entry-dependent, i.e. s_{ij}. Another relevant shrinkage operator is the adaptive one, where $s_{ij} = \tau(m_{ii}m_{jj})^{1/2}$

([18]). A generalized shrinkage function which encompasses the described ones was defined in [100].

Coming back to the covariance estimation problem, Bickel and Levina establish a contact point between the class of "thresholdable" and "bandable" matrices, in order to be able to exploit the results of [14].

They define for $0 \leq q < 1$ the uniformity class of matrices invariant under permutations:

$$\left\{\Sigma^* : \sigma^*_{ii} \leq M, \qquad \sum_{j=1}^{p} |\sigma^*_{ij}|^q \leq c_0(p), \qquad \forall i\right\}, \tag{2.11}$$

where $c_0(p)$ is a constant not depending on p.

Note that if $q = 0$, the condition becomes $\sum_{j=1}^{p} |\sigma^*_{ij}|^q = \sum_{i,j} \mathbf{1}(\sigma^*_{ij} \neq 0)$. Here we can consider M as a constant. In paragraph (5.1) we will relax this assumption.

In [15], the authors prove that, if a matrix Σ^* satisfies (2.11) for $q > \frac{1}{\alpha+1}$, which is equivalent to $1 - q > \frac{\alpha}{\alpha+1}$, then Σ^* satisfies also (2.9) and belongs to the class of approximately bandable matries (2.8).

We mention a technical result (in bold), which will be crucial for the discussion of our contributions in Chapter 5. The sample covariance matrix $\hat{\Sigma}_n$ satisfies the following property:

$$\max_{\mathbf{i,j}} |\hat{\sigma}_{\mathbf{ij}} - \sigma^*_{\mathbf{ij}}| = \mathbf{O}\left(\sqrt{\frac{\log \mathbf{p}}{\mathbf{n}}}\right). \tag{2.12}$$

under $\frac{\log p}{n} \to 0$.

As a consequence, under the condition $q > \frac{1}{\alpha+1}$ the loss of the thresholded matrix $T_s(\hat{\Sigma}_n)$ is bounded and vanishes asymptotically when $\frac{\log p}{n} \to 0$:

$$||T_s(\hat{\Sigma}_n) - \Sigma^*|| \leq O\left(\left(\frac{\log p}{n}\right)^{(1-q)/2}\right). \tag{2.13}$$

The banding and the thresholding methods are non-likelihood ones. The Frobenius norm as reference loss gives two advantages respect to a likelihood function. First, the Frobenius norm is the analogous for matrices of the l_2 norm for vectors. Second, Frobenius loss is model free, as the covariance matrix. These methods allow to ignore the underlying distribution for the data, which can be an advantage in high dimensions.

In addition, [80] and [19] describe two very effective non likelihood methods employing sparsity for precision matrix estimation in the multivariate Gaussian setting, where the likelihood is known. However, likelihood methods are still useful for the precision matrix especially, for their connection to graphical modelling (see [31]).

To sum up, sparsity models are useful tools to improve covariance estimation. In fact, in high dimensions we often have few pairs of variables showing a particularly large (idiosyncratic) covariance. On the other hand, a sparsity assumption may not be enough, especially in high dimensions, since the covariances are too large to be modelled by a purely sparse matrix, for the reasons outlined in paragraph (2.2) and because our target is probably not sparse. This is why factor-analysis and PCA based covariance estimators play a relevant role, for their ability to significantly reduce the problem dimension, as we are about to describe.

2.5 Factor analysis based estimator

This paragraph is devoted to the analysis and description of the factor model approach to covariance matrix estimation. This topic assumes a particular relevance in a large dimensional context, when the dimension p is very large, because p/n may be difficult to keep negligible, as enough n could not be available.

The first who defined the concept of factor model was Spearman (1904) ([102]), in a psycometric study about the measurement of intelligence. The main problem was: "how to explain most of the variance of a set of correlated variables by approximating them with a smaller set of uncorrelated variables?" In this specification, the covariance matrix resulted in the sum of a lower ranked matrix and a diagonal residual matrix, where all the covariances are explained by the factors, while the presence of the error term implies that there are residual variances unexplained by the factors.

A general factor model setting for Σ^* can be described as follows:

$$\Sigma^* = L^* + S^*. \tag{2.14}$$

We can write $L^* = BB'$, with $B = UD^{1/2}$, where U is a $p \times r$ matrix, D is a $r \times r$ diagonal matrix $d_{jj} > 0, \forall j = 1, \dots, r$, $r \ll p$.

A generalized static factor model for a $p-$dimensional vector x_i, $i = 1, \dots, n$, is the following:

$$\begin{aligned} x_i = Bf_i + \epsilon_i = l_i + \epsilon_i, \\ E(f) = 0, V(f) = I_r; \\ E(\epsilon) = 0, V(\epsilon) = S^*; \\ E(\epsilon' f) = \mathbf{0}. \end{aligned}$$

where f_i is a $r \times 1$ vector, and x_i, l_i, ϵ_i are $p \times 1$ vectors.

In this framework, $\hat{\Sigma}_n$ is the $p \times p$ sample covariance matrix computed on the n generated data. For $i = 1, \dots, n$, $l_i = Bf_i$ is called **common** part of x_i, ϵ_i is called **idiosyncratic** part.

Note that L^* has rank r and is positive semidefinite, while S^* and Σ^* are full rank and positive definite.

The reason why a factor model assumption for the data is effective in this context is two-fold:

- model (2.14) prescribes for the covariance matrix a decomposition into a $r-$ranked matrix ($r \ll p$) and a full rank residual matrix. Specifying a low rank matrix means reconditioning the eigenvalues, since we replace a p-ranked probably ill-conditioned matrix (see section (2.2)) with a semidefinite positive $r-$ ranked matrix, well conditioned by definition. At the same time, the full rank rank residual component ensures that the estimate is positive definite. So, imposing this structure to a large covariance matrix means reconditioning its eigenvalues, not using the identity matrix as a shrinkage target (as in [75]), but removing the strongest correlations from the raw (sample) estimate, thus shrinking down its condition number.

- model (2.14) significantly reduces the number of parameters, by replacing $p(p+1)/2$ parameters with $p(r+1)$ only. This approach overcomes the problem of identifiability in the large dimensional context, by relevantly reducing the parameter space dimension.

Anyway, model (2.14) is the most general definition. Different model settings impose different assumptions on L^* and S^*. Key assumptions for our purpose, which is to exploit effectively a factor model structure for covariance matrix estimation, mainly concern the eigenvalues of L^*, which reflect upon the eigenvalues of Σ^*.

We are going to briefly recall the historical path of factor modelling by the description of three main steps (for an extended overview, see [59]):

- the classical factor model, with p fixed, $n \to \infty$. This specification was due to [102], and its development was then possible thanks to the pioneeristic work on Principal Component Analysis by Hotelling [65]. Its main characteristic is the imposition of a diagonal structure to the residual covariance matrix S^* (paragraph (2.5.1)).

- the approximate factor model, where nonzero residual correlation is allowed, i.e. S^* is no longer diagonal. This advance was due to Chamberlain and Rothschild ([29]), and is based on the assumption of limitedness for λ_{r+1} (the $(r+1)-$th eigenvalue of Σ^*) as n goes to ∞ (p here is still fixed). This approach allowed to effectively use PCA for recovering Σ^* (paragraph (2.5.3)).

- factor modeling with sparse residual ([45]), where specific assumptions on L^* and S^* are made. The eigenvalues of L^* are assumed to be pervasive while p as well as n) tends to ∞ (spikiness property). On

the contrary, the largest eigenvalue of S^* vanishes asymptotically. This approach impacts on the covariance matrix estimate allowing to reduce even more the parameter space dimension, still employing the PCA of $\hat{\Sigma}_n$ together with a thresholding algorithm for the sparse component (paragraph (2.5.4)).

2.5.1 Strict factor model

We are now going to explore this first specification, which is called strict or classical factor model. In this specification, we have that

$$X = Bf + \epsilon, \tag{2.15}$$

where X and ϵ are $p \times 1$ random vectors, B is a p times r matrix also called loading matrix, f is the $r \times 1$ random vector of latent variables.

Under all previously outlined assumptions, $E(X) = 0$. Defining $\Sigma^* = E(XX')$, this model leads to the following model on the covariance matrix:

$$\Sigma^* = L^* + S^* \tag{2.16}$$

with $L^* = BB'$. The identifiability condition imposes $B'S^{*-1}B$ to be diagonal. It is necessary because the strict factor model is equivariant under orthogonal transforms, and this results in an identifiability issue. Note that the condition $E(f\epsilon') = B$ holds. Bf is the common part, while ϵ is the idiosyncratic (or unique, or specific) part of the model.

For each component X_i,$i = 1, \dots, p$, $Var(X_i)$ can be disentangled in two components. $\frac{\sum_j B_{ij}^2}{\Sigma_{ii}^*}$ is the portion of variance of X_i, $i = 1, \dots, p$ explained by the factors. It is also called **communality** of X_i. $\frac{S_{ii}}{\Sigma_{ii}^*}$ is the portion of variance of X_i unexplained by the factors. It is also called **idiosyncratic** component of X_i.

The ratio between communality and total variance for each X_i, $i = 1, \dots, p$ is very important for the interpretation of factor models (FM), as well as, if S^* is not diagonal, the ratio between the sum of residual covariances and the total sum of covariances. The proportion of variance explained by the model describes the goodness of fit and the covariance matrix between the factors and the observed variables, equal to B, outlines the most relevant variables in the composition of factors.

As explained, if we impose S^* diagonal we impose all the covariances to be explained by the factors. This assumption is clearly inappropriate in a large dimensional context. Specifying a pure factor model structure is there quite far from being effective. We have already explained that if p is large the sample covariance matrix is likely to be bad-conditioned. For this reason, it is likely that factors are not enough to explain covariances, and that the diagonal assumption for the residual covariance matrix is too strict. For an overview of factor analysis in large dimensions, see [7].

FM estimation has been a relevant problem in the literature. It is well known that factor analysis moves out from principal component analysis (PCA), but PCA without further assumptions is not a consistent estimator for the factor model, as we are going to explain.

2.5.2 PCA and factor analysis

Let us A be a $p \times p$ matrix, with $||A||_{Fro} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^2$. Its spectral decomposition is

$$A = U\Lambda U' = \sum_{i=1}^{p} \lambda_i u_i' u_i = \sum_{i=1}^{p} (\sqrt{\lambda_i} u_i')(\sqrt{\lambda_i} u_i),$$

where $\sqrt{\lambda_i} u_i$, $i = 1, \ldots, p$ are the principal directions, ordered respect to the magnitude of associated eigenvalues. The first to address PCA was Pearson (1901) ([92]), and the idea was then refined by Hotelling (1933) ([65]). They found out that the best approximation property is possessed by principal components, that is, the linear combinations of observed variables which maximize the explained variance are subsequently the first, the second, ..., the last principal component. In formula,

$$\min_{Z, rank(Z) \leq r} ||X - Z||_{Fro}, Z = AX \longleftrightarrow Z = PCA_r(X),$$

where $PCA_r(X)$ is the (2.1) truncated to the $r-$th eigenvalue.

The underlying approximation problem comes from linear algebra. If

$$z_i = u_{i1} F_1 + u_{i2} F_2 + \ldots + u_{ir} F_r,$$

with $F = [F_1, \ldots, F_r]'$, $E(F) = \mathbf{0}$, $V(F) = I_r$, $r \ll p$, we can write:

$$\min_{u_i, z_i} \frac{1}{n} \sum_{i=1}^{n} ||x_i - z_i||_2 = \frac{1}{n} ||X - Z||_{Fro}^2 =$$

$$= \frac{1}{n} ||X' - U'F||_{Fro}^2 = \frac{1}{n} ||X - F'U||_{Fro}^2,$$

where X is our $n \times p$ data matrix, $U = [u_{.1} \ldots u_{.r}]$ is a $r \times p$ matrix and $F = [F_{.1} \ldots F_{.r}]$ is $r \times n$. If we post-multiply all terms by X', we obtain $\frac{1}{n} \min_{F,U} ||X'X - X'F'U||_{Fro}^2$, which can also be viewed as $\min_{F,U} ||\hat{\Sigma}_n - X'FU||_{Fro}^2$.

As we can understand from one of the expressions above, since orthogonal projections have the best approximation property, $||X - F'U||_{Fro}^2$ is minimum if $F'U$ is the principal component set of X truncated to the $r-$th one. Under the condition $r = p$, $Z = X$. Since X and $X'X$ have the same column (and row) spaces, the same holds also using the first r PCs of $\hat{\Sigma}_n$. This is why if

we want to approximate $\hat{\Sigma}_n$ with a $r < p$ matrix, the first solution we think about is the extraction of its principal components up to the $r-$th.

Unfortunately, the approximation problem in the FM setting is different from the PCA one, because in the factor model setting there are also relevant issues concerning identifiability and estimation. In fact, we immediately encounter relevant problems using this method to estimate strict factor models (SFM), because we would have $\hat{S} = \hat{\Sigma}_n - \sum_{j=1}^{r} \hat{\lambda}_i \hat{u}_i \hat{u}_i'$ which cannot estimate S^* since it is exactly the sum of residual principal components (from $r+1-$th to $p-$th), and so will never be diagonal. This is coherent with the fact that PCA subsequently maximizes the **variance** explained by the factors, and not the covariances. Therefore, without further assumptions, extracting r of p components means that the residual matrix will be non-diagonal, and so that our SFM estimator will be inconsistent ([5]).

For this reason, lots of factor model specifications and estimation methods have been proposed. Some methods using iteratively PCA for FM estimation, like the principal factors method, have been developed. Unfortunately, they require an a priori choice of the number of factors to be included in the model, and they usually are very inefficient for large scale problems. In addition, the principal factors method is not scale-equivariant, that is, it is not equivariant under linear transforms of the data. As an alternative, Maximum Likelihood methods can be used, requiring the assumption of multivariate normal distribution for the data.

Hence, a natural question arises: how can we establish an asymptotic convergence between PCA and factor analysis (FA)? Which assumptions are needed? Identifying a factor model structure via PCA requires specific assumptions on the eigenvalues of Σ^*, which can be imposed as a result of appropriate assumptions on L^* and S^* .

2.5.3 Approximate factor model

The above mentioned problem problem was first faced by Chamberlain & Rothschild in [29]. They were the first to define an approximate factor structure, i.e. a structure where the residual matrix is allowed to be non-diagonal. Model (2.14) with this assumption is called approximate factor model. In this context, the key condition is a bound on the $(r+1)-$th eigenvalue of matrix Σ^*, which results in a bound for the largest eigenvalue of S^*. This condition is necessary to establish the asymptotic equivalence between PCA and FA. Therefore, the two main points discussed so far, i.e. the need to over come the diagonal structure of S^* and the need of estimating consistently a factor model via a standard method as PCA, can find a common solution.

This theory was born in the field of portfolio pricing theory. When S^* is diagonal, model (2.14) is a **strict** factor model (SFM) structure. Ross ([99]) derived the SFM structure in the context of capital asset pricing. He showed that if Σ^* is a covariance matrix referred to asset prices and has such

a structure, the mean expected return is linear (i.e. a linear combination of factors) because of the absence of arbitrage opportunities, that is, $E(\epsilon) = 0$. He proved that the SFM structure can be asymptotically recovered when $n \to \infty$ with bounded error by Principal Components. However, if we impose a diagonal structure for S^*, the number of factors needed to ensure that S^* is diagonal would increase too much when $n \to \infty$.

Suppose Σ_n^* is a sequence of matrices for $n \to \infty$. If Σ^* is positive semi-definite and $\sup_n \lambda_{\Sigma_n^*, r+1}$ (the $(r+1)-$th eigenvalue of Σ_n^*) is finite, we refer to (2.14) across n as an **approximate** factor model (AFM) structure.

Chamberlain and Rothschild proved in [29] that the main characterization of the approximate factor structure needed to perform FM estimation via PCA is:

$$\sup_n \lambda_{\Sigma_n^*, r+1} \qquad \text{finite},$$

i.e. r of p eigenvalues of Σ^* diverge when $n \to \infty$. This result means that under these assumptions the error between the PCA truncated to the $r-$ th component and the theoretical mean (the deterministic part of the model) is asymptotically bounded by $\lambda_{\Sigma_n^*, r+1}$. The proof exploits these assumptions and the properties of the matrix $B'B + I$.

The outlined assumption works as an identification condition for the approximate factor model: the authors showed that this condition is sufficient for the existence of an approximate factor model structure. More, they showed that the approximate factor structure is uniquely identified extracting the top r principal components of Σ_n, and that the error is bounded by a function of $\lambda_{\Sigma_n^*, r+1}$ (and a parameter controlling the trade-off between mean and variance of the process).

This pioneeristic work opened the path for a wide literature on FM estimation exploiting PCA as an asymptotic estimator. It is an asymptotic approach where $n \to \infty$, differently from the following ones (as the POET approach), where p varies together with n. We also highlight that a similar condition to the sufficient condition here reported is essential to the estimation of dynamic factor models, as explained in [52].

For sake of completeness we mention two other famous factor model specifications in the economic context: the three factor model by Fama and French ([42]) and the approximate dynamic factor model by Stock and Watson ([105]) (used for economic forecasts).

By the way, the work by Chamberlain and Rothschild allows for the presence of residual covariances, but does not specify any structure for the matrix S^*. As explained, in large dimensional real data analysis the assumption of diagonal residual matrix is not acceptable. The data generating process becomes so complex that assuming no idiosyncratic correlation among variables is very unrealistic. However, setting specific assumptions on the residual component, defining a particular structure, has become a central topic in the recent statistical literature. This is why the concept of sparsity

for the residual matrix (i.e. the presence of non-zero elements in selected positions) came out.

At the same time, the number of parameters becomes so large that identifiability issues arise, especially when n is not so large. Allowing for the presence of non-zero residual covariance, preserving model identifiability, is one of the major challenges in this field, as we will study in deep in Chapter 4.

Sparsity assumptions are motivated by the following two reasons:

- a strong interpretability issue supports this approach. Factor model approach finds out a small set of uncorrelated latent (unobserved) variables able to explain most of the correlations among a large set of observed variables. It means that, by removing the correlations due to some common explaining factors, we are able to identify those pairs of variables which are so correlated that their residual covariance is still non-zero. This can be particularly helpful in a few application contexts, such as hypothesis testing, portfolio analysis, and macroeconomic analysis. We are thus able to identify also block-wise correlation structures. The sparsity pattern of the covariance matrix becomes a key to data interpretation, as well as the covariance between variables and factors, in order to understand the nature of variables and their relationship.

- an identifiability issue. The number of parameters is now $p(r+1)+s$, $s \ll p(p+1)/2$, which is still pretty fewer than $p(p+1)/2$, allowing a more flexible interpretation and a better conditioning (a sparse estimate is better conditioned than the sample covariance matrix, since it is further from collinearity).

However, exploring conditions ensuring identification of FM with specific sparsity assumptions on the residual component is a really hard task.

2.5.4 POET estimator

We are now going to describe a very recent contribution to covariance matrix estimation. Fan, Liao and Micheva in their paper ([45]) provide in the time series setting a large covariance matrix estimator which plays a central role for our dissertation. Their estimator, called POET (Principal Orthogonal complEment Thresholding estimator), is a PCA-based estimator, but it also has the characteristics of a sparsity-based estimator. The underlying model assumptions prescribe an approximate factor model for the data, thus allowing to reasonably use the truncated PCA of the sample covariance matrix. Furthermore, at the same time, they impose sparsity in the sense of [15] (see paragraph (2.4)) to the residual matrix.

If we refer to (2.14), S^* is approximately sparse in the sense of (2.9), while L^* has a small number of very spiked eigenvalues, growing with p at rate $O(p)$, and the rest of eigenvalues are asymptotically negligible. This feature, i.e. the pervasiveness of a few spiked eigenvalues, is the distinctive trait of their model, which allows to consistently recover L^* via PCA. At the same time, they recover the sparse component imposing a bound on the approximate sparsity parameter (2.11), which allows them to recover S^* applying a thresholding algorithm to the orthogonal complement of the truncated PCA.

Deriving the performance of the most recent numerical estimator we will describe in Chapter 4 under the outlined conditions of POET estimator, comparing both performances, is one of the main goals of our thesis. A related one is the attempt to relax in some way the assumption of spikiness for the eigenvalues of L^*, developing an appropriate estimator.

We immediately outline that rank choice in this context is a relevant issue, which is typical for rank minimization programs, like PCA. Rank minimization allows to improve conditioning, reduce the number of parameters and compress information, thus improving interpretability, which is crucial in high dimensions. However, we know that covariance estimators based on pure rank minimization suffer from rank deficiency (see for example [119] and [11]). What is more, rank is a non-convex function, and this causes the impossibility to give any mathematical guarantee for model recovery. In POET setting, the authors select the latent rank of truncated PCA using standard criteria from Bai and Ng (2002) ([6]). We will show in our simulations (Chapter 5) that POET can suffer from rank deficiency in high dimensions. Another relevant application exploiting PCA structure is [71], where the authors impose the presence of one leading principal component and select a subset of variables by a method called sparse PCA. Recovery is performed given that $\frac{p_n}{n} \to 0$, but p_n can be much larger than n. Even if this model is effective for some time series data (like ECG data), imposing the latent rank equal to 1 is not usually appropriate.

We now describe in detail the model setting of POET, keeping model structure (2.14) in mind. Here we will use T instead of n, since we are in a time series model setting.

We report the two main features of POET setting. The spectral decomposition of Σ^* (positive definite symmetric squared p-dimensional matrix) is $U\Lambda U'$. The columns of U and B (both $p \times r$ matrices) are denoted by u_j and $\tilde{b}_j$, $j = 1, \dots, r$, respectively.

Proposition 2.5.1 ([45] 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 (canonical condition of SFM) we have:*

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

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

This proposition prescribes that the eigenvalues of the low rank component L^* (equal to BB') are pervasive, i.e. they grow at rate $O(p)$ while $p \to \infty$. This entails that the top r eigenvalues of Σ^* are pervasive, while the remaining $p - r$ asymptotically vanish. The largest eigenvalue of S^* is the relevant bound for the top r eigenvalues of Σ^* minus the corresponding ones of L^* as well as for the remaining $p - r$ eigenvalues of Σ^* . Note that in the setting of AFM ([29]), differently from here, p is fixed.

Proposition 2.5.2 ([45] Proposition 2)**.** *Under the assumptions of Proposition 1, if* $||\tilde{b}_j||_{j=1}^r$ *are distinct, then* $||u_j - \tilde{b}_j/||\tilde{b}_j|||| = O(p^{-1}||S^*||)$.

This proposition states that if the columns of B are distinct, the distance between the top r eigenvectors of Σ^* and the normalized eigenvectors of L^* are bounded by a rate proportional to $p^{-1}||S^*||$.

Proposition 1 and 2 together state that matrix U and matrix B are approximately the same if $||S^*|| = o(p)$.

Now, the thresholding estimator by Bickel and Levina ([15]) described in (2.4) comes into play. The outlined bound is ensured imposing an approximate sparse structure on S^*. Sparsity parameter (2.11) is defined for some $q \in [0, 1]$ as follows:

$$m_p = \max_{i\leq p} \sum_{j\leq p} |\sigma_{ij}|^q. \tag{2.17}$$

For standard properties of matrix norms, we have:

$$||S^*|| \leq ||S^*||_1 \leq \max_i \sum_{i=1}^{p} |s_{ij}|^q (s_{ii}s_{jj})^{(\frac{1-q}{2})} = O(m_p), \tag{2.18}$$

given that s_{ii} are bounded $\forall i$. So, $||S^*|| \leq O(m_p)$.

It is now clear that if $m_p = o(p)$, the PCA of $\hat{\Sigma}_n$ allows to perfectly identify the eigenvalues and the eigenvectors of Σ^* under these assumptions. In particular, the first r principal components of Σ^* are approximately the same as the factor loadings. We emphasize the relevance of this point, which represents one of the most important innovations in [45]. Here the asymptotic equivalence between PCA and factor analysis is established by applying a conditional (to factors) sparsity model to the residual matrix, provided that p is enough large. The assumption $m_p = o(p)$ will be modified in order to study the case of generalized spiked eigenvalues.

The key point in their proof is that under these assumptions the eigenvalues of $B'\Sigma^{-1}B$ are bounded. Thus, the relative norm of $||\hat{\Sigma} - \Sigma||$, defined as $||\hat{\Sigma} - \Sigma||_\Sigma = p^{-1/2}||\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_p||_{Fro}$, is bounded, cancelling out the curse of high dimensionality introduced by B (see paragraph (2.2), Theorem 2.2.1).

As in [15], the sparse component S^* is then consistently estimated by thresholding, relying on the results described in section (2.4). They define for each $i \neq j$ an adaptive threshold ([18]) of the form

$$\tau_{ij} = C\omega_T\sqrt{\theta_{ij}},$$

where

$$\omega_T = \frac{1}{\sqrt{p}} + \sqrt{\frac{\log(p)}{T}}$$

and

$$\hat{\theta}_{ij} = \frac{1}{T}\sum_{t=1}^{T}(\hat{s}_{it}\hat{s}_{jt} - \hat{s}_{ij})^2,$$

with

$$\hat{s}_{ij} = \frac{1}{T}\sum_{t=1}^{T}\hat{s}_{it}\hat{s}_{jt}$$

and

$$\hat{s}_{it} = x_{it} - \hat{l}_{it},$$

where $\hat{l}_{it} = \hat{b}_i^r \hat{f}_t^r$ is estimated via the PCA of $\hat{\Sigma}_n$ up to the $r-$th component.

This approach holds for sufficiently large $C > 0$. ω_T (which is the uniform rate of convergence of $\max_{i\leq p, j\leq p}|\hat{s}_{ij} - s_{ij}|$, as in [15] and [18]) is a decreasing sequence in p and T. Note that term $\frac{1}{\sqrt{p}}$ is due to the estimation of the unknown factors and is usually unavoidable.

Any generalized thresholding function $h(z)$ (including the soft-thresholding operator) such that $h_{ij}(z) = 0$ when $|z| \leq \tau_{ij}$ and $|h_{ij}(z) - z| \leq \tau_{ij}$ otherwise (see ([3])) can be effectively used. Note that thresholding is applied only on the off-diagonal elements. The thresholded estimate of the residual matrix S^* is thus $\hat{S}_{\hat{r}}^{\mathbf{T}} = h_{ij}(\hat{s}_{ij})$.

The sequential approach to compute POET estimator is the following. First, perform PCA on $\hat{\Sigma}_n$, extracting the top r components (eigenvalues and eigenvectors). So, $\hat{L}_{\hat{r}} = U_r\Lambda_r U_r'$, where Λ_r is a $r \times r$ diagonal matrix containing the top r eigenvalues of $\hat{\Sigma}_n$, and U_r is the $p \times r$ matrix containing the associated eigenvectors. $\hat{l}_{it} = \hat{b}_i^r \hat{f}_t^r$ is thus simply the $i \times t$ entry of $\hat{L}$. S^* is estimated by applying as described an adaptive thresholding step on the matrix $\hat{S} = U_{p-r}\Lambda_{p-r}U_{p-r}'$ (the principal orthogonal complement of $\hat{\Sigma_n}$), where Λ_{p-r} contains the remaining $p-r$ eigenvalues, and U_{p-r} the associated eigenvectors. This is why POET contains in its name the thresholding of the principal orthogonal complement. Here is the expression of POET:

$$\hat{\Sigma}_{POET,\hat{r}} = \hat{L}_{\hat{r}} + \hat{S}_{\hat{r}}^{\mathbf{T}}.$$

As pointed out in the introduction to this paragraph, the rank choice is a relevant issue. The number of diverging eigenvalues, i.e. the latent rank r is

determined in a data-driven way minimizing appropriate penalty functions which were first described in [6]. These functions of p and T must satisfy the following conditions: $g(T,p) = o(1)$ and $\min_{p,t} g(T,p) \to \infty$. In this way, POET is estimated with a data-driven rank $\hat{r}$. We refer to [45], paragraph (2.4), for the details.

POET is a non-parametric estimator. At the same time, it requires some distributional assumptions to perform consistent recovery. We now list for sake of completeness the most relevant technical assumptions on factors and residuals:

1. Strictly Stationarity of $(\epsilon_t, f_t)_{t\geq 1}$.
2. Non-correlation between ϵ_t and f_t, $\lambda_{min}(S^*) > c_1$, $||S^*||_1 < c_2$, $\min var(\epsilon_{it}\epsilon_{jt}) > c_1$.
3. Tails of f_t and ϵ_t:

$$P(|\epsilon_{it}| > s) \leq \exp(-s/b_1)^{r_1}, \qquad i \leq n$$

$$P(|f_{jt}| > s) \leq \exp(-s/b_2)^{r_2}, \qquad j \leq r.$$

We note that bounds on the minimum eigenvalue and the l_1 norm of S^* are needed. Further assumptions include strong mixing between the sigma-algebras generated by $[(f_t, \epsilon_t) : t \leq 0]$ and $[(f_t, \epsilon_t) : t \geq T]$ and some regularity conditions to estimate loadings and factor scores.

Most of these assumptions will not be necessary in our numerical context. Anyway, we will use part of them to study the behaviour of our numerical estimator in the POET context. Part of the technical conditions were derived in a previous paper by Fan, Fan and Lv ([44]). There, the authors analyze the same setting deriving the correspondence between PCA and factor analysis without thresholding the residual component. Another paper by Fan, Fan and Lv ([43]) studied the same setting but with observable factors.

The two main theorems of [45] state that, under all described assumptions and supposing $\gamma^{-1} = 3r_1^{-1} + 1.5r_2^{-1} + r_3^{-1} + 1$, $log(p) = o(T^{\gamma/6})$ and $T = o(p^2)$, we have:

$$\begin{aligned} ||\hat{S}_{\hat{r}}^{\mathbf{T}} - S^*|| &= O_p(\omega_T^{1-q} m_p) \\ ||\hat{\Sigma}_{POET,\hat{r}} - \Sigma^*||_\Sigma &= O_p\left(\frac{\sqrt{p}\log p}{T} + m_p \omega_T^{1-q}\right) \\ ||\hat{\Sigma}_{POET,\hat{r}} - \Sigma^*||_{max} &= O_p(\omega_T) \end{aligned} \tag{2.19}$$

If $m_p w_T^{1-q} = o(1)$, $\hat{S}_{\hat{r}}^{\mathbf{T}}$ and $\hat{\Sigma}_{POET,\hat{r}}$ are non-singular with probability approaching 1:

$$\begin{aligned} ||\hat{S}_{\hat{r}}^{\mathbf{T}-1} - S^{*-1}|| &= O_p(\omega_T^{1-q} m_p) \\ ||\hat{\Sigma}_{POET,\hat{r}}^{-1} - \Sigma^{*-1}||_{=} &O_p(\omega_T^{1-q} m_p) \end{aligned}$$

The assumption $T = o(p^2)$ is necessary to estimate the rT factor loadings. It means that recovery is effective until $p \log p \gg T$. The assumption $log(p) = o(T^{\gamma/6})$ is necessary for recovering the sparse component.

For the following of our dissertation we report two technical results of [45] describing model-based relationships (in bold). The first one, which was proved in [44], prescribes, under all described assumptions, that the following claims hold:

$$\max_{\mathbf{i,j \le r}} \left| \frac{\mathbf{1}}{\mathbf{T}} \sum_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{f_{it} f_{jt}} - \mathbf{E}(\mathbf{f_{it} f_{jt}}) \right| = \mathbf{O_p}\left(\frac{\mathbf{1}}{\sqrt{\mathbf{T}}} \right) \tag{2.20}$$

$$\max_{\mathbf{i,j \le r}} \left| \frac{\mathbf{1}}{\mathbf{T}} \sum_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{s_{it} s_{jt}} - \mathbf{E}(\mathbf{s_{it} s_{jt}}) \right| = \mathbf{O_p}\left(\frac{\log \mathbf{p}}{\sqrt{\mathbf{T}}} \right) \tag{2.21}$$

$$\max_{\mathbf{i,j \le r}} \left| \frac{\mathbf{1}}{\mathbf{T}} \sum_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{f_{it} s_{jt}} \right| = \mathbf{O_p}\left(\frac{\log \mathbf{p}}{\sqrt{\mathbf{T}}} \right). \tag{2.22}$$

Thanks to this result, it is possible to prove that, under all described assumptions, $||\hat{\Sigma}_n - \Sigma^*|| = o(p)$ with a rate proportional to $O(\frac{p}{\sqrt{T}})$, i.e. the $r-$th largest eigenvalue of $\hat{\Sigma}_n$ grows at rate $O(p)$ with probability approaching 1:

$$||\hat{\mathbf{\Sigma}}_\mathbf{n} - \mathbf{\Sigma}^*|| = \mathbf{O}\left(\frac{\mathbf{p}}{\sqrt{\mathbf{T}}} \right). \tag{2.23}$$

For the following of our study, we here define the generalized pervasiveness context for $\alpha \in (0, 1]$ as follows ([45], p. 656):

Definition 2.5.1. *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 into the POET setting.

Applications of POET are very wide. We explicitly mention applications on financial data. In Chapter 5, we will show an application to banking supervisory data where the performance of our numerical estimator will be compared to the one of POET.

We shall use repeatedly these results on the sample covariance matrix for proving the rates of our numerical estimator under POET assumptions and in the generalized spikiness context. Non-asymptotic large covariance matrix recovery under generalized assumptions for the eigenvalues of the low rank matrix is one of the goals of the rest of our thesis. In fact, POET approach is elegant and effective, but spikiness in real applications is not so usual. What is more, in this way it is difficult to catch the proportion of variance explained by the factors, since the model does not provide any attention to that. In addition, when p is not enough large, the errors could be still correlated (as pointed in the discussion of [45] by Montanari).

To conclude, we note that rank selection also 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. Using IC criteria from Bai and Ng (2002), as pointed out in the discussion of [45] by Yu and Samworth, if the eigenvalues are not really spiked at rate $O(p)$, the probability to underestimate the latent rank does not asymptotically vanish. This is why we are going to derive a method intrinsically detecting the latent rank, without applying any existing criterion. We are going to do that in the non-asymptotic context, where the absolute losses are bounded given finite values for relevant parameters.

Chapter 3

Covariance regularization and convex analysis: numerical and computational aspects

The aim of the present chapter is to explain the rationale behind the numerical methods needed to estimate the covariance matrix under the assumption of approximate factor model with sparse residual for the data.

Such a data structure has become very popular in recent years and has found relevant applications in various fields like, among others, image reconstruction, MRI (Magnetic Resonance Imaging) data, financial portfolio selection and electrical engineering. The statistical challenge lies in the need to estimate a latent structure summarizing a huge number of variables, even starting from a number of observations comparable or smaller.

Let us suppose the population covariance matrix of our data is the sum of a low rank and a sparse component. Suppose we have a data matrix $X = [x_{ij}]$, where $i = 1, \ldots, n$ are the observations and $j = 1, \ldots, p$ are the variables. The $p-$ dimensional random vector x has a **low rank plus sparse structure** if its covariance matrix Σ^* satisfies the following relationship:

$$\Sigma^* = L^* + S^*, \tag{3.1}$$

where:

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

Let us suppose $L^* = UDU' = BB'$, where $B = UD^{1/2}$, U is a $p \times r$ matrix, D is a $r \times r$ diagonal matrix, with $d_{jj} > 0$, $\forall j = 1, \ldots, r$. Suppose that our $p \times 1$ random vector X_i, $i = 1, \ldots, n$, has the following structure:

$$X_i = Bf_i + \epsilon_i, \tag{3.2}$$

with

$$f_i = N_r(0, I_r); \tag{3.3}$$
$$\epsilon_i = N_p(0, S^*), \tag{3.4}$$

where f_i is a $r \times 1$ random vector, and ϵ_i is $p \times 1$ random vector.

X_i is assumed to be a zero mean random vector, without loss of generality. $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' = \frac{1}{n} X'X$ is the $p \times p$ sample covariance matrix, where X is the $n \times p$ data matrix.

If we set $x = X_i$, it is easy to observe that x follows a low rank plus structure:

$$\begin{aligned} E(xx') = E((Bf + \epsilon)(Bf + \epsilon)') = \\ = E(B'f'fB) + E(Bf\epsilon') + E(\epsilon B'f') + E(\epsilon\epsilon') = \\ = BB' + S^* = \Sigma^* \end{aligned} \tag{3.5}$$

under the usual assumption $f \perp \epsilon$, i.e. $cov(f, \epsilon) = E(f\epsilon') = E(\epsilon f') = \mathbf{0}$ ($r \times p$ null matrix).

If we assume a normal distribution for f and ϵ, we know that the matrix $W := \hat{\Sigma}_n - (BB' + S^*)$ is a re-centered Wishart noise, i.e. it is distributed as a zero-mean Wishart (refer to Chapter 2 paragraph (2.1) for detailed explanations on the Wishart distribution). However, the normality assumption is not essential in the finite sample context.

The main aim of this Chapter and of the entire work is to provide an alternative approach to covariance matrix estimation respect to POET under a similar data structure, deriving the necessary assumptions to perform identifiability and recovery. This approach is based on numerical analysis, and exploits the theory of non-smooth convex optimization provided by [98] and [28].

As suggested by the data structure, the method we are going to describe should at the same time consistently estimate the covariance matrix and catch sparsity and spikiness in the best possible way. The starting point for our study is offered by numerical analysis, which summarizes the problem of our interest in a natural way. As discussed in the previous chapter, this approach has several advantages, like a better conditioning (for the presence of the low rank component), a smaller number of parameters ($pr + s$ against $\frac{p(p-1)}{2}$), a better interpretability of the output, both on the low rank side (degree of covariance explained by the factors) and on the sparse side (the sparsity pattern maps the most relevant relationships among variables).

However, even if the numerical problem can be efficiently solved by using subgradient techniques, it is not straightforward to investigate the statistical properties of these estimators. Non standard tools of algebraic geometry

([60]) are required to derive identifiability conditions, as well as relevant results of random matrix theory are necessary to establish consistency ([39]). It is relevant that the statistical performance in terms of the covariance matrix as a whole and in terms of the two components (rank and sparsity pattern) separately considered are not necessarily aligned. As we will see, the loss function here depends on the Lagrangian dual theory of non-smooth function, thus implying that the loss function of the two components (sparse and low rank) separately considered is referred to the sum (i.e. the estimated covariance matrix), thus differing from the usual (Frobenius) loss of the estimated covariance matrix.

Our problem can essentially be stated as

$$\min_{L,S} \frac{1}{2}||(L+S) - \Sigma_n||^2_{Fro} + \lambda rank(L) + \rho||S||_0, \tag{3.6}$$

where $||S||_0$ is the number of nonzero elements, and $rank(L)$ can be seen as $||diag(D)||_0$, given that $L^* = UDU'$. This is a combinatorial problem, which is known to be NP-hard, since both $rank(L)$ and $||S||_0$ are not convex. A very well known convex relaxation of problem (3.6) is

$$\min_{L,S} \frac{1}{2}||(L+S) - \Sigma_n||^2_{Fro} + \lambda||L||_* + \rho||S||_1, \tag{3.7}$$

where λ and ρ are non-negative **threshold** parameters. $||S||_1 = \sum_{i=1}^{n}\sum_{j=1}^{n}|s_{ij}|$ is the l_1 norm of S, while $||L||_* = \sum_{i=1}^{r}|d_i| = \sum_{i=1}^{r} d_i = ||diag(D)||_1$ is the nuclear norm of L^*. Basic references are [108] for the former and [46] for the latter.

More in detail, the study and implementation in statistics of the nuclear norm l_* is due to [49]. Problem (3.7) is a penalized least squares program, where the penalty is composite and non-smooth. For the reasons explained before, problem (3.7) is also often referred to as a regularization problem. From a numerical point of view, it is an approximate unconstrained inverse matrix problem with two unknowns, L and S. The key to its solution will be to disentangle the problem in two easier related problems, one in L and the other in S. We will deal with the constrained version of (3.7), imposing that S and $L+S$ are positive definite, and L is positive semidefinite.

In this Chapter, we are going to describe the genesis of problem (3.6), showing how the l_1 and l_* heuristics came out. [36] proved that for most underdetermined systems the l_1 norm detects the sparsest solution, while [97] proved that the nuclear norm solution is the one with minimum guaranteed rank. In section (3.1) the rationale behind both problems is analyzed from the numerical point of view. In section (3.2) the computational aspects related to solving problem (3.7) are shown.

3.1 Nuclear norm and l_1 norm regularization: an historical review

In this section we are going to describe the numerical approach to covariance matrix estimation. The key argument for this approach rises from the need of regularizing the covariance matrix. Respect to the PCA based approach of [45], this alternative provides a way to numerically estimate the two components and their sum, without imposing the pervasive condition on the eigenvalues of L^* (and Σ^*). The other main issue of POET approach is that the rank is chosen according to some information criteria, while we would like an approach automatically detecting BOTH the low rank and the sparsity pattern.

Combinatorial problem (3.6) is the most natural way to formalize this search. However, (3.6) is computationally intractable, and can be approached replacing the composite non convex penalty $\lambda rank(L)+\rho||S||_0$ with the composite non smooth penalty $\lambda||L||_* + \rho||S||_1$. We can say that the numerical approach here essentially consists in model selection via convex optimization, where convexity is needed to achieve a unique minimum. The statistical properties of estimates will be derived using the tools of non-smooth mathematical analysis and random operator theory (functional analysis).

