High Dimensional Portfolio Selection

Estimation of the Inverse Covariance Matrix

Èrik Avilés Castilla

Pau Gimeno Perramon

June 15, 2017

Code: EWI06

Universitat Pompeu Fabra, 2016-2017

Supervisor: Piotr Zwiernik

Acknowledgement

We would like to thank our supervisor, Professor Piotr Zwiernik, for his patience, support and constant feedback. We could have not imagined a better advisor for our Bachelor Thesis.

Contents

Abstract

Estimating covariance matrices and their inverses are a crucial problem in statistics and in many other applied areas. In particular, it is a fundamental element for portfolio selection models in finance. The conventional covariance estimator, the sample covariance matrix, is invertible and unbiased when the number of observations, n, is larger than the number of parameters to be estimated, p. However as the dimensional setting grows, estimating the covariance matrix and its inverse becomes very challenging. The sample covariance matrix no longer performs well and alternative robust estimators are needed. To overcome this problem, different methods, stemming from high-dimensional statistics, are required. The aim of this project is to study these different high-dimensional techniques, with special focus on the estimation of the inverse covariance matrix. A real-world example with stocks return data is carried out in the statistical software R to apply such techniques and obtain the desired sparse inverse covariance matrix. Moreover, an additional step is taken to attain sparse and stable portfolios through a ℓ_1-regularization penalty on the vector of portfolio weights.

Keywords: Inverse Covariance Matrix, sparsity, high-dimensional techniques, ℓ_q-regularization, Graphical Models, semiparametric Gaussian copula, Portfolio Selection, Markowitz Theory.

1 Introduction

In the recent years, the study of high dimensional data has been increasingly gaining importance in several areas of economics. This fact comes from the necessity to find new techniques in order to deal with massive amounts of data. Concretely, different methods have been developed to solve the problem when the number of parameters to be estimated p is significantly larger than the available sample size n. These techniques can be organised in three groups: (1) Subset selection, (2) Shrinkage and (3) Dimension reduction techniques. In the context of linear regression model, the first approach identifies a subset of the p covariates which is believed to be more related to the response variable. The shrinkage approach fits the model using all the covariates but it imposes a regularization to shrink the estimated coefficients towards zero. And, finally, the third approach reduces the dimension of the predictors by computing their optimal projections. The main aim of this project is to explain and apply the second approach which, as stated above, imposes certain regularization on the coefficient estimates in order to drive some of the estimated parameters towards zero or, in other words, impose certain levels of sparsity. The imposition of these regularization parameters is meant to effectively reduce the variance of the estimates [6].

The different shrinkage techniques are presented in this project from a financial perspective, specifically from the Markowitz theory framework. The introduction of the Markowitz theory is the opening of our paper. Then we explain the foundations and applications of the different shrinkage techniques. A particular focus will be placed on the graphical lasso which performs a penalized maximum likelihood with a lasso penalty on the inverse covariance matrix, also known as precision matrix, to reduce the effective numbers of parameters to be estimated. This technique will be applied to stocks return data so as to obtain the optimal sparse precision matrix. However, it is typically assumed that data follow a multivariate Gaussian distribution when applying the graphical lasso. Instead, we will use a semiparametric Gaussian copula, known as nonparanormal distribution, to efficiently estimate the precision matrix as Gaussianity is violated. Then, we will represent through undirected graphical models the estimated sparse inverse covariance matrix in order to assess an evaluation of the potential conditional independences among the NASDAQ stocks. Moreover, as we work with time series data, the stocks return data will be transformed to correct for difference stationarity and we will represent its optimal undirected graph and compare it to the optimal graph with the untransformed data.

Finally, we conduct the Markowitz optimization problem, with the estimated inverse covariance matrix, with a ℓ_1-regularization penalty on the vector of portfolio weights to obtain a sparse and stable portfolio. By construction, such regularization should only penalize (i.e. cause sparsity) short positions and shrinkage the active-positive weights.

The remainder of this project is organized as follows: In Section 2, we review the Markowitz's Modern Portfolio Theory. In Section 3, we present the different shrinkage techniques with a special focus on the sparse inverse covariance matrix estimation. In section 4, we apply these models to estimate the precision matrix of the NASDAQ components returns in R. We analyse the conditional independences among the stocks through the visualization of high-dimensional undirected graphical models. Because of the nature of our data, the Gaussian distribution assumption is relaxed. In section 5, we transform the data to correct for difference stationarity and compare the optimal undirected graph obtained with the previous one. In the final section, the Markowitz optimization problem is solved with a ℓ_1-regularization parameter on the vector of portfolio weights using the software Matlab.