We are now going to briefly describe the history of this minimization (or optimization or regularization) problems, showing the various context where l_1 and nuclear norm regularization problems arose. We start with l_1 norm (3.1.1) and we proceed with l_* norm (3.1.2). In (3.1.3) we then describe how the combined use of both heuristics came out.

3.1.1 Cardinality minimization problem: l_1 norm heuristics

As outlined also in Chapter 2, a central role in numerical analysis is played by ill-posed inverse problems (paragraph (2.2)). The genesis of the l_1 norm problem dates back to the problem of recovering a sparse vector from an observed full vector. The most famous appearance comes probably from [108] in the context of regression modelling.

In that famous paper by Robert Tibshirani (1996), the problem of selecting significant regressors in the "Big Data" context, when $p > n$, is effectively solved by shrinking towards zero the irrelevant regression coefficients. The resulting estimator of regression coefficients is called LASSO (Least Absolute Shrinkage and Selection Operator). The LASSO problem can be formalized in the following terms:

$$(\hat{a}, \hat{b}) = \min_{a,b} \sum_{i=1}^{n} (y_i - a - \sum_{j}^{p} b_j x_{ij})^2 \quad (3.8)$$
$$\text{subject to} \sum_{j}^{p} |b_j| \leq t.$$

where t is a tuning parameter.

Assuming without loss of generality that $\bar{x}_j = 0$ for all $j = 1, \dots p$ and that $\bar{y} = 0$, a can be omitted. The same problem is substantially equivalent (see [22], note 1) to

$$\min_{b \in R^p} \frac{1}{2} ||y - Xb||_{Fro} + \rho ||b||_1, \quad (3.9)$$

where $||b||_1 = \sum_j |b_j|$, ρ is a regularization parameter depending on t, and $\frac{1}{2}$ is an arbitrary scale term chosen for computational convenience.

In the language of numerical analysis, problem (3.9), i.e. the l_1 regularization problem, can be intended as a quadratically constrained linear problem (QCLP) or a quadratic program (QP).

The l_1 heuristics was born in the context of signal/image recovery. Tibshirani's contribution was of fundamental importance in the regression context, since it provided a substantial improvement not only upon OLS (in terms of prediction accuracy and interpretability) but also upon ridge regression (which is simply (3.9) with $||b||_2^2$ in place of $||b||_1$, also known as Tikhonov regression or l_2 regularization problem) and upon subset selection techniques. In fact, the LASSO is more stable and interpretable.

Tibshirani showed that, under the condition $X'X = I_p$,

$$\hat{b}_j = \text{sign}(\hat{b}_j^0)|\hat{b}_j^0 - \gamma|, \; j = 1, \dots, p,$$

where $\hat{b}^0$ is the usual OLS estimate, γ is determined by the condition $\sum |b_j| = t$ and X is the $n \times p$ design matrix. However, this is a very special circumstance, and the strength and amplitude of the conditions on X under which model selection is effective is still under investigation, as well as the validity of solution algorithms. A very well known algorithm for LASSO estimation is LARS (Least Angle Regression, [41]).

After Tibshirani's contribution, the literature on model selection via l_1 minimization grew up. In [22] the problem of model selection via l_1 optimization was formalized very elegantly.

Let us consider the linear model $y = Xb + z$, where $y = (y_1 \dots y_n)'$, b is the $p-$ dimensional vector of coefficients and $z = (z_1 \dots z_n)'$ is a vector of independent errors, $z_i \sim N(0, \sigma^2)$.

In the $p > n$ setup, it is difficult to detect which are the coefficients b_i, $i = 1, \ldots, p$ representing the "right" variables to determine the values in y. A standard approach to find $\hat{b}$ is

$$\min_{b \in R^p} \frac{1}{2}||y - Xb||_2^2 + \rho\sigma^2||b||_0, \tag{3.10}$$

where $||b||_0$ is the number of non-zero components in b.

A number of model selection criteria in the form (3.10) has been developed. However, (3.10) is computationally intractable (NP-hard) because it requires exhaustive search over all subsets of columns of X, thus having a complexity of 2^p (if $p \sim n$).

The most popular convex relaxation of (3.10) is the LASSO:

$$\min_{b \in R^p} \frac{1}{2}||y - Xb||_2 + \rho\sigma^2||b||_1, \tag{3.11}$$

where $||b||_1 = \sum_{i=1}^{p} |b_i|$ and λ is a regularization parameter which controls the complexity of the model. We will see why problem (3.11) is the most appropriate convex relaxation of problem (3.10).

The most important condition for recovery, as outlined in [22], is that the predictors are not highly correlated. This is summarized in the notion of coherence, which is the maximum correlation between unit-norm variables and is defined here as

$$\mu(X) = \sum_{1 \le i < j \le p} | < X_i, X_j > |, \tag{3.12}$$

i.e. the maximum inner product between pairs of predictor variables. When the vector b has only s non-zero components, it is said to be s- sparse. In [22] it is proved that assuming appropriate bounds for the values of μ and s and for appropriate values of λ, the error distance is bounded with rate $O(\log p)$. It is remarkable that we need to enforce the maximum inner product among the columns of X, i.e. the maximum correlation between predictors, for identifying the model. The bound on μ is an example of restricted isometry property, which will be necessary to bound the error for all covariance matrix models taken into account.

The l_1 minimization, as explained in [27], was first used for sparse signal reconstruction. This technique can be effectively used in a large number of fields, among which we mention the very recent applications of gene expression data. This setting also includes relevant applications on system control, digital image reconstruction, sparse graphs. Suppose we want to recover a $n \times 1$ signal x_0, from an incomplete set of measurement $y = \phi x_0$, where y is $m \times 1$, ϕ is $m \times n$, with $m \ll n$. Φ represents the coefficient sequence of the signal in the appropriate basis.

The most immediate approach is by solving the l_0 minimization problem:

$$\min_{x \in R^n} ||x||_{l_0} \tag{3.13}$$

under $y = \phi x$, where $||x||_0 = \sum_i \mathbf{1}(x_i \neq 0)$.

Even if this problem would be identified if $||x_0||_{l_0} \leq m/2$, problem (3.13) is intractable because $||x||_{l_0}$ is non convex. Therefore, the most used convex relaxation of problem (3.13) for signal detection is again the l_1 regularization problem:

$$\min_{x \in R^n} ||x||_{l_1} \qquad \text{under } y = \phi x. \tag{3.14}$$

This application is relevant, not only historically, but also because it shows that l_1 heuristics started to be used far from the context of statistical modelling.

Before going on with our brief historical description, it is worth underlining why convex relaxations make problems tractable. A standard theorem of calculus states that a sufficient condition for X to be a minimum of $f(X)$ is that the second derivative of $f(X)$ is strictly positive in an open domain. Since (strictly) convex functions always have a (strictly) positive second derivative, convexity is essential for optimization because it ensures that we find a global optimum. If the function is strictly convex, the minimizer is also unique.

In the case of a matrix function $f(X)$, the sufficient condition becomes the positive definiteness of the Hessian matrix of f. If the function has two or more arguments, it must be convex respect to all arguments in order to have a global minimum. In this way, critical points, i.e. points satisfying $df = 0$, are also minima. We can thus exploit the Lagrangian dual theory.

Another important application of l_1 heuristics, which is exactly the opposite respect to the signal detection problem, is the recovery of a sparse signal representation from overcomplete dictionaries in the harmonic context. Here, the signal y $(n \times 1)$ must be recast from an overcomplete representation (overcomplete dictionary) x having dimensions $m \times n$, with $m > n$. The model in this case is: $y = \Phi x$, where Φ is $n \times m$. The challenge is to recast the orthogonal basis closest to signal y. In linear algebra, these are underdetermined linear systems, i.e. linear systems with infinite solutions. David Donoho ([36]) was the first to prove that among the infinite solutions, l_1 minimization recovers the sparsest one. The fundamental necessary condition is the following restricted isometry property:

$$(1 - \epsilon)||x||_2 \leq \sqrt{\frac{\pi}{2n}}||\Phi x||_1 \leq (1 + \epsilon)||x||_2.$$

Relevant results in this field show that a number of non zero elements in x proportional to $\sqrt{\frac{n}{\log(m)}}$ is usually enough to find a unique solution. Surprisingly, the recovery can be successfully done for a wide range of problems having a relatively small number of samples, until $n = O(m^{1/4} \log^{5/2}(m))$ ([37]), if y is sparse and the observations are selected uniformly at random.

A relevant application described by Candes and Tao in ([23]) deals with the problem of recovering an input vector from corrupted measurements.

Their problem is $y = Af + e$, where f is the unknown $m \times 1$ input vector, y is the observed $n \times 1$ vector, e is the $n \times 1$ error and A is the $m \times n$ coding matrix. Their solution to recover f is

$$\min_{g \in R^n} ||y - Ag||_1. \tag{3.15}$$

This problem is also called error correction problem.

We note that here we have both approximation and recovery from highly incomplete measurements. The recovery is effective with overwhelming probability if the size of the support of e is bounded. Theorems are proved using the concept of restricted isometry, which impose a bound to the incoherence (intended as the distance from being an orthonormal system) of the n input vectors f_i, where f is $m \times 1$ and $i = 1, \ldots, n$. Their problem can be rewritten as

$$\min \mathbf{1}'t, \qquad -t \leq y - Ag \leq t, \tag{3.16}$$

where $t \in R^m$ and $g \in R^n$, and can be recast as a linear problem with inequality constraints and solved efficiently using standard algorithms ([16]). Formulation (3.16) will be very useful for our purpose.

We finally mention an important lemma ([23], Lemma 3.1) which describes the necessary conditions for obtaining a unique minimizer from problem (3.15).

To sum up, l_1 heuristics allowed the rise of a new sampling theory (much fewer samples necessary than before), which results in a new data acquisition protocol. As pointed out in [22], l_1 regularization can be defined as the modern least squares method, for the variety of applications and the capability of providing solutions in the Big Data context.

To conclude this section, we give a remark on solution methods for l_1 minimization problems. An exhaustive review of existing algorithms for the l_1 regularization problem (with specific reference to the face recognition context) is provided in [115]. We want to emphasize here the importance of Iterative Shrinkage Thresholding Algorithms (IST). These algorithms were born in the vector denoising context. The first approach to solve this issue was to set to zero too small entries (which is exactly the shrinkage approach). This could be done solving the usual problem:

$$\min_{x \in \mathbb{R}^n} \phi(x) = \frac{1}{2}||Ax - y||^2 + \rho\tau C(x) \tag{3.17}$$

If C is proper and convex and ϕ is strictly convex, then there is a unique minimizer. If $A = I$, we are in the pure denoising context, and $\phi(x)$ is always strictly convex provided that $C(x)$ is.

This approach moves from the work of a French mathematician, J.J Moreau, who first proposed the concept of proximal mapping ([81], [83]). Problem (3.17) has not to be necessarily solved using $C(z) = ||z||_1$ (l_1 norm).

It can be solved using $C(z) = ||z||_2$ (ridge), $C(z) = ||z||_\infty$ or $C(z) = ||z||_0$ (l_0 norm).

The solution to problem (3.17) was shown to be

$$x_{k+1} = \Phi_{\frac{\rho}{\alpha}} \left(x_k - \frac{1}{\alpha} A'(AX_k - y) \right) \tag{3.18}$$

where $A'(AX_k - y)$ is simply the gradient $\nabla \frac{1}{2}||Ax - y||_2^2$ in x_k, and Φ is the thresholding operator with parameter $\frac{\rho}{\alpha}$. This is the proximal mapping method (recently been proved to be equivalent to the projected gradient approach, see [50]). If $c(z) = ||z||_1$ (3.18) is called soft-thresholding operator, if $c(z) = ||z||_0$ (3.18) is called hard-thresholding operator. The basic shrinkage solution algorithm is called ISTA (Iterative Shrinkage Thresholding Algorithm, see [35]). This approach has been easily extended to the nuclear norm regularization problem.

This algorithm can be equivalently seen in four different ways: as an Expectation-Maximization (EM), a Minimum-Maximum (MM), a Forward-Backward Splitting and a Separable Approximation algorithm. For details we refer to [50].

Finally, we mention Augmented Lagrangian Methods and proximal gradient algorithms, which will be crucial for the solution of our problem (3.7). We note that in this context the ALM and the proximal gradient solution coincide. The fastest solution algorithm, as we will see, is FISTA (Fast Iterative Shrinkage Thresholding Algorithm, [10]).

3.1.2 Rank minimization problem: nuclear norm heuristics

We now move to briefly explain the history of l_* heuristics. Its genesis and use in statistics is much more recent than the one of l_1 heuristics. This topic was first deeply studied in the PhD thesis of Maryam Fazel ([49]). That work explains widely how l_* heuristics can be used for matrix rank minimization problems.

The first relevant application in statistics can be found in [24]. There l_* is effectively used for exact matrix completion. The underlying problem moves from a very well-known case study: the Netflix prize problem. The Netflix prize was an award given to those who were able to set up the best prediction model for movie rating. The Netflix dataset was composed by a large number of movies and a large number of movie ratings. The statistical problem was that most of ratings concerned a small number of movies. This resulted in a matrix where around 99 per cent of entries were missing, since many ratings were empty.

In this context, it is natural to seek for the low rank matrix which best approximates the observed matrix. This equals to recover a matrix from a sample of its entries. Suppose we have a squared $p \times p$ matrix M with rank r having a fraction of entries missing or corrupted. Call Ω the set of locations

corresponding to the observed entries such that $i, j \in \Omega$ if and only if M_{ij} is observed. The original problem to solve is

$$\min \operatorname{rank}(X) \text{ subject to } X_{i,j} = M_{ij}, (i,j) \in \Omega. \tag{3.19}$$

This problem is known to be NP-hard ($\operatorname{rank}(X)$ is nonconvex). Even if some good algorithms exist ([47]), they are of very little practical use, since they require doubly exponential computational times in p.

As Fazel shows in her thesis ([49]) an effective convex relaxation to solve this problem is

$$\min ||X||_* \text{ subject to } X_{i,j} = M_{ij}, (i,j) \in \Omega, \tag{3.20}$$

where $||X||_* = \sum_{i=1}^{p} ||\sigma_i(M)||$ and $\sigma_i(M)$ is the i-th largest singular values of M. This is why for positive semidefinite X, problem (3.20) becomes:

$$\min \operatorname{trace}(X) \text{ subject to } X_{i,j} = M_{ij}, (i,j) \in \Omega,\ X \succeq 0, \tag{3.21}$$

where the symbol $\succeq$ denotes positive semi-definiteness ($\succ$ will denote positive definiteness).

In [16], problem (3.21) is shown to be recast as a semidefinite program (SDP) exploiting the fact that the dual norm of the nuclear norm is the spectral norm. In particular, it can be written as:

$$\min_{L,W_1,W_2} \frac{1}{2}(trace(W_1) + trace(W_2)), \text{ s.t.} \tag{3.22}$$

$$\begin{bmatrix} W_1 & L \\ L' & W_2 \end{bmatrix} \succeq 0.$$

In [24], Candes and Recht define a very intuitive characterization of the matrices it is possible to recover via this method. Consider a real-valued rectangular matrix M. Let its singular value decomposition (SVD) be $\sum_{i=1}^{r} \sigma_i u_i v_i'$. If $u_i, i = 1, \ldots, r$ (left singular vectors of M) and $v_i, i = 1, \ldots, r$ (right singular vectors of M) are selected uniformly at random from all sets of r orthonormal vectors, the SVD is called random orthogonal model. Note that no condition is placed on the singular values $\sigma_i(M)$, since their magnitude is not relevant for recovery.

Candes and Recht show that under the random orthogonal model, if the number of samples $n \leq Cp^{5/4} r \log n$, M is recovered by (3.20) with very high probability. If $r \leq n^{1/5}$, the condition becomes $n \leq Cp^{6/5} r \log n$. The strength of bound is proved to depend not only on p, r and n, but also on the singular vectors of M. If the singular vectors are highly concentrated in few positions, it becomes impossible to recover M from a sample of its entries. This is why they define the coherence quantity for the $p \times r$ matrix of left singular vectors U respect to the standard basis:

$$\mu(U) = \frac{p}{r} \max_{1 \leq i \leq p} ||P_U e_i||. \tag{3.23}$$

$\mu(U)$ ranges from 1 (if v_i are spanned by vectors whose entries are $\sqrt{1/p}$) to n/r (if the basis of U contains any standard basis element). In the same way they define $\mu(V)$ for the matrix of right singular vectors. They then prove that under a bound on $\max(\mu(U), \mu(V))$ and on the maximum entry of $\sum_{1 \leq i \leq r} u_k v_k'$, the previous bound can be generalized.

The concept of coherence, which is also referred to as incoherence (which is the opposite) will play a central role in our covariance matrix recovery. In [24] it is also showed which matrix subspaces satisfy these conditions and which analytical conditions on the subgradient of $||X||_*$ are necessary to ensure that (3.20) is the unique minimizer (Lemma 3.1). This result, together with the analogous one holding for the l_1 norm, will be a key proof tool in the covariance matrix context.

From a mathematical point of view, we are dealing for both heuristics (l_1 and l_*) with underdetermined linear systems. The task is to fill missing entries, in a situation where a large fraction of entries (or elements in the vector case) are missing. This fraction must be not too large in order to identify the unknowns and perform an effective recovery. We note here that the l_1 norm of a vector is simply the nuclear norm l_* of the diagonal matrix containing the same vector as the main diagonal.

In the matrix case, beyond the Netflix problem, this need finds wide application in the field of collaborative filtering, of which reccommender system is a relevant application, as well as genomic data and image processing. All these applications require to estimate a low rank $r \ll p$ matrix to compress information. More widely, as we have seen for the decoding linear program, we may also be interested to relax the reconstruction problem, i.e. to relax the assumption which leaves observed entries unaltered. In a statistical perspective, the approximation problem is much more interesting, since it implicitly assumes a model behind.

Let us call $P_\Omega(X) = X_{i,j}$ if X_{ij} is observed and 0 otherwise. Problem (3.20) can easily be rewritten as

$$\min ||X||_*$$

subject to $P_\Omega(X)_{i,j} = P_\Omega(M)$.

At the same time, we could also be interested in:

$$\min ||X||_* \tag{3.24}$$

subject to $||P_\Omega(X) - P_\Omega(M)||_F \leq \delta$, where
$\sum_{\text{Observed}(i,j)} ||X_{ij} - M_{ij}||^2 = ||P_\Omega(X) - P_\Omega(M)||_F^2$.
Problem (3.24) is equivalent to

$$\min_{\text{Observed}(i,j)} ||X_{ij} - M_{ij}||^2$$

subject to $||X||_* \leq \tau$.

The first form is a quadratically constrained semidefinite program (SDP), the second one is a quadratic program (QP). As explained in [116], we explicitly note that the two problems are strictly related, since the values δ and τ are related. These parameters reflect the level of noise present in the input matrix. Solving one of the two, it is possible to determine the level of noise for which the other problem shares the same solution.

There is an important difference between the reconstruction and the approximation problem. Both problems can be recast as semidefinite problems. We will discuss computational aspects in paragraph (3.2.2). In the former, the constraint is a linear equality, while in the latter the constraint is a quadratic inequality. For this reason, as we will discuss, the latter one requires more than one sparse SVD to be solved, differently from the former one. In [61] there is a wide discussion on large-scale SVD methods which can be effectively used for matrix completion problems.

The same occurs in the l_1 context. The reconstruction problem is a linearly constrained linear program, the approximation problem is a quadratically constrained linear program.

All in all, low rank approximation is the key ingredient of problem (3.20) and (3.24). The underlying combinatorial problem is

$$\min_L \sum_{i,j} (\Sigma_{ij} - L_{ij}) \text{ under } \text{rank}(L) \leq r,$$

which is computationally intractable (NP-hard).

In spite of that, basic theorems from linear algebra state that

$$\min_{B, rank(B)=r} ||A - B||_2$$

and

$$\min_{B, rank(B)=r} ||A - B||_{Fro}$$

are both solved for

$$B = \sum_{i=1}^{r} \lambda_i u_i u_i',$$

which is the SVD truncated to the r-th summand ([40]), when r is known. This is why SVD is the key computational ingredient of recent algorithms.

As explained, if we replace $rank(L)$ with $||L||_* = \sum_{j=1}^{r} \lambda_j(L)$, the problem is made convex ([46]) and assumes the form

$$\min_L \sum_{i,j} (\Sigma_{ij} - L_{ij}) \text{ under } ||L||_* \leq \tau.$$

A natural question arises: is problem (3.20) really minimizing the latent rank? This crucial passage was proved in [97]. There the authors define the general affine rank minimization problem:

$$\min rank(X) \text{ under } \mathbb{A}(X) = b \tag{3.25}$$

where $\mathbb{A}$ is a linear matrix operator. The attribute "affine" means that the rank is minimized under a system of equality constraints. This problem is known to be NP-hard, and has lots of applications, including low rank matrix completion and image compression problems. There is a strict parallelism between compressed sensing (i.e. the cardinality minimization problem) and rank minimization. In particular it is proved that, as l_1 heuristics provides the sparsest solution of an underdetermined linear system, l_* heuristics provides the lowest rank solution of underdetermined system (3.25). This holds if and only if the following restricted isometry property (RIP) holds:

$$(1 - \delta_r)||X|_F \leq ||\mathbb{A}(X)|| \leq (1 + \delta_r)||X|_F, \tag{3.26}$$

where δ_r is the restricted isometry constant, i.e. the smallest scalar satisfying (3.26). The relaxed l_* version of (3.25) is shown to give the minimum rank under suitable conditions on δ_r ($\delta_{5r} < \frac{1}{10}, r \geq 1$).

These results ensure that nuclear norm heuristics recovers the minimum rank solution. We will show in paragraph (3.2.1) why l_* (and l_1) are undoubtedly the most effective proxies of $rank(L)$ (and $||S||_0$).

An exhaustive overview of the algorithms for l_* minimization is provided in [118] with specific reference to image analysis. We mention proximal gradient algorithms ([90]), Augmented Lagrangian (ALM) and Alternating Direction methods (ADM) ([116]). These algorithms will be crucial for our purposes.

In addition, we point out that matrix factorization issues can be effectively exploited also for the rank minimization problem (by the so called UV parametrization). That tool becomes very convenient when dealing with positive semidefinite matrices (PSD). In that case, the nuclear norm becomes the trace norm, and UU' parametrization is very easy-to-implement ([73]). In [4], the consistency of trace norm regularization for PSD was proved very elegantly, respect to the relationship between the regularization threshold λ and the sample dimension n.

However, we will use proximal gradient algorithms, which are more convenient for the particular shape of our composite problem.

3.1.3 Composite penalisation: combined use of l_1 norm and nuclear norm

The nuclear norm minimization approach just described can be extended. In order to make problem (3.24) robust to the presence of outliers, we can assume that the input M can be approximated by $L + S$, where L is a low rank matrix with rank r and S is a sparse matrix, i.e. a matrix with only a fraction of nonzero entries. (3.24) thus becomes

$$\min_{L,S} \frac{1}{2}||(L + S) - M||^2_{Fro} + \lambda||L||_* + \rho||S||_1, \tag{3.27}$$

where $||S||_1 = \sum_{i=1}^{p} \sum_{j=1}^{p} |s_{ij}|$, and is surrogate of $||S||_0$, the number of nonzero elements in S. This problem is called robust convex matrix completion, as pointed out in [61], where this example was mentioned as an application of large-scale SVD methods.

We define our composite (convex non-smooth) penalty $P(L, S)$ as

$$P(L, S) = \lambda||L||_* + \rho||S||_1. \tag{3.28}$$

Problem (3.27) is effective for matrix completion. It comes from the analogous matrix reconstruction problem, which aims at recovering exactly L and S (without any quadratic penalty term). It can be thought of as a robust principal component problem, resulting in a data compression which is robust against corrupted or missing entries. Here we allow for a small matrix S to perturb the low rank matrix L, such that incomplete data matrix reconstruction can be performed. Applications include video surveillance, face recognition, latent semantic indexing, ranking and collaborative filtering.

Suppose now we have a matrix $M = L + S$, where L is a low rank matrix and S is the sparse matrix. M does not need to be squared: this technique was born to reconstruct data matrices.

Classical Principal Component analysis solves the problem:

$$\min ||M - L||, \; rank(L) \leq r, \text{ under } L + S = M.$$

As we described before, this can be solved using classical principal component pursuit (PCP), i.e. taking

$$L = \sum_{i=1}^{r} \lambda_i u_i u_i',$$

where u_i and λ_i, $i = 1, \ldots, r$, are respectively the r eigenvalues and eigenvectors of M.

In [25], the Robust Principal component framework is described. The described problem is

$$\min_{L,S} ||L||_* + ||S||_1, \text{ under } L + S = M.$$

This is a non-smooth minimization problem, since both penalties (and thus their sum) are not convex. In the next paragraph we will analyze the numerical problem, and describe possible approaches for numerical solution. In [25], an effective and relatively fast recovery is shown to be possible only under specific bounds for the rank of L, the number of non-zeros of S, and the singular vectors of L. In particular, $\max_i ||Ue_i||_2 \max_i ||Ve_i||_2 \max_i ||UV||_\infty$ must be bounded, where e_i, $i = 1, \ldots, p$, are the standard basis vectors.

In following works, as we describe in Chapter 4, these conditions have been weakened. Anyway, the approach for ensuring identifiability and recovery comes from the same proof strategy. We will show how this method

can be effectively applied to covariance matrix estimation, i.e. to the noisy context, when one additional noise term is inserted for modelling M.

Now we move to the discussion of the mathematical aspects of the low rank plus sparse decomposition problem, coming back to our key matrix approximation problem (3.7).

3.2 Nuclear norm and l_1 norm minimization: analytical and algorithmic aspects

Our aim is to perform covariance matrix estimation under the assumption of low rank plus sparse decomposition. Such an assumption is equivalent to assume a sparse approximate factor model for the data.

Chapter 4 will be devoted to modelling aspects behind these assumptions. As we pointed out in previous paragraphs, applying (3.7) to the covariance matrix setting requires several assumptions on key parameters, in order to guarantee identifiability, recovery, positive definiteness and invertibility.

In this section, we describe the nature of problem (3.7) from the point of view of numerical analysis (paragraph (3.2.1)) and computational analysis (paragraph (3.2.2)). The structure of the l_1 norm plus nuclear norm regularization problem is described in detail, with reference to mathematical aspects.

3.2.1 Numerical context: a semidefinite program

Let us suppose we have a random vector x with covariance matrix Σ^* following a low rank plus sparse structure (3.1). Let us call X the $n \times p$ data matrix. Suppose $\Sigma_n = \hat{\Sigma}_{n-1}$ is the $p \times p$ unbiased sample covariance matrix computed on the observed data X.

Our combinatorial problem (rank minimization problem (RMP) plus cardinality minimization problem (CMP)) is:

$$\min_{L,S} \frac{1}{2}||(L+S) - \Sigma_n||_{Fro}^2 + \lambda \text{rank}(L) + \rho||S||_0, \\ \text{under } L \succeq 0, S \succ 0, L+S \succ 0. \tag{3.29}$$

Problem (3.29) is NP-hard, since rank(L) and $||S||_0$ are not convex. In fact, both rank(L) and $||S||_0$ have jumps, s.t. they are not continuous (hence not differentiable). The constraints are for ensuring that our covariance matrix and residual matrix estimates are positive definite, as well as the low rank estimate is positive semidefinite. This is the algebraic counterpart of (3.6).

According to section (3.1), the CMP can be approached via the l_1 heuristics, the RMP via the nuclear norm heuristics.

So, problem (3.29) can be rephrased in this way:

$$\min_{L,S} f(L,S) = \frac{1}{2}||(L+S) - \Sigma_n||^2_{Fro} + \lambda||L||_* + \rho||S||_1,$$
$$\text{under} L \succeq 0, S \succ 0, L+S \succ 0. \tag{3.30}$$

where λ and ρ are **threshold** parameters.

- $||S||_0$ has been replaced by the $\mathbf{l_1}$ norm of S, i.e. $\sum_{i=1}^{n}\sum_{j=1}^{n}|s_{ij}|$ (Tibshirani, 1996 [108]);
- rank(L) has been replaced by the **nuclear** norm of L, i.e. $||L||_* = \sum_{i=1}^{r}|d_i|$ (Fazel et al., 2001 [46]).

Since L^* is a PSD (Positive Semidefinite Matrix), $||L||_* = \sum_{i=1}^{r} d_i = ||diag(D)||_1 = trace(D)$.We can thus talk about trace norm heuristics. More specifically, our analysis is restricted to symmetric positive semidefinite matrices.

On a mathematical point of view, $f(L,S)$ is a non-smooth convex function. It is composed by a least squares penalty ($\frac{1}{2}||(L+S)-\Sigma_n||^2_{Fro}$), which is a quadratic function, convex and smooth (differentiable), and a composite penalty which is the sum of two non-smooth convex functions.

The l_1 norm $||S||_1 = \sum_{i=1}^{p}\sum_{j=1}^{p}|s_{ij}|$ is convex if $||tS_1 + (1-t)S_2||_1 \leq t||S_2||_1 + (1-t)||S_2||_1$. This property descends from the properties of the absolute value, which satisfies the Cauchy-Schwarz inequality as it is a norm in the R^1 space.

The nuclear norm can be alternatively defined as $||L||_* = \text{trace}(\sqrt{L'L})$ ([63]). In order to prove it is a convex function, we have to show:

$$\text{trace}\sqrt{(tL_1 + (1-t)L_2)(tL_1 + (1-t)L_2)'}$$
$$\leq t\text{trace}\sqrt{L_1L_1'} + (1-t)\text{trace}\sqrt{L_2L_2'}$$

We develop the first term of the inequality as:

$$\text{trace}\sqrt{(tL_1 + (1-t)L_2)(tL_1 + (1-t)L_2)'} =$$
$$= \text{trace}\sqrt{t^2L_1L_1' + (1-t)^2L_2L_2' + 2t(1-t)L_1L_2'} = A$$

For Cauchy-Schwarz inequality,

$$A \leq \text{trace}\sqrt{t^2L_1L_1'} + \text{trace}\sqrt{(1-t)^2}L_2L_2' + \text{trace}\sqrt{2t(1-t)L_1L_2'} =$$
$$= (t||L_1||_*)^2 + ((1-t)||L_2||_*)^2 + 2t(1-t)||L_1'L_2||_*.$$

Now, we recall a theorem proving that $||L_1'L_2||_* \leq ||L_1||_*||L_2||_*$ ([63]). This result relies on the fact that the nuclear norm is unitarily invariant by definition, i.e. $||UXV|| = ||X||$, for any U, V unitary matrices.

So,

$$A \leq ||(tL_1 + (1-t)L_2)^2||_* \leq ||tL_1||_* + ||(1-t)L_2||_*,$$

where the last step depends again on Cauchy-Schwarz inequality, thus proving the claim.

It is easy to show that the l_1 norm and the nuclear norm are not differentiable. If we think $||.||_*$ as $\sum_{i=1}^{r} \lambda_i = ||\lambda_L||_1$ (where λ_L is the vector of eigenvalues of L), it is straightforward that $||.||_*$ is not smooth if some of the eigenvalues are 0, from the properties of the absolute value. The same holds for $||.||_1$.

In terms of differential, we have $\frac{\delta||x||_1}{\delta x_k} = x_k|x_k|^{-1}$. So, for $x_k = 0$, $||.||_1$ does not exist. The same holds for $||.||_*$: $\frac{\delta||X||_*}{\delta X} = X(X'X)^{-1/2}X$, which means that $||X||_*$ is not smooth if X is not invertible.

We can now explain in detail why l_1 and l_* are the best possible convex relaxations of l_0 and rank respectively. The reason lies in a mathematical argument. Relaxation (3.30) is the tightest convex relaxation of problem (3.29). This is due to the fact that $||X||_*$ is the convex envelope of rank(X), and $||.||_1$ is the convex envelope of $||.||_0$. This fundamental result was proved in Maryam Fazel's PhD thesis. The convex envelope of a non convex function is defined as the largest convex function being smaller or equal to the original one. She was able to prove that the nuclear norm is the lower bound of the solution of the original rank minimization problem ([49], p.55).

The proof is based on the so called conjugate functions. Essentially, Fazel proves that the conjugate of the conjugate of the rank over the set of all matrices having spectral norm less or equal to one ($||X||_2 \leq 1$) is the nuclear norm. Since the conjugate of the conjugate is known to be the convex envelope of the function, the theorem is proved. This result is also extended to $||.||_1$, since the l_1 norm of a vector is simply the rank of a diagonal matrix containing the same entries. Analogously, $||.||_1$ is the convex envelope of $card(x)$ over all vectors x s.t. $||x||_\infty \leq 1$.

This result holds for any matrix X and vector x. In our case, our search is restricted to symmetric PSD for L, and to symmetric positive definite matrices for S and $\Sigma = L + S$.

Therefore, problem (3.30) can be recast as a SDP (SemiDefinite Program).

$$\min_{L,S} \frac{1}{2}||(L+S) - \Sigma_n||_{Fro}^2 \text{ subject to } ||L||_* \leq \lambda \text{ and } ||S||_1 \leq \rho, \qquad (3.31)$$

where S and $L+S$ are positive definite and L is positive semidefinite. This is the PRIMAL problem, and is a quadratically constrained quadratic problem. The least square penalty is a quadratic function. The composite penalty is a non smooth function: the nuclear norm constraint involves the square root of squared entries, thus imposing a quadratic constraint, while the l_1 norm imposes a linear constraint.

Reversely, the problem can be thought of as the following quadratically constrained quadratic SDP program:

$$\min_{L,S} \lambda||L||_* \leq +\rho||S||_1 \text{ subject to } \frac{1}{2}||(L+S) - \Sigma_n||^2_{Fro} \leq \tau. \tag{3.32}$$

It is possible to prove that (3.31) and (3.32) are equivalent.

Since the nuclear norm is the dual of the spectral norm, i.e.

$$||M||_* = \max \text{trace}(M'Y), ||Y'|| = 1$$

(see [16]), exploiting the SDP characterization of the nuclear norm and putting together (3.16) and (3.22) it is possible to write:

$$\min_{S,L,E,W_1,W_2} \gamma 1_n' Z 1_n + \frac{1}{2}(\text{trace}(W_1) + \text{trace}(W_2)) + \frac{1}{p}\text{trace}(E'E), \tag{3.33}$$

$$\begin{bmatrix} W_1 & L \\ L' & W_2 \end{bmatrix} \succeq 0.$$

$$-Z_{ij} \leq S_{ij} \leq Z_{ij}, \forall i,j$$

$$L + S + E = C.$$

As additional constraints, we want that S and $L+S$ are positive definite, and L is positive semidefinite. This formulation was obtained by an appropriate use of slack variables.

Form (3.33) is the SDP characterization of problem (3.30). It is a convex problem; numerically, it is defined as a quadratically constrained quadratic problem (QCQP, [16]). More in detail, it is composed by a linear program (the l_1 part), a quadratic (the l_* part) and a least squares program (the Frobenius loss term). As explained, the least squares penalty is a quadratic function and thus is smooth, differently from the other two components.

Let us now introduce the algebraic matrix context. From an algebraic point of view, the objects we need to identify are the following algebraic **matrix varieties**:

$$\mathscr{L}(r) = \{L \in \mathbb{R}^{p\times p} \mid L = UDU', U \in \mathbb{R}^{p\times r}, D \in \mathbb{R}^{r\times r}\}, \tag{3.34}$$

$$\mathscr{K}(s) = \{S \in \mathbb{R}^{p\times p} \mid |support(S)| \leq s\}, \tag{3.35}$$

where $\mathscr{L}(r)$ is the variety of matrices with **at most** rank r, and $\mathscr{K}(s)$ is the variety of (entrywise/elementwise) sparse matrices with **at most** s nonzero elements. $support(S)$ is the orthogonal complement of $ker(S)$.

The **tangent** spaces to (3.34) and (3.35) are:

$$T(L^*) = \{UY_1' + Y_2V' \mid Y_1, Y_2 \in \mathbb{R}^{p\times r}\}, \tag{3.36}$$

$$\Omega(S^*) = \{N \in \mathbb{R}^{p\times p} \mid support(N) \subseteq support(S)\}. \tag{3.37}$$

As pointed out in [97], the characteristics of the two varieties show a contrastive analogy. They are both Hilbert spaces of matrices: for (3.35) the Hilbert norm is the Euclidean one, for (3.34) is the Frobenius one. The sparsity inducing norm is l_1 for (3.35) and l_* for (3.34). As we will describe in the next section, norm additivity ($||A+B|| = ||A|| + ||B||$)is a key condition for our spaces, since we need them to be as close as possible to this condition to perform identification. Norm additivity requires disjoint support for (3.35) and orthogonal row and column spaces for (3.34).

In [97], it was also showed that a dual formulation for the SDP characterization holds. For the affine minimum nuclear norm problem (3.25), we can write

$$\max b'z \text{ subject to } ||\mathbb{A}^*(X)|| \leq 1 \tag{3.38}$$

as well as

$$\max b'z \tag{3.39}$$

s.t.

$$\begin{bmatrix} I_m & \mathbb{A}^*(z) \\ \mathbb{A}^*(z)' & I_n \end{bmatrix} \succeq 0,$$

where $\mathbb{A}^*$ is the dual operator of $\mathbb{A}$. The first formulation is the convex one, while the second is the numerical one which exploits the SDP charcterization of (3.25).

We note that it is straightforward to obtain the dual version of the l_1 problem (3.9) by simply reshaping formulation (3.38) accordingly. In particular, the dual norm of the operator $\mathbb{A}$ becomes the l_∞ norm. The same holds for the least squares problem in Frobenius norm. It is only necessary to replace $||\mathbb{A}^*(X)||$ with $||\mathbb{A}^*(X)||_{Fro}$, because the dual norm of $||.||_{Fro}$ is $||.||_{Fro}$. Therefore, in order to obtain the dual characterization of our general problem (3.30), it is sufficient to aggregate the characterizations of all sub-problems properly reshaping the operator $\mathbb{A}$ for each term. The same holds for formulation (3.39) too.

3.2.2 Solution methods

The SDP characterization of our problem allows us to apply all standard optimization methods. These include interior point methods (with logarithmic barrier function) and penalty methods. A detailed review of these methods can be found in [16]. The standard MATLAB tool to perform optimization is called SDPT 3, and computes the optimum via infeasibile path-following algorithm (see [101]). This method is effectively used to approach standard low rank plus sparse recovery in the noiseless setting (see [30]). However, in the noisy setting, the presence of the least squares term renders these standard differential methods computationally inefficient, for the use of second derivatives in a large scale context ([16], p.54).

In order to apply standard second-order minimization methods, we should define and solve the Lagrangian dual problem i.e. minimize the Lagrangian function of (3.33). This could be done using the classical method of multipliers, formulating and solving the system of KKT (Karush-Kuhn-Tucker) conditions ([106]). Unfortunately, this method would require to solve an underdetermined non-linear system, by using for instance Newton methods or logarithmic barrier functions, which can be computationally hard. More efficiently, the Lagrangian method could be adapted to include constraints and penalties (Augmented Lagrangian method). Alternatively, the Alternating Direction method (ADM) could also be effective. In order to simplify the nuclear norm minimization and avoid iterative computations of SVD, another solution implies the use of UV-parametrization. Further details can be found in [116], where possible gradient solutions of the affine rank minimization problem (3.24) are analyzed. Alternative methods like interior point methods, penalty methods and barrier methods ([16]) can also be implemented ([101]). In any case, all these methods are not particularly suitable in the large-scale context, because minimizing the quadratic loss requires the computation of a second derivative in large dimensions, which is computationally expensive.

For this reason, recent solutions proposed in the literature are based on first-order method approaches (exploiting only first derivatives), which combine the use of standard differential for the smooth part and a procedure based on the non-smooth properties of the composite penalty.

Proximal accelerated algorithms developed by Yurii Nesterov (see [88] [87]) are the key for our problem. They are essentially augmented Lagrangian methods (ALM) where the first order derivative of the smooth part is augmented by the composite penalty (an overview for this kind of methods is in [90]). In order to solve the non-smooth part, iterative shrinkage solution (IST) methods are used. A very well known method developed for l_1 linear inverse problems is FISTA (Fast Iterative Shrinkage Thresholding Algorithm, [10]). FISTA is an accelerated algorithm derived from the previous one (named ISTA) using Nesterov's acceleration scheme ([86]). This approach was extended to the l_* case in [17] and was named singular value thresholding (SVT). The SVT can be accelerated using the same scheme too.

Talking about non-smooth methods, the subgradient (or subderivative) was first defined for convex functions by Moreau and Rockafellar ([82], [98]) and was then generalized to non convex functions by Clarke ([28]). For the use of subgradient for minimization purposes (subgradient approach) a wide historical and methodological review is in [12].

Given a function $f : I \in R^n \to R$ at point x_0 in the open interval I, the subderivative of f is any vector $v \in \mathbb{R}^n$ satisfying

$$f(x) - f(x_0) \leq < v, x - x_0 > .$$

The set of subderivatives is called subdifferential and is denoted by $\delta f(x_0)$.

$\delta f(x_0)$ is always a non-empty compact set.

The definition of subderivative and subdifferential is analogous in the univariate context, where I and v lie in $\mathbb{R}$. There, it is possible to show that the subdifferential is always a closed set $[a, b]$ where $a = \lim_{x \to x_0^-} \frac{f(x)-f(x_0)}{x-x_0}$ and $b = \lim_{x \to x_0^+} \frac{f(x)-f(x_0)}{x-x_0}$. a and b always exist with $a \leq b$. A typical example very useful to us is the case $f(x) = |x|$. That function is convex (even if not strictly) but non-differentiable at the origin, where the subdifferential is equal to [-1,1]. For negative x_0 the subdifferential coincides with the differential and is equal to -1, for positive x_0 it is the same but equal to 1.

The subdifferential is

$$\delta f(x) = \{d \in R^n : f(y) \geq f(x) + < d, y - x >, y \in R^n\}.$$

For our optimization problem, the subdifferentials of l_* and l_1 are relevant ([112]). We report both:

$$\delta ||X||_* = \left\{UU' + W : W \text{ and } X \text{ have orthogonal row and column spaces, } ||W|| \leq 1\right\} \tag{3.40}$$

$$\delta ||x||_1 = \left\{d \in R^1 : d_i = sign(d_i) \text{ for } i \in T, |d_i| \leq 1, i \in T, T = \{1, \ldots, n\}\right\}. \tag{3.41}$$

Note that both subdifferentials share a common structure. They are both composed by the differential at smooth points (UV' or $sign(d_i)$) and a possible contraction (W or the complement to $1/-1$ as the case). In [97] the optimality conditions for the affine rank minimization problem (3.25) are described:

1. Feasibility condition ($\mathbb{A}(X) = b$)
2. Unimprovability of the subdifferential at any feasible direction $\mathbb{A}^*(z) \in \delta f(x)$,

where $\mathbb{A}^*$ is the adjoint operator such that $< \mathbb{A}x, y >=< y, \mathbb{A}x >$. These conditions ensure that problem (3.25) is solved and the nuclear norm achieves its minimum in the feasible set (which is the set of all candidate matrices Y). In fact it holds:

$$||Y||_* \geq ||X||_* + < \mathbb{A}^*(z), Y - X >= ||X||_* + < z, \mathbb{A}(Y - X) >= ||X||_*.$$

The same considerations hold for the l_1 case with the appropriate changes.

The principles of proximal gradient method are the following. Suppose we have a function $\Phi(x) = f(x) + \Psi(x)$ where f is smooth and Ψ is non-smooth, both convex. Our problem is to minimize $\Phi(x)$ over its feasible set Z ($x \in Z$). This minimization problem can be approached by the composite prox-mapping ([88]):

$$Prox_{\Phi,z}(\xi) = \arg\min_{w \in Z} \left[< \xi, w > + \frac{L_f}{2} ||z - \xi||^2 + \Psi(w) \right], \tag{3.42}$$

where z belongs to the set of points in the domain of Ψ having non-empty subdifferential, and ξ belongs to the domain of f. The procedure works under the condition of Lipschitz continuity for f ($||\Delta f(x) - \Delta f(y)||_2 \leq L_f||x-y||^2$, where L_f is the Lipschitz constant).

We will approach the solution of (3.30) by minimizing (3.42). Following [76], we are going to employ proximal gradient methods, based on the subgradient approach ([88]). The problem of additional constraints will be solved theoretically, showing that problem (3.30) with or without the constraints is geometrically the same. Therefore we now focus on the unconstrained problem (3.7).

Recalling (3.42), we note that the composite prox-mapping equals to finding out the point in the subgradient of the composite penalty which is as close as possible to the gradient of the smooth part at each feasible point. In this respect, this approach is also a projected gradient method. It is also a gradient method, in particular, it is a first order approximation methods because it exploits first derivatives only. It is also a Min-Max (MM) approach, in the sense that proximal gradient is minimized under the assumption that the composite gradient mapping maximizes the gain in terms of iterative minimization of our objective. For this reason, the method works only under the assumption of Lipschitz continuity for the gradient of our objective, i.e. under the assumption of limited variation for our objective.

Our **objective** function is:

$$F(L,S) = \frac{1}{2}||(L+S) - \Sigma_n||^2_{Fro} + \lambda||L||_* + \rho||S||_1. \tag{3.43}$$

The **differentiable** part of (3.43) is

$$f(L,S) = \frac{1}{2}||(L+S) - \Sigma_n||^2_{Fro}, \tag{3.44}$$

where Σ_n is the input of our procedure.

The matrix **gradients** of f are $\nabla_L f = \nabla_S f = L + S - \Sigma_n = W$.

The (matrix bivariate) gradient $\nabla_{L,S}$ is **Lipschitz** continuous with Lipschitz constant $l = 2$:

$$||\nabla_{L,S} f(L_1,S_1) - \nabla_{L,S} f(L_2,S_2)||_2 \leq l\sqrt{|L_1 - L_2|^2_F + |S_1 - S_2|^2_F},$$

$l = 2$.

The **first-order** approximation of (3.43) is:

$$\begin{aligned} Q_{l=2}((L,S),(L_{t-1},S_{t-1})) &= f(L_{t-1},S_{t-1}) + \\ + < \nabla_L f_{(t-1)}, L - L_{(t-1)} > + < \nabla_S f_{(t-1)}, S - S_{(t-1)} > + \\ + \frac{l}{2}|L - L_{(t-1)}|^2_{Fro} + \frac{l}{2}|S - S_{(t-1)}|^2_{Fro} + \lambda||L||_* + \rho|S|_1. \end{aligned}$$

The matrix inner product <> here is the standard $< A, B >= tr(A'B)$.
Note that our composite prox-gradient mapping is:

$$< \nabla_L f_{(t-1)}, L - L_{(t-1)} > + < \nabla_S f_{(t-1)}, S - S_{(t-1)} > + \\ + |L - L_{(t-1)}|^2_{Fro} + |S - S_{(t-1)}|^2_{Fro} + \lambda ||L||_* + \rho |S|_1.$$

This formulation exploits a previous work ([69]), which develops a proximal gradient method for trace norm minimization (i.e. the nuclear norm for PSD matrices). The key is that the gradient step needed to minimize $F(L, S)$:

$$L_k = L_{k-1} - \frac{1}{2}\nabla_{L_{k-1}}, S_k = S_{k-1} - \frac{1}{2}\nabla_{S_{k-1}}$$

is the same minimizing

$$Q_2((L, S), (L_{t-1}, S_{t-1})).$$

In this respect, this method is also and augmented Lagrangian method.

Another relevant aspect is that here we have two matrix variables (L and S). In order to perform minimization, $\nabla_{L_{k-1}}$ must belong to the subdifferential of $\lambda ||L_{k-1}||_*$ and $\nabla_{S_{k-1}}$ must belong to the subdifferential of $\rho ||S_{k-1}||_1$. The problem would be hard to solve via subgradient methods if these two related problems could not be approached somehow separately.

We report the step-size assumption ensuring that the optimization of Q_l is effective.

Lemma 3.2.1. *Let* $(\check{L}, \check{S}) = d_l(\tilde{L}, \tilde{S}) = \min_{L,S} Q_l((L, S), (\tilde{L}, \tilde{S}))$.
If the following stepsize assumption is satisfied for some $l > 0$*:*

$$F(\check{L}, \check{S}) \leq Q_l((L, S), (\tilde{L}, \tilde{S})),$$

then *for any* (L, S)*, we have*

$$F(L, S) - F(\check{L}, \check{S}) \geq l < \tilde{L} - L, \check{L} - \tilde{L} > +$$

$$+ l < \tilde{S} - S, \check{S} - \tilde{S} > + \frac{l}{2}|\check{L} - \tilde{L}|^2_F + \frac{l}{2}|\check{S} - \tilde{S}|^2_F.$$

This passage highlights the nature of **min-max** approach for the method.

Standard subgradient methods have an optimal convergence rate of $O(\frac{1}{\sqrt{t}})$ ([69]). This can be very low for large scale problems. Another feature of this extended gradient approach is that it substantially improves convergence. The key is the separability of our problem in two ones, one in variable L

and the other one in variable S. In fact, our first-order approximation Q_2 is **separable** in L and S:

$$L(t) = \min_L |L - (L_{(t-1)} - \frac{1}{2}\nabla f_{(t-1)})|^2_{Fro} + \lambda||L||_* \tag{3.45}$$

$$S(t) = \min_S |S - (S_{(t-1)} - \frac{1}{2}\nabla f_{(t-1)})|^2_{Fro} + \rho||S||_1 \tag{3.46}$$

These two subproblems can be solved easily, by simple algebraic operations, making this algorithm suitable for large-scale problems. The problem in L (3.45) can be solved by applying the SVT (Singular Value Thresholding, [17]) to $L_{(t-1)} - \frac{1}{2}\nabla f_{(t-1)}$.

Lemma 3.2.2. *(Ji and Ye (2009), Cai et al.(2010))* $\tau_\lambda(Y) = \min_M \frac{1}{2}||M - Y||^2_F + \lambda||M||_*$ *is given by* $\tau_\lambda(Y) = UD_\lambda V'$, *where* $(T_\lambda)_{ii} = \max\{0, D_{ii} - \lambda\}$. T_λ *is called* ***SVT (Singular Value Thresholding*** *operator). The unique solution of (3.45) is thus the SVT of* $L_{(t-1)} - \frac{1}{2}\nabla f_{(t-1)}$.

In [17] it is proved that the SVT operator is the unique minimizer of the l_* minimization problem (3.24), because (3.24) is strictly convex and the SVT of L is proved to belong to the subdifferential of $||L||_*$. Even if the SVT was first developed for the matrix completion problem, it can be effectively used for all nuclear norm approximation problems.

Since we can express a vector as a diagonal matrix having the same vector as the main diagonal, Lemma 3.2.2 holds as well for the l_1 case:

Lemma 3.2.3. *([35])* $T_\rho(Y) = \min_M \frac{1}{2}||M - Y||^2_F + \rho||M||_*$ *is given elementwise by* $(T_\rho(Y))_{ij} = sign(Y_{ij}) \max\{0, |Y_{ij} - \rho|\}$.

$T_\rho(Y)$ is called **Soft-Thresholding** operator. Therefore, (3.46) is solved by applying soft-thresholding to $S_{(t-1)} - \frac{1}{2}\nabla f_{(t-1)}$.

In origin, this algorithm was proposed in [35] to solve the LASSO regularization problem ([108]). The extension to the l_1 matrix norm problem is straightforward. This algorithm has been effectively used in a number of situations, like for instance in the graphical lasso context for sparse inverse covariance matrix estimation ([53]).

Due to the separability property and to the use of trace norm heuristics, our minimizer can now converge at a rate $O(t)$ ([69]). As this cost can be still expensive in the large-scale context, Nesterov's acceleration scheme for composite gradient mapping minimization problems ([87]) is applied. As a consequence, the algorithm assumes the form ([77]):

- **repeat**
- set $(L_0, S_0) = (diag(\Sigma_n), diag(\Sigma_n))/2$
- Initialize $L = L_0 = Y_1$ and $S = S_0 = Z_1$

- Apply SVT operator to the SVD of $(Y_{(t-1)} - 1/2\nabla(Y_{(t-1)}, Z_{t-1})$ and set $L_t = UT_\lambda U'$
- Apply soft-thresholding operator to $M = (Z_{(t-1)} - 1/2\nabla(Y_{(t-1)}, Z_{t-1})$ and set $S_t = T_\rho(M)$.
- Set $(Y_{(t+1)}, Z_{(t+1)}) = (L_t, S_t) + \frac{\alpha_t - 1}{\alpha_{t+1}}[(L_t, S_t) - (L_{t-1}, S_{t-1})]$ where $\alpha_{t+1} = \frac{1+\sqrt{1+4\alpha_t^2}}{2}$.
- **until** Convergence criterion $\frac{||L_t - L_{t-1}||_F}{||1+L_{t-1}||_F} + \frac{||S_t - S_{t-1}||_F}{||1+S_{t-1}||_F} \leq \epsilon$.

This algorithm has also been effectively used for dynamic Magnetic Resonance Imaging (MRI) data ([89]). More generally speaking, also the Lipschitz constant l can be linearly updated during the algorithm, when there is some suspect that $l = 2$ is not appropriate ([76]).

The described algorithm is proved to converge at rate $O(t^2)$ ([77]):

Theorem 3.2.1. *Let (L_t, S_t) be the update produced by the algorithm at iteration t. Then for any $t \leq 1$, we have the following computational accuracy bound:*

$$F(L_{(t)}, S_{(t)}) - F(\hat{L}, \hat{S}) \leq 8\frac{||L_0 - \hat{L}||_{Fro} + ||S_0 - \hat{S}||_{Fro}}{(t+1)^2}$$

where $(\hat{L}, \hat{S})$ minimizes (3.7).

This results allows to highlight another advantage of this approach concerning computational cost. Standard methods for SDPs like interior point methods (IPMs) require $O\left(\frac{p^6}{\log(\epsilon)}\right)$ operations, which is too expensive for large-scale problems. This algorithm requires only $O\left(\frac{p^4}{\sqrt{\epsilon}}\right)$ operations. This can be obtained multiplying the number of computations for full SVD $O(p^3)$ (which is the one of standard least squares problems because it requires at each iteration to solve p quadratic systems) times the square root of the bound in Theorem 3.2.1 (at most $O(p^2)$), divided by the square root of the computational precision ϵ. This cost is $O(p^2)$ smaller than the one of IPMs given that the precision requirementis not high. This rate could be further improved by using partial (soft) SVD methods like soft-impute, which require, if there are no missing entries, only $O(p^2)$ computations (otherwise, in the pure l_* context, even fewer: see [61], slide 15).

Chapter 4

Covariance estimation via low rank plus sparse decomposition: statistical performance

The main topic of this chapter is covariance matrix estimation under the assumption of low rank plus sparse structure (3.1). Here we discuss recovery and identifiability conditions for Σ^* under various model assumptions. The unifying feature of all these models is that the estimation is carried out by composite minimization problems including (3.28), which is our composite (convex non-smooth) penalty.

In section (4.1), existing works on matrix reconstruction or approximation using composite penalty (3.28) are discussed.

In paragraph (4.1.1), we discuss the approach to matrix reconstruction by Chandrasekaran et al. (2011) ([30]), which minimizes a composite penalty in the form (3.28) (apart from appropriate re-scaling of regularization parameters). Therein, the exact decomposition is performed, in a noiseless context.

In paragraph (4.1.2), we describe the approach to matrix approximation by Agarwal et al. (2012) ([1]), which provides a first (approximate) solution to the problem of approximate decomposition (in the noisy context) into approximately low rank and sparse matrices. There, both components (and consequently their sum) are recovered by minimizing (3.7) under specific assumptions on $||L||_\infty$.

In paragraph (4.1.3), we show the exact solution of the approximate decomposition problem for a latent variable graphical model proposed by Chandrasekaran et al. (2012) ([31]). In that paper the precision matrix is estimated under model structure (3.1) by minimizing a regularized likelihood problem including a Gaussian log-likelihood term and the composite penalty

(3.28). It is the first exact solution to the recovery problem of both components and their sum in the noisy context, and provides the mathematical context for identification and exact recovery (for the inverse covariance matrix). Therein, the error rates for the covariance matrix were obtained as a consequence.

In paragraph (4.1.4), we describe the most recent covariance estimator obtained minimizing (3.30), which is called LOREC (LOw Rank and sparsE Covariance estimator, [77]). We provide recovery and identifiability conditions for a covariance matrix (as well as its inverse) under model (3.1), following the results appeared in Luo (2013) ([77]). These results were obtained adapting the mathematical setting of [31], thus giving an exact solution to the approximate recovery problem.

4.1 Low rank plus sparse decomposition: identification and recovery

This section is devoted to the description of existing estimators based on the composite minimization of nuclear norm and l_1 norm, under the assumption of low rank plus sparse decomposition for the covariance matrix. We have widely described in previous chapters why the need for a regularized estimate of the covariance matrix comes out. We keep in mind two keywords: reconditioning and model parsimony.

We now distinguish two cases: the noiseless context and the noisy context. In the former, we want to recover a squared $p \times p$ matrix

$$C = A^* + B^*, \tag{4.1}$$

where A^* is sparse having at most s nonzero elements and B^* is low rank with rank $r < p$. This is the context of paragraph (4.1.1), derived by [30], and is for us an unavoidable preliminary step, because identifiability and recovery were first established in that context. Here C is simply an input matrix.

Then we have the noisy context, where we start from an input estimate:

$$\hat{\Sigma} = L^* + S^* + W, \tag{4.2}$$

which contains an error term (**noise**) W distributed as a centered zero-mean Wishart. S^* is **sparse** having at most s nonzero elements and L^* is **low rank** with rank $r < p$. This is the context of all the following paragraphs and models we will describe. We usually have $\hat{\Sigma} = \hat{\Sigma}_{n-1}$, that is, the unbiased sample covariance matrix. This point is a relevant one because this choice implies the condition $n \leq p+1$, which can be not appropriate in a large dimensional context, as explained in paragraph (2.1). We will try to overcome this issue in paragraph (5.1).

The first attempt to identify both the low rank and the sparse component was made in the noiseless context. The problem was set into the context of algebraic geometry, as a deterministic (exact) recovery for general complex non-symmetric matrices. It is easy to see that strong identifiability issues arise, for the simultaneous recovery of the two matrices under the sum constraint. The identifiability issue is central in our discussion. We now start to define the setting we are working on.

4.1.1 Exact recovery: rank-sparsity incoherence

Let us suppose we have an input matrix $C \in \mathbb{R}^{p\times p}$. We suppose that C is the sum of a low rank matrix B^* and sparse matrix A^*, both unknown. Which classes of low rank and sparse matrices allow to perform exact decomposition? The aim of this paragraph is to show how to disentangle C in the two underlying components, following the approach in [30]. This is a decomposition problem: sufficient conditions for fundamental identifiability and recovery are needed. We face a deterministic (purely numerical) problem, which is to find out A^* and B^* as well as the number and the location of non-zeros in A^* (sparsity pattern) and the rank of B^*. This is why here we have no sample dimension n: the parameters are only the the dimension p, the number of non-zeros s and the latent rank r.

In order to perform this task, we need first to properly define the objects to identify. As explained, the tools of algebraic geometry (a reference book is [60]) are very useful to us. In particular we are going to exploit the basic concept of matrix algebraic variety. Matrices A^* and B^* are assumed to come from the following set of matrices:

$$\mathscr{L}(r) = \{B \in \mathbb{R}^{p\times p} \mid B = UDU', U \in \mathbb{R}^{p\times r}, D \in \mathbb{R}^{r\times r}\} \tag{4.3}$$

$$\mathscr{K}(s) = \{A \in \mathbb{R}^{p\times p} \mid |support(A)| \leq s\}. \tag{4.4}$$

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

The decomposition problem (4.1) is fundamentally ill-posed, that is, it is not possible to find out a unique decomposition without further assumptions. In fact, two natural identifiability problems arise:

- the low rank matrix may be itself very sparse;
- the sparse matrix may have itself very low rank.

In order to obtain a unique disentanglement, an upper bound on the degree of sparsity of the low rank component as well as a lower bound on the rank of the sparse component are needed. For this purpose, in [30] 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. In particular, quantities involving tangent spaces to algebraic varieties (4.3) and (4.4) are needed.

Matrix sets (4.3) and (4.4) can be seen as differentiable manifolds (away from their singularities) or as algebraic varieties, as they essentially are set of polynomial equations. The variety of rank-constrained matrices (4.3) is characterized by the vanishing of all $(r+1) \times (r+1)$ minors of B. For this reason, since the (unknown) parameters are p^2 and the equations are $(p-r)^2$, the dimension of this variety is $r(2p-r)$. This variety is nonsingular everywhere except at those matrices with rank less than or equal to $r-1$. This happens because the tangent space at those points has zero measure (and thus it is not uniquely identified). The tangent space to r - ranked matrices is:

$$T(B) = \{UY_1' + Y_2V' \mid Y_1, Y_2 \in \mathbb{R}^{p \times r}\}, \tag{4.5}$$

where UDV' is the SVD decomposition of B.

The tangent space $T(B)$ is the space of all the matrices having the same row or column space of B. For this reason, the dimension of $T(B)$ is again $r(2p-r)$ (if B has rank r). $T(B)$ is a subspace of $\mathbb{R}^{p \times p}$, because it is closed under addition and scalar multiplication.

The variety of sparse matrices (4.4) is the set of all the matrices having a limited size of their support. If the number of non zero elements is equal to $s \ll p^2$, the dimension of the support is constrained by s. This is due to the properties of null spaces and homogenous systems: since the support is the orthogonal complement of $ker(S)$, if $ker(S)^{\perp}$ has dimension s, $ker(S)$ has dimension $p^2 - s$ and S has exactly s zeros. Analogously to the low rank case, this variety is singular everywhere except from those matrices having a dimension of their support less than or equal to $s-1$, because in that case $ker(S)$ has measure 0 (and thus it is not uniquely identified) in $\mathbb{R}^s$.

The tangent space to (4.4) is:

$$\Omega(A) = \{N \in \mathbb{R}^{p \times p} \mid support(N) \subseteq support(A)\}. \tag{4.6}$$

It is the variety of all the matrices having a support contained in the one of A. It has dimension s and it is a subspace of $\mathbb{R}^{p \times p}$.

In this algebraic context, it is easy to understand why the authors of [30] chose to estimate A^* and B^* solving the following optimization problem:

$$(\hat{A}, \hat{B}) = \min_{A,B} f(A,B) = \gamma ||A||_1 + ||B||_* \text{ under } C = A^* + B^*. \tag{4.7}$$

For the discussion on the opportunity of using this problem for rank-sparsity recovery we refer to Chapter 3. This is a deterministic (recovery) problem. Note that γ is a tuning parameter depending on the relative size of $||A||_1$ respect to $||B||_*$.

Identifiability conditions depend on relevant quantities referred to tangent spaces $T(B^*)$ and $\Omega(A^*)$. In particular, the relevant quantity is the product of two quantities, one for each space, describing the degree of rank-sparsity incoherence between the rank of B^* and the sparsity pattern of A^*.

We define the following rank-sparsity incoherence measures between $\Omega(A^*)$ and $T(B^*)$:

$$\xi(T(B^*)) = \max_{N \in T(B^*), ||N||_2 \leq 1} ||N||_\infty, \tag{4.8}$$

$$\mu(\Omega(A^*)) = \max_{N \in \Omega(A^*), ||N||_\infty \leq 1} ||N||_2. \tag{4.9}$$

Note that $\xi(T(B^*)) \leq 1$, $\mu(\Omega(A^*)) \leq \sqrt{p}$.

These quantities are the maximum infinity norm among the matrices belonging to $T(B^*)$ and the maximum spectral norm among the matrices belonging to $\Omega(A^*)$. They arise naturally from the study of the relationship between the rank and the sparsity pattern of one matrix. In fact, a relevant result on $\mu(M)$ and $\xi(M)$, holds for any matrix $M \in \mathbb{R}^{p \times p}$:

Theorem 4.1.1. *For any matrix $M \neq \mathbf{0}$, we have that $\xi(M)\mu(M) \geq 1$.*

This results describes the deep meaning of the concept of rank-sparsity incoherence: it is not possible for one matrix to have $T(M)$ with all diffuse elements and to have diffuse spectra for $\Omega(M)$. The uncertainty principle states that a matrix M cannot have $\mu(M)$ and $\xi(M)$ simultaneously small.

Another relevant result involving μ and ξ arises analyzing the conditions ruling the intersection between (4.5) and (4.6). If we could assume to know the tangent spaces, a necessary and sufficient condition for exact decomposition would be

$$\Omega(A^*) \bigcap T(B^*) = \mathbf{0},$$

i.e. the condition of transverse intersection between the two spaces. This conditions involves crucially quantities (4.8) and (4.9), as outlined in the following proposition:

Proposition 4.1.1. *Given two matrices A^* and B^*, we have that*

$$\mu(A^*)\xi(B^*) < 1 \Rightarrow \Omega(A^*) \bigcap T(B^*) = 0.$$

The smallest $\mu(A^*)$ and $\xi(B^*)$, the closer to the condition of perfect transversality we are, and so the easiest is the decomposition. In this case, since we are in the noiseless context, we need perfect transversality. From the next paragraph (4.1.2), as we set into the noisy context, we will relax this assumption, allowing a small degree of intersection, since we allow random perturbations for A^* and B^*. However, in order to perform recovery, this degree shall be suitably bounded.

From these results we can argue that, in order to perform recovery, we need to **control** the spikiness of the eigenvalues of A^* and the sparsity pattern of B^*. In fact, If B^* is nearly sparse, A^* cannot be recovered, as well as, if A^* is nearly low rank, B^* cannot be recovered. An uncertainty principle between the rank of B^* and the sparsity pattern of A^* holds, i.e. too sparse low rank matrices as well as sparse matrices with too low rank cannot be recovered. It is interesting that the magnitude of the eigenvalues of the low rank component as well as the number of nonzeros in the sparse component play no role for identification. The product $\mu(A^*)\xi(B^*)$ is the rank-sparsity incoherence measure and bounding it controls for that.

In light of Proposition 4.1.1, the two identifiability issues can be described in a more technical way as follows:

- The low rank component is not too sparse if its row/column spaces are NOT closely aligned to the standard basis vectors, i.e. if the maximum projection of a standard basis vector onto the vector subspace spanned by the columns of U is as small as possible.

- The sparse component is not low rank if it does not have too concentrated support, i.e. if its spectrum (set of eigenvalues) is bounded. In other words, we want that the maximum number of non-zeros per column to be bounded.

These technical conditions naturally arise from the geometric algebraic setting and from the minimization context using (4.7) under the sub-gradient approach. In fact, (4.8) and (4.9) are the dual norms of tangent spaces (4.5) and (4.6) respectively. Optimality conditions are derived using the projected gradient method. In that approach, a (Lagrangian) dual candidate Q which belongs at the same time to the subgradient of A^* and B^* is sought for:

$$Q \in \gamma\partial||A^*||_1 \text{ and } Q \in \partial||B||_*.$$

Two duals, Q_A and Q_B, are defined, and the conditions proving they minimize (4.7) are derived. For the expression of the subgradients we refer to (3.40) and (3.41).

In principles, this method consists in projecting onto Ω and $\Omega^\perp$ the subgradient of Q_A and onto T and $T^\perp$ the subgradient of Q_B, where (Q_A, Q_B) is a subgradient of (4.7). Differently from here, in the noisy context (paragraphs (4.1.3), (4.1.4)) we will project the dual candidate augmented by the gradient of the differentiable part of the objective.

We can now report the following key proposition which displays necessary conditions for obtaining a unique minimizer via (4.7) in the noiseless context.

Proposition 4.1.2. *Suppose $C = A^* + B^*$. Then, $(\hat{A}, \hat{B}) = (A^*, B^*)$ is the unique optimizer if the following conditions are satisfied:*

1. $\Omega(A^*) \bigcap T(B^*) = 0$

2. There exists a Lagrangian dual $Q \in \mathbb{R}^{n\times n}$ such that:

- $P_{T(B^*)} = UV'$
- $P_{\Omega(A^*)} = \gamma sign(A^*)$
- $||P_{(T(B^*)^\perp)}|| < 1$
- $||P_{\Omega(A^*)^\perp}||_\infty < \gamma$.

We note that the second claim describes necessary conditions on Q for belonging to both subgradients simultaneously (two for each subgradient), which is equivalent to ensure that $(\hat{A}, \hat{B})$ is an optimum. The first condition, instead, is necessary to guarantee uniqueness.

This proposition is of fundamental importance. It basically proves that only one dual $\hat{Q} \in \Omega \bigoplus T$ may exist satisfying the subgradient conditions, such that $(\hat{A}, \hat{B})$ is the only optimum of the convex program (because only one point provides $\Omega(A^*) \bigcap T(B^*) = 0$).

Therefore, $\mu(A^*)\xi(B^*) < 1$ is a necessary condition for performing recovery. However, a stronger necessary condition for exact recovery respect to the one of Proposition 4.1.2 can be derived. The proof technique builds a dual $\hat{Q} \in \Omega \bigoplus T$, under which the conditions of Proposition 4.1.2 for recovery are satisfied, and finds out the range of γ for which $\hat{Q}$ satisfies all conditions simultaneously. This proof results in the following statement:

Theorem 4.1.2. *Given (4.1), if*

$$\mu(A^*)\xi(B^*) < \frac{1}{6},$$

the unique *optimum for $(\hat{A}, \hat{B})$ is (A^*, B^*), for $\gamma \in \left[\frac{\xi(B^*)}{1-4\mu(A^*)\xi(B^*)}, \frac{1-3\mu(A^*)\xi(B^*)}{\mu(A^*)}\right]$, where $\gamma = \sqrt{\frac{3\xi(B^*)}{2\mu(A^*)}}$ is always inside the range as it is the geometric mean of the extremes, and thus guarantees exact recovery of (A^*, B^*).*

We have identified a sufficient condition for exact recovery, which is $\mu(A^*)\xi(B^*) < \frac{1}{6}$. However, in reality we do not have any knowledge on $\mu(A^*)$ and $\xi(B^*)$. In order to make this condition somehow verifiable, in [30] two nice more operative concepts about rank-sparsity incoherence are formalized, with the aim of providing useful proxies of μ and ξ. The first is the *degree* of a matrix, which is defined as the maximum (deg_{max}) or minimum (deg_{min}) number of non zero entries per row/column. It is proved that

$$deg_{min}(A) \leq \mu(A) \leq deg_{max}(A). \tag{4.10}$$

The second is the concept of incoherence of a vector subspace S of $\mathbb{R}^n$. Define $\beta(S) = \max_i ||P_S e_i||_2$, where e_i is the i -th standard basis vector. $\beta(S)$ is the maximum norm of the projection of any standard basis vector onto S. It is proved that $\sqrt{r/n} \leq \beta(S) \leq 1$, where the maximum (which is 1)

is reached for any basis containing a standard basis vector, and the minimum is reached for an Hadamard matrix, which is a matrix having entries $+1/-1$ and mutually orthogonal rows (see [56]). The *incoherence* of a matrix is defined as:

$$inc(B) = \max(\beta(row - space(B)), \beta(column - space(B))).$$

This quantity satisfies the following property:

$$inc(B) \leq \xi(B) \leq 2inc(B). \tag{4.11}$$

Therefore, a small $deg_{max}(A^*)$ implies a small $\mu(A^*)$ and small $inc(B^*)$ implies a small $\xi(B^*)$. As a consequence, the deterministic sufficient conditions on exact decomposability $\mu(A^*)\xi(B^*) < \frac{1}{6}$ can be rephrased as

$$deg_{max}(A^*)inc(B^*) < \frac{1}{12},$$

as well as the range for γ in Theorem (4.1.2). The central value in that range becomes $\gamma = \sqrt{\frac{3inc(B^*)}{deg_{max}(A^*)}}$.

Finally, the authors provided in [30] a random analysis of their setting. They define A^* to follow a *random sparsity model* if $support(A^*)$ is selected uniformly at random from all collections of supports of size s. In that case, the following relevant property holds:

$$deg_{max}(A^*) \leq \frac{s}{p}\log(p)$$

with high probability. Analogously, a r-ranked squared matrix B of dimension p is said to follow a *random orthogonal model* (see also [24]) if the singular vectors $U, V \in \mathbb{R}^{p\times r}$ are chosen among all partial isometries in $\mathbb{R}^{p\times r}$, where a partial isometry is an isometry on the orthogonal complement of the kernel. Under this hypothesis, we have

$$inc(B^*) \preceq \sqrt{\frac{\max(r, \log(p))}{p}}$$

with very high probability (the symbol $\preceq$ is used to denote rates, with the meaning of the "smaller or approximately equal to", as well as the symbol $\succeq$ will be used with the opposite meaning). Given (4.1), if A^* is drawn from a random sparsity model and B^* is drawn from a random orthogonal model, the conditions of Theorem 4.1.2 hold provided that $s \preceq \frac{p^{1.5}}{\log(p)\sqrt{max(r,\log(p))}}$.

We signal that this approach comes from the one by Candes and Recht ([24]) described in paragraph (3.1.2). There, the degree of coherence between singular vectors and the standard basis is bounded using the quantities

$$||UU' - \frac{r}{p}I_p||_\infty,\ ||VV' - \frac{r}{p}I_p||_\infty,\ ||UV'||_\infty.$$

For symmetric matrices, only the first quantity is relevant. Contrastively, the approach of [30] allows for a more unified condition, taking into account **simultaneously** the row and the column spaces (that is, left and right singular vectors).

This approach is overall very elegant, effective and algebraically founded and provides a new environment for matrix reconstruction analysis. However, the sufficient condition provided by Theorem 4.1.2 is local, i.e. it is not robust to perturbations of B^* and A^* along varieties $T(B)$ and $\Omega(A)$. Tangent space transversality is a linearized identifiability condition around (A^*, B^*), but does not provide any guarantee even for slightly perturbed inputs, because it only guarantees an exact solution in the noiseless context.

This is why we are now going to explore numerical methods providing solutions to the matrix approximation problem in the noisy context.

4.1.2 Approximate recovery: a functional approach

The topic of this paragraph is the purely mathematical approach to matrix approximation by Agarwal et al. (2012) ([1]). This is a numerical approach based on pure functional analysis, in the general setting of complex rectangular matrices. Before describing it in detail, we outline the relevant characteristics for our purpose.

First of all, the reference matrix setting is the noisy setting (4.2), from here towards the end of our thesis. In [1], L^* is allowed to be exactly or approximately low rank and S^* is allowed to be exactly or approximately sparse. Their setting thus includes a wide set of matrix classes, including our reference model (3.1) as a particular case. Their model is the following

$$X = \aleph(L^* + S^*) + W,$$

where $\aleph$ is called observation operator, and is a linear mapping operator from $(S^* + L^*)$ to $\aleph(L^* + S^*)$ (we define $\Omega = L^* + S^*$).

In our case, $\aleph = I$ (identity mapping). If $W = 0$, we fall back into the noiseless setting. The noise W can be either deterministic or stochastic. This setting includes a wider class of sparsity assumptions, including the cases of element-wise and column-wise sparsity. In our reference model (4.2), we have exact element-wise sparsity and exact low rankness with stochastic noise. The matrix to recover, Σ^*, is a squared $p \times p$ real matrix in $\mathbb{R}^{p \times p}$.

The input $\hat{\Sigma}$ is the sample covariance matrix $\hat{\Sigma}_n$. We underline again the statistical centrality of this passage, which is relevant for our purpose also in the approach we are describing. Whenever $\hat{\Sigma} = \hat{\Sigma}_n$, the related condition $p \leq n$ comes out, even if (here and in the following paragraphs) the estimation method via regularization allows $p \sim n$.

As we explain in paragraph (5.1), there are essentially two solutions to this drawback: using a regularized input (for instance $\hat{\Sigma}_{LW}$, see (2.7)), allowing to drop the technical condition $p \leq n$, or using a method which allows

to consistently use $\hat{\Sigma}_n$ without the need of specifying $p \leq n$. In this respect, POET approach ([45]) is central, and we will show how it is possible to use the POET estimation context in order to avoid the condition $p \leq n$ even if p and n are finite.

In light of this, we go on explaining the proposal of [1]. This method consists in estimating Σ^* by program (3.7) (we set aside for the moment the three additional constraints) under specific conditions. The most relevant one is the following: $||L^*||_\infty \leq \frac{\alpha}{p}$, that is, a bounded infinity norm for L^*, which controls the spikiness of the singular values of L^*. This assumption prescribes, from our point of view, that the maximum communality across variables must be bounded. It is an analytical assumption in nature, differently from the algebraic approach aimed at bounding the degree of coherence between singular vectors and canonical basis ([24]):

$$\| UU' - \frac{r}{p} I_r \|_\infty, \qquad \| VV' - \frac{r}{p} I_r \|_\infty, \qquad ||UV'||_\infty.$$

Here the imposed condition is $||UDV'||_\infty \leq \frac{\alpha}{p}$, which uses the singular values of L^* as weights in the l_∞ bound. We note that here a bound on singular values (the eigenvalues for covariance matrices) is implicitly posed, which is equivalent to bound the condition number of L^*, differently from the approach described in paragraph (4.1.1). This condition is weaker: no condition is imposed on the row/column spaces of L^* (only its maximum element must be bounded) and allows for wider classes of matrices.

It is relevant that no explicit condition is placed on the sparse component: in this purely analytical approach, recovery is performed imposing regularity conditions on the objective function (3.7), with particular reference to the convexity properties of the smooth and the non-smooth part jointly. So, the sparsity pattern of S^* is involved only in contrast to the spikiness pattern of L^*, by imposing a lower bound to quantity

$$\Phi(\Delta) := \inf_{S+L=\Delta} Q(S, L), \tag{4.12}$$

where

$$Q(S, L) := ||L||_* + \frac{\rho}{\lambda} ||S||_1$$

is a weighted combination of the regularizers (ρ and λ are non-negative regularization parameters).

However, this approach has a relevant drawback: the approximate recovery of the approximately low rank and sparse components is itself approximate, because it provides not an identifiability condition, but a bound on the radius of non-identifiability (in our setting, $||L^*||_\infty \leq \frac{\alpha}{p}$). The larger α, the broader is the class of identifiable models, but the more difficult is the recovery, especially of the sparse component. Indeed, in [1], paragraph 4, the authors provide mini-max optimality properties for their method over

the classes of approximately low rank and approximately sparse matrices (which are broader than those we need).

This method descends from the previous work of a subset of the same authors ([85]) where weighted matrix completion (respect to rows/columns) is performed into the same mathematical setting using only the nuclear norm. On that path, [1] represents a direct extension.

The sense of their mathematical approach is now described. The regularization problem is:

$$\min_{L,S} f(L,S) = \frac{1}{2}||(L+S) - \Sigma_n||^2_{Fro} + \lambda||L||_* + \rho\mathscr{R}(S^*), \tag{4.13}$$

where $\mathscr{R}(S^*)$ is a regularizer. For us, $\mathscr{R} = ||.||_1$, i.e. we want to recover exactly low rank matrices with rank $r \ll p$ and exactly entry-wise sparse matrices with at most $s \ll p^2$ nonzero elements. Here, non-asymptotic error rates are given for a wider class of regularizers. For example, a related heuristics imposes to S^* columnwise (blockwise) sparsity, which is recovered using $\mathscr{R}(S^*) = ||S^*||_{2,1} = \sum_{k=1}^{p} ||S_k||_2$, where S_k denotes the k-th column of S^*.

In general, $\mathscr{R}$ can be any decomposable regularizer, which is defined respect to the pair of subspaces $(M, M^\perp)$ as:

$$\mathscr{R}(U+V) = \mathscr{R}(U) + \mathscr{R}(V),$$

for all $U \in M$ and $V \in M^\perp$. Our reference norm, $\mathscr{R} = ||.||_{l_1}$, is decomposable respect to $(M(T), M(T)^\perp)$, where

$$M(T) = \{U \in R^{d_1 \times d_2} | U_{jk} = 0 \, \forall (j,k) \notin T\}$$

$$M^\perp(T) = M(T)^\perp$$

and $T \in \{1,\dots,p\} \times \{1,\dots,p\}$ is an arbitrary collection of indices. In fact, $||U + U'||_1 = ||U||_1 + ||U'||_1$, for all $U \in M$ and $U' \in M^\perp$.

With respect to subspace M, they defined a compatibility constant between the regularizer $\mathscr{R}$ and the Frobenius norm:

$$\Phi(M, \mathscr{R}) := \sup_{U \in M, U \neq \mathbf{0}} \frac{\mathscr{R}(U)}{||U||_{Fro}}.$$

In our case, we have $\Phi(M, ||.||_1) = \sqrt{s}$.

The following norm-related quantity is then defined:

$$\kappa_d(\mathscr{R}) := \frac{||V||_F}{R(V)},$$

as well its associated dual norm:

$$R^*(U) := \sup_{R(V) \leq 1} < V, U >$$

with $< V, U >:= trace(V'U)$. The quantity describing the interaction between the low rank and the sparse component, equivalent of $\mu(A^*)\xi(B^*)$ in paragraph (4.1.3), is the following:

$$\varphi(L^*) = \kappa_d(\mathscr{R}^*)\mathscr{R}^*(L^*).$$

Thus, the interaction between the low rank and the sparse component is here constrained using the dual norm of $\mathscr{R}$ computed on L^*, rescaled by the norm-related constant $\kappa_d(\mathscr{R})$. The general bound on the radius of non-identifiability is thus $\varphi(L^*) \leq \alpha$. Note that, analogously to POET approach, the spikiness of the low rank component is bounded starting from the sparsity features of the sparse components. This feature is at the same time the most relevant weakness of this approach for our purpose, because there is not an intrinsic bound for the dual norm of the nuclear norm assessed in S^*, which is $||S^*||_2$. For us, $\kappa_d(\mathscr{R}^*) = p, \mathscr{R}^* = ||.||_\infty$, from which the previously described condition $||L^*||_\infty \leq \frac{\alpha}{p}$ follows.

A decomposable regularizer is a norm penalizing deviations from the model subspace M as much as possible. Using first-order Taylor series approximation, we can derive a quadratic lower bound on the quadratic error.