2 The Markowitz Theorem

In this section, we briefly explain the Markowitz model and why high-dimensional techniques are needed to solve the optimization problem under certain circumstances.

In 1952, Harry Markowitz introduced the modern portfolio theory [14], which greatly revolutionized the finance field. He was the first who proposed to apply simple mathematical ideas to the problem of formulating optimal investment portfolios. As mentioned in his paper, a strategy only aimed at obtaining high returns is poor and inefficient. Instead, he suggested, the desirable strategy for an investor is to maximize the portfolio's return while minimizing its volatility, measured by the variance. Thus, a rational investor should balance the portfolio's risk and return by choosing the portfolio weighting factors optimally.

Consider a portfolio composed of N risky assets whose rates of returns at time t are given by the random variables $X_{1,t}, \ldots, X_{N,t}$ and are assumed to be stationary over t. Let $\mathbf{X}_t = (X_{1,t}, \ldots, X_{N,t})^T$ be the $N \times 1$ vector of the returns at time t, $\mu = (\mu_1, \ldots, \mu_N)^T$ be the vector $N \times 1$ vector of the means, $\mathbf{\Sigma}$ be the $N \times N$ variance-covariance matrix with σ_i^2 on the diagonal and σ_{ij} otherwise and ω be the $N \times 1$ vector of portfolio weights. Then, for a given portfolio ω, the expected return and the variance are equal to $\omega^T \mu$ and $\omega^T \mathbf{\Sigma} \omega$, respectively, where for a matrix A, A^T denotes its transpose.

In the Markowitz optimization problem, the objective is to find a portfolio which has minimal volatility for a given expected return, x^*. This is, the investor needs to solve the following quadratic program with respect to ω:

$$\underset{w}{\text{minimize}} \quad \omega^T \mathbf{\Sigma} \omega \quad \text{subject to} \quad \omega^T \mathbf{1} = 1 \qquad \text{and} \quad \omega^T \mu = x^*. \tag{1}$$

For the purpose of the project, we do not add any short sale restrictions (i.e. $\omega \geq 0$) since we will show in section 6 that adding a ℓ_1- regularization parameter will be equivalent to penalizing short sales. The problem (1) can be rewritten in the so-called Lagrangian form:

$$\mathcal{L}(\omega, \lambda_1, \lambda_2) = \frac{1}{2}\omega^T \mathbf{\Sigma} \omega + \lambda_1 (1 - \omega^T \mathbf{1}) + \lambda_2 (x^* - \omega^T \mu). \tag{2}$$

The KKT conditions for constrained minimization problem are the following:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \omega} &: \mathbf{\Sigma}\omega - \lambda_1 \mathbf{1} - \lambda_2 \mu = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda_1} &: 1 - \omega^T \mathbf{1} = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda_2} &: x^* - \omega^T \mu = 0. \end{aligned} \tag{3}$$

In practice the covariance matrix $\mathbf{\Sigma}$ is unknown and has to be estimated from the data. When

p is smaller than n, the sample covariance matrix is the natural candidate as it is unbiased and invertible with probability 1 (i.e. $\mathbf{\Sigma}$ is positive definite). It is the Maximum Likelihood Estimator (MLE) under Gaussian distribution and consistent under very mild conditions. However, as the number of assets n increases, estimating a $p \times p$ covariance matrix becomes very challenging since the number of parameters to be estimated p grows quadratically in n. The sample covariance matrix is no longer robust for moderate or large dimensionality. In fact, when $p > n$, the sample covariance is rank-deficient and thus its inverse does not exist, making unfeasible to solve the optimization problem stated above.

In the following section, we study the most popular shrinkage methods which we next use in the software R to estimate the precision matrix necessary to solve the Markowitz optimization problem.

3 High-Dimensional Techniques

In this section we provide a brief overview of the high-dimensional methods. We begin our discussion in the context of the linear regression model with the introduction of the penalized ordinary least squares techniques and its advantages compared to the Ordinary Least Squares (OLS). Later, we study the penalized maximum log-likelihood estimator and its application on the inverse covariance matrix estimation.