Defining $Loss(\Omega) = \frac{1}{2}||\hat{\Sigma} - \aleph(\Omega)||_{Fro}$, we have

$$Loss(\Omega + \Delta) - Loss(\Omega) - \triangle Loss(\Omega)^T\Delta = \frac{1}{2}||\aleph(\Delta)||^2_{Fro}.$$

The Strong Convexity condition provides us a lower bound on $\frac{1}{2}||\aleph(\Delta)||^2_{Fro}$, stating:

$$\frac{1}{2}||\aleph(\Delta)||^2_{Fro} \geq \frac{\gamma}{2}||\Delta||_{Fro},$$

where $\gamma > 0$ is the strong convexity constant.

The Restricted Strong Convexity (RSC) condition prescribes:

$$\frac{1}{2}||\aleph(\Delta)||^2_{Fro} \geq \frac{\gamma}{2}||\Delta||^2_F - \tau_n\Phi^2(\Delta),$$

where $\gamma > 0$, τ_n depends on the mapping operator $\aleph$ (and decreases as $n \to 0$), $\Phi(\Delta)$ is defined in (4.12), and

$$Q(S, L) := ||L||_* + \frac{\rho}{\lambda}\mathscr{R}(S^*).$$

The sample size n is not a problem until τ_n is sufficient large (large as long as $\gamma > 0$). We underline the particular role of n: since this approach provides deterministic guarantees, n serves to improve the approximation of $\frac{1}{2}||\aleph(\Delta)||^2_{Fro}$, That is, the larger n, the more precise is the observation model, and the smaller can be τ_n. However, in our particular case we have $\tau_n = 0$ (identity operator), and $\gamma = 1$. Note that $\Phi^2(\Delta)$ is a measure of relative importance of the regularizer respect to the nuclear norm.

The RSC condition is the key to provide non-asymptotic error bound rates, bounding the absolute losses provided that $\varphi(L) \leq \alpha$. If $\mathscr{R}$ is a decomposable regularizer, it was proved in [84] that the associated statistical models satisfy the RSC condition. Therefore, in that case the authors proved that it is straightforward to obtain non-asymptotic error bounds, and that the M-estimators minimizing a composite regularizer (the loss term plus a decomposable regularizer) converge fast. In this sense, [1] is an extension of [84], where both the nuclear norm (which is also a decomposable regularizer) and a general decomposable regularizer represent the composite penalty. Roughly speaking, we can say that [1] represents the meeting point of [85] and [84].

Another key element of this approach concerns the error composition. Let us define $\Delta_\Sigma = \hat{\Sigma} - \Sigma^*$, $\Delta_L = \hat{L} - L^*$, $\Delta_S = \hat{S} - S^*$. For Cauchy-Schwartz inequality, $||\Delta_\Sigma||^2_{Fro} \leq ||\Delta_L||^2_{Fro} + ||\Delta_S||^2_{Fro}$. Therefore, in the noisy setting, under the numerical approach, the quantity to lower bound is

$$e^2(\hat{L}, \hat{S}) = ||\Delta_L||^2_{Fro} + ||\Delta_S||^2_{Fro}.$$

This choice has to be discussed. It is intuitive that bounding $||\Delta_L||^2_{Fro} + ||\Delta_S||^2_{Fro}$ can be quite different from bounding $||\Delta_\Sigma||^2_{Fro}$. More details and a proposal on this topic can be found in paragraph (5.1).

Given our observation model $\hat{\Sigma} = \aleph(S^* + L^*) + W$, under $\varphi_R(S^*) \leq \alpha$ and the RSC condition, the error $e^2(\hat{L}, \hat{S})$ is bounded by three terms: one in L^*, one in S^*, one depending on τ_n. Each term is composed by two summands: an estimation error term, measuring the error on the subspace M, and an approximation error term, due to the fact that approximately low rank and sparse matrices are allowed. The second one, which was absent in previous approaches, measures the error on the orthogonal complement $M^\perp$ (these terms include $\lambda_j(L^*)$, $r+1, \ldots, p$ for L^*, and the regularizer of the projection of S^* in the orthogonal complement, for us equal to $\sum_{j,k \notin supp(S^*)} ||S^*_{jk}||$). Since $\tau_n = 0$ and we seek for exactly low rank and sparse matrices, in our case we do not have the third error component and we do not allow for approximation errors.

Their general theorem states that under two specific regularity conditions involving r, $\Psi(M, R)$, λ, ρ proportionally to τ_n and γ, and under lower bounds for λ and ρ, there are three limiting universal constants limiting each of the three error terms. The strength on the bound depends on the strength of the RSC condition respect to the curvature of $Loss(\Omega)$. For the entire statement, we refer to [1], p. 1182.

The bound on the curvature will be relevant also in the approach we are going to present in paragraph (4.1.3). While here the convexity structure of $Loss(\Omega)$ is enforced via the l_∞ norm (dual of the l_1 norm) of the low rank component, there the curvature of the low rank matrix variety is bounded, and the Lagrangian dual subgradient approach is applied. The method we

will present allows to identify the model, since it prescribes, following [30], symmetric assumptions respect to BOTH components controlling entirely the interaction of the two spaces. Here, an analytical control based only on the regularizer (thus asymmetric) is imposed to the low rank component. In [31], a bound on the norm of the projection onto the orthogonal complement is given for BOTH matrix spaces simultaneously. This allows perfect identification.

If $\tau_n = 0$ and in the exact matrix setting:

$$e^2(\hat{L}, \hat{S}) \preceq \lambda^2 r + \rho^2 \Psi(M, R)^2$$

up to constant factors. In our case $\Psi(M||.||_1) = \sqrt{s}$. In this approach r and s are chosen adaptively. If we choose $r = rank(L^*)$ and $s = |supp(S^*)|$, we have

$$e^2(\hat{L}, \hat{S}) \preceq \lambda^2 r + \rho^2 s.$$

If $W = 0$ (noiseless setting), for specialization we have

$$e^2(\hat{L}, \hat{S}) \preceq \alpha^2 \frac{s}{p^2}.$$

This rate is weaker respect to the one in [30], but requires weaker conditions on L^*. Anyway, mini-max properties show that in the noiseless setting the rate $\alpha^2 \frac{s}{p^2}$ cannot be improved if $s \leq p$. In addition, we have to consider that the allowed classes of low rank and sparse matrices are much wider.

The lower bounds for threshold parameters here depend on functional norms $||\aleph^*(W)||_{op}$, and $||\aleph^*(W)||_\infty$, as well as on γ, p and α. $||\aleph^*(W)||_{op}$ is here simply the spectral norm of the dual operator at W.

Suppose now we have a stochastic error W generated with normal entries $N(0, \frac{\sigma^2}{n})$. If we set $\aleph = I$, specific threshold values can be found. Under the described conditions, using large deviation theory and some non-asymptotic random matrix theory results to bound $||W||_{op}$ and $||W||_\infty$, we have that for specific threshold parameters, with very high probability, an error rate composed by the noise variance times the usual two error components, function of p, r, s and α, holds.

If we allow W to be a zero-mean Wishart, we fall back into the pure sparse factor analysis case (3.1), which is relevant for our purpose. We now recall it.

Let us suppose $L^* = UDU' = BB'$, where $B = UD^{1/2}$, U is a $p \times r$ matrix, D is a $r \times r$ diagonal matrix, with $d_{jj} > 0$, $\forall j = 1, \ldots, r$. Suppose that our $p \times 1$ random vector X_i, $i = 1, \ldots, n$, has the following structure:

$$X_i = Bf_i + \epsilon_i,$$

with

$$f_i = N_r(0, I_r),$$
$$\epsilon_i = N_p(0, S^*),$$

where f_i is a $r \times 1$ random vector, and ϵ_i is $p \times 1$ random vector.

X_i is assumed to be a zero mean random vector, without loss of generality. The observation matrix is the sample covariance $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$. The error term $W := \frac{1}{n}\sum_{i=1}^{n} X_i X_i' - (BB' + S^*)$ is a zero-mean re-centered Wishart matrix noise.

The Corollary relative to this case is the following.

Corollary 4.1.1. *Consider the factor analysis with $n \geq p$ samples, and regularization parameters*

$$\lambda = ||\sqrt{\Sigma^*}||_2 \sqrt{\frac{p}{n}} \; and \; \rho = 32\rho(\Sigma^*) + \frac{4\alpha}{p} \tag{4.14}$$

where $\rho(\Sigma^*) = max_j \Sigma^*_{jj}$. Then with probability greater than $1 - c_2 exp(-c_3 \log(p))$, any optimal solution $(\hat{L}, \hat{S})$ satisfies

$$e^2(\hat{L}, \hat{S}) \leq c_1 \left\{ ||\Sigma^*||_2 \frac{rp}{n} + \rho(\Sigma^*) \frac{s \log p}{n} \right\} + c_1 \frac{\alpha^2 s}{p^2}. \tag{4.15}$$

This result is derived using large deviation theory and some non-asymptotic random matrix theory results, which allow (under the Wishart assumption) to translate norms of W into norms of Σ^*. It states that under specific threshold choices, involving the spectral norm and the maximum diagonal term of Σ^*, the error is bounded with very high probability by three terms, one representing the degrees of freedom of L^*, $||\Sigma^*||\frac{rp}{n}$ (rp is the number of loadings), one representing all possible sparsity patterns of S^*, $\rho(\Sigma^*)\frac{s \log p}{n} \approx \frac{s \log p}{n}$ (number of subsets of size s from $\mathbb{R}^{p \times p}$), and a term deriving from the non-identifiability issue $\frac{\alpha^2 s}{p^2}$. As usual, the condition $n \geq p$ is necessary in order to obtain consistent estimates in factor analysis model using $\hat{\Sigma}_n$.

Note that now we find again the usual rates $||\Sigma^*||\frac{p}{n}$ and $\rho(\Sigma^*)\frac{s \log p}{n}$ described in [39] and [15]. Terms $||\Sigma^*||_2$ and $\rho(\Sigma^*)$ are present for probabilistic reasons, using standard tail bounds for random Gaussian matrices and their product (see supplementary material to [1], p.35). This is why the threshold parameters λ and ρ have the shape of (4.14). The two terms are weighted by r and s respectively: this is a major difference with the algebraic approach, where r and s have no impact, because there, differently from here, they are implicitly incorporated in the threshold parameters. On the contrary, the probabilistic argument depends in that context on $\hat{\Sigma}_n$ as a whole. The condition $n \geq p$ is the same: however, it is easier trying to overcome it working on $\hat{\Sigma}_n$ under specific model assumptions than using probability assumptions on matrix W.

Finally, we mention a very interesting approach to the same problem, in our case (exact low rank/sparse matrices, identity operator $\aleph = I$): Hsu et al. (2012) [67]. That work is based on rank-sparsity incoherence, and uses the standard singular vector incoherence conditions of [24] deriving

non-asymptotic rates depending on those quantities using the sub-gradient method. In particular, since they employ the orthogonal singular vector incoherence bound $||UV||_\infty$, they need to impose a bound on the product rs:

$$rs \preceq \frac{p^2}{2 \log p}.$$

This bound is not present in the algebraic approach we are going to describe in paragraph (4.1.3), since rank-sparsity incoherence is enforced bounding quantities related to the tangent spaces to the reference varieties (see paragraph (4.1.1)). For a comparison between this approach and the analytical one see [1], p.1188.

We now introduce the algebraic approach by Chandrasekaran et al. (2012) for approximate matrix recovery.

4.1.3 Approximate recovery: an extended algebraic approach

The method we are going to describe now is the core of our thesis. This approach, by Chandrasekaran et al. (2012) ([31]), provides a numerical heuristics for inverse covariance matrix estimation under the Gaussian assumption, exploiting the tools of graphical modelling. From a certain point of view, we could say this is the extension of the graphical lasso for sparse inverse covariance estimation by Friedman, Hastie and Tibshirani ([53]). The affinities are in the estimation target (the precision matrix), in the nature of the minimization target (they both are likelihood methods), in the Gaussian assumption for the data and in the use of the l_1 heuristics (sparsity assumptions).

In contrast, while the graphical lasso imposes sparsity on the overall covariance matrix, the approach in [31] uses the same assumption on the residual component of the model. This solution is based on the strong link between Gaussian random variables and graphical modelling such that the Schur complement of Σ^* is directly modelled. The chosen conditioning block is a vector of $r \ll p$ latent variables which are assumed to explain a large part of the covariances among variables, and the residual covariance is supposed to be sparse. Since the Schur complement of the covariance matrix of a Gaussian random vector is the covariance matrix of the variables conditioned to the the variables belonging to the conditioning block, this model results in a low rank plus structure for the inverse covariance matrix, which is a latent variable graphical model with sparse residual for the data (allowing for missing edges given the latent graphical structure). The problem is solved minimizing the log-likelihood (parameterized in the low rank and in the sparse component) augmented by a composite penalty in the form (3.28), where the nuclear norm regularizes the low rank component and the l_1 norm the sparse one. This is a regularized maximum likelihood program, a convex program tractable via off-the-shelf algorithms ([111]).

What is new, the approach in [31] is algebraic in nature, while the one in [53] is mainly algorithmic (data approximation method). This one provides an algebraic setting for model identifiability and consistent recovery. In addition, this method provides a double notion of consistency: an algebraic one, which describes the correspondence between estimated and theoretical rank and sparsity pattern, and a parametric one, which provides finite bounds for the error rate taking into account simultaneously the low rank and the sparse component. Finally, both consistencies allow (theoretically) $r, p \sim n$, even if there is still the usual problem concerning the use of $\hat{\Sigma}_n$. Here, the condition $n \geq 2p$ is imposed in order to obtain sharper rates.

We now present the model in detail. Consider we have a finite collection of Gaussian random variables $X_O \cup X_H$, where X_O are observed variables and X_H are hidden variables. Call $\Sigma_{O,H}$ the covariance matrix of $X_O \cup X_H$ (in this case we remove * to avoid cluttered notation). $K_{O,H} = \Sigma_{O,H}^{-1}$ is the concentration matrix of the full model. The marginal covariance matrix Σ_O is simply a submatrix of $\Sigma_{O,H}$. Suppose we parameterize the model starting from the concentration matrix $K = \Sigma_{O,H}^{-1}$. The marginal concentration matrix $\tilde{K}_O = \Sigma_O^{-1}$ is given by the Schur complement with respect to block K_H:

$$\tilde{K}_O = \Sigma_O^{-1} = K_0 - K_{O,H} K_H^{-1} K_{H,O}. \tag{4.16}$$

This is a low rank plus sparse structure, where $\Sigma^{-1} = S - L$. The graphical model holds because the covariance matrix of $X_O | X_H$ is Σ_O^{-1}. For $i, j \in O$, due to the joint Gaussian property, $\Sigma_{O,ij}^{-1}$ describes the strength of the relationship between X_i and X_j conditional to X_H. The following relationship holds:

$$cov(X_i, X_j | X_{O \setminus \{i,j\}}) = 0 \Leftrightarrow \Sigma_{O,ij}^{-1} = 0,$$

that is, the is edge between X_i and X_j is missing if the two variables are conditionally independent. Differently from [53], the sparse graphical model is not imposed directly to Σ_O^{-1}, because (conditional) independence is often a too strong assumption in high dimensions. This is why here it is assumed that a number of latent variables X_H, $|H| \ll |O|$, explains most of the observed covariances among the variables in X_O. So, $\tilde{K}_O$ is not sparse in general due to extra-correlations induced from marginalization over the latent variables X_H. The latent variables X_H are also referred to as hidden components. The additional low rank term $K_{O,H} K_H^{-1} K_{H,O}$ summarizes the covariances induced by the marginalization over X_H. Then, it is possible to set up a sparse graphical model on the residual concentration matrix K_0, which summarizes the covariances among the variables in X_O conditioned on the hidden components. From this model framework, a natural low rank plus sparse decomposition for the precision matrix of the observed variables $\tilde{K}_O$ arises, in the form:

$$\tilde{K}_O = S - L,$$

where $S = K_0$ and $L = K_{O,H}K_H^{-1}K_{H,O}$. This framework combines dimensionality reduction (to identify latent variables) and graphical modelling (to catch the residual covariance structure). For us, $|O| = p$, $|H| = r$.

Under these model assumptions, the problem of identifying two matrix varieties, one low-rank (4.3) and one sparse (4.4) naturally arises. We need to uniquely decompose the low rank and the sparse component starting from their sum. This problem is similar to the one presented in (4.1.1), even if random perturbations on the data are allowed. The identification requires to exploit the notion of geometric transversality between tangent spaces $\Omega(K_O)$ $T(K_{O,H}K_H^{-1}K_{H,O})$. We will show that, analogously to [30], if the sparse component has a small number of nonzero elements and the low rank component has row/column spaces not closely aligned to coordinate axes, then the latent variable model is identifiable. However, there is one more problem to face: in the noisy context, the curvature of the low-rank variety (i.e. its local sensitivity to perturbations) plays a relevant role. If we think the two tangent spaces as algebraic systems, we note that the one tangent to the low-rank variety is non-linear, while the other one is linear. For this reason, if $T(K_{O,H}K_H^{-1}K_{H,O})$ is very curve, it may be impossible to identify L in the noisy context, since the tangent space can vary locally very fast. Therefore, a bound on this curvature is necessary. Note that the approach by Agarwal et al. ([1]) does not provide identifiability just because it does not pay attention to this aspect, enforcing assumptions via a pure analytical approach.

The regularized likelihood problem is the following:

$$\hat{S}_n, \hat{L}_n = \arg\min_{S,L} -l(S-L;\hat{\Sigma}_n) + \lambda_n(\gamma||S||_1 + tr(L)) \tag{4.17}$$
$$\text{s.t. } S - L \succ 0 \; L \succeq 0,$$
$$l(K;\Sigma) = \log\det(K) - tr(K\Sigma) \tag{4.18}$$
$$K \succ 0.$$

It is composed by a Gaussian log-likelihood term ($-l(S-L;\hat{\Sigma}_n)$) and the composite penalty (3.28), where the trace is the nuclear norm heuristics over the cone of Positive SemiDefinite matrices (PSD). γ is a trade-off parameter between the trace and the l_1 norm. (4.17) is a regularized max-det problem (a discussion on these problems is in [49]). Note the presence of constraints $S - L \succ 0$ and $L \succeq 0$, which are tractable in this algebraic framework. This is a variational formulation of the problem, which provides also a model selection heuristics: the error term (log-likelihood) is penalized by the model complexity in terms of sparsity of S and spectrum of L. The problem can be easily solved using standard off-the-shelf solvers ([111]).

Due to the log-likelihood term, another identification problem arises. If the log-likelihood is too curve, i.e. if the Fisher information behaves poorly respect to the tangent spaces $T(L)$ and $\Omega(S)$ (and their sum), errors in

the data are amplified too much, creating an additional identification issue. Functional operator theory is crucial in this context. The curvature of Fisher information $\mathbb{I}^*$ as well as the curvature of the low rank variety are described and bounded as functional operators.

A formal statement of the latent variable model selection problem is reported below ([31]).

Definition 4.1.1. *A pair of symmetric matrices (S, L) with $S, L \in R^{|O|\times|O|}$ is an algebraically consistent estimate of a latent-variable Gaussian graphical model given by the concentration matrix K_{OH} if the following conditions hold:*

1. *The sign pattern of S is the same of K_O: $sign(S_{ij}) = sign((K_O)_{i,j})$, $\forall i, j$. Here we assume that $sign(0) = 0$.*
2. *The rank of L is the same as the rank of $K_{O,H}K_H^{-1}K_{H,O}$.*
3. *The concentration matrix $S - L$ can be realized as the marginal concentration matrix of an appropriate latent-variable model: $S - L \succ 0$, $L \succeq 0$.*

Model consistency here is defined according to the following three estimation features:

1. correct structural estimate of the conditional graphical model (given by K_0) of the observed variables conditioned on the hidden components. This feature is called "sparsistency" of standard graphical model selection.
2. number of hidden components correctly estimated.
3. the model is realizable: $|O \cup H| = |O| + |H|$.

It is also defined the usual parametric consistency, which holds if the estimates of (S, L) are close to $(K_O, K_{O,H}K_H^{-1}K_{H,O})$ in some norm with high probability. Parametric consistency does not imply algebraic consistency and vice versa. Besides, the model suffers from the usual model indeterminacy coming from a latent variable context: there are infinite $K_H \succ 0$, $K_{O,H} = K'_{H,O}$ giving rise to the same low-rank matrix $K_{O,H}K_H^{-1}K_{H,O}$.

Consistently to their geometric approach (and to identifiability conditions), the reference norm to assess parametric consistency is nothing but the dual norm of the composite penalty. Given the norm

$$f_\gamma(S, L) = \gamma||S||_1 + ||L||_*, \tag{4.19}$$

$\gamma > 0$, where $||L||_* = tr(L)$ (since L is over the cone of PSD), the dual norm of $f_\gamma(S, L)$, which is used to bound the error, is

$$g_\gamma(S, L) = \max\big\{\frac{||S||_\infty}{\gamma}, ||L||_2\big\}. \tag{4.20}$$

Identification and recovery: technical aspects

Suppose we have n samples $(X_O^i)_{i=1}^n$ of the observed variables X_O. $X_i, i = 1, \ldots, n$, are jointly Gaussian zero-mean p - dimensional random variables. The latent variable model holds on the marginal concentration matrix K.

We define the induced operator norm of a linear bounded operator $Z: R^{p\times p} \to R^{p\times p}$ as:

$$||\mathbb{Z}||_{q\to q} = \max_{N\in R^{p\times p}, ||N||_q \leq 1} ||\mathbb{Z}(N)||_q.$$

The covariance matrix is the usual $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n X_{O,i}X'_{O,i}$. The log-likelihood of K is

$$l(K; \hat{\Sigma}_n) = \log\det(K) - tr(K\hat{\Sigma}_n),$$

function of K.

Applying Jacobi's formula we have ([120] [121])

$$\frac{\delta^2}{dK^2} tr(K\hat{\Sigma}_n) = \frac{\delta}{dK}\hat{\Sigma}_n I = \mathbf{0}.$$

As a consequence,

$$\frac{\delta^2}{dK^2} dK \log\det(K) = \frac{\delta}{dK} tr(K^{-1}) = -K^{-1}K^{-1},$$

which results in

$$\frac{\delta^2}{dK^2} l(K; \hat{\Sigma}_n) = -K.$$

This result means that $l(K; \hat{\Sigma}_n)$ is strictly concave for $K \succ 0$, i.e. $-l(K; \hat{\Sigma}_n)$ is strictly convex.

Consider now the latent variable model (4.16) for $\tilde{K}_O = (\Sigma_O)^{-1}$, where $S = K_0$ represents the conditional statistics of X_O given some extra variables X_H, and $L = K_{O,H}K_H^{-1}K_{H,O}$ summarizes the effect of marginalization on X_O over X_H. Respect to (S, L), $\bar{l}(S, L, \Sigma_n) = l(S - L, \Sigma_n)$ is jointly concave whenever $S - L \succ 0$.

We know that Fisher information is the negative Hessian of the likelihood function and thus controls the curvature of Fisher information operator $\mathbb{I}$. Its formulation is

$$\mathbb{I}(K) = -\Delta_K^2 \log\det(K)|_K = K \otimes K$$

for $K \succ 0$. If K is $p \times p$, $\mathbb{I}(K)$ is $p^2 \times p^2$.

Considered that $\tilde{K_O}^* = (\Sigma_O^*)^{-1}$, for model (4.16) we have:

$$\mathbb{I}(\tilde{K}_O) = \tilde{K}_O^{-1} \otimes \tilde{K}_O^{-1} = \Sigma_O \otimes \Sigma_O. \tag{4.21}$$

This matrix is precisely the $|O|^2 \times |O|^2$ sub matrix of

$$\mathbb{I}(\tilde{K}_{(OH),(i,j)(k,l)}) = [\Sigma_{(O,H)} \otimes \Sigma_{(O,H)}]_{(i,j)(k,l)},$$

which is $|O \cup H|^2 \times |O \cup H|^2$, given that $\tilde{K}_{(OH)} = (\Sigma_{(O,H)})^{-1}$.

Bounding $\mathbb{I}(\tilde{K}_O)$ is crucial for obtaining consistent estimates with high probability from (4.17).

As previously explained, the tangent space T to the low-rank matrix variety is locally curved at any smooth point. Results from perturbation matrix theory are needed in order to bound the curvature of T, which may affect the identification of the unknown varieties. The curvature of T at any smooth point M (symmetric and having rank less or equal to r) can be described in terms of projection onto the row space $U(M)$ (denoted by $P_{U(M)}(N)$) as follows (see [9] p.15):

$$\mathbb{P}_{T(M)}(N) = P_{U(M)}N + NP_{U(M)} - P_{U(M)}NP_{U(M)}$$

where operator $\mathbb{P}$ is the (bounded) projection operator and N is any squared matrix. $T(M)$ is curved because the projection changes locally around M (differently from $\Omega(M)$, which has curvature 0 at any smooth point). The curvature is the "angle" between the tangent space at any smooth point and the tangent space at a neighboring point.

It is therefore necessary to bound the curvature. The twisting between two subspaces of matrices T_1 and T_2 is defined as:

$$\rho(T_1, T_2) = ||P_{T_1} - P_{T_2}||_{2\to 2} = \max_{||N||_2 \leq 1} ||P_{T_1} - P_{T_2}(N)||_2.$$

It is proved that perturbing a rank-r matrix M with a matrix Δ such that $||\Delta||_2 \leq \frac{\sigma}{8}$ and $M + \Delta$ has rank r, the following two results which bound the twisting between tangent spaces at nearby points hold:

$$\rho(T(M+\Delta), T(M)) \leq \frac{2}{\sigma}||\Delta||_2 \tag{4.22}$$

$$||P_{T(M)^\perp}||_2 \leq \frac{||\Delta||_2^2}{\sigma}, \tag{4.23}$$

where σ is the smaller singular value of M. So, lower bounding σ, which is for covariance matrices simply the smallest eigenvalue, means controlling the curvature of T. The closer σ is to 0, the more curved T is locally.

Analogously to [30], quantities $\mu(K_0)$ and $\xi(K_{O,H}K_H^{-1}K_{H,O})$ play a key role for identification. A useful Lemma links the twisting between two subspaces $\rho(T_1, T_2)$ (if smaller than 1) and parameters $\xi(T_1)$, $\xi(T_2)$ as follows:

$$\xi(T_2) \leq \frac{1}{1 - \rho(T_1, T_2)}[\xi(T_1) + \rho(T_1, T_2)].$$

This allows to conclude that we consider all the neighbour subspaces T' satisfying $\rho(T', T) \leq \frac{\xi(T)}{2}$ as close to T.

We can now approach the problem of local identifiability of the sparse and the low rank component from their observed sum. Define the addition operator $\mathbb{A}(S, L) = S + L$, its adjoint $\mathbb{A}^\dagger$ s.t. $< \mathbb{A}x, y >=< x, \mathbb{A}^\dagger y >$ for all $x, y \in H$, H Hilbert space ($<>$ is the standard Euclidean inner product). $\mathbb{A}^\dagger(S, L) = (S + L)' = S + L$ (since both components are symmetric). $\mathbb{A}$ and $\mathbb{A}^\dagger$ are both linear bounded (hence continuous) operators.

The identifiability of tangent spaces $T(L)$ and $\Omega(S)$ is possible if and only if they have a sufficient degree of transverse intersection, which means they are sufficiently distinct. This condition depends, as described in paragraph (4.1.1), on quantities $\xi(T)$ and $\mu(\Omega)$; in this context, since transversality is not perfect, we need also to quantify and bound the level of transversality between the two spaces with reference to the Cartesian product $\mathbb{Y} = \Omega \times T$. This is unavoidable to provide necessary and sufficient conditions for identifiability from the Maximum Likelihood (ML) regularized program (4.17).

The minimum gain with respect to some norm $||.||_q$ on $\mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}$ of the addition operator $\mathbb{A} : \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ restricted to the cartesian product $\mathbb{Y} = \Omega \times T$ is defined as:

$$\epsilon(\Omega, T, ||||_q) = \min_{(S,L)\in\Omega\times T, ||(S,L)||_q=1} ||\mathbb{P}_\mathbb{Y}\mathbb{A}^\dagger\mathbb{A}\mathbb{P}_\mathbb{Y}(S, L)||_q,$$

where $\mathbb{P}_\mathbb{Y}$ is the projection operator onto $\mathbb{Y}$ and the products are Cartesian products.

Quantity $\epsilon(\Omega, T, ||.||_q)$ measures the level of transversality. The large it is, the more transverse $T(L)$ and $\Omega(S)$ are. The tangent spaces have a transverse intersection if and only if

$$\epsilon(\Omega, T, ||.||_q) > 0.$$

Since we have $A^\dagger A(S, L) = (S + L, S + L)$ and $P_\mathbb{Y}A^\dagger AP_\mathbb{Y}(S, L) = (S+P_\Omega(L), P_T(S)+L)$, this condition is equivalent to bound the projection of each component onto the other space, in order to avoid the misidentification of each component. This why we want $\epsilon(\Omega, T, ||.||_q)$ to be as large as possible.

As the subdifferential of the regularization function (4.19) is specified in terms of its dual norm (4.20), the natural norm $||.||_q$ to measure transversality is the dual norm of the regularization function (4.20).

Given Ω and T, tangent space to varieties S, L and their Cartesian

product $\mathbb{Y} = \Omega \times T$, the following bounded linear operator properties hold:

$$\begin{aligned} ||P_\Omega||_\infty &\leq ||M||_\infty \\ ||P_{\Omega^\perp}||_\infty &\leq ||M||_\infty \\ ||P_T(M)||_2 &\leq 2||M||_2 \\ ||P_{T^\perp}(M)||_2 &\leq ||M||_2 \\ g_\gamma(P_\mathbb{Y}(M,N)) &\leq 2g_\gamma(M,N) \\ g_\gamma(P_{\mathbb{Y}^\perp}(M,N)) &\leq g_\gamma(M,N). \end{aligned}$$

These properties are used for the subgradient minimization process. Note that the projection rule for the $||.||_2$ norm of the projection doubles the corresponding norm of the argument, differently from other norms. See [96] for more explanations.

Defining

$$\chi(\Omega, T, \gamma) = \max\left\{\frac{\xi(T)}{\gamma}, 2\mu(\Omega)\gamma\right\},$$

we can study the transversality respect to g_γ, obtaining the following crucial result:

Lemma 4.1.1. *Given, $S \in \Omega$, $L \in T$, with $||S||_\infty = \gamma$ and $||L||_2 = 1$, and $\mathbb{Y} = \Omega \times T$, we have:*

$$g_\gamma(P_\mathbb{Y} A^\dagger A P_\mathbb{Y}(S, L)) \in [1 - \chi((\Omega, T, \gamma), 1 + \chi((\omega, T, \gamma)].$$

In particular:

$$1 - \chi(\Omega, T, \gamma) \leq \epsilon(\Omega, T, g_\gamma).$$

This is a stochastic joint (matrix bivariate) isometry property, and is the Restricted Isometry Property (RIP) of this model setting. It allows to lower bound $\epsilon(\Omega, T, g_\gamma)$ and to link transversality to parameters $\mu(\Omega)$ and $\xi(T)$ even in the noisy context. For instance, if $\mu(\Omega)\xi(T) < 1/2$ then $\gamma \in (\xi(T), \frac{1}{2\mu(\omega)})$ implies Ω and T have a transverse intersection.

It is easy to note that the smaller are $\mu(\Omega)$ and $\xi(T)$, the more transverse are Ω and T, exactly as in the noiseless context of paragraph (4.1.1).

Tangent spaces in this framework are precisely defined as

$$\Omega = \Omega(K_O) = \Omega(S) \text{ and } T = T(K_{O,H} K_H^{-1} K_{H,O}) = T(L),$$

where $K_{H,O} = K'_{O,H}$. They both lie in a functional space where the inner product is the Fisher information operator $\mathbb{I}^*$, which is a map between $\mathbb{R}^{p \times p}$ and $\mathbb{R}^{p^2 \times p^2}$. We want that S and L are distinguishable respect to $\mathbb{I}^*$, i.e. to study the behaviour of $\mathbb{I}^*$ restricted to $\Omega \oplus T$, in order to identify and recover S and L by $l(S - L; \hat{\Sigma}_n)$.

In order to do that, we need to study the gains of $\mathbb{I}^*$ restricted to Ω and T separately, as well as their orthogonal complements $\Omega^\perp$ and $T^\perp$, such that

elements in both spaces are identifiable under the map $\mathbb{I}^*$. Finally, conditions to control $\mathbb{I}^*$ restricted to the direct sum $\Omega \oplus T$, in conjunction with bounds on μ and ξ, are provided.

The minimum gain of $\mathbb{I}^*$ restricted to Ω and $\Omega^\perp$ is given by the following quantities:

$$\alpha_\Omega = \min_{M \in \Omega, ||M||_\infty = 1} ||P_\Omega \mathbb{I}^* P_\Omega(M)||_\infty \tag{4.24}$$

$$\delta_\Omega = \min_{M \in \Omega, ||M||_\infty = 1} ||P_{\Omega^\perp} \mathbb{I}^* P_\Omega(M)||_\infty \tag{4.25}$$

$\mathbb{I}^*$ is injective on Ω if $\alpha_\Omega > 0$. The irrepresentability condition, which is a sufficient identification condition for graphical lasso using l_1 regularization problem, is $\frac{\delta_\Omega}{\alpha_\Omega} \leq 1 - \nu$, and is sufficient for consistent recovery of graphical model structure using lasso ([53]). More, the local behaviour of $\mathbb{I}^*(M)$ respect to Ω is described by

$$\beta_\Omega = \max_{M \in \Omega, ||M||_2 = 1} ||\mathbb{I}^*(M)||_2.$$

The same holds for functional operators $P_T^\perp \mathbb{I}^* P_T(M)$ and $P'^\perp_T \mathbb{I}^* P_T(M)$, which describe the behaviour of $\mathbb{I}^*$ restricted to T and T' respectively. Their minimum gain is respectively given by:

$$\alpha_T = \min_{\rho(T,T') \leq \frac{\xi(T)}{2}} \min_{M \in T', ||M||_2 = 1} ||P'_T \mathbb{I}^* P'_T(M)||_2 \tag{4.26}$$

$$\delta_T = \min_{\rho(T,T') \leq \frac{\xi(T)}{2}} \min_{M \in T', ||M||_2 = 1} ||P'^\perp_T \mathbb{I}^* P'_T(M)||_2. \tag{4.27}$$

$\mathbb{I}^*$ injective on all tangent spaces T' such that $\rho(T, T') \leq \frac{\xi(T)}{2}$ if $\alpha_T > 0$. An analogous irrepresentability condition holds for the recovery of T (solely considered): $\frac{\delta_T}{\alpha_T} \leq 1 - \nu$.

The local behaviour of $\mathbb{I}^*(M)$ respect to Ω is described by

$$\beta_T = \max_{\rho(T,T') \leq \frac{\xi(T)}{2}} \max_{M \in T', ||M||_\infty = 1} ||\mathbb{I}^*(M)||_\infty$$

Quantities β_Ω and β_T control the behaviour of $\mathbb{I}^*$ restricted to $\Omega \oplus T$, together with conditions on $\xi(T)$ and $\mu(T)$ coming from Lemma 4.1.1.

Let us now define:

$$\alpha = \min(\alpha_\Omega, \alpha_T) \tag{4.28}$$

$$\beta = \min(\beta_\Omega, \beta_T) \tag{4.29}$$

$$\delta = \min(\delta_\Omega, \delta_T). \tag{4.30}$$

The main assumption on $\mathbb{I}^*$, which summarizes both sets of conditions, is the following:

Lemma 4.1.2. *There exists a* $\nu \in (0, \frac{1}{2}]$ *such that* $\frac{\gamma}{\alpha} \leq 1 - 2\nu$.

We can now report the Proposition of [31] describing the necessary assumptions on parameters for model identification. This statement recaps identifiability conditions related to the curvature of $T(L)$, to Fisher information $\mathbb{I}^*$ and to $\epsilon(\Omega, T, g_\gamma)$.

Proposition 4.1.3 (Chandrasekaran et al. (2012) [31])**.** *Let* T *be as in (4.3),* Ω *be as in (4.4), and let* $\mathbb{I}^*$ *be the Fisher information matrix evaluated at the true* $K = \Sigma_O^{-1}$. *Suppose that*

$$\mu(\Omega)\xi(T) \leq \frac{1}{6}\left(\frac{\nu\alpha}{\beta(2-\nu)}\right)^2,$$

and γ *is in the following range:*

$$\gamma \in \left[\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}\right].$$

Then we have the two following conclusions for $\mathbb{Y} = \Omega \times T'$, *with* $\min \rho(T, T') \leq \frac{\xi(T)}{2}$:

- *The minimum gain of* $\mathbb{I}^*$ *restricted to* $\mathbb{Y} = \Omega \oplus T$ *is bounded below:*

$$\min_{(S,L)\in\mathbb{Y}, ||S||_\infty=\gamma, ||L||_2=1} g_\gamma(P_\mathbb{Y} A^\dagger \mathbb{I}^* A P_\mathbb{Y}(S, L)) \geq \frac{\alpha}{2}.$$

  Specifically this implies for all $(S, L) \in \mathbb{Y}$:

$$g_\gamma(P_\mathbb{Y} A^\dagger \mathbb{I}^* A P_\mathbb{Y}(S, L)) \geq \frac{\alpha}{2} g_\gamma(S, L).$$

- *The minimum effect of elements in* $\mathbb{Y} = \Omega \oplus T$ *on the orthogonal complement* $\mathbb{Y}^\perp = \Omega^\perp \oplus T'^\perp$ *is bounded above:*

$$||(P_{\mathbb{Y}^\perp} A^\dagger \mathbb{I}^* A P_\mathbb{Y}(S, L))(P_\mathbb{Y} A^\dagger \mathbb{I}^* A P_\mathbb{Y}(S, L))^{-1})||_{g_\gamma \to g_\gamma} \leq 1 - \nu$$

  Specifically this implies for all $(S, L) \in \mathbb{Y}$:

$$g_\gamma(P_{\mathbb{Y}^\perp} A^\dagger \mathbb{I}^* A P_\mathbb{Y}(S, L)) \leq (1 - \nu) g_\gamma(P_\mathbb{Y} A^\dagger \mathbb{I}^* A P_\mathbb{Y}(S, L))$$

Another necessary condition to ensure probabilistic consistency is a bound on ψ, the spectral norm of Σ ($\psi = ||\Sigma||_2$). ψ controls also $\mathbb{I}^*$, since it can be noted that here $||\mathbb{I}^*||_{2\to 2} = \psi^2$ (see (4.21)).

We now describe consistency properties of (4.17) in the high dimensional setting, where p, r, n are allowed to grow simultaneously ($n, r \sim p$). For us, $p = |O|$ is the number of observed variables, $r = |H|$ is the number of latent variables, n is the number of samples of the observed variables X_O. $K_{O,H}$ gives the latent variable graphical model whose complexity is explained by

$\mu(\Omega(K_O))$ and $\xi(T(K_{O,H})K_H^{-1}K_{H,O})$, describing the sparsity pattern of the conditional graphical model among the observed variables and the diffusivity of the extra correlations due to marginalization over the hidden variables. Parameters α, β, ν, ψ do not scale with other parameters and are bounded.

There is a natural trade-off between $\mu(\Omega)$ and $\xi(T)$. The classes of latent-variable graphical models which can be identified by (4.17) depend on their relationship, and on corresponding scalings of p, r, n.

In (4.17), γ is a trade-off parameter between rank and sparsity terms, and λ_n is a regularization parameter, which must be suitably chosen to ensure consistency. Since $\xi(T)$ and $\mu(\Omega)$ are not known a priori, a numerical choice for γ must be done too.

We now report the main result on model selection consistency.

Theorem 4.1.3 ([31]). *Let K_{OH} denote the concentration matrix of a Gaussian model. We have n samples $X_i, i = 1, \ldots, n$ p of the observed variables denoted by O. Let $\Omega = \Omega(K_O)$ and $T = T(K_{O,H}K_{O,H}K_H^{-1}K_{H,O})$ denote the tangent spaces at K_O and at $K_{O,H}K_{O,H}H^{-1}K_{H,O}$ with respect to the sparse and low-rank matrices respectively.*

***Assumptions**: Suppose the following conditions hold:*

1. *The quantities $\mu(\Omega)$ and $\xi(T)$ satisfy the assumption of Proposition 4.1.3 for identifiability, and γ is chosen in the range specified by Proposition 4.1.3.*

2. *The number of samples n available is such that*

$$n \succeq \frac{p}{\xi(T)^4}.$$

3. *The regularization λ_n is chosen as*

$$\lambda_n \asymp \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}.$$

4. *The minimum nonzero singular value σ of $K_{O,H}K_H^{-1}K_{H,O}$ is bounded as*

$$\sigma \succeq \frac{1}{\xi(T)^3}\sqrt{\frac{p}{n}}.$$

5. *The minimum magnitude nonzero entry θ of K_O^* is bounded as*

$$\theta \succeq \frac{1}{\xi(T)\mu(\Omega)}\sqrt{\frac{p}{n}}.$$

***Conclusions**: Then with probability greater than $1 - 2\exp(p)$ we have:*

1. *Algebraic consistency: The estimate* $(\hat{S}_n, \hat{L}_n)$ *given by (4.17) is algebraically consistent, i.e., the support and sign pattern of* $\hat{S}_n$ *is the same as that of* K_O*, and the rank of* $\hat{L}_n$ *is the same as that of* $K_{O,H}K_H^{-1}K_{H,O}$*.*

2. *Parametric consistency: The estimate* $(\hat{S}_n, \hat{L}_n)$ *given by the convex program (4.17) is parametrically consistent:*

$$g_\gamma(\hat{S}_n - K_O, \hat{L}_n - K_{O,H}K_H^{-1}K_{H,O}) \preceq \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}.$$

We can note that both components are algebraically and parametrically consistent, given a number of conditions involving the minimum nonzero entry of K_O and the minimum singular value of $K_{O,H}K_H^{-1}K_{H,O}$, the number of samples n (which are lower bounded) and the regularization parameter λ_n (which follows a precise scale). $(\hat{S}_n, \hat{L}_n)$ are thus ensured not to have smaller support size/rank than $(K_O, K_{O,H}K_H^{-1}K_{H,O})$. The condition on the minimum singular value is more stringent than the one on the minimum non zero elements, because it plays a crucial role to bound the curvature of $T(L)$ around $K_{O,H}K_H^{-1}K_{H,O}$. Relevant parameters for consistency are p, n, μ, ξ. This result will be the key to prove consistency of the low rank plus sparse covariance estimator by Luo (2013) [77] we will describe in paragraph (4.1.4).

All the results hold under the conditions of Proposition 4.1.3, especially under the condition $\gamma \in [\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}]$. Theorem 4.1.3 is derived using the lower end of the range for γ.

If this assumption is weakened, we have the following Corollary.

Corollary 4.1.2. *Consider the same setup and notation as in Theorem 4.1.3. Suppose that the quantities* $\mu(\Omega)$ *and* $\xi(T)$ *satisfy the assumption of Proposition 4.1.3 for identifiability. Suppose that we make the following assumptions:*

1. *Let* γ *be chosen to be equal to* $\frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}$ *(the upper end of the range specified in Proposition 4.1.3), i.e.* $\gamma \asymp \frac{1}{\mu(\Omega)}$*.*

2. $n \succeq \mu(\Omega)^4 p$*.*

3. $\lambda_n \asymp \mu(\Omega)\sqrt{\frac{p}{n}}$*.*

4. $\sigma \succeq \frac{\mu(\Omega)^2}{\xi(T)}\sqrt{\frac{p}{n}}$

5. *The minimum magnitude nonzero entry* θ *of* K_O^* *is bounded as* $\theta \succeq \sqrt{\frac{p}{n}}$*.*

Then with probability greater than $1 - 2\exp(p)$ *we have estimates* $(\hat{S}_n, \hat{L}_n)$ *that are algebraically consistent, and parametrically consistent with the error bounded as*

$$g_\gamma(\hat{S}_n - K_O, \hat{L}_n - K_{O,H}K_H^{-1}K_{H,O}) \preceq \mu(\Omega)\sqrt{\frac{p}{n}}.$$

Theorem 4.1.3 and Corollary 4.1.2 describe the extremes of matrix classes recoverable using program (4.17). In practice, a range of values for γ is necessary in order to ensure the stability of the sparsity pattern and the rank, while λ_n is usually taken in a range of values proportional to $\sqrt{\frac{p}{n}}$.

Recalling results (4.10) and (4.11), we can define $d = deg(K_O)$, degree of the conditional graphical model among the observed variables, and $i = inc(K_{O,H}(K_H)^{-1}K_{H,O})$, incoherence of the covariances due to the marginalization over the latent variables. The following relations hold:

$$\mu \leq d, \qquad \xi \leq 2i.$$

Since α, β, ν, ψ are assumed to be bounded, from Proposition 4.1.3 we have

$$di = O(1).$$

These conditions include non-trivial classes of latent-variable graphical models. In particular, we mention the case of constant degree $d = O(1)$ and maximum incoherence $\sqrt{r/p}$, with $r \sim p$. In this setting, the effect of marginalization over latent variables is diffuse almost across ALL variables. Consistent recovery is allowed also from $n \sim p$ samples, even if condition $n \geq 2p$ is here specified following [39] in order to ensure finite bounds for $\hat{\Sigma}_n$.

From this results, rates for the covariance matrix (i.e. the inverse of the precision matrix) can be easily derived as follows.

Corollary 4.1.3. *Under the same conditions of Theorem 4.1.3, we have with probability greater than* $1-2\exp(p)$ *that* $g_\gamma(A^\dagger[(\hat{S}_n-\hat{L}_n)^{-1}-\Sigma_O^*]) \preceq \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}$. *Specifically, this implies that*

$$||(\hat{S}_n - \hat{L}_n)^{-1} - \Sigma_O^*||_2 \preceq \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}. \tag{4.31}$$

Rates for $\hat{\Sigma} = \hat{S}_n - \hat{L}_n$ and $\hat{\Sigma}^{-1}$ coincide, and are proportional to $\sqrt{\frac{p}{n}}$. However, using the (inverse) sample covariance matrix as an input, these results hold if and only if $n \geq 2p$.

We finally give some basic notes on the proof strategy. These concepts will be recalled while showing the analogous proof from [77] in paragraph (4.1.4). Standard results from [98] state that $(\hat{S}_n, \hat{L}_n)$ is a minimum for (4.17) if the zero matrix belongs to the subdifferential of the objective function evaluated at $(\hat{S}_n, \hat{L}_n)$. The subdifferential structure of $||.||_1$ and $||.||_*$ is the following. The subdifferential of the l_1 norm at a symmetric matrix M is:

$$N \in \delta||M||_1 \Leftrightarrow P_{\Omega(M)}(N) = sign(M),\ ||P_{\Omega(M)^\perp}(N)||_\infty \leq 1.$$

Let $M = UDU'$ be a symmetric positive semidefinite matrix M. The subdifferential of the trace function restricted to the cone of positive semidefinite matrices (i.e. the nuclear norm over this set) is:

$$N \in \delta[tr(M) + I_{M \succeq 0}] \Leftrightarrow P_{T(M)}(N) = UU', P_{T(M)^\perp}(N) \preceq I_p,$$

where $I_{M \succeq 0}$ evaluates to 0 over the cone of PSD and to ∞ otherwise, and the condition on $T(M)^\perp$ indicates that the spectral norm of $P_{T(M)^\perp}(N)$ is smaller or equal to 1.

The key point for proving Theorem 4.1.3 is that elements of the subdifferential decompose with respect to the tangent spaces $\Omega(M)$ and $T(M)$.

In order to solve (4.17), it is necessary to add the non-convex constraints $S \in \mathscr{K}(s)$ and $L \in \mathscr{L}(r)$. The pair $(\tilde{S}, \tilde{L})$ solution of this problem is proved to be composed by smooth points of $\mathscr{K}(s)$ and $\mathscr{L}(r)$ respectively. The first-order optimality condition state that the Lagrange multipliers corresponding to the additional variety constraints must lie in $\Omega(S)^\perp$ and $T(L)^\perp$, such that the first part of the subgradient optimality conditions of (4.17) is respected. Then, the idea is to prove that the variety-constrained program is algebraically equivalent to the tangent-space constrained program, where $S \in \Omega(S)$ and $L \in T(L)$. Finally, it is proved that tangent-space constraints are locally inactive, such that the original problem (4.17) has the same solution.

Therefore, the second part of the subgradient conditions (relative to the components in $\Omega^\perp$ and $T^\perp$) is also satisfied and the solution of the original problem shares the same algebraic and parametric consistency properties with the variety-constrained program.

This approach is valid if and only if the twisting between $T(\tilde{L})$ and $T(K^*_{O,H} K_H^{-1} K_{H,O})$ is bounded. This why the minimum singular value of $K^*_{O,H} K_H^{-1} K_{H,O}$ is lower bounded, thus providing the local identifiability of $T(L^*)$. The entire proof exploits the basic matrix property $||M||_\infty \leq ||M||_2$.

We will give details on the steps needed to prove the analogous of Theorem 4.1.3 into the covariance matrix context in paragraph (4.1.4).

We now outline the optimality conditions of our problem (4.17). Our convex objective at the optimum $(\hat{S}_\Omega, \hat{L}_{T'})$ satisfies, for some Lagrangian multipliers $Q_{\Omega^\perp}$ and $Q_{T'^\perp}$, the following conditions:

$$\hat{S}_\Omega + \hat{L}_{T'} - \hat{\Sigma}_n + Q_{\Omega^\perp} \in -\lambda_n \gamma \delta ||\hat{S}_\Omega||_1,$$

$$\hat{S}_\Omega + \hat{L}_{T'} - \hat{\Sigma}_n + Q_{T'^\perp} \in -\lambda_n \delta ||\hat{L}_T||_*.$$

The key to derive the solution is to project $\hat{S}_\Omega + \hat{L}_{T'} - \hat{\Sigma}_n$ onto $\mathbb{Y} = \Omega \times T'$ and to define

$$P_\Omega(\hat{S}_\Omega + \hat{L}_{T'} - \hat{\Sigma}_n) = Z_\Omega,$$

$$P_T(\hat{S}_\Omega + \hat{L}_{T'} - \hat{\Sigma}_n) = Z_{T'},$$

with $||Z_\Omega||_\infty = \lambda_n \gamma$ and $||Z_{T'}|| \leq 2\lambda_n$. The bi-dimensional projection is

$$P_{\mathbb{Y}} \mathbb{A}^\dagger (\hat{S}_\Omega + \hat{L}_{T'} - \hat{\Sigma}_n) = Z = (Z_\Omega, Z_{T'}),$$

where $\mathbb{Y} = \Omega \oplus T'$. This is the projected gradient method, and provides the mathematical base to algebraically solve the numerical problem (3.30).

4.1.4 Approximate recovery: LOREC approach

This section deals with covariance matrix estimation via low rank plus sparse decomposition. Here we describe the numerical approach of Luo (2013) ([77]) which recovers the covariance matrix via low rank plus sparse decomposition in the noisy setting. This approach moves from the one of [31] described in paragraph (4.1.3), and provides rates and identifiability conditions under the same algebraic setting.

The underlying structure for Σ^* is model (4.2), and the data structure is the one described in (3.1). Model (4.2) can be thought of as a general approximate factor model in the form

$$\Sigma^* = BVar(f)B' + \Sigma_\epsilon,$$

where $Var(f) = I_r$ and $\Sigma^* - \Sigma_\epsilon$ has exactly rank r. The low rank matrix $L^* = BVar(f)B'$ and the sparse matrix $S^* = \Sigma_\epsilon$ are symmetric (as well as their sum Σ^*) . Our sample estimate $\hat{\Sigma}$ is drawn from the noisy model

$$\hat{\Sigma} = L^* + S^* + W$$

where W is an error term.

At present, the recovery of the loading matrix B via the method we are going to describe has not be discussed. This can be partially done only if $r = 1$, where the loadings is recovered up to a constant. The factor model assumption is here used as a useful tool to estimate the covariance matrix in a large dimensional context.

The usual matrix spaces $\mathscr{L}(r)$, $\mathscr{K}(s)$, $T(L)$ and $\Omega(S)$, as well as quantities $\mu(\Omega)$ and $\xi(T)$, are defined as in (4.3), (4.4), (4.5), (4.6), (4.9) and (4.8) respectively. The objective function is (3.43), which is composed by a Frobenius loss term and composite penalty (3.28). For a discussion of mathematical properties of (3.43), see section (3.2). Here, we explicitly note that the composite penalty (3.28) is simply a re-scaled version of the composite penalty used in program (4.17) ($\lambda_n(\gamma||S||_1 + tr(L))$), where $\gamma = \frac{\rho}{\lambda}$ and $\lambda_n = \lambda$. Version (3.43) is useful to choose threshold parameters in empirical applications. Parameter γ is again the relative size of the subdifferential of $||.||_1$ respect to $||.||_*$. We note also that the original problem (3.30), which is our true objective, is solved in this context via (3.43), because it is proved that the three constraints $L \succeq 0, S \succ 0, L + S \succ 0$ are inactive at the optimum of (3.43), such that the two problems are algebraically equivalent.

First of all, we set the basic definitions of algebraic and parametric consistency into the covariance matrix context.

Definition 4.1.2. *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 (4.2) for the covariance matrix Σ^* 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: $L + S \succ 0$, $S \succ 0$, $L \succeq 0$.*

Model consistency here is defined according to the following three estimation features:

1. correct structural estimate of the residual covariance matrix of X conditioned on the latent variables f (given by S). This feature is called "sparsistency" of low rank plus sparse model selection.
2. number of latent variables correctly estimated.
3. the model is realizable as a covariance matrix model: $L + S$ is positive definite and L is positive semi-definite. We add the condition $S \succ 0$, which prescribes that also the sparse component can be interpreted as a covariance matrix. This last condition is not necessary to ensure a consistent estimate for Σ^*.

Parametric consistency is defined analogously to the approach described in paragraph (4.1.3). It holds if the estimates of (S, L) are close to (S^*, L^*) in some norms with high probability. The used norms are $||.||_2$ for L, $||.||_\infty$ for S, $g_\gamma(S, L)$ (4.20) for $L + S$, in application of the dual principle. Rates in spectral and Frobenius norm are also derived for $L + S$. We recall that parametric consistency does note imply algebraic consistency and vice versa.

We discuss now the main theorem ensuring identifiability and consistency. This theorem is a direct application of Theorem 4.1.3, with an important difference: in order to apply a sparsity model of the type of Bickel and Levina (2008b) (see paragraph (2.4)) on the sparse component S^*, Σ^* is imposed to be in the following matrix class:

$$\Sigma^*(\epsilon_0) = \{M \in \mathbb{R}^{p\times p} : 0 < \epsilon_0 \leq \Lambda_i(M) \leq \epsilon_0^{-1} \forall i = 1, \ldots, p\} \qquad (4.32)$$

which is the class of positive definite matrices having uniformly bounded eigenvalues ($\Lambda_i(M), i = 1, \ldots, p$, are the eigenvalues of M).

This assumption is worth some reflection. 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 (4.32) for Σ^* may cause S^* to have an high degree, and simultaneously the row/column space of L^* to have high values of incoherence (we have no spiked eigenvalues). This may result in possible non-identifiability issues. To be clear, the merge between the transversality conditions and the sparsity assumptions of [15] is possibly dangerous for model identifiability.

We report now Luo's main theorem ([77]).

Theorem 4.1.4 (Luo's Theorem 1 [77]). *Let $\Omega = \Omega(S^*)$ and $T = T(L^*)$. Suppose $\Sigma^* \in$ (4.32), $\mu(\Omega(S^*))\xi(T(L^*)) \leq 1/54$, and for $n \geq p$*

$$\lambda = C_1 \max\left(\frac{1}{\xi(T)}\sqrt{\frac{log(p)}{n}}, \sqrt{\frac{p}{n}}\right),$$

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 $C_2\lambda/\xi^2(T)$ and the smaller absolute value of the nonzero entries of S^* is greater than $C_3\frac{\lambda}{(\mu(\Omega))}$. THEN, with probability greater than $1-C_4p^{-C_5}$, the LOREC estimator $(\hat{L}, \hat{S})$ (minimizing (3.43)) 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 $1-C_4p^{-C_5}$, the matrix losses for each components are bounded as follows:

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

We call $\hat{\Sigma}_{LOREC} = \hat{L} + \hat{S}$.

The key model-based results for deriving consistency rates are **Bickel and Levina** (2008b) ([15]) for the sample loss in infinity norm:

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

and **Davidson, K. R. and Szarek, S. J.** (2001) ([39]) for the sample loss in spectral norm:

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

where $\Sigma_n = \hat{\Sigma}_{n-1}$ is the $p \times p$ unbiased sample covariance matrix computed on the observed data X.

Using the conclusions of Theorem 4.1.4, which are $||\hat{L} - L^*||_2 \leq C\lambda$, $||\hat{S} - S^*||_\infty \leq C\rho$, it is possible to derive the following overall rate for

$$e(\hat{L}, \hat{S})^2 = ||\Delta_L||^2_{Fro} + ||\Delta_S||^2_{Fro}$$

(where $\Delta_L = \hat{L} - L^*$,$\Delta_S = \hat{S} - S^*$,$\Delta_\Sigma = \hat{\Sigma}_{LOREC} - \Sigma^*$):

$$e(\hat{L}, \hat{S})^2 \leq C \left[\frac{rp}{n} \max\left(\frac{\log p}{r}, 1 \right) + \frac{s}{n} \max\left(\log p, r \right) \right], \tag{4.33}$$

where s is the usual number of non-zero elements in S^*. If $r \sim \log p$ (as it is for exactly low rank matrix recovery), this rate coincides with the one under the Agarwal's approach (4.15), where $\alpha = 0$, since we no longer have non-identifiability issues. This is obtained using the lower bound $\xi(T) = O(\sqrt{\frac{r}{p}})$ (see (4.11)).

From Theorem (4.1.4), Luo derives the following **rates** for $\hat{\Sigma}_{LOREC}$:

$$||\hat{\Sigma}_{LOREC} - \Sigma^*||_2 \leq C(s\xi(T) + 1)\lambda = \phi$$

$$||\hat{\Sigma}_{LOREC} - \Sigma^*||_{Fro} \leq C(\sqrt{ps}\xi(T) + \sqrt{r})\lambda$$

with probability larger than $1 - C_1 p^{-C_2}$, if and only if $\lambda_{min}(\Sigma^*) \geq \phi$.

The same rates hold for the inverse covariance estimate $\hat{\Sigma}^{-1}_{LOREC}$,

$$||\hat{\Sigma}^{-1}_{LOREC} - \Sigma^{-1*}||_2 \leq C(s\xi(T) + 1)\lambda = \phi$$

$$||\hat{\Sigma}^{-1}_{LOREC} - \Sigma^{-1*}||_{Fro} \leq C(\sqrt{ps}\xi(T) + \sqrt{r})\lambda$$

with probability larger than $1 - C_1 p^{-C_2}$, if and only if $\lambda_{min}(\Sigma^*) \geq 2\phi$. Here r is the true latent rank of L^*, while s, differently from (4.33), is defined as the maximum number of non zero elements per column (which is the induced $||.||_1$ norm). This is done to further improve error rates. From now to the end of Chapter, parameter s will change its meaning as explained: $s = \max_j \sum_{i=1}^{p} \mathbf{1}(s_{ij} \neq 0)$, $j = 1, \ldots, p$. Both results are reported as Corollaries in [77]. We will show proof details in next paragraph (5.1).

We now describe the meaning of needed assumptions. Since (3.43) contains a Frobenius loss term instead of the log-likelihood, this method is no longer a likelihood method. For this reason, there is no need here to bound the curvature of Fisher information $\mathbb{I}^*$, since $\mathbb{I}^* = I_p$. So, referring to Proposition 4.1.3, parameters α, β, and γ (see (4.28) (4.29) (4.30)) are now all equal to 1, with $\nu = \frac{1}{2}$ (see Lemma 4.1.2). On the contrary, the analogous of Proposition 4.1.3 is still needed, because the tangent space $T(L^*)$ is still curve, and transversality between T and Ω still needs to be bounded (even if $\mathbb{I}^*$ has no longer impact).

Proofs are contained in [76], which is a previous version of [77]. There it is possible to find (at page 26) the analogous of Proposition 4.1.3, where $\mathbb{I}^*$ has

no longer impact. The identifiability assumption here becomes $\mu(\Omega(S^*)\xi(T(L^*)) \leq 1/54$, which can also be rewritten, using (4.11) and (4.10), as

$$deg_{max}(S^*)inc(L^*) \leq \frac{1}{108}.$$

The range $\gamma \in [9\xi(T), 1/(6\mu(\Omega))]$ is obtained by Proposition 4.1.3 setting α, β, and γ equal to 1, ν equal to $\frac{1}{2}$. Note that, $\gamma = \sqrt{9\xi(T) * \frac{1}{(6\mu(\Omega))}}$, geometric mean of the two ends, is always inside the range, and using (4.11) and (4.10), we can write $\gamma = \sqrt{2 * 9inc(B)\frac{1}{(6deg_{max}(A))}} = \sqrt{\frac{3inc(B)}{(deg_{max}(A))}}$. The minimum magnitude of the non-zero entries of S^* and the minimum eigenvalue of L^* ($\lambda_r(L^*)$) are lower bounded, in order to ensure consistent recovery, and also identifiability in the case of $\lambda_r(L^*)$. The use of $\hat{\Sigma}_{n-1}$ is responsible for the usual assumption $p \leq n$.

There is one major difference with the approach of [1] explained in (4.1.3): here, the sparsity assumption on S^* imposes that the parameter λ, coming from probabilistic analysis, must take into account both probabilistic frameworks, the one from $||\hat{\Sigma}_n - \Sigma^*||_2$ (represented by $\sqrt{\frac{p}{n}}$) and the one from $||\hat{\Sigma}_n - \Sigma^*||_\infty$ (represented by $\frac{1}{\xi(T)}\sqrt{\frac{log(p)}{n}}$).

The parameter $\rho = \gamma\lambda$ has this shape to re-scale accordingly the subdifferential of the sparse component. The parameter λ has this shape because, even if we are in a deterministic context, the need of a probabilistic bound for $g_\gamma(\mathbb{A}^\dagger E_n)$, where $E_n = \hat{\Sigma} - \Sigma^*$, rises throughout the proof. If the input is the unbiased sample covariance matrix ($\hat{\Sigma} = \hat{\Sigma}_{n-1}$), the rates are the ones above written, and the condition $p \leq n$ is unavoidable. We will make some effort to overcome this issue in paragraph (5.1), providing statistical rates under POET assumptions and in the generalized spikiness context.

It is now easier to understand which are the possible non-identifiability issues coming out. Differently from POET approach, where the sparsity assumption (4.32) is imposed to the sparse component S^*, LOREC approach imposes it directly to the covariance matrix Σ^*.

So, two conditions must hold which may be in contradiction: if the minimum eigenvalue of L^* is too large, it is unlikely that Σ^* is into the matrix class (4.32). This makes the matrix class for which recovery is effective quite unclear. In addition, the product $\mu(\Omega)\xi(T)$ is affected by this trade-off, such that, if $\lambda_r(L^*)$ is too large, S^* must be very sparse in order to respect the upper bound for $\mu(\Omega)\xi(T)$. We will find confirmation of that in our simulation study (Chapter 5).

Another aspect of Theorem 4.1.4 is that the two losses (in L^* and S^* respectively) are bounded separately. This may result in some issues concerning the overall performance represented by the loss $||\hat{\Sigma} - \Sigma^*||_{Fro}$, as our simulation study confirms (see (5.3.1)), because here $||\Delta_\Sigma||_2$ is simply derived using triangle inequality $||\Delta_\Sigma||_2 \leq ||\Delta_L||_2 + ||\Delta_S||_2$, as well as $||\Delta_\Sigma||_{Fro}$.

More explanations and a proposal to improve LOREC estimation process on this side is given in paragraph (5.1).

We now describe the steps used in [76] to prove Theorem (4.1.4). They directly descend from the proof of Theorem 4.1.3 in [31], set into our context, where the reference model is (3.1).

The chain of programs to be solved and the mathematical rationale are showed. We start from the brief explanations given at the end of paragraph (4.1.3). First, we need to bound the curvature of T. So, for the equivalent of Proposition 4.1.3, we restrict our analysis to tangent spaces satisfying $\rho(T, T') \leq \xi/2$. We can then solve problem (3.43) with additional tangent space constraints:

$$\min_{L,S} \frac{1}{2}||(L+S) - \Sigma_n||^2_{Fro} + \lambda||L||_* + \rho||S||_1, \quad \text{s.t. } S \in \Omega, L \in T', \tag{4.34}$$

where $T = T(L^*)$ s.t. $\rho(T, T') \leq \xi(T)/2$.

We know that $||L||_*$ and $||S||_1$ are **non differentiable**. In order to bound the Loss function: $g_\gamma(\Delta_S, \Delta_L) = g_\gamma(\hat{S}_\Omega - S^*, \hat{L}_{T'} - L^*) = \max\{|\hat{S}|_\infty/\gamma, ||\hat{L}||\}$, where $(\Delta_S, \Delta_L) = (\hat{S}_\Omega - S^*, \hat{L}_{T'} - L^*)$, the needed tools are:

- the projected gradient method;
- Brouwer's fixed point theorem (see [76], p.27).

We start recalling the subgradient conditions for $||L||_*$ and $||S||_1$. Our CONVEX objective at the optimum $(\hat{S}_\Omega, \hat{L}_T)$ satisfies, for some Lagrangian multipliers, $Q_{\Omega^\perp} \in \Omega^\perp$ and $Q_T \in T^\perp$ the following optimality conditions:

$$\hat{S}_\Omega + \hat{L}_T - \Sigma_n + Q_\Omega^\perp \in -\lambda_n \gamma \delta|\hat{S}_\Omega|_1$$

$$\hat{S}_\Omega + \hat{L}_T - \Sigma_n + Q_T\perp \in -\lambda_n \delta|\hat{L}_T|_*,$$

where δ denotes the **subdifferential**.

Lagrangian duality theory is a first order method. So, we need to bound the second-order Taylor rest of Σ^*. The key is to project $\hat{S}_\Omega + \hat{L}_{T'} - \hat{\Sigma}_n$ onto $\mathbb{Y} = \Omega \times T'$ (where $\times$ represents here the Cartesian product), and to define

$$P_\Omega(\hat{S}_\Omega + \hat{L}_{T'} - \hat{\Sigma}_n) = Z_\Omega,$$

$$P_T(\hat{S}_\Omega + \hat{L}_{T'} - \hat{\Sigma}_n) = Z_{T'},$$

with $||Z_\Omega||_\infty = \lambda_n \gamma$ and $||Z_{T'}|| \leq 2\lambda_n$. The bi-dimensional projection is

$$P_{\mathbb{Y}} \mathbb{A}^\dagger(\hat{S}_\Omega + \hat{L}_{T'} - \hat{\Sigma}_n) = Z = (Z_\Omega, Z_{T'}),$$

where $\mathbb{Y} = \Omega \oplus T'$ (i.e. Z is a feasible point).

This is the projected gradient method. Then, the application of Brouwer's fixed point theorem allows to bound $g_\gamma(\mathbb{P}(\Delta_S, \Delta_L))$, which in turn serves as a limit for the error $g_\gamma(\Delta_S, \Delta_L)$, thus satisfying the first half of optimality conditions (recall (3.40) and (3.41)). This error bound is needed to prove there is a unique minimizer, and establish parametric consistency.

Then, imposing $g_\gamma(\mathbb{A}^\dagger E_n) \leq \frac{\lambda_n}{18}$, it is possible to prove that the tangent-space constrained problem (4.34) is equivalent to the following variety-constrained problem

$$\min_{L,S} \frac{1}{2}||(L+S) - \Sigma_n||^2_{Fro} + \lambda||L||_* + \rho||S||_1, \qquad (4.35)$$
$$\text{s.t. } S \in \Omega, L \in T_{\mathscr{M}},$$

where $T_{\mathscr{M}} = T(\hat{L}_{\mathscr{M}})$, and $(\hat{S}_{\mathscr{M}}, \hat{L}_{\mathscr{M}})$ is the solution of

$$\mathscr{M} = \{(S,L) \mid s \in \Omega(S^*), rank(L) \leq rank(L^*),$$
$$||\mathscr{P}_{T^\perp}(L - L^*)||_2 \leq \xi(T)\lambda, g_\gamma(\hat{S}_\Omega - S^*, \hat{L}_{T'} - L^*) \leq 11\lambda\}.$$

This serves for ensuring algebraic consistency, and holds under all the assumptions of Theorem 4.1.4. It also allows to solve the non-convex problem (4.35) as a convex one, linearizing the constraints.

Finally, under the same assumptions, the solution of problem (4.35) is shown to be solution of the original problem (3.43) without any constraints. In [31], another bound on the Taylor rest of Σ^{*-1} is needed, since they are dealing with the inverse. For us, the condition $g_\gamma(\mathbb{A}^\dagger E_n) \leq \frac{\lambda_n}{18}$, limiting the g_γ norm of $E_n = \Sigma^* - \Sigma_n$, is sufficient.

Another important quantity to bound during the proof is $g_\gamma(\mathbb{A}^\dagger C_{T'})$, where $C_{T'} = P_{T'^\perp}(L^*)$. This is needed to bound the curvature of T, as well as the constraint $||\mathscr{P}_{T^\perp}(L - L^*)||_2 \leq \xi(T)\lambda$.

During this last step, probabilistic bounds come into play. Since we need to bound $g_\gamma(\mathbb{A}^\dagger E_n)$, large deviation theory must be applied to $||E_n||_2$ and $||E_n||_\infty$. This is done using the outlined results from **Bickel and Levina** (2008b) and **Davidson, K. R. and Szarek, S.J**. (2001). The strength of the probabilistic bound depends on the relationship between p and n. In particular, key ratios $\frac{p}{n}$ and $\frac{\log p}{n}$ come from the probabilistic bounds of $||E_n||_2$ and $||E_n||_\infty$ respectively. This is why $\lambda = C_1 \max\left(\frac{1}{\xi(T)}\sqrt{\frac{log(p)}{n}}, \sqrt{\frac{p}{n}}\right)$. The condition $p \leq n$ is unavoidable in order to obtain finite probabilistic bounds.

We have already pointed out the possible weakness of this approach respect to identifiability issues, due to the need of imposing matrix class (4.32) directly to Σ^*, and not to S^*. This choice causes, jointly with the identifiability assumptions, uncertainty on the underlying structure of Σ^*. Another difficulty of Luo's approach is that (2.8) is only partially imposed to Σ^*, leaving out the conditions on limited correlations. On the contrary, no matrix class is actually imposed to S^*, whose sparsity is recovered algebraically

(deterministically) using the standard property $||M||_\infty \leq ||M||_2$ exploiting the scale parameter γ.

We find a key difference between the approaches of Luo (2013) and Chandrasekaran et at (2012). In the latter, ONLY the probabilistic bound for $||E_n||_2$ is used, and the one for $||E_n||_\infty$ is simply derived as a consequence using the basic relationship $||E_n||_\infty \leq \ ||E_n||_2$. For this reason, there we have the following parametric rate:

$$g_\gamma(\hat{S}_n - K_O, \hat{L}_n - K_{O,H}K_H^{-1}K_{H,O}) \preceq \frac{1}{\xi(T)}\sqrt{\frac{p}{n}}. \tag{4.36}$$

The two components are bounded jointly, exactly as in Agarwal's approach. In the former, the two components are approached separately, and the shape of λ_n reflects this choice.

Therefore, Luo should have imposed matrix class (4.32) together with the covariance assumptions (see (2.8)) to S^*, in order to have the desired sparsity model. However, this would have been useless for the mathematical proof, which requires that Σ^* belongs to (4.32), in order to derive the probabilistic bound of $||E_n||_\infty$. On the other side, in absence of specification of that matrix class, he would have left the infinity norm rate dependent on the spectral one, with no progress respect to Chandrasekaran et al. (2012).

The number of samples n can be $O(p)$, thanks to probabilistic results contained in [110], provided that $n \leq p$. In contrast, the condition $n \leq 2p$ is needed for Chandrasekaran et al.(2012), and $p = O(\frac{p}{\xi^4(T)})$, which corresponds to $O(\frac{p^3}{r^2})$ in the worst case (see Theorem 4.1.3). Starting from (4.31), it is easy to show (using the lower bound $n = O(\frac{p^3}{r^2})$) that the overall Frobenius rate for the covariance matrix estimate in [31] is $O(r^{1/2}pn^{-1/2})$. This occurs because the rate is there determined only by the low rank component. The analogous rate for the low rank component under Luo's approach is $O(r^{1/2}p^{1/2}n^{-1/2}\max(\log p, r^{1/2}))$, which is lower (for explanations see (4.33)). This rate can be even lower under different model specifications using the same low rank plus sparse decomposition, as the so called spiked covariance model of Johnstone and Lu (2009) [71] (for more details see [76] and [77]).

To conclude this paragraph, we give some terms of comparison among probabilistic rates respect to alternative PCA-based approaches recovering Σ^* under similar assumptions. In our numerical context, the strength of probabilistic bounds depends on the relationship between the finite values of p and n.

In [43], factors are observable and the residual component is diagonal. There, the rate for $\hat{\Sigma}$ (and $\hat{\Sigma}_n$) is $O(n^{-1/2}pr)$, while LOREC under the same conditions shows $O(n^{-1/2}(p + p^{1/2}r^{1/2}))$ (see (4.33)). For the eigenvalue convergence rate, [43] has the same $O(n^{-1/2}pr)$, while LOREC shows $O(n^{-1/2}p^{1/2})$. Only LOREC provides spectral bounds. Concerning the in-

verse, [43] has a Frobenius rate of $O(n^{-1/2}pr^2 \log p^{1/2})$, while LOREC shows again $O(n^{-1/2}(p + p^{1/2}r^{1/2}))$, which is lower. The difference occurs because an additional error term $O(p^{-1/2})$ comes out when the residuals are unobservable.

In the approximate sparse factor model context, it is hard to provide absolute rates, as the spectral or the Frobenius ones, using a PCA-based approach. This is due to the fact that the necessary pervasiveness assumption requires large p (see paragraph (2.5)). What is more, an additional error term $O(p^{-1/2})$ comes out when the residuals are unobservable (as in [44]). When also the factors are unobservable, as explained in [76], there is an unavoidable additional error term $O(\log p)$. In POET setting ([45]) we find both. More, for the just explained reasons, the rate for $\hat{\Sigma}$ is provided only in relative norm (see (2.19)), exactly as in [44].

This is why we will compare extensively the performance of $\hat{\Sigma}_{POET}$ and $\hat{\Sigma}_{LOREC}$ in a wide simulation study (Chapter 5). As a comparison term, we now list the main differences in the theoretical **assumptions** bewteen POET and LOREC approaches:

- For POET the spectral bound is provided only on $||S^*||$, while for LOREC is provided both on $||S^*||$ and $||\Sigma^*||$.
- In POET setting, the r eigenvalues of $p^{-1}B'B$ are bounded away from 0 and ∞ as p increases (**pervasiveness** condition). In LOREC setting, there is only a lower bound on the minimum eigenvalue of L^*.
- In LOREC setting, ALL the eigenvalues of Σ^* are bounded away from 0 and infinity. In POET setting, the smallest $p - r$ are upper bounded by $||S^*||$, the largest r are approximately equal to the ones of $B'B$.
- In LOREC setting, Λ_{max} controls for the strongness of the probability bound, Λ_{min} controls for the **positive definiteness** of $\hat{\Sigma}$ (necessary to estimate the **inverse**).
- The latent rank r is exactly recovered automatically by LOREC without the need for any external tool. In contrast, POET selects r using the well known rank selection criteria by Bai and Ng ([6]).
- Concerning the sparsity pattern, LOREC needs only a lower bound on the smaller absolute value of the non-zero entries of S^*, while POET requires
$$m_p = \max_{i \leq p} \sum_{j \leq p} |s_{ij}|^q = o(p)$$
for some $q \in [0, 1)$.
- Statistical performance is assessed asymptotically for POET, non-asymptotically for LOREC. In the first case the reference norm is

the relative norm (2.19), in the second is the Frobenius norm (relative VS absolute rates).

As a final remark, we note that both LOREC and POET procedures are not scale-equivariant, that is, the estimates are not equivariant under linear transforms. For POET, this is due to the use of PCA, depending on the sample eigenvalues (which are not scale-equivariant), and also depends on the use of thresholding for the recovery of the sparse component. For LOREC, this is due to the singular value thresholding of the low rank component and to the soft thresholding of the off-diagonal elements of the sparse component. We recall that also the factor model estimates by the principal factors method are not scale-equivariant, still for the use of sample eigenvalues.

We are now ready to introduce a set of novelties improving upon LOREC approach exploiting features of some of the methods we have shown throughout our thesis. First, in the pure LOREC setting, we propose a solution to the approximation problem caused by the separate bounding of the errors in L^* and S^*. This solution involves the unshrinkage of the estimated eigenvalues at the end of the solution algorithm (composed by the singular value thresholding of the low rank component and the soft threhsolding of the sparse component, see (3.2.2)). This proposal is proved to be algebraically meaningful for improving the original LOREC on the side of the overall loss $||\Delta||_{\Sigma}^{Fro}$, and to better catch the proportion of variance explained by the low rank component.

The other advances concern the number of necessary samples n respect to p. In order to do that, we want to exploit the theory of approximate factor model. So, we abandon the hypothesis $\Sigma^* \in$ (4.32), which is not coherent with the presence of few spiked eigenvalues. We thus link the infinity norm of E_n to the spectral one as in the approach by Chandrasekeran et al. (2012). We show that using the POET spikiness assumption (Proposition 2.5.1) and imposing a sparse model for S^* in the spirit of Bickel and Levina (2008b) ($S^* \in$ (2.8)) we can prove, using (2.23), that the described algebraic setting holds with rate $O(\frac{p}{\sqrt{n}})$, and simultaneously the probabilistic bound is guaranteed until $p \log p \ll n$. Finally, we extend this result into the generalized spikiness context of Proposition 2.5.1. We prove an updated version of (2.23) in the α-spiked context, such that the described algebraic setting holds with rate $O(\frac{p^\alpha}{\sqrt{n}})$, and simultaneously the probabilistic bound is guaranteed until $p^\alpha \log p \ll n$, with $\alpha \in (0, 1]$.

The results we need are:

$$P\left(||\Sigma_n - \Sigma|| > \frac{p}{\sqrt{n}}\right) \leq C_1 \exp\left(-C_2 p^2\right),$$

if all the assumptions under Theorem (2.19) hold, and

$$P\left(||\Sigma_n - \Sigma|| > \frac{p^\alpha}{\sqrt{n}}\right) \leq C_1 \exp\left(-C_2 p^{2\alpha}\right),$$

if all the assumptions under Theorem (2.19) hold, with the difference that Definition 2.5.1 replaces Proposition 2.5.1, for $\alpha \in (0, 1]$.

Chapter 5

Improving LOREC: empirical and theoretical results

In this chapter, original advances and extensions to LOREC approach are described, with particular reference to the estimation performance and to different assumptions for the eigenvalues of the low rank component, in respect to the ones of POET ([45]).

In paragraph (5.1), Luo's approach ([77]) is completed with the rates for the sparse component, its inverse and its positive definiteness conditions. A more operative identifiability condition is also derived from [30]. The quality of the overall solution is improved performing the unshrinkage of the estimated eigenvalues of the low rank component. The rates of convergence under the spikiness assumptions of [45] and under the setting of α - generalized spikiness structure (Definition 2.5.1) are derived using the key tools of [45] and [15] described in paragraphs (2.5.4) and (2.4) respectively.

Then, we show simulated and real data analysis results in support of the proposals contained in paragraph (5.1). In particular, we focus on the approximation improvement offered by $\hat{\Sigma}_{New}$ respect to $\hat{\Sigma}_{LOREC}$, and on the comparison between the performance of $\hat{\Sigma}_{New}$ and $\hat{\Sigma}_{POET}$ in the POET setting.

In paragraph (5.2.1), we describe an original simulation algorithm created for this purpose, which is enough flexible to catch all the different situations we need in a unique framework. The comparison quantities needed to assess the performance of estimators are described in (5.2.2). In paragraph (5.2.3), we show a model selection criterion specifically thought for our estimation method.

Simulated data analysis is reported in paragraph (5.3.1). A number of simulated data settings, particularly useful for assessing the performance of $\hat{\Sigma}_{New}$ and compare it to the one of $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{POET}$, are described, with the aim of testing the theoretical advances described in paragraph (5.1). Simulations are performed with MATLAB.

Real data analysis is then offered in paragraph (5.3.2) with the aim of comparing the performance of $\hat{\Sigma}_{POET}$ and $\hat{\Sigma}_{New}$. Two real data-sets are taken into account: one on UK market data (publicly available) which was used by Fan and colleagues to assess the performance of POET ([45], paragraph 7) and a supervisory banking data-set which collects balance sheet data for some of the most relevant Euro Area banks. For the last one, we deeply acknowledge the Supervisory Statistics Division of the European Central Bank, where the author spent a semester as a PhD trainee, for the allowance to use these data in anonymous form for research purpose.

5.1 Theoretical advances

We start showing in detail the algebraic steps which allow to derive the Frobenius rates for Σ^* from the Conclusions in Theorem 4.1.4. The reference is here [76], paragraph 6.

We set $\Sigma_n = \hat{\Sigma}_{n-1}$, estimation input. For the triangular inequality we have:

$$||\hat{L} + \hat{S} - (L^* + S^*)|| \leq ||\hat{L} - L^*|| + ||\hat{S} - S^*||.$$

Using standard matrix norm properties, we obtain

$$||\hat{L} + \hat{S} - (L^* + S^*)|| \leq ||\hat{L} - L^*|| + ||\hat{S} - S^*||_1,$$

and then

$$||\hat{L} + \hat{S} - (L^* + S^*)|| \leq ||\hat{L} - L^*|| + s||\hat{S} - S^*||_\infty,$$

where s is there the maximum number of non zeros per column in S^*. This result is derived using $sign(\hat{S}) = sign(S^*)$, which allows to improve upon the standard constant p.

Setting $\gamma = 9\xi(T)$ (its minimum), we obtain

$$||\hat{\Sigma}_{LOREC} - \Sigma^*||_2 \leq C(s\xi(T) + 1)\lambda = \phi. \tag{5.1}$$

An analogous triangular inequality holds for the Frobenius rate:

$$||\hat{L} + \hat{S} - (L^* + S^*)||_{Fro} \leq ||\hat{L} - L^*||_{Fro} + ||\hat{S} - S^*||_{Fro}.$$

Exploiting the fact that the algebraic sum $A + B$, when A and B have rank r, has at most rank $2r$ (see [62]), and using previous results for $\hat{S}$ together with the standard inequality $||A||_F \leq \sqrt{ps}||A||_{max}$, we obtain

$$||\hat{L} + \hat{S} - (L^* + S^*)||_{Fro} \leq \sqrt{2r}||\hat{L} - L^*|| + \sqrt{ps}||\hat{S} - S^*||_\infty.$$

Setting $\gamma = 9\xi(T)$ (its minimum), we obtain

$$||\hat{\Sigma}_{LOREC} - \Sigma^*||_{Fro} \leq C(\sqrt{ps}\xi(T) + \sqrt{r})\lambda. \tag{5.2}$$

For simplicity of notation, we now remove all $*$. Recalling Theorem 2.2.1, we know that $\hat{L} + \hat{S}$ **is positive definite if and only if the minimum eigenvalue of** Σ^* **is larger than the spectral bound** ϕ. We give a further justification of this basic result. Weyl's Theorem (see [45] Appendix C) prescribes that, for any matrix Σ, we have

$$|\hat{\lambda}_i - \lambda| \leq ||\hat{\Sigma} - \Sigma|| \; \forall i = 1, \ldots, p,$$

where $\hat{\lambda}_i$, $i = 1, \ldots, p$ are the sample eigenvalues. This result relates the rate of sample eigenvalues to the matrix spectral loss rate. The triangular inequality gives

$$\begin{aligned} &|\lambda_{min}(\hat{L} + \hat{S}) - \lambda_{min}| \leq \\ &\leq |\lambda_{min}(\hat{L} + \hat{S})| + |-\lambda_{min}| = \\ &= |\lambda_{min}(\hat{L} + \hat{S})| + \lambda_{min}, \end{aligned}$$

because Σ is positive definite. Thus,

$$|\lambda_{min}(\hat{L} + \hat{S})| \geq |\lambda_{min}(\hat{L} + \hat{S}) - \lambda_{min}| - \lambda_{min}.$$

Since for the Weyl's theorem $|\lambda_{min}(\hat{L} + \hat{S}) - \lambda_{min}| \leq \phi$ we have

$$\lambda_{min}(\hat{L} + \hat{S}) > 0 \Longleftrightarrow \lambda_{min} > \phi. \tag{5.3}$$

This proves the claim.

In order to achieve the same rate ϕ **for the inverse spectral rate** $||(\hat{L} + \hat{S})^{-1} - \Sigma^{-1}||$**, it is necessary that** $\lambda_{min} \geq 2\phi$**.**

In fact, the triangular inequality gives

$$||(\hat{L} + \hat{S})^{-1} - \Sigma^{-1}|| \leq ||(\hat{L} + \hat{S})^{-1}|| + \lambda_{min}^{-1} \tag{5.4}$$

By summing and subtracting Σ and using triangular inequality

$$\begin{aligned} ||(\hat{L} + \hat{S})^{-1}|| &= ||(\hat{L} + \hat{S} - \Sigma + \Sigma)^{-1}|| \leq \\ &\leq ||(\hat{L} + \hat{S} - \Sigma)^{-1}|| + ||\Sigma^{-1}|| \leq \\ &\leq ||(\hat{L} + \hat{S})^{-1} - \Sigma^{-1}|| + ||\Sigma^{-1}|| = \\ &\leq ||(\hat{L} + \hat{S} - \Sigma)^{-1}|| + \lambda_{min}^{-1}. \end{aligned}$$

For the Weyl's theorem, we have

$$||(\hat{L} + \hat{S} - \Sigma)^{-1}|| \leq |\lambda_{min}((\hat{L} + \hat{S})^{-1}) - \lambda_{min}(\Sigma)^{-1}|.$$

For triangular inequality, we have

$$\begin{aligned} &|\lambda_{min}((\hat{L} + \hat{S})^{-1}) - \lambda_{min}(\Sigma)^{-1}| \leq \\ &\leq |\lambda_{min}((\hat{L} + \hat{S})^{-1})| + |-\lambda_{min}^{-1}| \leq \end{aligned}$$

$$\leq |\lambda_{min}((\hat{L}+\hat{S})^{-1})| + \lambda_{min}^{-1}$$

since Σ is positive definite.

At the same time, for (5.1), we have

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

Hence, inequality (5.4) becomes

$$\phi^{-1} \leq |\lambda_{min}((\hat{L}+\hat{S})^{-1})| + 2\lambda_{min}^{-1}.$$

We can write

$$|\lambda_{min}((\hat{L}+\hat{S})^{-1})| \geq \phi^{-1} - 2\lambda_{min}^{-1},$$

which allows to conclude that

$$||\hat{\Sigma}_{LOREC}^{-1} - \Sigma^{-1}||_2 \leq \phi \Longleftrightarrow \phi^{-1} \geq 2\lambda_{min}^{-1}. \tag{5.5}$$

Using this assumption, it is possible to derive the rate for $(\hat{L}+\hat{S})^{-1}$, by property $||(A+E)-A^{-1}|| \leq ||A^{-1}|| \cdot ||E|| \cdot ||(A+E)^{-1}||$ (see [76], p. 31-32):

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

$$\leq ||(\hat{L}+\hat{S})^{-1}|| \cdot ||[\hat{L}+\hat{S}-\Sigma]|| \cdot ||(\Sigma)^{-1}|| \leq \tfrac{2}{\lambda_{min}^2}||[\hat{L}+\hat{S}-\Sigma]||.$$

Hence, we have

$$||\hat{\Sigma}_{LOREC}^{-1} - \Sigma^{-1}||_2 \leq C(s\xi(T)+1)\lambda = \phi \tag{5.6}$$

By property $||M_1M_2||_{Fro} \leq ||M_1|| \cdot ||M_2||_{Fro}$, it is straightforward to derive

$$||\hat{\Sigma}_{LOREC}^{-1} - \Sigma^{-1}||_{Fro} \leq C(\sqrt{ps}\xi(T)+\sqrt{r})\lambda. \tag{5.7}$$

Using the same framework, we can complete Luo's analysis with the rates for $\hat{S}$. From $||\hat{S}-S^*|| \leq s||\hat{S}-S^*||_\infty$, we obtain

$$||\hat{S}-S^*||_2 \leq Cs\xi(T)\lambda = \phi_S. \tag{5.8}$$

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

$$||\hat{S}-S^*||_{Fro} \leq C\sqrt{ps}\xi(T)\lambda. \tag{5.9}$$

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

Unshrinking the eigenvalues of the low rank component

We now approach the approximation problem due to the separate bounds for the two components. The problem is that the combined shrinkage approach gets closer to each component separately, but in such a way it goes further from the overall solution, as we will show in Chapter 5. The need rises to correct for this drawback, re-shaping $\hat{\Sigma}_{LOREC}$, because the overall Loss function used in the algebraic setting, g_γ, derives the overall performance as a consequence of the two separate bounds. That means that LOREC approach can be somehow sub-optimal for the whole covariance matrix.

We will describe a finite sample analysis, which could be referred to as a re-optimization least squares method. From now, we will refer to the usual objective function (3.43) where $||S||_1 = ||S||_{1,off} = \sum_{i=1}^{p-1}\sum_{j=i+1}^{p}|s_{ij}|$, i.e. the l_1 norm excluding the diagonal. This approach is coherent with the sparse approximate factor model (3.1) and with POET (see (2.5.4)), which will be our reference competitor in Chapter 5.

We start from a standard result: the PCA of M truncated to the r-th component is the r-ranked matrix best approximating M. In fact,

$$\min_{B,rank(B)=r} ||A-B||_2$$

and

$$\min_{B,rank(B)=r} ||A-B||_{Fro}$$

are both solved for

$$B = \sum_{i=1}^{r} \lambda_i u_i u_i',$$

which is the SVD truncated to the r-th summand ([40]), when r is known.

Suppose now that $\hat{\mathscr{L}}(\hat{r})$ and $\hat{\mathscr{K}}(\hat{s})$ are the varieties ensuring the algebraic consistency of (3.30). A natural **question** comes out: which is the solution (say $(\hat{L}_{New}, \hat{S}_{New})$) of the problem

$$\min_{L\in\hat{\mathscr{L}}(\hat{r}),S\in\hat{\mathscr{K}}(\hat{s})} ||(\Sigma_n - (L+S)||^2_{Fro}? \tag{5.10}$$

We know that, the sample covariance matrix follows the model $\Sigma_n = L^* + S^* + W$, where $W \sim Wishart(\mathbf{0}_{p\times p}, n)$, given a sample $X_i, i = 1,\dots,n$.

We define the **total loss** for the generic pair $L \in \hat{\mathscr{L}}(\hat{r}), S \in \hat{\mathscr{K}}(\hat{s})$ as:

$$TL(L,S) = ||(\Sigma_n - (L+S)||^2_{Fro}.$$

In other words, we face the following question: which pair $L \in \hat{\mathscr{L}}(\hat{r}), S \in \hat{\mathscr{K}}(\hat{s})$ satisfying algebraic consistency shows the best approximation properties of Σ_n? We prove the following original result.

Theorem 5.1.1. ***Suppose*** *that $\hat{L}_{LOREC}$ and $\hat{S}_{LOREC}$ are the LOREC solutions satisfying Theorem 4.1.4, with $\hat{\Sigma}_{LOREC} = \hat{L}_{LOREC} + \hat{S}_{LOREC}$.* ***Suppose*** *that $\hat{\mathscr{L}}(\hat{r})$, $\hat{\mathscr{K}}(\hat{s})$ are the recovered matrix varieties, and that $\hat{L} = \hat{U}\hat{D}\hat{U}'$ is the eigenvalue decomposition of $\hat{L}_{LOREC}$.* ***Assume*** *that the off-diagonal elements of $\hat{S}_{New}$ are the same as the ones of $\hat{S}_{LOREC}$ as well as the diagonal elements of $\hat{\Sigma}_{New}$ are the same as the ones of $\hat{\Sigma}_{LOREC}$.* ***Then,*** *the minimum $\min_{L\in\hat{\mathscr{L}}(\hat{r}),S\in\hat{\mathscr{K}}(\hat{s})} ||(\Sigma_n - (L+S)||^2_{Fro}$ is achieved if and only if $\hat{L}_{New} = \hat{U}(\hat{D}+\lambda I_r)\hat{U}'$ and if $diag(\hat{S}_{New,ii}) = diag(\hat{\Sigma}_{LOREC,ii}) - diag(\hat{L}_{New,ii})$, where λ is the threshold parameter.* ***In addition,*** *the gain in terms of spectral loss is strictly positive and bounded by λ.*

We now prove Theorem 5.1.1. Given finite p and n we have

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

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

(the signs are put in a convenient form).

The LOREC solution is $\hat{\Sigma}_{LOREC} = \hat{L} + \hat{S}$, $L \in \hat{\mathscr{L}}(\hat{r}), S \in \hat{\mathscr{K}}(\hat{s})$, with

$$\hat{L} = \hat{U}\hat{D}\hat{U}', \tag{5.11}$$

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

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

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

$$\begin{aligned}
&\min_{S\in\hat{\mathscr{K}}(\hat{s})} ||S - S^*||^2_{Fro} = \\
&= \min_{L\in\hat{\mathscr{L}}(\hat{r})} ||(\hat{\Sigma} - L) - (\Sigma^* - L^*)||^2_{Fro} = \\
&= \min_{L\in\hat{\mathscr{L}}(\hat{r})} ||(\hat{\Sigma} - \Sigma^*) - (L - L^*)||^2_{Fro} \leq \\
&\sum_{i=1}^{p}(\hat{\sigma}_{ii} - \sigma_{ii})^2 + \sum_{i=1}^{p}(\hat{l}_{ii} - l_{ii})^2 \qquad (5.12) \\
&= B' + B''.
\end{aligned}$$

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

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

Consequently:

$$\hat{S}_{New,ii} = \hat{\Sigma}_{ii} - \hat{L}_{New,ii},\ \forall i.$$
$$\hat{S}_{New,ij} = \hat{S}_{ij},\ \forall i \neq j.$$

We can thus define $\hat{\Sigma}_{New} = \hat{L}_{New} + \hat{S}_{New}$. We call $\hat{L}_{Orig}$ and $\hat{S}_{Orig}$ the original LOREC estimates. We know that $||\hat{L}_{New} - \hat{L}_{Orig}||_2 = \lambda$.

Recalling that $\hat{L}_{New} = \min_{L \in \hat{\mathscr{L}}(\hat{r})} ||L - L^*||^2_{Fro}$, we have

$$0 < ||\hat{L}_{Orig} - L^*||_2 - ||\hat{L}_{New} - L^*||_2 \leq \lambda, \tag{5.13}$$

because $||\hat{L}_{Orig} - L^*||_2 \leq ||\hat{L}_{New} - \hat{L}_{Orig}||_2 + ||\hat{L}_{New} - L^*||_2$. As a consequence, $||\hat{L}_{New} - \hat{L}_{Orig}||_{Fro} = \sqrt{2r}\lambda$ and

$$0 < ||\hat{L}_{Orig} - L^*||_{Fro} - ||\hat{L}_{New} - L^*||_{Fro} \leq \sqrt{2r}\lambda. \tag{5.14}$$

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

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

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

$\sum_{i=1}^{p}(\hat{l}_{Orig,ii} - l_{ii})^2 \forall i$ depends on the statistical properties of $\hat{L}_{LOREC}$. $\sum_{i=1}^{p}(\hat{l}_{New,ii} - \hat{l}_{Orig,ii})^2 = r\lambda^2$, for basic algebraic considerations on the trace. It is also straightforward that $||diag(\hat{L}_{New} - L_{Orig})||_2 = \lambda$. So, recalling that $\hat{S}_{New} = \min_{S \in \hat{\mathscr{K}}(\hat{s})} ||S - S^*||^2_{Fro}$, we can write $||\hat{S}_{New} - \hat{S}_{Orig}||_{Fro} = \sqrt{r}\lambda$ and

$$0 < ||\hat{S}_{Orig} - S^*||_2 - ||\hat{S}_{New} - S^*||_2 \leq \lambda. \tag{5.15}$$

$$0 < ||\hat{S}_{Orig} - S^*||_{Fro} - ||\hat{S}_{New} - S^*||_{Fro} \leq \sqrt{r}\lambda. \tag{5.16}$$

We can now analyze the performance of $\hat{\Sigma}_{New}$. Since we have no gain from $diag(\hat{\Sigma}_{New})$, we have to subtract from $||\hat{L}_{New} - \hat{L}_{Orig}||_{Fro}$ the gain from diagonal elements. At the same time, no gain comes from the diagonal elements of $\hat{S}_{New}$. Hence, we can write

$$||\hat{\Sigma}_{New} - \hat{\Sigma}_{Orig}||_{Fro} \leq \sqrt{r}\lambda.$$

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

$$0 < ||\Sigma_n - \hat{\Sigma}_{LOREC}||_2 - ||\Sigma_n - \hat{\Sigma}_{New}||_2 \leq \lambda. \tag{5.17}$$

$$0 < ||\Sigma_n - \hat{\Sigma}_{LOREC}||_{Fro} - ||\Sigma_n - \hat{\Sigma}_{New}||_{Fro} \leq \sqrt{r}\lambda. \tag{5.18}$$

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

To sum up, we pay the price of accepting a non-optimal solution in terms of nuclear norm (we allow to increment $||_{nuc}$ by $r\lambda$) but we have a best fitting performance for the whole covariance matrix, decrementing the squared Frobenius loss by a quantity bounded from $r\lambda^2$. Note that $||\hat{S}||_{off}$ is invariant. $||S||_1$ (considering also the diagonal) is decreased by a quantity bounded from $\sqrt{r}\lambda$.

We can easily write

$$||\hat{\Sigma}_{New} - \Sigma||^2_{Fro} = ||\hat{L}_{New} + \hat{S}_{New} - (L+S)||^2_{Fro} =$$

$$0 < ||\hat{\Sigma}_{New} - \Sigma_n + \Sigma_n - \Sigma|| \leq ||\hat{\Sigma}_{New} - \Sigma_n||^2_{Fro} + ||\Sigma_n - \Sigma||^2_{Fro}. \tag{5.19}$$

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

Therefore, by (5.18) and (5.19), it is straightforward that

$$0 < ||\hat{\Sigma}_{LOREC} - \Sigma||^2_{Fro} - ||\hat{\Sigma}_{New} - \Sigma||^2_{Fro} \leq r\lambda^2. \tag{5.20}$$

Analogously, it is easy to prove that

$$0 < ||\hat{\Sigma}_{LOREC} - \Sigma||_2 - ||\hat{\Sigma}_{New} - \Sigma||_2 \leq \lambda. \tag{5.21}$$

Now we recall the following expression:

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

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

From (5.20) we can conclude that

$$0 < ||(\hat{L}_{LOREC}+\hat{S}_{LOREC})^{-1} - \Sigma^{-1}||^2_{Fro} - ||(\hat{L}_{New}+\hat{S}_{New})^{-1} - \Sigma^{-1}||^2_{Fro} \leq r\lambda^2. \tag{5.22}$$

Analogously, it is straight forward that

$$0 < ||(\hat{L}_{LOREC} + \hat{S}_{LOREC})^{-1} - \Sigma^{-1}||_2 - ||(\hat{L}_{New} + \hat{S}_{New})^{-1} - \Sigma^{-1}||_2 \leq \lambda. \tag{5.23}$$

Our study has allowed us to improve the estimation performance in a finite sample analysis. However, the rates for $\hat{L}_{New}$, $\hat{S}_{New}$ and $\hat{\Sigma}_{New}$ are

exactly the same as $\hat{L}_{LOREC}, \hat{S}_{LOREC}$ and $\hat{\Sigma}_{LOREC}$. Our new estimate improves the statistical performance of LOREC given the sample, inheriting all its algebraic and parametric consistency properties.

In spite of that, the un-shrinkage of the estimated eigenvalues of L relaxes the necessary condition for positive definiteness and invertibility of $\hat{S}$ and $\hat{\Sigma}$. In empirical analysis, one can consider that parameters ϕ and ϕ_S can be decreased by a quantity bounded from λ.

LOREC and spiked eigenvalues: a relaxed sampling theory

Suppose now that the eigenvalues of L^* are pervasive in the sense of Proposition 2.5.1, and that all propositions and assumptions of POET approach hold in our finite sample context.

For instance, we suppose that

$$\lambda_{1,\dots,r}(\Sigma^*) \geq \epsilon p,$$
$$\lambda_{r+1,\dots,p}(\Sigma^*) \leq \epsilon p,$$

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

Suppose that the relationship between p and n is such that all the necessary conditions to prove the consistency of POET described in paragraph (2.5.4) hold (see Theorem 2 in [45]), included the assumptions on the sparsity structure of S^*. As already said, we drop the assumption (4.32).

In particular, suppose that (2.20), (2.21), (2.22) hold, such that (2.23) can be proved, that is,

$$||\mathbf{\Sigma_n} - \mathbf{\Sigma}|| = \mathbf{O}\left(\frac{\mathbf{p}}{\sqrt{\mathbf{n}}}\right) \tag{5.24}$$

holds. This is a key model-based result (outlined in bold), because it is necessary to prove the consistency of POET. It is proved as Lemma 5 in [45].

(5.24) is equivalent to state that

$$P\left(||E_n|| \geq C_1\frac{p}{\sqrt{n}}\right) \leq 1 - C_2e^{-C_3p^2}. \tag{5.25}$$

Since we have dropped the assumption (4.32), we can simply write, using the standard norm property $|||.||_\infty \leq ||.||_2$ as in [31] (see paragraph (4.1.3)),

$$P\left(||E_n||_\infty \geq C_1\xi(T)\frac{p}{\xi(T)\sqrt{n}}\right) \leq 1 - C_2e^{-C_3p^2}, \tag{5.26}$$

because $\rho = \gamma\lambda$ and γ has the same shape as in Theorem 4.1.4. We explicitly note that in this way we also drop the assumption of normality for the data, implicit in the result of [110] used by Luo to bound the spectral loss of the unbiased sample covariance matrix.

So, if we plug-in this expression in the proof of Theorem 4.1.4, and we use (5.26), we can write

$$\lambda_n = \left(\frac{1}{\xi(T)} \frac{p}{\sqrt{n}} \right) = \lambda. \tag{5.27}$$

Hence, we can exploit (4.36) to conclude

$$g_\gamma(\hat{S}_n - S^*, \hat{L}_n - L^*) \preceq \frac{1}{\xi(T)} \frac{p}{\sqrt{n}}, \tag{5.28}$$

given that all the necessary conditions (with particular attention to the identifiability ones) of Theorem 4.1.4.

Theorem 5.1.2. *Under all the assumptions of Theorem 2 in [45] (see paragraph (2.5.4)) and all the assumptions of Theorem 4.1.4, the LOREC estimate* $(\hat{L}, \hat{S})$ *satisfies*

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

It is straight forward that the success of this approach depends on the coherence between the assumptions in both settings (POET and LOREC). We will give specific attention to that in paragraph (5.3.1), widely describing the necessary setup conditions for ensuring this coherence.

Consistently to POET approach, here we can overcome the problem of the restrictive condition $p \leq n$. In fact, we know that the probabilistic bound is finite until $p \log(p) \gg n$, because Theorem 2 in [45] prescribes $p = o(n^2)$.

Note that all the described rates for $\hat{S}$ and $\hat{\Sigma}$ still hold, simply updating λ accordingly to (5.27). Also the described results on the un-shrinkage and the consequences on the requisites for positive definiteness and invertibility still hold.

In particular, since in this context $||\Sigma_n - \Sigma^*||_2$ is $o(p)$ with rate $O(\frac{p}{\sqrt{n}})$, we have

$$\phi = C(s\xi(T) + 1) \frac{1}{\xi(T)} \frac{p}{\sqrt{n}},$$

$$\phi_S = Cs\xi(T) \frac{1}{\xi(T)} \frac{p}{\sqrt{n}}.$$

In order to relax the strong assumption of pervasiveness of latent eigenvalues (Proposition 2.5.1), we set into the generalized spikiness context of Definition 2.5.1, where $\alpha \in (0, 1)$. In order to obtain an error rate for our numerical program under these conditions, since the nature of this approach comes from a non-asymptotic (finite sample) analysis, we only need to study the behaviour of the model-based quantity $P(||\Sigma_n - \Sigma||)$ under these assumptions, because the only probabilistic component derives from $P(||E_n||_2)$. In particular, we want to generalize (5.25) showing that

$$P\left(||\Sigma_n - \Sigma|| > C_1 \frac{p^\alpha}{\sqrt{n}} \right) \leq 1 - C_2 e^{-C_3 p^{2\alpha}}, \tag{5.29}$$

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

In order to do that, the relevant argument to take into account is Lemma 5 in Fan et al. (2013), the conclusion of which is (5.24). Since Lemma 5 (as it is) is the key to prove that under Fan's condition (5.28) holds, the updated version of Lemma 5 in the α - spiked context is the key to prove that

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

We remark again the difference with Luo's approach. In his setting, he proved that, given $E_n = \Sigma_n - \Sigma^*$,

$$P(||E_n||_2) \leq O_p\left(\sqrt{\frac{p}{n}}\right)$$

$$P(||E_n||_\infty) \leq= O_p\left(\sqrt{\frac{\log p}{n}}\right)$$

separately for $P(||E_n||_2)$ and $P(||E_n||_\infty)$.

The key to prove (5.29) is to adapt claims (2.20), (2.21), (2.22) (coming from [44]) to this setting, where the pervasiveness of latent eigenvalues has been relaxed, applying the proof technique in [45], Appendix C, Lemma 5, page 639.

From the fact that $||B'\Sigma^{-1}B|| \leq |cov(f)^{-1}|$ (page 194 Fan (2008) [43], Assumption (B)), (2.20) follows. This claim is unaffected by the relaxing of Proposition 2.5.1. So, from the proof of Lemma 5, we can argue that, under the α - spiked context, $||D_1|| \leq O(p^\alpha\sqrt{\frac{1}{n}})$, because now $||BB'|| = O(p^\alpha)$. This happens also because $r \log p = o(n)$.

In order to show how (2.21) changes in this context, we need to recall the key results of Bickel and Levina (2008b). Differently from Luo's approach, in this setting (as in the POET one) the sparsity assumption is imposed to S^*, and not to Σ^*.

The relevant quantity m_p (2.17) in Fan's setting is $o(p)$, in order to have $||S|| = o(p)$, which allows to identify the low rank component via PCA.

Here, since Definition 2.5.1 holds, we have that $m_p = o(p)$ is no longer appropriate. We impose, in order to preserve the correspondence between the rates 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 (2.11).

$$\left\{ S^* : s^*_{ii} \leq M, \qquad \sum_{j=1}^{p} |s^*_{ij}|^q \leq c_0(p), \qquad \forall i \right\}. \tag{5.30}$$

We have residual variances uniformly bounded by M. This assumption here is no longer valid, because M is no longer negligible respect to p.

Here we can no longer write (see [15] 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 $m_p = o_p(p^\alpha)$ causes that M cannot longer be considered as a constant when $p \to \infty$. 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. Plugging-in $M = O(p^{\alpha-1})$ in the proof deriving the sample covariance rate of a matrix under class (5.30) (see [15] page 2582) we can prove:

$$||\mathbf{\Sigma_n} - \mathbf{\Sigma}||_\infty \leq \mathbf{O}\left(\mathbf{p^{\alpha-1}}\sqrt{\frac{\log \mathbf{p}}{\mathbf{n}}}\right), \tag{5.31}$$

which is outlined in bold as a key technical result.

Now, using (5.31), we can apply the proof tools of Lemma 5 ([45], Appendix C) to matrix D_2, obtaining

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

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

$$||D_2|| \leq pO_p(p^{\alpha-1})O\left(\sqrt{\frac{\log p}{n}}\right) = O_p\left(p^\alpha\frac{1}{\sqrt{n}}\right). \tag{5.32}$$

To conclude, we analyze (2.22):

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

Note that here Assumption 2b) $||S^*||_1 < const$ in Theorem 2 of [45], necessary to ensure the consistency of POET, is no longer necessary, because rank consistency is ensured via the numerical method.

Since $r = O(\log(p))$ and $n = o(p^2)$, we can set $n = O(p^\alpha)$ and we obtain $O(\sqrt{\frac{r}{n}}) = O(p^{-\frac{\alpha}{2}})$, because $\log(p) = o(p^\alpha)$. This method works if and only if $p = o(n^{2\alpha})$. The rate thus becomes $O\left(p^{\frac{\alpha}{2}}\sqrt{\frac{\log p}{n}}\right)$.

Applying the tools of Lemma 5 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 $log(p) = o(n)$ leads to:

$$||D_3|| \leq O\left(\frac{p^\alpha}{\sqrt{n}}\right). \tag{5.33}$$

Rate (5.29) is consequently proved, and we have

$$||\Sigma_n - \Sigma|| = O\left(\frac{p^\alpha}{\sqrt{n}}\right). \tag{5.34}$$

The argument follows from the combined use of tools from Fan et al. (2013), Fan et al. (2011), Fan et al. (2008) and Bickel and Levina (2008b).

This is equivalent to state that

$$P\left(||E_n|| \geq C_1\frac{p^\alpha}{\sqrt{n}}\right) \leq 1 - C_2e^{-C_3p^{2\alpha}}.$$

Since we have dropped the assumption (4.32) for Σ^*, we can simply write, using $||.||_\infty \leq ||.||_2$ and the minimum for γ in Theorem 4.1.4,

$$P\left(||E_n||_\infty \geq C_1\xi(T)\frac{p^\alpha}{\sqrt{n}}\right) \leq 1 - C_2e^{-C_3p^{2\alpha}}. \tag{5.35}$$

By the outlined plug-in in the proof of Theorem 4.1.4 and (5.35), exploiting Chandrasekaran et al. (2012) ([31]) (see paragraph (4.1.3)), it is possible to prove that under these assumptions we have:

$$g_\gamma(\hat{S} - S^*, \hat{L} - L^*) \preceq \frac{1}{\xi(T)}\frac{p^\alpha}{\sqrt{n}}, \tag{5.36}$$

given that all the necessary conditions (with particular attention to the identifiability ones) of Theorem 4.1.4 hold.

Theorem 5.1.3. *Under all the assumptions of Theorem 2 in [45], assuming that the latent eigen-structure of Proposition 1 and 2 (see paragraph (2.5.4)) is replaced by the one of Definition 2.5.1, and under all the assumptions of Theorem 4.1.4, the LOREC estimate* $(\hat{L}, \hat{S})$ *satisfies*

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

It is straight forward that the success of this approach depends on the coherence between the relaxed spikiness assumption (Proposition 2.5.1, see the discussion of [45] by Yu and Samworth on that) as well as all the assumptions in Theorem 2 of Fan et al. (2013) and the assumptions of Theorem 4.1.4.

Consequently, we can write here

$$\lambda_n = \left(\frac{1}{\xi(T)}\frac{p^\alpha}{\sqrt{n}}\right) = \lambda. \tag{5.37}$$

We can again overcome the problem of the restrictive condition $p \leq n$. In this relaxed setting, the probabilistic bound is finite until $p^\alpha \log(p) \gg n$, because (5.34) holds until $p^\alpha = o(n^2)$.

Note that if $\alpha = 0$, we have $\log(p) \gg n$, which means $p = o(n)$. So, in the case of no latent eigenvalues (no spikiness), the convergence rate of the sample covariance matrix simply becomes $O(\sqrt{\frac{1}{n}})$. Note that Theorem 2.2.1 gives the same result imposing $p = o(n)$. Therefore, we can say that (5.34) holds for $\alpha \in [0, 1]$, thus encompassing also the classic sampling context (small and fixed data dimension). In addition, (5.36) holds also under the no-spikiness case of Theorem 4.1.4.

All the described rates for $\hat{S}$ and $\hat{\Sigma}$ still hold, simply updating λ accordingly to (5.37). The described results on the un-shrinkage and the consequences on the requisites for positive definiteness and invertibility still hold too, consequently updated.

In particular, since in this context $||\Sigma_n - \Sigma^*||_2$ is $o(p)$ with rate $O(\frac{p^\alpha}{\sqrt{n}})$, we have

$$\phi = C(s\xi(T) + 1)\frac{1}{\xi(T)}\frac{p^\alpha}{\sqrt{n}},$$

$$\phi_S = Cs\xi(T)\frac{1}{\xi(T)}\frac{p^\alpha}{\sqrt{n}} = Cs\frac{p^\alpha}{\sqrt{n}}.$$

This approach offers an original proof setting to recover consistently a more relaxed (and wider) spikiness context. By plugging-in into the proof of Luo (2013), it allows to overcome the condition $p \leq n$ even using $\hat{\Sigma}_{n-1}$ as estimation input. It offers a recovery context where the rate directly depends on the spikiness of latent eigenvalues, because the larger α, the further are the identifiability and invertibility conditions from being satisfied, as well as the worse is the error rate. We underline that our rates are in absolute norms, and reflect the underlying degree of spikiness.

However, this approach works if and only if the identifiability and consistency assumptions of LOREC and POET 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 note that this theory is specifically addressed to the Big Data context, where $p \gg n$. Sparse factor model assumptions together with the numerical approach are the key to provide recovery in a relaxed sampling setting, particularly useful when p is very large compared to n. This result is obtained by a combined use of numerical analysis (finite sample) and probabilistic convergence theory of the sample covariance matrix under sparse factor model assumptions, linking the sample dimension to the spikiness of latent eigenvalues.

We are going to verify the strength and the width of all these assumptions as well as the validity of our theories on the performance of numerical esti-

mators, with particular reference to the statistical advances just described, in a wide original simulation study and in a real data analysis context.

5.2 Simulation setting

5.2.1 Simulation algorithm

Let C, S, L and W be real-valued symmetric $p \times p$ matrices. Let us consider a framework where C is a $p \times p$ unbiased sample covariance matrix, L is the latent low rank covariance matrix (i.e. factor-driven covariance), S is a sparse residual covariance matrix with $2s$ ($s \ll p(p-1)/2$) non-zero elements and W is an error term.

Our aim is to decompose the matrix C (which is for us the unbiased covariance matrix estimator) into the sum of S, L and W, satisfying the extended "lasso" condition (5.38):

$$\begin{aligned} \min_{S,L} \ & \rho||S||_1 + \lambda||L||_{nuc} + ||W||_{Fro} \\ & sub \qquad C = S + L + W, \end{aligned} \tag{5.38}$$

where $||\cdot||_1$ is the l_1 matrix norm, and $||\cdot||_{nuc}$ is the nuclear norm, i.e. the trace of the vector of singular values, λ and ρ are non-negative scalars. For us, the l_1 norm is here excluding the diagonal elements, that is $||S||_1 = ||S||_{1,off} = \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} |s_{ij}|$, according to POET approach.

The matrices C and S are positive definite, the matrix L is positive semidefinite. The parameters ρ and λ are the sparsity and spikiness thresholds respectively. Our aim is to obtain the estimate $\hat{\Sigma} = \hat{L} + \hat{S}$ of the true covariance matrix Σ minimizing (5.38).

With this purpose in mind, we now describe the data generation framework. First, we set to $r = \beta p$, $\beta \in [0, 0.1]$, the rank of L. We can thus apply to L (real-valued and symmetric) the spectral theorem:

$$L = U_L \Lambda_L U_L', \tag{5.39}$$

where:

1. U_L is a $p \times r$ matrix with orthonormal columns, i.e $U_L' U_L = I_r$;
2. $\Lambda_L = diag(\lambda_{L,1}, \ldots, \lambda_{L,r})$ is a $r \times r$ diagonal matrix, where $\lambda_{L,1}, \ldots, \lambda_{L,r}$ are real and positive, since L is positive semidefinite.

For our purpose, we immediately need to set the proportion $\alpha \in [0, 1]$ of the total variance explained by the factors. So, in the generation framework we can set $tr(\Lambda_L) = \tau \alpha p$, where $\tau \in [0, \infty[$ allows to control for the magnitude.

The matrix U_L is generated applying the Gram-Schmidt algorithm to any basis of R^p and extracting r random $p-$dimensional columns from the obtained matrix. This is performed pre-multiplying by a positive definite permutation matrix the matrix I_p, and then applying Gram-Schmidt algorithm. The matrix Λ_L is generated by an algorithm (see [48]) which returns a diagonal matrix with fixed trace $\tau\alpha p$ and condition number exactly equal to c.

The sparse symmetric matrix S, which is a $p \times p$ sparse matrix with $2s$ off-diagonal nonzero elements ($s \ll p(p-1)/2$), is generated as follows.

First of all, we need to split the residual variance $\tau(1-\alpha)p$ among the diagonal elements of S. This problem can be solved by using the Dirichlet probability distribution. It is sufficient to set $s^*_{ii} = \frac{s_{ii}}{\tau(1-\alpha)p}$. Then, $(s^*_{11}, \ldots, s^*_{pp}|(1-\alpha, \ldots, 1-\alpha))$ is a Dirichlet distribution. We can generate s^*, and consequently compute s. We permute the elements in $diag(S)$ associating the $i-$th largest element in $diag(L)$ with the $i-$th largest element in $diag(S)$.

The off-diagonal elements of S are generated as follows. For each entry i, j a number $\theta_{ij} = Unif(0, \delta\sqrt{s_{ii}}\sqrt{s_{jj}})$ is generated, where $\delta \in [0,1]$ is a parameter controlling for the positive definiteness of S. The larger the dimension p is, the smaller δ has to be in order to ensure positive definiteness. Then, s_{ij} is generated as $sign(L(i,j))Unif(0, \theta_{ij})$ for each i, j.

Once we have generated L, we compute $inc(L)$ (see (4.11) for the definition). Using the identifiability inequality $deg_{max}(S)inc(L) \leq \frac{1}{108}$, we set $deg_{max}(S) = \frac{1}{108 \cdot inc(L)}$. Using the lower bound on the minimum eigenvalue of L $\lambda_r(L)$ (Theorem 4.1.4), we derive the minimum allowed non zero element $thr_{min} = \frac{\sqrt{\frac{p}{n}} \cdot inc(L)^2 \lambda_r(L)}{deg_{max}(S)}$,where $\sqrt{\frac{p}{n}}$ comes from the shape of λ. From thr_{min} we can derive s_{min} as the position occupied by the lowest element larger than thr_{min} in the sorted vector of the off diagonal entries of S (in descending order). Then, a threshold thr_{prop} is proposed as $\delta_{bis} \in [0,1]$ times the maximum off-diagonal element of S, from which we can derive the proposed number of nonzero elements s_{prop} in the same way. The number of non zeros is then set to $s = \min(s_{min}, s_{prop})$, and the lowest allowed element of S is derived accordingly as $thr = \max(thr_{prop}, thr_{min})$.

Note that s_{min} is an approximate indication. It represents a control procedure respect to the correspondence with the theoretical assumptions of Theorem 4.1.4. In any case, it may happen that the maximum eigenvalue of Σ_n is actually more than proportional or less than proportional to $\sqrt{\frac{p}{n}}$. In that case, s_{min} can give a too restrictive or a too generous indication, and this may result in partial recovery or non-recovery of non-zeros. In addition, the choice of δ_{bis} is also arbitrary, and is limited by s_{min} only. This procedure is an attempt to deal with the alignment between the number of non zeros and the magnitude of non zeros (which is relevant for recovery). The model

selection criterion we are going to describe will appropriately signal problems on that, recovering in case more or less non-zeros than expected.

In light of this, we can generate n replicates of our data. Given the generated $L = U_L \Lambda_L U_L'$ and S, the data generation process is:

$$z_i = Bu_i + \epsilon_i,\ i = 1, \ldots, n,$$

where:

1. z_i is a $p \times 1$ vector;
2. $B = U_L D$ is a $p \times r$ matrix, with $D = \sqrt{\Lambda_L}$;
3. $u_i \sim N(0, I_r)$;
4. $\epsilon_i \sim N(0, S)$;
5. $u_i \perp \epsilon_i,\ i = 1, \ldots, n$.

Once n replicates have been generated, we can compute the matrix C as the unbiased sample covariance estimator of our n replicates of z.

The spikiness threshold λ is initially set to the mean eigenvalue of C (say $\bar{\lambda}_C$), while the sparsity threshold ρ is initially set to the average of the off-diagonal elements of C ($\rho_C = (\frac{p(p-1)}{2})^{-1} \sum_{i=1}^{p-1} \sum_{j=i}^{p} |c_{ij}|$).

5.2.2 Simulated settings and comparison quantities

After the description of our generation framework, we come back to our statistical problem. Let us suppose that $\Sigma = L + S$ is a $p \times p$ covariance matrix, where L is a r - ranked matrix ($r < p$) and S is a sparse matrix with s non zero elements as in model (3.1). We set $C = \Sigma_n$, where Σ_n is now the unbiased covariance matrix estimator $\hat{\Sigma}_{n-1}$.

We take as reference setting the following one:
setting 1:

$$p = 100, n = 1000, \beta = 0.04, r = 4, \tau = 1, \alpha = 0.7, c = 2,$$

$$\delta = 0.1, \delta_{bis} = 0.2, s = 118, s_{max} = 732, \rho_{corr} = \frac{\rho_S}{\rho_\Sigma} = 0.045,$$

where $\rho_S = \frac{p(p-1)}{2})^{-1} \sum_{i=1}^{p-1} \sum_{j=i}^{p} |s_{ij}|$ and $\rho_\Sigma = \frac{p(p-1)}{2})^{-1} \sum_{i=1}^{p-1} \sum_{j=i}^{p} |\sigma_{ij}|$.

The dimension p is fixed to 100 and the sample dimension n is set to 1000. A data-set with a larger dimension will be explored in paragraph (5.3.2). These settings are good for comparing the performance of our NEW method to the LOREC method. The latent rank is 4, the magnitude parameter τ is fixed to 1. The proportion of non-zeros is $(\frac{p(p-1)}{2})^{-1} s$ is 2.38%.

The other settings we have explored are the following:
setting 2:

$$p = 100, n = 1000, \beta = 0.03, r = 3, \tau = 3, \alpha = 0.8, c = 4,$$

$$\delta = 0.1, \delta_{bis} = 0.1, s = 580, s_{max} = 1604, \rho_{corr} = 0.0072,$$

setting 3:

$$p = 100, n = 1000, \beta = 0.04, r = 4, \tau = 1, \alpha = 0.7, c = 4,$$

$$\delta = 0.1, \delta_{bis} = 0.1, s = 335, s_{max} = 892, \rho_{corr} = 0.0048.$$

In **setting 2**, the magnitude is increased by three times (τ passes from 1 to 3). The rank is 3, the proportion of latent variance is increased to 0.8. The proportion of non zeros is increased to 11.72%. The condition number c is increased to 4. This setting has quite more spiked eigenvalues.

In **setting 3**, the condition number c is 4, and the number of non-zeros is increased respect to the reference setting. The proportion of non-zeros here is 6.77%. This setting is something between **setting 1** and **setting 2**.

The spikiness threshold λ is initially set to the mean eigenvalue of Σ_n, $\bar{\lambda}_{\Sigma_n}$. The sparsity threshold ρ is initially set to the average of the absolute values of the off-diagonal elements of Σ_n, $\rho_{\Sigma_n} = (\frac{p(p-1)}{2})^{-1} \sum_{i=1}^{p} \sum_{j=i}^{p} |\sigma_{n,ij}|$.

In **setting 1** we have:

$$\lambda = \left[\frac{2i}{10} \lambda_{\Sigma_n} \frac{\sqrt{n}}{p} \right], i = 1, \ldots, 20; \tag{5.40}$$

$$\rho = \left[4i \frac{\log(p)}{n} \rho_{\Sigma_n} \right], i = 1, \ldots, 20. \tag{5.41}$$

These formulations are adapted in each setting by succesive approximations.

Lots of quantities are computed in order to describe comparatively the performance of our NEW approach, of LOREC (Luo, 2013) and POET (Fan et al., 2013) on the same data. The computation algorithm is described in Section 3 (paragraph (3.2.2)), and is applied to the generated 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 the following:

- $Loss = ||\hat{S} - S||_{Fro} + ||\hat{L} - L||_{Fro}$,
- $TotalLoss = ||\hat{\Sigma} - \Sigma||_{Fro}$,
- $SampleTotalLoss = ||\hat{\Sigma} - \Sigma_n||_{Fro}$.

The estimated proportion of total variance $\hat{\alpha}$ and the residual covariance proportion $\hat{\rho}_{corr}$ are computed.

The performance of $\hat{S}$ is assessed by using 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 can define:

- the estimated proportion of non-zeros $perc_{nz} = nz/numvar$,
- the *error* measure: $err = \frac{fp+fn}{numvar}$, where $numvar = p(p-1)/2$ is the number of off-diagonal elements,
- $errplus = \frac{fpos+fneg}{s}$, which is the same as err but computed for non-zeros only, distinguishing between positive and negative in the usual way.

Sensitivity and specificity measures are then derived, as the correct classification rates of (true) non-zeros and zero elements respectively. Sensitivity and specificity rates are derived also for positive, zero and negative elements separately, disentangling the error rates computing the elements classified by mistake in each of the other two classes.

The overall error rate err_{tot} using the number of false zeros, false positive, and false negative elements is also computed as $err_{tot} = \frac{fpos+fneg+fn}{numvar}$.

The condition numbers of $\hat{\Sigma}$, $\hat{S}$, $\hat{L}$ are computed and compared to the ones of Σ and S and L. We compute the error rates for the estimated eigenvalues of L, S, and Σ, and provide a comparative analysis of the gains on the three indicated losses coming from the unshrinkage procedure for all threshold parameters.

The vector of the eigenvalues of Σ_n and its Euclidean distance from the vector of eigenvalues of Σ are computed, as well its condition number. The spectral and the Frobenius losses of Σ_n from Σ are calculated too.

The performance of $\hat{\Sigma}^{-1}$ for all estimators in terms of Frobenius loss from Σ^{-1} is also investigated: $InvTotalLoss = ||\hat{\Sigma}^{-1} - \Sigma^{-1}||_{Fro}$.

All these statistics are computed and averaged over $N = 100$ replicates.

5.2.3 A new model selection criterion

We now develop a model selection criterion specifically thought for our estimation method. The inspiration rises from the reference norm g_γ used by Luo (see (4.20)), which is the starting point of our analysis:

$$g_\gamma = \max\left(\frac{||\hat{S} - S||_\infty}{\gamma}, ||\hat{L} - L||_2\right) \tag{5.42}$$

From (5.42), the **need** of **rescaling** both arguments of g_γ rises in order to raise informative power and to detect the optimal point in the spiki-

ness/sparsity trade-off. For exploiting (5.42) with model selection purposes, we need to make the two terms comparable.

How can we compare the goodness of fit of the sparse term by the estimated l_1 norm of the sparse component and of the low rank term by the estimated nuclear norm of the low rank component? How it is possible to establish if their equilibrium is intrinsically balanced? In order to perform a sample comparison between $||\hat{L}||_2$ and $\frac{||\hat{S}||_\infty}{\gamma}$, we need to find a unique comparison ground for them.

Considered that $\frac{||\hat{S}||_\infty}{\gamma}$ contains a maximum norm, we can re-scale it to the trace of $\hat{S}$. Given that in our simulation setting

$$trace(S^*) = (1-\alpha)trace(\Sigma^*),$$

$trace(\hat{S})$ is estimated by $(1-\hat{\alpha})trace(\Sigma_n)$. Similarly, in order to compare the magnitude of the two quantities, we multiply $||\hat{L}||_2$ by r, which is the bound for the maximum norm of $\hat{L}$, and then divide it by the trace of $\hat{L}$, estimated by $\hat{\alpha}trace(\Sigma_n)$.

Our maximum criterion MC can be therefore defined as follows:

$$MC = \max\left\{\frac{\hat{r}||\hat{L}||_2}{\hat{\alpha}trace(\Sigma_n)}, \frac{||\hat{S}||_\infty}{\hat{\gamma}(1-\hat{\alpha})trace(\Sigma_n)}\right\}, \tag{5.43}$$

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

This 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. If τ is not large enough to ensure that the largest eigenvalue of S is not too small, there are possible problems for non zeros recovery (identifiability problems). τ must be large enough to guarantee the lower bound on the minimum non zero entry of S and that its maximum eigenvalue scales with $\sqrt{\frac{p}{n}}$. Analogously, if δ_{bis} is too small, that is if we allow for very small non zero off-diagonal entries in S, the method is not able to recover them. In addition, also α and c can influence the nonzero choice, controlling the spikiness of the low rank component.

We note that the MC method performs considerably better than the usual cross-validation using H-fold Frobenius Loss (used in (Luo, 2013)), since minimizing a loss based on sample approximation like the Frobenius one causes that the parameter $\hat{\alpha}$ is shrunk too much. Quantities ρ_{corr} and nz are also usually underestimated in that way, unless the true s is really low. Simulating $N = 100$ samples, we have that $Loss$, $SampleTotalLoss$ and $TotalLoss$ are considerably higher using the thresholds obtained by Frobenius cross validation, both for $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{NEW}$.

On the contrary, 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, taking into account the latent low rank and sparse structure simultaneously. MC criterion thus offers a unique comparison ground for both penalties simultaneously considered. Selecting thresholds focusing on the fitting performance highlights that cross-validation is worse than using MC criterion, also because the un-shrinkage procedure corrects itself for the fitting performance. In addition, MC criterion takes into account rank and sparsity pattern detection simultaneously.

For selecting the thresholds for POET estimation, the cross validation method described in paragraph 4 of [45] is used. There, the set of residuals from PCA is divided in a training and a validation set. On the first, POET method is applied. On the second, the sample residual covariance matrix is computed. The optimal threshold is then chosen minimizing the average Frobenius Loss of the estimated sparse component. The training set dimension is $n_{training} = n(1 - \log(n)^{-1})$, the validation set dimension is $n_{validation} = n - n_{training}$. For us, $n_{training} = 855$ and $n_{validation} = 145$.

For rank selection, POET procedure exploits the classical Bai and Ng criteria, as indicated in paragraph 2.4 of [45]. The risk of underestimating the latent rank if the eigenvalues of Σ do not scale with p were pointed out in the discussion of [45] by Yu and Samworth. We note that the authors used there the Relative Error measure $||\hat{\Sigma} - \Sigma||_{\Sigma} = p^{-1/2}||\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_p||_{Fro}$ as a reference norm, which will also be computed for LOREC and NEW.

We note that POET systematically overestimates the proportion of variance explained by the factors (given the true rank) because the eigenvalues of Σ_n are more spiky than the true ones (see Theorem 2.3.1, by Ledoit and Wolf). The shrinkage approach corrects for that.

The condition number of $\hat{L}$ is usually smaller than c. This drawback depends on Theorem 2.3.1, and unfortunately holds also for LOREC and NEW (not only for POET). It depends on the eigenvalues of Σ_n. The ratio between the first and the r-th largest eigenvalue of Σ_n tends to be smaller than the true one, even if it can vary a lot across replicates, for large values of r, c and τ too. In fact, we note that the r-th largest eigenvalue of Σ_n is usually larger than the r-th largest eigenvalue of Σ.

5.3 Data analysis results

In this section we describe the results of the application of our method respect to the competitors under various situations. In paragraph (5.3.1), we describe the performance of $\hat{\Sigma}_{NEW}$ in the simulated settings described in section (5.2.1), comparing it with the one of $\hat{\Sigma}_{LOREC}$. Particular emphasis is given to the advantages and the performance of unshrinkage, on which clear indications are given. Threshold selection is performed using the model selection criterion described in (5.2.3). All the relevant quantities described in (5.2.2) are computed and averaged over $N = 100$ simulated settings. Sim-

ulations are performed with MATLAB. The results are reported in form of technical report.

In paragraph (5.3.2) two real examples are reported. The first is drawn from [45], and is a UK market data-set. The second is a supervisory banking data-set, which is derived from the balance sheet data of a list of relevant Euro Area banks. The calculations here reported treat these data only on the variable side, in fulfillment of confidentiality obligations. We deeply acknowledge for that the Supervisory Statistics Division of the European Central Bank, where the author spent a semester as a PhD Trainee, for the allowance of these data for research purposes.

5.3.1 Simulation results

We now start analyzing the performance of $\hat{\Sigma}_{New}$ in comparison to the one of $\hat{\Sigma}_{LOREC}$ on our reference setting (**setting1**), which is contained in the following table:

p	100
c	2
tau	1
alpha	0.7
r	4
s	118
s_max	732
delta	0.1
delta_bis	0.2

First of all, we simulate one draw and compute $\hat{\Sigma}_{n-1}$. In figures (5.1), (5.2) and (5.3) we can see the most important features of the generated setting. Figure (5.1) shows the top $r = 4$ eigenvalues of Σ, L and S respectively. Σ and L have spiked eigenvalues linearly distributed, almost overlapped. S has much lower eigenvalues. Figure (5.2) shows the sorted diagonal elements of L and S. Only the first three variances of S are larger than the ones in L. Figure (5.3) shows the sorted eigenvalues of Σ and Σ_n. We note a jump in correspondence of $r = 4$. The sorted eigenvalues from the fifth to the last of both matrices are much lower. This setting is consistent to POET assumptions too.

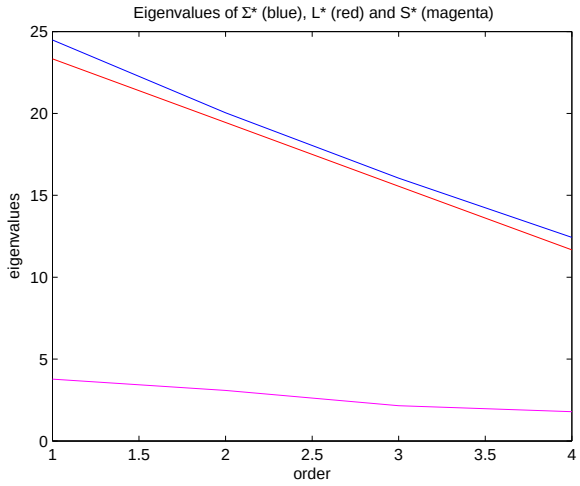

Figure 5.1: Eigenvalues of L, S, Σ

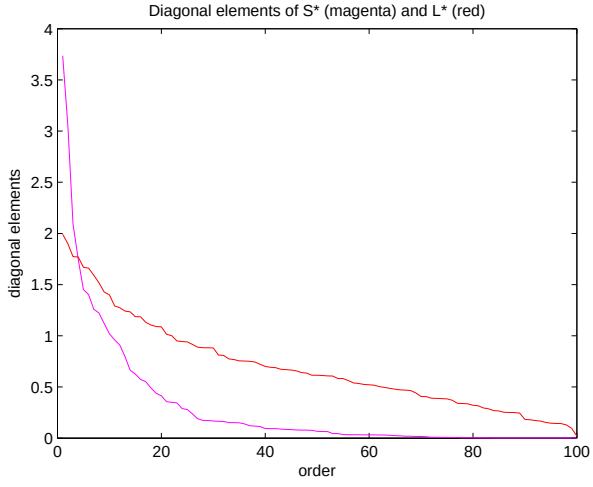

Figure 5.2: Sorted diagonal elements of L and S

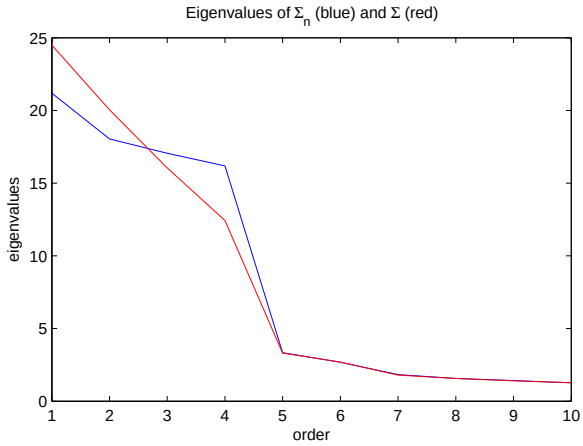

Figure 5.3: Eigenvalues of Σ_n and Σ

The thresholds ρ and λ are computed using formulas (5.40) and (5.41):

rho	lambda
0.0047908	0.062663
0.0095817	0.12533
0.014373	0.18799
0.019163	0.25065
0.023954	0.31332
0.028745	0.37598
0.033536	0.43864
0.038327	0.50131
0.043118	0.56397
0.047908	0.62663
0.052699	0.6893
0.05749	0.75196
0.062281	0.81462
0.067072	0.87729
0.071863	0.93995
0.076654	1.0026
0.081444	1.0653
0.086235	1.1279
0.091026	1.1906
0.095817	1.2533

We perform estimation for all the threshold pairs we can obtain from these two grids (i.e. $20 \times 20 = 400$).

We then compute the model selection criterion MC. The results are shown in figure (5.4) for $\hat{\Sigma}_{NEW}$ and in figure (5.5) for $\hat{\Sigma}_{LOREC}$.

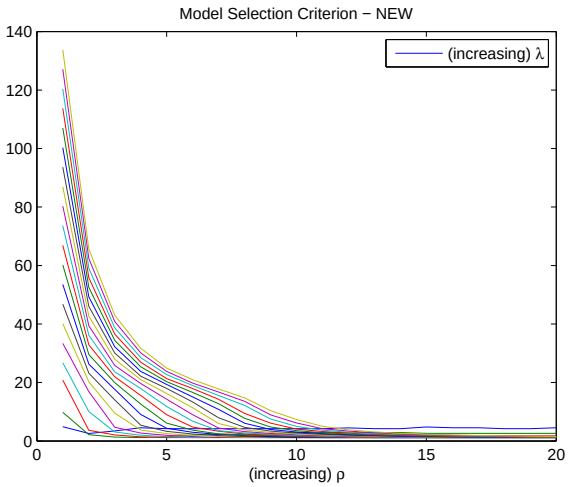

Figure 5.4: Model selection criterion - $\hat{\Sigma}_{NEW}$

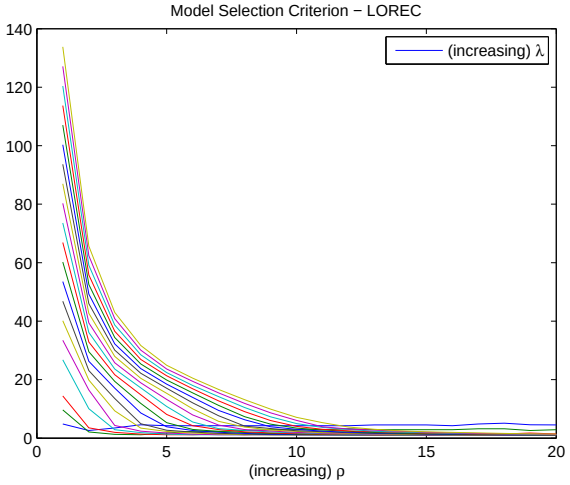

Figure 5.5: Model selection criterion - $\hat{\Sigma}_{LOREC}$

We can see that MC criterion (5.43) is usually increasing in ρ and λ, with the exception of the very first thresholds in both grids. For $\hat{\Sigma}_{NEW}$, the selected thresholds are $\rho(4) = 0.0192$ and $\lambda(2) = 0.1253$, for $\hat{\Sigma}_{LOREC}$ $\rho(6) = 0.0287$ and $\lambda(3) = 0.1880$.

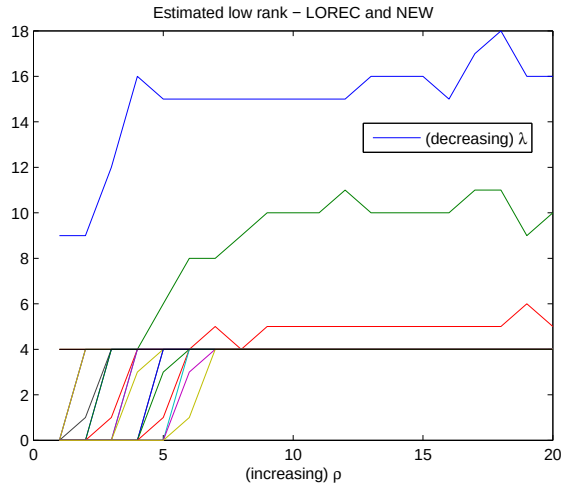

Figure 5.6: Estimated rank - $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{NEW}$

In figure (5.6) we have the distribution of the estimated rank for both methods. For very small λ, we have very large estimated ranks, for very large λ we have that the rank is sometimes shrunk to 0. For the central values of λ, the rank is correctly recovered. The sparsity parameter ρ also plays a role: if it is large enough, it can counterbalance the effect of λ, thus correctly estimating the true rank ($r = 4$, black line) even if λ is large.

In figure (5.7) and (5.8) we report the differences between the Total Losses and the Sample Total Losses of LOREC and NEW. We have 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 4.1.4). This pattern is more remarkable for Sample Total Loss than for Total Loss. For both losses and each λ, we note that the gain across ρ never overcomes its maximum $\sqrt{r}\lambda$ (plotted for each λ).

The dynamics of the difference between the Losses of LOREC and NEW, reported in figure (5.9), is quite more controversial. There we have some negative values even for central threshold values. This is due to the differences between the losses of the sparse component for incorrect thresholds (see figure (5.10)) which is better for $\hat{\Sigma}_{LOREC}$ when the latent rank is not exactly recovered or the estimated number of non-zeros is null. On the contrary, the difference between the losses of the low rank components is always better for $\hat{\Sigma}_{NEW}$ than for $\hat{\Sigma}_{LOREC}$ (see figure (5.11)).

The settings for which we have negative differences are characterized by a very large ρ which makes the sparse component too sparse. In that case, LOREC is underestimating the number of non-zeros in the sparse component, such that the unshrinkage gets the situation even worse. Anyway, for the thresholds selected by MC criterion, the gains obtained via unshrinkage are largely positive for all losses. Besides, the Loss relative to the low rank component is always much more relevant in absolute terms respect to the one relative to the low rank component.

We note also that if we linearly add a quantity to the eigenvalues of L estimated via the LOREC method, we usually improve the Total Loss. This is true even if we add a quantity larger than λ (unless λ is very high); however, the proportion of variance explained by the factors α and the number of nonzeros are in that case completely missed. In fact, the strength of our method is in the fact that the unshrinkage corrects for the underestimation of α when LOREC method exactly recovers rank and sparsity pattern. Given that the rank and the sparsity pattern are correctly recovered, the unshrinkage provides the closest solution to the true Σ and the closest proportion of latent variance to the true α. This happens while POET overestimates and LOREC underestimates α. Ad-hoc simulations provide a confirmation. The best estimate $\hat{\alpha}$ is reached for the thresholds which best recover rank and sparsity pattern. This it the same reason why the usual cross validation method based on sample total loss has a poorer performance.

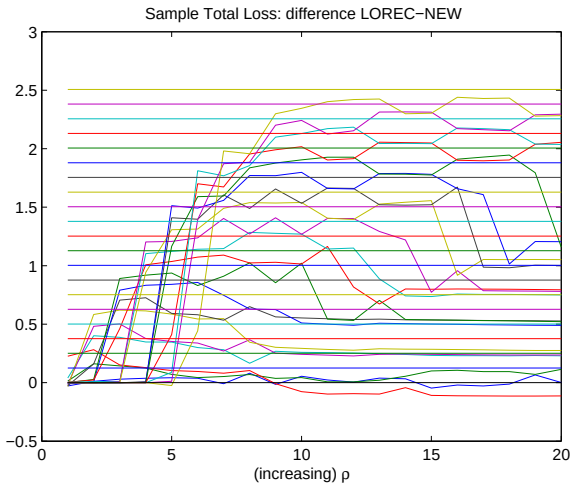

Figure 5.7: Sample Total Loss difference - $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{NEW}$

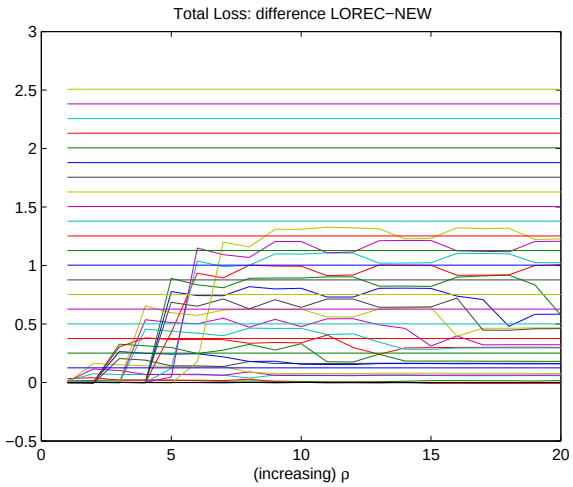

Figure 5.8: Total Loss difference - $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{NEW}$

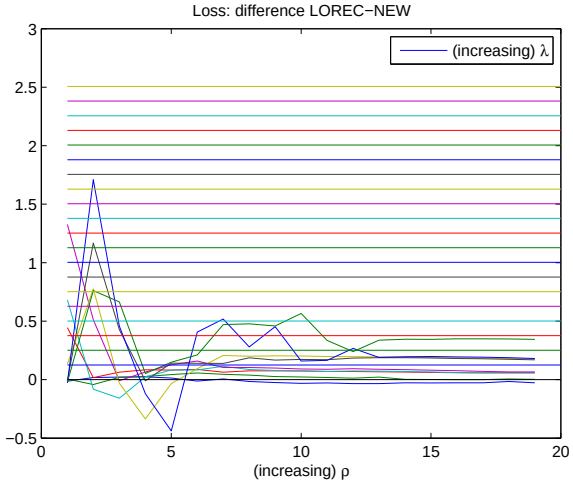

Figure 5.9: Loss difference - $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{NEW}$

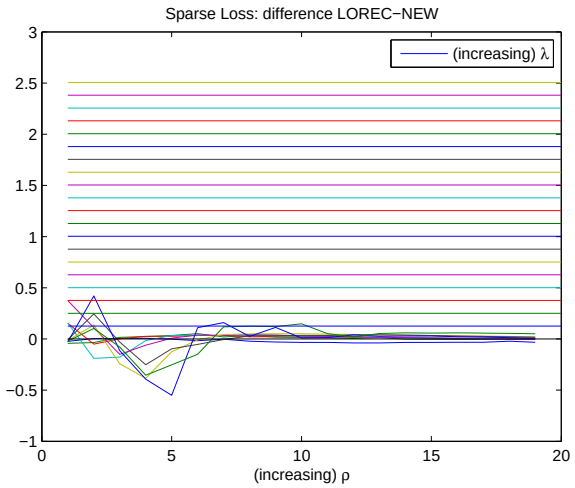

Figure 5.10: Sparse Loss difference - $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{NEW}$

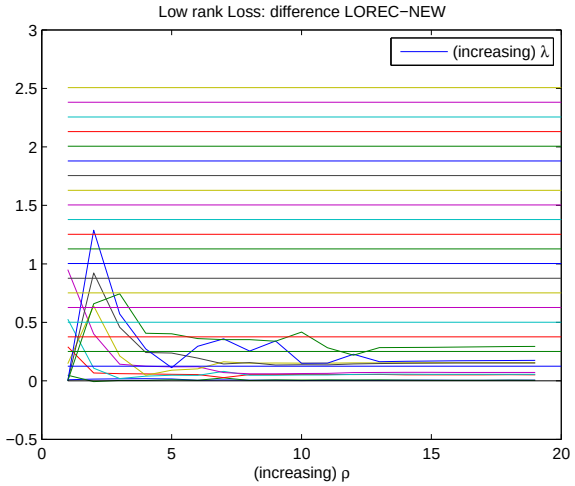

Figure 5.11: Low rank Loss difference - $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{NEW}$

In figure (5.12) we report the plot of the estimated proportion of latent variance α across thresholds for $\hat{\Sigma}_{NEW}$ (in black the true $\alpha = 0.7$). We note that for each λ, $\hat{\alpha}$ usually increases and then gets stable across ρ. The larger λ, the smaller $\hat{\alpha}$. We point out that in correspondence to the smallest values of ρ the estimated α is 0, provided that λ is enough large.

In figure (5.13) the proportion $\hat{\alpha}$ is shown for $\hat{\Sigma}_{LOREC}$. The shape is exactly the same as for $\hat{\Sigma}_{NEW}$, the only difference is that all the patterns are negatively shifted.

In figure (5.14) we report the plot of the estimated proportion of residual covariance $\hat{\rho}_{corr}$. We have inserted only the ten largest values of ρ. We note that the larger is λ, the lower is $\hat{\rho}_{corr}$ across sparsity thresholds. In black we have the true $\rho_{corr} = 0.045$.

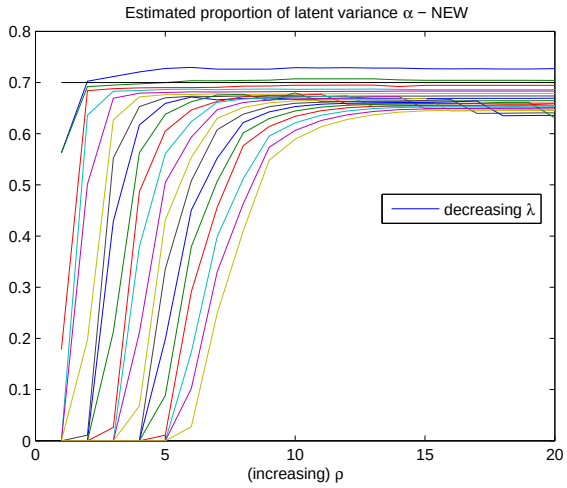

Figure 5.12: Estimated proportion of latent variance - $\hat{\Sigma}_{NEW}$

In figure (5.15) we report the estimated number of non-zeros across thresholds (in black the true $s = 118$). In general, we have that the larger is ρ, the lower is nz. The spikiness parameter λ impacts on the rate of the decay across ρ: the larger it is, the slower the decay.

The error measure err, reported in figure (5.16) shows a minimum across ρ for each λ. The larger λ, the larger is the ρ in correspondence to which the minimum is attained.

The specificity measure (figure (5.17)) is larger for small λ. It reaches 1 for completely diagonal sparse estimates.

The sensibility measure (5.18) is persistently larger for larger λ. The larger λ, the smaller is the value of ρ in correspondence to which the sensibility is 0, because in that case we have diagonal sparse solutions.

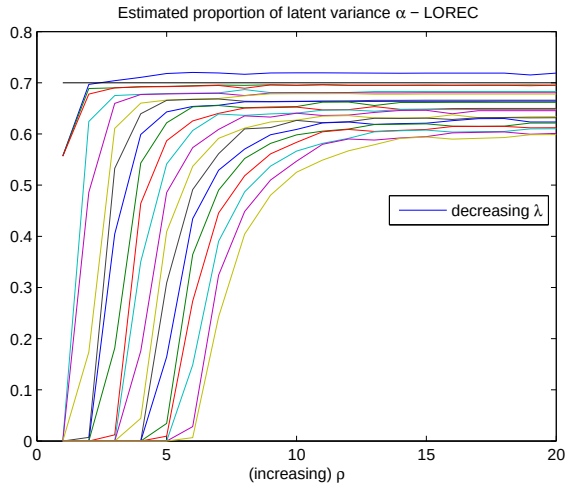

Figure 5.13: Estimated proportion of latent variance - $\hat{\Sigma}_{LOREC}$

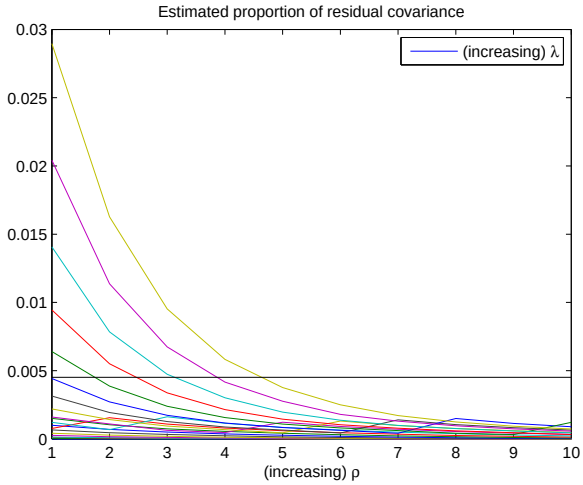

Figure 5.14: Estimated proportion of residual covariance - $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{NEW}$

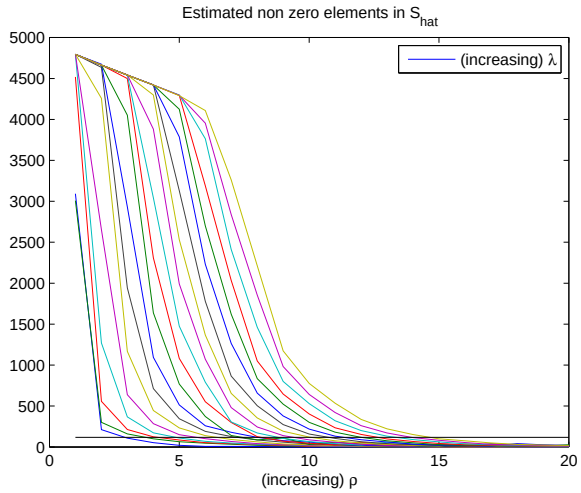

Figure 5.15: Estimated number of nonzero elements - $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{NEW}$

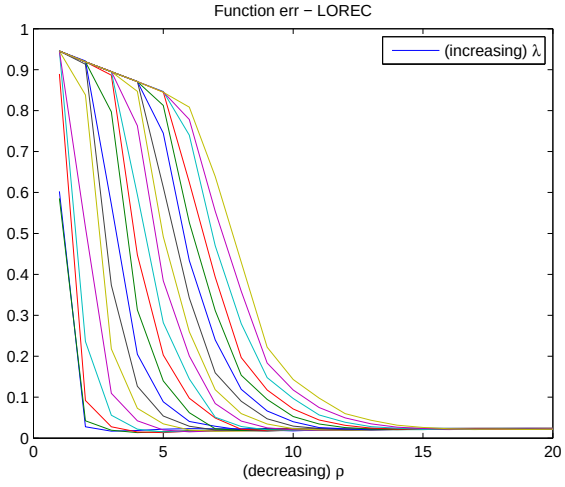

Figure 5.16: Error measure *err* - $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{NEW}$

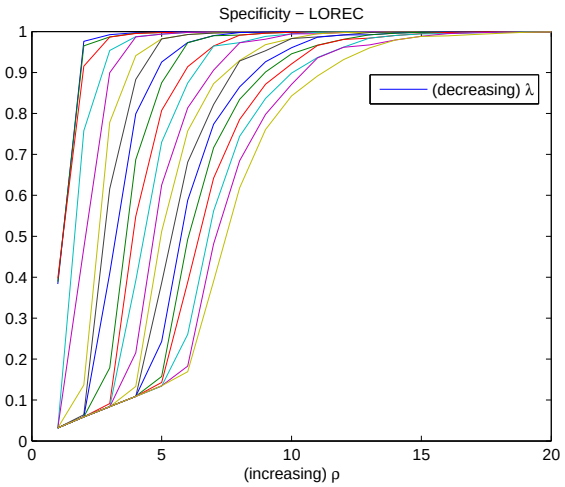

Figure 5.17: Specificity - $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{NEW}$

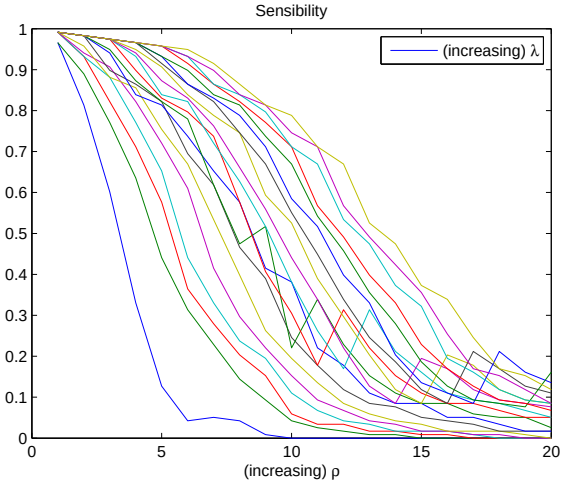

Figure 5.18: Sensibility - $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{NEW}$

We can now compare the sample results of NEW and LOREC obtained selecting the thresholds by the MC criterion (5.43).

We give some explanations about the acronyms used in the reported tables. *lambda* is the vector of spikiness thresholds, *sparse* is the vector of sparsity thresholds. $fin1$ is the indicator of the optimal ρ selected via MC, $fin2$ is the indicator of the optimal λ selected via MC, TL is the Total Loss, TL_s is the Sample Total Loss, *rappvar* is $\hat{\alpha}$, *rappcorr* is $\hat{\rho}_{corr}$, *rapptrue* is ρ_{corr}. *spec* is the specificity of zero elements, *sens* is the sensitivity of non-zero elements.

In figure (5.19) the losses obtained (using $\hat{\Sigma}_{n-1}$ as an input for our procedure) are shown. The thresholds selected by MC are $\rho = 0.0192$, $\lambda = 0.1253$ for $\hat{\Sigma}_{NEW}$ and $\rho = 0.0287$, $\lambda = 0.188$ for $\hat{\Sigma}_{LOREC}$. The table shows that our unshrinkage approach prevails for Loss, Total Loss and Sample Total Loss on LOREC approach. The new method shows best fitting properties, going closer to the estimation target.

	NEW	LOREC
sparse(fin1)	0.0192	0.0287
lambda(fin2)	0.1253	0.188
fin1	4	6
fin2	2	3
Loss(fin1,fin2)	7.217	7.3564
TL(fin1,fin2)	6.6899	6.71
TL_s(fin1,fin2)	0.7631	1.0808

Figure 5.19: Sample statistics - Losses

In figure (5.20) we can see that the NEW approach is better also for the estimated proportion of common variance $\hat{\alpha}$ (closer to α) and the estimated proportion of total residual covariance $\hat{\rho}_{corr} = \frac{\rho_{\hat{S}}}{\rho_G}$. It shows a better performance also for the recovery of the true number of non-zeros s. Better results are achieved also for the err rate, for specificity and sensibility. Anyway, we note that there is in general a specific problem on the recovery of non-zero elements. For NEW, the 63.56% are recovered, which has to be considered a good result. Both LOREC and NEW are particularly effective for this aspect only for very sparse matrices.

In figure (5.21) we report the condition number and the Euclidean errors of the estimated eigenvalues for the three components (the low rank, the sparse and the whole covariance matrix). For conditioning, the NEW approach does worse: this is price to pay to improve fitting properties ($condA, condB, condSigma_{hat}$ are the condition numbers of $\hat{S}, \hat{L}, \hat{\Sigma}$ respectively). NEW is on this side between the Sample covariance matrix and LOREC estimate. Concerning the errors of estimated eigenvalues, NEW

	NEW	LOREC
rappvar(fin1,fin2)	0.6973	0.6935
rapptrue	0.0045	0.0045
rappcorr(fin1,fin2)	0.0025	9.89E-04
nz(fin1,fin2)	99	46
s	118	118
spec(fin1,fin2)	0.995	0.9981
sens(fin1,fin2)	0.6356	0.3136
err(fin1,fin2)	0.0135	0.0182
errplus(fin1,fin2)	0.0085	0

Figure 5.20: Sample statistics - rank/sparsity measures

does better for the low rank component only ($errA, errB, errSigma$ are the Euclidean distance of the eigenvalues of $\hat{S}, \hat{L}, \hat{\Sigma}$ from the ones of S, L, Σ respectively). On the other side, the unshrinkage has a positive impact on the maximum estimated eigenvalue of Σ ($maximum_{eig}$ in figure).

	NEW	LOREC	Sigma
condB(fin1,fin2)	1.2904	1.2956	2
condA(fin1,fin2)	2.75E+04	1.19E+04	2.26E+07
condSigma_hat(fin1,fin2)	6.42E+04	5.97E+04	9.49E+07
errB(fin1,fin2)	5.497	5.5181	
errA(fin1,fin2)	0.1681	0.2324	
errSigma(fin1,fin2)	5.5383	5.5144	
maximum_eig	21.04	20.8601	24.4886

Figure 5.21: Sample statistics - conditioning properties

In figure (5.22) we can note that the sample covariance matrix has a slightly smaller Euclidean error of estimated eigenvalues ($errC$) and Total Loss ($TLInput$), but too large condition number ($condC$). $\hat{\alpha}_C$ ($rappvarC = 0.7314$) is much larger that the true 0.7. Parameter c_C is the ratio between the largest and the 4-th eigenvalue of Σ_n. The maximum eigenvalue of Σ_n is 21.1821, the 4-th is 16.1900. $Diff_C$ shows the difference in Total Loss respect to NEW and LOREC respectively.

	Sample	
TL_Input	6.6765	
rappvarC	0.7314	
errC	5.4893	
c_C	1.3083	
condC	9.19E+07	
	NEW	LOREC
Diff_C	0.0133	0.0334

Figure 5.22: Sample statistics - Σ_n

In figure (5.23) we extensively report some measures relative to sparsity detection. The sensitivity of positive elements ($senspos$) and the specificity of negative elements ($specpos$) are reported. For positive elements, the misclassification rate to null elements is $posnnrate$ and to negative elements is $posnegrate$. The same is done for negative elements (the misclassification rate to positive elements is $negposrate$, to null elements is $negnnrate$) and for null elements (the misclassification rate to positive elements is $possens$, to negative elements is $negsens$) respectively. Quantities $posrate = posnnrate + posnegrate$, $negrate = negposrate + posnegrate$ and $nnrate = possens + negsens$ are the total misclassification rates derived from the previous rates (three sums of two elements). There is a specific problem: positive (in particular) and negative elements are too often classified as zeros. On the contrary, it is very rare that a positive element is classified as a negative and viceversa. The error classification rates of positive and of negative elements is lower for NEW than for LOREC. Also err_{tot} ($totrate$ in figure) is lower for NEW.

In figure (5.24) we start showing some statistics across $N = 100$ simulations. In figures, the subscript m stands for mean across all the N replicates, the subscript $m2$ stands for standard error. We immediately note that for NEW the rank is systematically overestimated, differently from LOREC. The proportion of correct rank recovery is 25% against 97% (in figures $rank_{Thr}$ stands for Thresholded Rank, $rank_{exact_{perc}}$ as the percentage of ranks exactly recovered). Simultaneously, in figure (5.25), we see that NEW is better concerning all the Losses (Total Loss, Sample Total Loss and Loss). In ad-

	NEW	LOREC
senspos(fin1,fin2)	0.5094	0.2642
specpos(fin1,fin2)	0.7231	0.3538
posnegrate(fin1,fin2)	0	0
posnnrate(fin1,fin2)	0.4906	0.7358
negposrate(fin1,fin2)	0.0154	0
negnnrate(fin1,fin2)	0.2615	0.6462
possens(fin1,fin2)	0.0027	0.001
negsens(fin1,fin2)	0.0023	8.28E-04
posrate(fin1,fin2)	0.4906	0.7358
negrate(fin1,fin2)	0.2769	0.6462
nnrate(fin1,fin2)	0.005	0.0019
totrate(fin1,fin2)	0.0137	0.0182

Figure 5.23: Sample statistics - Sparsity measures

dition (figure (5.26)), NEW beats LOREC concerning the detection of the proportion of latent variance, of residual covariance and of the number of non zeros. Only on the error measure *err* NEW is slightly worse.

These findings, given that our sample estimate has rank $r = 4$, suggest some considerations about the nature of our improvement. These results show that the unshrinkage is a sample technique. Indeed, we improve upon LOREC for all fitting measures. The fact that the estimated rank is sometimes 5 or 6 means that our technique is able to optimize the sample, finding the ultimate cut-off before non-recovery. This allows to optimize as much as possible fitting properties.

N=100	NEW	LOREC	Sigma
lambda	0.1253	0.188	
sparse	0.0192	0.0287	
rank_Thr_m	4.82	4.03	4
rank_Thr_m2	0.539	0.1714	
rank_exact_perc	0.25	0.97	

Figure 5.24: N=100 - Statistics

N=100	NEW	LOREC
TL_m	6.8335	6.864
TL_m2	0.0326	0.013
TL_m_s	0.8204	1.1703
TL_m2_s	0.7646	0.7667
Loss_m	7.4941	7.5418
Loss_m2	0.7749	0.7776

Figure 5.25: N=100 - Statistics

N=100	NEW	LOREC	Sigma
rappvar_m	0.6945	0.6849	0.7
rappvar_m2	0.0048	0.0049	
rappcorr_m	0.0036	0.0021	0.045
rappcorr_m2	3.09E-04	2.32E-04	
err_m	0.0178	0.0164	
err_m2	0.0016	0.0011	
nz_m	130.87	69.71	118
nz_m2	8.1942	4.9935	

Figure 5.26: N=100 - Statistics

In figure (5.27) we can see that our NEW estimate has not an average number of negative eigenvalues equal to 0, differently from LOREC estimate ($defSpSigma$ is the number of negative eigenvalues of $\hat{\Sigma}$). The same holds for the estimate of the sparse component ($defSpS$ is the number of negative eigenvalues of $\hat{S}$). Since our NEW estimates of the whole covariance matrix and of the sparse component are positive definite in the sample, we have one more argument for the effectiveness of NEW as a sample technique. On the other side, we can see that NEW better recovers on average the eigenvalues of the three matrices L, S, Σ.

In figure (5.28), we can see that NEW is worse for conditioning, but better recovers the maximum eigenvalue of Σ. The NEW procedure here has a larger number of iterations respect to LOREC (Arr_m is the averaged number of iterations).

In figure (5.29) we report some statistics about the detection of the sparsity pattern. We note that NEW is particularly effective for recovering both positive and negative elements respect to LOREC in correspondence of the chosen thresholds. The quantity *senspos* is the rate of correct classification of positive elements, the quantity *specpos* is the rate of correct classification of negative elements.

We explicitly note that this pattern does not depend on the criterion used to select the thresholds. Even using the Frobenius Loss, the relationship between LOREC and NEW performance does not change. The performance is only worse for both methods in terms of sparsity pattern (nonzero detection) and proportion of latent variance.

N=100	NEW	LOREC
defSpSigma_m	3.46	0
defSpS_m	4.4	0
defSpSigma_m2	2.2893	0
defSpS_m2	2.8674	0
errB_m	1.5085	5.261
errA_m	0.3144	0.3503
errSigma_m	5.2182	5.2584
errB_m2	2.4084	0.9299
errA_m2	0.0769	0.0703
errSigma_m2	0.7007	0.7158

Figure 5.27: N=100 - Statistics

N=100	NEW	LOREC	Sigma
condA_m	3.49E+05	9.21E+03	2.26E+07
condB_m	113.9113	1.3882	2
condSigma_hat_m	5.85E+06	4.49E+04	9.49E+07
condA_m2	1.61E+06	2.91E+03	
condB_m2	65.7137	2.23E-16	
condSigma_hat_m2	6.55E+05	575.4263	
Arr_m	58.82	44.87	
Arr_m2	1.6659	1.1604	
maximum_eig_m	20.9901	20.7542	24.4886
maximum_eig_m2	0.8463	0.8468	

Figure 5.28: N=100 - Statistics

N=100	NEW	LOREC
spec_m	0.9896	0.9966
spec_m2	0.0013	6.48E-04
sens_m	0.6819	0.4524
sens_m2	0.041	0.0367
senspos_m	0.698	0.4901
senspos_m2	0.0198	0.0288
specpos_m	0.7144	0.4352
specpos_m2	0.0215	0.0283
totrate_m	0.0158	0.0167

Figure 5.29: N=100 - Statistics

In order to test the strength of results addition, we have tried to perform estimation using the thresholds of $\hat{\Sigma}_{NEW}$ for $\hat{\Sigma}_{LOREC}$ and the thresholds of $\hat{\Sigma}_{LOREC}$ for $\hat{\Sigma}_{NEW}$. While the results on sparsity detection are simply inverted, the estimated proportion of variance explained by the factors is still better for NEW: simulating $N = 100$ settings, the averaged $\hat{\alpha}$ is 0.6924 for NEW and 0.6885 for LOREC, in spite of the fact we have less favorable thresholds for fitting performance. In addition, Loss and Total Loss are still better for NEW, even if the performance is worse for both respect to the original thresholds in terms of fitting.

On the same data, we have applied also POET estimation procedure. First of all, we note that Bai and Ng criteria do not estimate the rank correctly. This is probably due to the fact the ratio $\frac{p}{n}$ is too low. Thus, we set the rank to the true one (4), and we then select the sparsity threshold applying the cross-validation procedure described in [45] with the hard thresholding rule.

The results are quite worse. Due to the natural bias of sample eigenvalues, the proportion of common variance is over estimated (0.7314). The estimated number of non-zeros is 432 (against the true 118). All the losses ($TL\ TL_s\ Loss$) are quite worse than for NEW and LOREC estimates. What is more relevant, the performance of the sparsity recovery is really low. This happens because POET approach does not provide any algebraic consistency framework, but only a parametric one. The relevant results for the POET estimate are reported in figure (5.30). In figure (5.31) we can note that POET is not able to catch the true non-zeros (the rates of correct classification of positive, negative and zeros are reported together with the measure err_{tot}).

	POET	Sigma
TL_C	7.0287	
TL_C_s	2.7323	
Loss_C	8.913	
rappvar_C	0.7314	0.7
rappcorr_C	3.99E-04	0.045
nz_C	432	118
err_C	0.1099	
cond_Sigma_C	3.50E+04	
cond_S_C	3.26E+03	
condL_C	1.3083	

Figure 5.30: POET Sample Statistics

	POET
senspos_C	0.0064
specpos_C	0
spec_C	0.9389
totrate_C	0.1244

Figure 5.31: POET Sample Statistics

Rel_Err	8.44E+03	NEW
	8.41E+03	LOREC
	3.47E+03	POET

Figure 5.32: Relative error: NEW, LOREC and POET

In figure (5.32), we outline the excellence of POET: the Relative Error measure, which is really better than for LOREC and NEW estimates. This happens because the parametric consistency of POET is ensured in the Relative norm $||.||_{\Sigma}$ (see paragraph (2.5.4)).

These results highlight that the two methods (the POET and the numerical one) differ for the application range. LOREC method works better for quite sparse targets. POET method allows for a larger number of non-zeros, given that they have a very low magnitude, because it does not provide any algebraic consistency for the sparsity pattern.

The other settings (**setting2** and **setting3**) show similar performances of $\hat{\Sigma}_{NEW}$ respect to $\hat{\Sigma}_{LOREC}$ and $\hat{\Sigma}_{POET}$. We signal that there are relevant differences concerning the control mechanism on the number of non-zeros and their recovery. If the smallest non-zero element of S is too small, s and ρ_{corr} are hardly recovered. The larger the rank r and the proportion α are, the smaller is the latent condition number c, the smaller must be the true number of non-zero s in order to perform recovery, and the more difficult is to recover s and ρ_{corr}. In addition, the parameter τ must be suitable for ensuring that the spectral norm of Σ_n scales to $\sqrt{\frac{p}{n}}$, in order to make the control mechanism work. At the same time, the higher is the rank r, and the smaller is α respect to c, the easier is to have non-positive definite estimates.

GIVEN that these conditions for the recovery of s are respected (obeying to Theorem 4.1.4), the same relative performances for NEW, LOREC and POET are observed, with particular reference to the Total Loss and the proportion of latent variance. The unshrinkage is proven to be still useful also for larger α and c and for smaller r. Relevant results for **setting2** and **setting3** are reported in figures (5.33) and (5.34) respectively.

The MC criterion for NEW and LOREC and the cross validation method of POET are observed to work effectively. For POET, Bai and Ng criteria are of some use only for the setting with $r = 3$, even if they overestimate the true rank. For all the other settings, the criteria are monotonically decreasing in r. For this reason, the true rank is directly imposed to POET.

r=3,c=4	NEW	Sigma
sparse(fin1)	0.0164	
lambda(fin2)	0.1892	
rank_Thr(fin1,fin2)	3	3
nz(fin1,fin2)	513	580
perczeros(fin1,fin2)	0.1036	0.1172
rappcorr(fin1,fin2)	0.003	0.0048
rappvar(fin1,fin2)	0.7994	0.8
TL_s(fin1,fin2)	1.3487	
TL(fin1,fin2)	9.3763	
Loss(fin1,fin2)	10.8465	

Figure 5.33: **setting2**: Sample Statistics

r=4,c=4	NEW	Sigma
sparse(fin1)	0.0113	
lambda(fin2)	0.0955	
rank_Thr(fin1,fin2)	4	4
nz(fin1,fin2)	263	335
perczeros(fin1,fin2)	0.0531	0.0677
rappcorr(fin1,fin2)	0.0043	0.0072
rappvar(fin1,fin2)	0.6976	0.7
TL_s(fin1,fin2)	0.6943	
TL(fin1,fin2)	13.2935	
Loss(fin1,fin2)	13.9186	

Figure 5.34: **setting3**: Sample Statistics

Up to now, we have fixed the dimension p in order to compare the performances of NEW and LOREC. Varying p does not modify significantly the contrastive performance of described estimators (except for computational times), in the sense that the key parameter in multivariate analysis is $\frac{p}{n}$.

This is why in paragraph (5.3.2) we provide covariance estimation on two real data-sets with two radically different ratios $\frac{p}{n}$. In the second example we have $p > n$, such that we explore the performance of described estimators also in a case somehow resembling the Big Data context.

5.3.2 Real data results

In this paragraph we show some applications of our method to two real data sets. The first is analyzed by Fan et al. in [45], and concerns UK market data. The second is a Euro Area supervisory banking data set, for which we thank the Supervisory Statistics Division of the European Central Bank. On both data sets, a direct comparison between POET and NEW is done, respect to performance and application range. We note that in real data analysis the relevant Loss is only the Sample Total Loss (that is, the distance from $\hat{\Sigma}_n$).

UK market data

In the first example, UK daily market data across the year 2010 are analyzed. The sample dimension is $T = 252$ days, such that we have 251 daily rates. A number of $p = 50$ asset prices are analyzed. These assets are naturally divided in five blocks of 10 firms (variables) corresponding to five economic sectors (see [45] paragraph 7.1 for more explanations). The problem here is to estimate the covariance matrix, taking into account if the different covariance structure among and within blocks may influence the estimate.

Applying POET method using hard thresholding (the sparsity threshold is selected via their cross-validation procedure), Fan et al. report that their POET estimate may have rank $r = 1, 2, 3$ indifferently, because the estimates share the same properties. We report the plot of sample eigenvalues in figure (5.35).

By Bai and Ng's criteria IC1 and IC2 (see [45] paragraph 2.4) we would select 9 or 13 factors according to the penalty used. In fact, in the view of a strict factor model estimation it would be necessary to have more than three components, as outlined in [94].

We signal that it is not straight forward to select low values for the latent rank using Bai and Ng's criteria unless the latent eigenvalues are very spiked. For example, in order to have $\hat{r} = 0$, it is necessary to have an approximately banded covariance structure. A simple experiment carried on the sample covariance matrix over $n = 1000$ samples drawn by a multivariate normal $N_p(0, I_p)$, $p = 100$, shows that in that extreme case we obtain $\hat{r} = 0$. Otherwise, we need that the latent eigenvalues are really spiked respect to

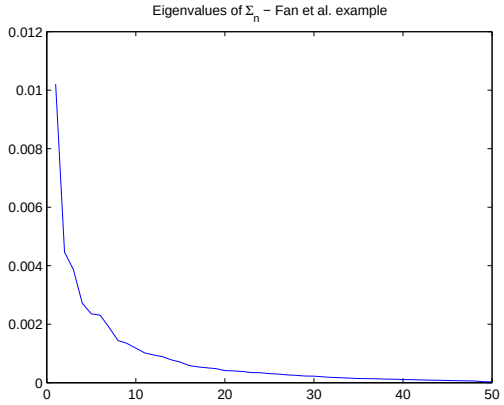

Figure 5.35: UK market data example: sample eigenvalues

the other ones and the latent eigenvectors are really incoherent respect to the standard basis.

However, for $\hat{r} = 2$, they report to have 25.8% of non-zeros within blocks, and 6.7% off-blocks. Among the surviving elements within blocks, they have that 100% of them are positive. In contrast, among the surviving off-blocks elements they obtain 60.3% positive and 39.7% negative.

In figure (5.36) some statistics for our unshrinkage estimate are reported. The solution chosen by our Maximum Criterion (always ensuring that the estimate is positive definite) is much more sparse than the POET one. The number of surviving elements is only $nz = 15$ out of 1225. In addition, the estimated rank is $\hat{r} = 1$. The proportion of common variance is 18.89%, the proportion of residual covariance is 0.92%. Conditioning properties are really good.

In figure (5.36) we can find also some statistics relative to the off-blocks and within-blocks elements. *rate* says that only 4.89% of the within blocks covariances are non-zeros. *rate*2 says that the same percentage for off-blocks covariances is 0.4%. All the surviving covariances within the blocks are positive (*ratepos*). In contrast, three fourth of the off-blocks elements are positive (*ratepos*2).

These results are worth some reflections. Using a strict factor model approach, the necessary number of factors would be larger. In [94], it is shown that the necessary number of factors would be seven. Using an approximate factor model approach (POET), a smaller number of factors is enough. In our thresholding approach, only one factor is surviving. This happens because our method is not PCA based, and does not select the number of factors according to fitting properties. On the contrary, it selects the latent rank and the number of surviving non zeros aiming at recovering the true rank and sparsity pattern. Thus, in our approach there is a non-negligible

UK market data	NEW
rank_Thr(fin1,fin2)	1
nz(fin1,fin2)	15
perczeros(fin1,fin2)	0.0122
rappvar(fin1,fin2)	0.1889
rappcorr(fin1,fin2)	0.0092
TL_s(fin1,fin2)	0.0023
sparse(fin1)	9.74E-05
lambda(fin2)	6.95E-04
rate(fin1,fin2)	0.0489
rate2(fin1,fin2)	0.004
ratepos(fin1,fin2)	1
rateneg(fin1,fin2)	0
ratepos2(fin1,fin2)	0.75
rateneg2(fin1,fin2)	0.25
condSigma_hat(fin1,fin2)	113.9172
condSparse(fin1,fin2)	56.5862
numvar	1225

Figure 5.36: UK market data: $\hat{\Sigma}_{NEW}$ statistics

proportion of covariance which is thrown away. This is done in order to recover exactly the low rank and the sparse components.

For this reason, two or three factors are maybe enough for fitting properties, but they are too many for rank/sparsity pattern detection. The thresholding algorithm returns that one factor is enough for that. In order to recover in the best possible way the two components, a relevant proportion of covariance is lost, as outlined in figure (5.37). The residual of the minimization procedure contains 21.15% of covariance, while $\hat{\Sigma}_{NEW}$ contains 78.85%. 78.13% of the total covariance belongs to the low rank component. Only 0.72% belongs to the sparse component. This is the reason why only one factor is enough.

By this minimization procedure, quite surprisingly, our method shows also a lower Sample Total Loss. We replicated POET procedure with 2 factors, and we obtained a Sample Total Loss equal to 0.028. In our case, the same indicator is equal to 0.023. Our rank/sparsity based estimation procedure is thus able to better approximate the sample covariance matrix.

In conclusion, we should wonder if the block structure is strong enough to really impact covariance estimation. In fact, this result is consistent to the single index factor model ([74]), and to the CAPM ([104]).

sumW/sumTOT	0.2115
sumSigmahat/sumTOT	0.7875
sumLow/sumTOT	0.7813
sumSparse/sumTOT	0.0072

Figure 5.37: UK market data: $\hat{\Sigma}_{NEW}$ statistics

Euro Area supervisory banking data

We are now ready to estimate the covariance matrix on the Euro Area supervisory banking data. We thank for the use of this data set the Supervisory Statistics Division of the European Central Bank, where the author spent a semester as a PhD trainee. 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 dimension) 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 assets 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 remaining variables were $p = 382$. So, here we are in the typical $p > n$ case, where the sample covariance matrix is completely ineffective. We report the plot of sample eigenvalues in figure (5.38).

Our estimation method selects a solution having a latent rank equal to 6. The number of surviving non-zeros in the sparse component is 328, i.e. the 0.45% of $numvar = 72772$. Conditioning properties are inevitably very bad. The results are reported in figure (5.39).

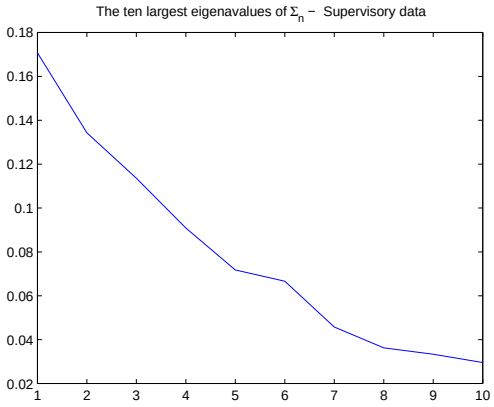

Figure 5.38: Supervisory data: sample eigenvalues

Supervisory data	NEW
rank_Thr(fin1,fin2)	6
nz(fin1,fin2)	328
rappvar(fin1,fin2)	0.3247
rappcorr(fin1,fin2)	0.1687
perczeros(fin1,fin2)	0.0045
TL_s(fin1,fin2)	0.0337
defSpSigma(fin1,fin2)	0
defSpS(fin1,fin2)	0
condSigma_hat(fin1,fin2)	6.35E+15
condSparse(fin1,fin2)	2.78E+15
condL	3.1335

Figure 5.39: Supervisory data: results for $\hat{\Sigma}_{NEW}$

Supervisory data	POET
TL_C_s	0.0645
nz_C	404
perczeros	0.0056
numvar	72771
rappvar_C	0.6123
rappcorr_C	0.0161
cond_S_C	1.11E+15
cond_C	6.68E+15
condL_C	2.5625
defSpSigma_C	0
defSpS_C	1

Figure 5.40: Supervisory data: results for $\hat{\Sigma}_{POET}$

We now pass to the POET procedure. Bai and Ng's criteria do not attain any minimum for $r = 0 : 20$. We thus decide to exploit the algebraic consistency of $\hat{\Sigma}_{NEW}$ setting the rank to 6. We perform the usual cross-validation and obtain a POET estimate (figure (5.40)). The number of non-zeros of POET estimate is 404 (0.56%).

Apparently, one could say that POET estimate is better: its estimated proportion of common variance is 0.6123, and its proportion of residual covariance is 0.0161. On the contrary, for NEW $\hat{\alpha} = 0.3247$ and $\hat{\rho}_{corr} = 0.1687$. However, a relevant question arises: how much is the true proportion of variance explained by the factors? In fact, a so high α, dependent on the use of PCA with 6 components, causes $\hat{\rho}_{corr}$ to be very low. This is why in the POET procedure a preference for the low rank part is given a priori. This pattern does not change even if we choose a lower value for the rank.

On the contrary, the NEW estimate, which depends on a double-step iterative thresholding procedure (8 iterations), 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 this rank/sparsity detection the NEW procedure shows better approximation properties respect to POET: the Sample Total Loss of the first procedure is relevantly lower than the one of the second (0.337 VS 0.645).

For our 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).

First of all, in figure (5.41) 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 (5.42) we report the top 6 variables by estimated degree. They are mainly credit-based variables: financial assets through profit and loss, central banks impaired assets, allowances to credit institutions and non-financial corporations, cash. These variables are related to the largest number of other variables.

In figure (5.43) we report the top 5 variables by estimated communality ($\frac{\hat{l}_{NEW,ii}}{\hat{\sigma}_{NEW,ii}} \forall i = 1, \ldots, 382$). The results are very meaningful: the most systemic variables are debt securities, loans and advances to households, specific allowances for financial assets, and advances which are not loans to central banks, which are all fundamental variables or banking supervision.

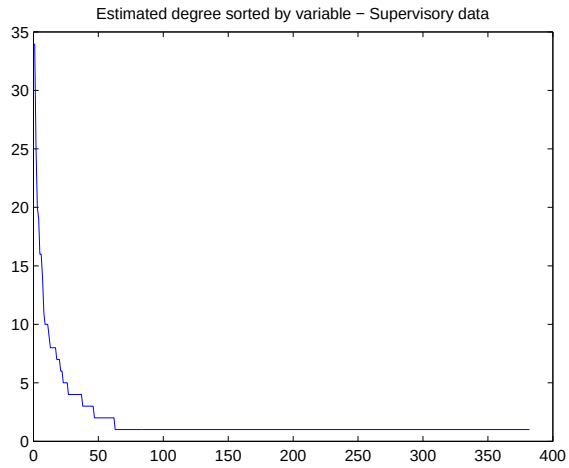

Figure 5.41: Supervisory data: sorted degree by variable

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 5.42: Supervisory data: top 6 by degree

In figure (5.44) we report the top 5 variables by estimated idiosyncratic covariance proportion ($\frac{\hat{s}_{NEW,ii}}{\hat{\sigma}_{NEW,ii}} \forall i = 1, \ldots, 382$). We note that those variables have a marginal power in the explanation of the common covariance structure. The first two are credit card debt and collateralized loans to other financial corporations. The others are equity instruments given to central banks, other financial corporations and general governments respectively. All these variables are less relevant for supervisory analysis than the previous five.

In conclusion, our NEW procedure offers here a realistic view of the underlying structure of variables, by allowing a largest part of covariance to

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 5.43: Supervisory data: top 5 by estimated communality

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 5.44: Supervisory data: top 5 by residual covariance proportion

be explained by the residual sparse component.

Chapter 6

Conclusions

The present work describes the numerical approach to covariance matrix estimation. The main focus is on a method based on convex non smooth optimization which assumes a low rank plus sparse decomposition for the covariance matrix.

In this framework, the estimation is performed solving a regularization problem where the objective function is composed by a smooth Frobenius loss and a non smooth composite penalty. The penalty is the sum of the nuclear norm of the low rank component and the l_1 norm of the sparse component.

The numerical nature and the algorithmic solutions to this problem are outlined highlighting the connections with sub-gradient minimization and semi-definite programming theory.

The study of the statistical properties of such a minimizer in the context of algebraic geometry, which involves necessary conditions for recovery and identifiability, is deeply explored, emphasising the non-asymptotic nature of the method. Recent solutions under different hypothesis are described, in order to understand how the exact recovery in the noisy context is possible. The key for the exact identification of the low rank and the sparse algebraic matrix varieties is proved to be the rank/sparsity incoherence principle between the two components.

We remark that the algebraic framework allows not only the usual parametric consistency but also the algebraic consistency of the estimate. As a consequence, the rank and the number of residual non zeros are simultaneously estimated by the solution algorithm. This automatic recovery is a crucial advantage respect to existing asymptotic methods, like the PCA-based POET (Principal Orthogonal complEment Thresholding) estimator. In the numerical framework, in fact, the latent rank is automatically selected and the sparsity pattern of the residual component is recovered considerably better, due to the algebraic consistency.

Two theoretical advances upon the most recent estimator of this family, LOREC (LOw Rank and sparsE Covariance estimator), are proved. First,

we prove 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. At the same time, the unshrinkage procedure improves fitting properties. Second, we prove that the numerical estimator can effectively recover the covariance matrix even in presence of spiked eigenvalues with rate $O(p)$, exactly as POET estimator does, requiring only $n = o(p^2)$ samples under POET assumptions. The loss from the target is bounded in absolute norm (in contrast to POET procedure). In addition, the recovery is effective even if we have an intermediate degree $\alpha \in [0, 1]$ of spikiness, and the loss is bounded accordingly to α with the need of $n = o(p^{2\alpha})$ samples only. Besides, our work completes LOREC approach deriving the rate of the inverse of the sparse component and an operative (feasible) identifiability condition.

The performance of these improvements 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, our NEW method 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 (as well as each of both) and the total loss.

Real data analysis 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 dissertation is the starting point for a number of possible research directions. We mention here the three most relevant in our view:

- in the time series context, this procedure can be potentially extended to covariance matrix estimation under dynamic factor models. Setting a low rank plus sparse structure on the auto-covariance matrix at a particular lag, or on the process fully considered under the co-integration

hypothesis, are two particularly promising options, in which the sparse component can be an additional flexibility tool useful for modelling large data sets;

- the extension of our procedure to the spectral matrix estimation context, under various definitions of stationary process;
- the adaptation of this procedure for clustering in high dimensions. Existing factor-based methods can be improved by the use of the nuclear norm and the relaxation offered by the sparse component.

In conclusion, our research provides a tool to automatically explore large data sets. This tool can be potentially used in the Big data context, where both the dimension and the sample dimension 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.

Bibliography

[1] Agarwal A., Sahand N. and Wainwright M.J. (2012). Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. The Annals of Statistics, Vol. 40, No. 2, 1171:1197.

[2] Anderson, T.W. (1958). An introduction to Multivariate Statistical Analysis, Third Edition. John Wiley & Sons, Inc.

[3] Antoniadis, A. and Fan, J. (2001). Regularized wavelet approximations. J. Am. Statist. Ass., 96, 939 : 967.

[4] Bach, F. R. (2008). Consistency of Trace Norm Minimization. Journal of Machine Learning Research 8, 1019 : 1048.

[5] Bai, J. (2003). Inferential Theory for Factor Models of Large Dimensions. Econometrica. Volume 71, Issue 1, pages 135:171, January 2003.

[6] Bai, J, Ng, S. (2002). Determining the number of factors in approximate factor models. Econometrica 70 (1), 191:221.

[7] Bai, J, Ng, S. (2008). Large dimensional factor analysis. Now Publishers Inc.

[8] Bai, Z. D. (1999). Methodologies in spectral analysis of large dimensional random matrices: a review. Statist. Sinica 9 611:677.

[9] Bajaj, C. and Gillette, A (2010). Polynomial Curves and Surfaces. September 8, 2010.

[10] Beck, A. and Teboulle, M. (2009). A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. Siam J. Imaging Sciences 2009 Society for Industrial and Applied Mathematics Vol. 2, No. 1, pp. 183:202.

[11] Berge, J. and Kiers, H. (1991). A numerical approach to the approximate and the exact minimum rank of a covariance matrix. Psychometrika, 56:309:315.

[12] Bertsekas, Dimitri P. (1999). Nonlinear Programming. Belmont, MA.: Athena Scientific.

[13] Blumensath, T., Davies, M. E. (2008). Iterative Thresholding for Sparse Approximations. The Journal of Fourier Analysis and Applications, vol. 14, n. 5, pp. 629:654, December 2008.

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

[15] Bickel, P. and Levina, E. (2008) Covariance regularization by thresholding. Ann. Statist., 36, 2577 : 2604.

[16] Boyd, S., Vandenberghe, L. (2004). Convex optimization. Cambridge University press.

[17] Cai J.-F., Candès, E. J. and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. SIAM J. on Optimization 20(4), 1956:1982.

[18] Cai, T., Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. J. Am. Statist. Ass., 106, 672:684.

[19] Cai, T., Liu, W., and Luo, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association, 106(494),594: 607.

[20] Cai, T., Zhang, C.-H., Harrison H., Zhou, H.H. (2010). Optimal rates of convergence for covariance matrix estimation. The Annals of Statistics 2010, Vol. 38, No. 4, 2118:2144.

[21] Candès, E. J. (2006). Compressive sampling. Proceedings of the International Congress of Mathematicians, Madrid, Spain.

[22] Candès, E. J. and Plan, Y. (2009). Near-ideal model selection by l_1 minimization. The Annals of Statistics, 2009, Vol. 37, No. 5A, 2145:2177.

[23] Candès, E., Tao, T. (2005). Decoding by linear programming. Journal IEEE Transactions on Information Theory archive Volume 51 Issue 12, December 2005 Page 4203:4215.

[24] Candès, E., Recht, B. (2009). Exact matrix completion via convex optimization. Foundations of Computational Mathematics, 9(6):717:772, 2009.

[25] Candès E. J., Li X., Ma Y., and Wright J. (2009). Robust Principal Component Analysis? Journal of ACM 58(1), 1:37.

[26] Candès, E., Romberg J., Tao, T. (2006). Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information. Transactions On Information Theory, VOL. 52, NO. 2, FEBRUARY 2006 489.

[27] Candes, Wakin, Boyd (2008). Enhancing Sparsity by Reweighted l_1 Minimization. J Fourier Anal Appl.

[28] Clarke, H. (1983), Optimization and Nonsmooth Analysis. New York et al., John Wiley & Sons 1983. XIII, 308.

[29] Chamberlain, G. and Rothschild, M. (1983) Arbitrage, factor structure and mean-variance analysis in large asset markets. Econometrica 51, 1305 : 1324.

[30] Chandrasekaran, V., Sanghavi S., Parrilo P. A., Willsky A. S. (2011). Rank-Sparsity Incoherence for Matrix Decomposition, SIAM Journal on Optimization. Vol. 21, No. 2, June 2011.

[31] Chandrasekaran V., Parrilo P. A., Willsky A. S. (2012). Latent Variable Graphical Model Selection via Convex Optimization. Annals of Statistics (with discussion), Vol. 40, No. 4, August 2012.

[32] Chen S., Donoho D., Saunders M. (1998). Atomic decomposition by basis pursuit. SIAM J. on Sci. Comp., vol. 20, no. 1, pp. 33:61, 1998.

[33] Chiani, M. (2012). Distribution of the largest eigenvalue for real Wishart and Gaussian random matrices and a simple approximation for the Tracy/Widom distribution. ArXiv.

[34] Cramer H., (1946). Mathematical methods of Statistics. Princeton University Press.

[35] Daubechies I. , Defrise M. , De Mol C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Communications on Pure and Applied Mathematics 57 (2004): pp. 1416 : 1457.

[36] Donoho D. L. (2006). For most large underdetermined systems of linear equations the minimal l_1 norm solution is also the sparsest solution. Communications on Pure and Applied Mathematics, volume 59, issue 6, pages 79 829, 2006.

[37] Donoho D. L. (2006). Compressed Sensing. Ieee Transactions On Information Theory, Vol. 52, N. 4, April 2006.

[38] Donoho D. L., Tsaig Y., Drori I., Starck J-L. (2006). Sparse Solution of Underdetermined Linear Equations by Stagewise Orthogonal Matching Pursuit Techinque Report. Stanford University USA.

[39] Davidson, K. R. and Szarek, S.J. (2001). Local operator theory, random matrices and Banach spaces. Handbook of the Geometry of Banach Spaces. I 317:366.

[40] Eckart C. and Young G. (1936). The approximation of one matrix by another of lower rank. Psychometrika, 1, 211:218, 1936.

[41] Efron B., Hastie T., Johnstone I., Tibshirani R. (2004). Least angle regression (2004). The Annals of Statistics 2004, Vol. 32, No. 2, 407:499 Institute of Mathematical Statistics, 2004.

[42] Fama, E. F., French, K. R. (1992). The Cross-Section of Expected Stock Returns. The Journal of Finance 47 (2): 427.

[43] Fan, J., Fan,Y. and Lv, J. (2008). High dimensional covariancematrix estimation using a factor model. J. Econmetr., 147, 186:197.

[44] Fan, J., Liao, Y. and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. Ann. Statist., 39, 3320:3356.

[45] Fan, J., Liao, Y. and Mincheva, M. (2013). Large Covariance Estimation by Thresholding Principal Orthogonal Complements (with discussion). Journal of Royal Statistical Society B , 75, 603:680.

[46] Fazel M., Hindi H., and Boyd S. (2001). A Rank Minimization Heuristic with Application to Minimum Order System Approximation. Proc. American Control Conference, Arlington, Virginia, June 2001.

[47] Fazel, M., Hindi H., Boyd S. (2004). Rank Minimization and Applications in System Theory. Proc. American Control Conference, Boston, Massachusetts, June 2004, pages 3273:3278.

[48] Farnè, M. (2014) An algorithm to simulate VMA processes having a spectrum with fixed condition number. Communications in Statistics - Simulation and Computation

[49] Fazel F. (2002). Matrix rank minimization in applications. PhD dissetertation. Stanford University.

[50] Figueiredo, M.A.T. (2009). Recent Developments (and Some History) in Iterative Thresholding Algorithms. SampTA Conference 2009.

[51] Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. Biometrika (Biometrika Trust) 10 (4), 507:521.

[52] Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000) The generalized dynamic factor model: identification and estimation. Rev. Econ. Statist., 82, 540:554.

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

[54] Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posteriori covariance matrices in Kalman filter variants. J. Multivariate Anal. 98 227:255.

[55] Groetsch C. W. (1984). The Theory of Tikhonov Regularization for Fredholm Integral Equations of the First Kind. Pitman. 1984.

[56] Hadamard, J. (1893), "Résolution d'une question relative aux déterminants", Bulletin des Sciences Mathématiques 17: 240:246

[57] Hadamard (1923). Lectures on Cauchy's Problem in Linear Partial Differential Equations. Dover Phoenix editions, Dover Publications, New York.

[58] http://www.unc.edu/ hannig/STOR655/handouts/Handout-asymptotics.pdf

[59] Harman H. H. (1976). Modern Factor Analysis. University of Chicago Press, 1976.

[60] Harris J. (1995). Algebraic Geometry: A First Course. Springer-Verlag, 1995.

[61] Hastie, T. joint work with Mazumder,R (2012). Matrix Completion and Large-scale SVD Computations. Keynote address MBCII, Catania, Sicily, September 2012.

[62] Horn, R. A., Johnson, C. R. (1990). Matrix analysis. Cambridge University Press.

[63] Horn, R. A., Mathias R. (1990). An Analog of the Cauchy-Schwarz Inequality for Hadamard Products and Unitarily Invariant Norms (1990). Siam J. Matrix Anal. Appl, 11(4), 481:498.

[64] Hotelling, H. (1931). "The generalization of Student's ratio". Annals of Mathematical Statistics 2 (3): 360:378.

[65] Hotelling, H. (1933). Analysis of a complex of Statistical Variables Into Principal Components. Journal of Educational Psychology, volume 24, pages 417:441 and 498:520.

[66] Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. Biometrika, 93:85:98.

[67] Hsu, D., Kakade, S. M. and Zhang, T. (2011). Robust matrix decomposition with sparse corruptions. IEEE Trans. Inform. Theory 57 7221:7234.

[68] James, W., Stein, C. (1961). Estimation with quadratic loss. Proc. Fourth Berkeley Symp. Math. Statist. Prob. 1, pp. 361:379.

[69] Ji, S. and Ye, J. (2009). An accelerated gradient method for trace norm minimization. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 457:464. ACM.

[70] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. Ann. Statist. 29 295:327. M

[71] Johnstone, I. M. and Lu, A. Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. J. Am. Statist. Ass., 104, 682:693.

[72] Jolliffe I. T. (2002). Principal component analysis. Springer verlag, 2002.

[73] Journee M., Bach F., Absil P.-A., Sepulchre R. (2010). Low-rank optimization on the cone of positive semidefinite matrices. Siam J. Optim. 2010 Society for Industrial and Applied Mathematics Vol. 20, No. 5, pp. 2327:2351.

[74] Ledoit, O., M. Wolf. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J. Empirical Finance 10 603:621.

[75] Ledoit, O., and Wolf, M., (2004). A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. Journal of Multivariate Analysis, 88, 365:411.

[76] Luo, X. (2011). High Dimensional Low Rank and Sparse Covariance Matrix Estimation via Convex Minimization. arXiv.

[77] Luo, X. (2013). Recovering Model Structures from Large Low Rank and Sparse Covariance Matrix Estimation. arXiv.

[78] Malioutov D M., Johnson J. K., Myung J. C. (2008). Low-Rank Variance Approximation in GMRF Models: Single and Multiscale Approaches. Ieee Transactions On Signal Processing, Vol. 56, No. 10, October 2008.

[79] Marchenko,V. A., Pastur, L. A. (1967) "Distribution of eigenvalues for some sets of random matrices", Mat. Sb. (N.S.) 72(114) : 4, 507 : 536.

[80] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. The Annals of Statistics, 34(3), 1436:1462.

[81] `http://www.inrialpes.fr/bipop/people/acary/Research/Moreau_CRAS1962.pdf`

[82] `http://www.inrialpes.fr/bipop/people/acary/Research/Moreau_CRAS1963.pdf`

[83] Moreau J.J. (1965), Proximitè and dualitè dans un espace hilbertien, Bulletin de la S.M.F., tome 93 (1965), p. 273:299.

[84] Negahban, S., Ravikumar, P., Wainwright, M. J., Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Statistical Science 2012, Vol. 27, No. 4, 538:557 Institute of Mathematical Statistics, 2012.

[85] Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. J. Mach. Learn. Res. 13 1665:1697.

[86] Nesterov, Y. (1983). A method for solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Math. Dokl., 27, 372:376.

[87] Nesterov, Y. (2007). Gradient methods for minimizing composite objective function CORE Discussion Paper 2007/76. September 2007.

[88] Nesterov Y., Nemirovski A.(2013). On first-order algorithms for l_1 nuclear minimization. Acta Numerica, 22, pp 509:575.

[89] Otazo, Candés and Sodickson (2013). Low-rank and sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components. Magnetic Resonance in Medicine, 2015 Mar, 73(3):1125:36.

[90] Parikh N., Boyd, S. (2013). Proximal Algorithms. Foundations and Trends in Optimization, Vol. 1, No. 3 (2013) 123:231.

[91] Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia, Philosophical Transactions of the Royal Society of London. $187, 253 : 318$.

[92] Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points is Space. Philosophical Magazine Series 6 2 (11): 559:572.

[93] Picard, E. (1930). Lecons sur quelques problemes aux limites de la theorie des equations differentielles. Paris: Gauthiers-Villars et cie.

[94] Marianna Poli. Modello fattoriale classico e modello fattoriale approssimato: ipotesi a confronto su un data set reale. Prova finale. Relatore: Prof.ssa Angela Montanari. Anno Accademico 2012/2013. Alma Mater Studiorum Università Di Bologna. Corso di Laurea in Scienze Statistiche

[95] Pourahmadi M. (2013). High-Dimensional Covariance Estimation: With High-Dimensional Data. June 2013, Wiley.

[96] Randrianantoanina, B. (2001). Norm-one projections in banach spaces. Taiwanese Journal Of Mathematics Vol. 5, No. 1, pp. 35-95, March 2001.

[97] Recht B., Fazel M., Parrilo P. A. (2010). Guaranteed Minimum Rank Solutions to Linear Matrix Equations via Nuclear Norm Minimization. SIAM Rev., 52(3), 471:501.

[98] Rockafellar R. T. (1970). CONVEX ANALYSIS (1970). Vol. 28 of Princeton Math. Series, Princeton Univ. Press, 1970 (470 pages).

[99] Ross, S. A. (1976). "The Arbitrage Theory of Capital Asset Pricing". Journal of Economic Theory, 13 (1976), 341:360.

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

[101] http://www.math.nus.edu.sg/ mattohkc/sdpt3.html

[102] Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. The American Journal of Psychology 15 (2): 201:292.

[103] Sharpe, William F. (1963). A Simplified Model for Portfolio Analysis. Management Science Vol. 9, No. 2 (Jan., 1963), pp. 277:293.

[104] Sharpe, William F. (1964). Capital Asset Prices. A Theory of Market Equilibrium Under Conditions of Risk. Journal of Finance XIX (3): 425:42.

[105] Stock, J. H. and Watson, M. (2002). Forecasting Using Principal Components From a Large Number of Predictors. Journal of the American Statistical Association December 2002, Vol. 97, No. 460, Theory and Methods.

[106] Sundaram R. K. (1996). A First Course in Optimization Theory. Cambridge University Press.

[107] Tracy, C., Widom, H. (1994). Level-spacing distributions and the airy kernel. Communications in Mathematical Physics 159, 151:174.

[108] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. J. or RSS B,58, 1 267:288.

[109] Tikhonov, Andrey Nikolayevich (1943). On the stability of inverse problems. Doklady Akademii Nauk SSSR.

[110] Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ell1 -constrained quadratic programming (lasso). 55(5) 2183:2202.

[111] Wang, C., Sun, D. and Toh, K. C. (2009). Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm. Preprint.

[112] Watson G. A. (1992), Characterization of the subdifferential of some matrix norms, Linear Algebra and Applications, volume 170, pages 1039:1053, 1992.

[113] Wishart, J. (1928), "The generalised product moment distribution in samples from a normal multivariate population". Biometrika 20A (1:2): 32:52.

[114] Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. Biometrika 90 831:844.

[115] Yang A.Y., Sastry S.S., Ganesh A., Ma Y. (2010). Fast l_1-minimization algorithms and an application in robust face recognition: A review. Image Processing (ICIP). 2010 17th IEEE International Conference on, 1849:1852.

[116] Yang J., Yuan X. (2013). Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization mathematics of computation. Volume 82, Number 281, January 2013, Pages 301:329.

[117] Zou H., Hastie T., Tibshirani R (2006). Sparse Principal Component Analysis JCGS 2006 15(2): 262:286.

[118] Zhou X., Yang C., Zhao H., Yu W. (2015). Low-Rank Modeling and Its Applications in Image Analysis. Journal ACM Computing Surveys (CSUR) Surveys Homepage archive Volume 47 Issue 2, January 2015 Article No. 36.

[119] Zhang, Z. and Wu, L. (2003). Optimal low-rank approximation to a correlation matrix. Linear Algebra and its Applications, 364:161:187.

[120] http://www.psi.toronto.edu/matrix/calculus.html

[121] http://matrixcookbook.com