Assume we are given a sample of size N of the form (y_i, x_i) where x_i is a p-dimension vector of predictors $x_i = (x_{i1}, \ldots, x_{ip})$, and y_i is the associated response variable. This linear relationship is modelled with an error term ϵ_i, which adds the random noise between the independent and the dependent variable relationship. The objective is to model the dependent variable y_i using a linear combination of the independent variables

$$\mathrm{y} = \beta_0 + \beta^T X + \epsilon \qquad \text{where, } \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p. \tag{4}$$

The most common estimator in this setting is the Ordinary Least Squares (OLS) which minimizes the sum of the squared residuals

$$\underset{\beta_0,\beta}{\text{minimize}} \sum_{i=1}^{N} \big(\mathbf{y_i} - \beta_0 - \sum_{j=1}^{p} \beta_i \mathbf{x}_{ij}\big)^2. \tag{5}$$

When $n > p$ the OLS is well-behaved and its properties are very well understood. However, when the number of parameters p is larger than n (i.e. high dimensionality), OLS is not an efficient option because the least-squares estimates are not unique. That is because the set of possible solutions becomes infinite. In this context, the high-dimensional techniques are known to outperform the OLS method mainly because of two reasons: accuracy and interpretability. The first reason comes from the fact that penalized regression methods negligibly increase the bias of the estimators but greatly reduce their variance, which potentially leads to an overall improvement of the accuracy of the estimators [7]. Moreover, these methods select a smaller subset of predictors, which are precisely the ones with the strongest effect on the model, and therefore the interpretation of those becomes easier than in the OLS case. The basic idea of the penalized regression methods is to impose some constraint on the coefficients vector. In comparison to the OLS case, this penalization enables to shrink the estimates of the coefficients and reduce the number of active ones in some cases

$$\underset{\beta_0,\beta}{\text{minimize}} \sum_{i=1}^{N} \big(\mathbf{y}_i - \beta_0 - \sum_{j=1}^{p} \beta_i \mathbf{x}_{ij}\big)^2 \text{ subject to } \|\beta\|_q^q \leq t \qquad \text{where, } \|\beta\|_q = \sum_{j=1}^{p} |\beta_i|^q. \tag{6}$$

We can interpret the constant t as the maximum size of the desired solution set. It is important to mention that the constant t is specified by the developer of the model, which is commonly established through the performance estimation of artificial training data generated (i.e. cross validation procedure). The expression above is commonly expressed in a Lagrangian form,

$$\text{minimize}\, \mathcal{L}(\beta_0, \beta_i, \lambda) \qquad \text{where, } \mathcal{L}(\beta_0, \beta, \lambda) = \left\{ \sum_{i=1}^{N} \big(\mathbf{y}_i - \beta_0 - \sum_{j=1}^{p} \beta_i \mathbf{x}_{ij}\big)^2 + \lambda \|\beta\|_q^q \right\}. \tag{7}$$

Lagrangian duality establishes equivalence of (6) and (7) for a particular value of λ [6]. In practical terms, the larger the λ is, the more restrictive the regularization is, which, in turn, causes a sparser solution on the coefficients vector.

Additionally, there are different penalized regression techniques that differ in the choice of the underlying norm. These methods have a common characteristic which is to penalize the ordinary least squares method by the imposition of a penalization on the estimators. The main difference among them is the norm of the constraint imposed. For a ℓ_1 norm (lasso or ℓ_1-constraint), the constraint is the sum of the absolute values of the coefficients. On the other hand, $q = 2$ (ridge regression or ℓ_2-constraint), constrains the squared Euclidean norm of the parameters. For $q \to 0$, best subset selection, it constrains the number of active parameters on the coefficients vector. Finally, there exists the elastic-net regularization which establishes as a constraint the linear combination of the ℓ_1 and ℓ_2 constraints

$$\mathcal{L}(\beta_0, \beta_i, \lambda) = \left\{ \sum_{i=1}^{N} (\mathbf{y}_i - \beta_0 - \sum_{j=1}^{p} \beta_i \mathbf{x}_{ij})^2 + \lambda\Big((1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1^1\Big) \right\}. \tag{8}$$

where $\alpha \in [0, 1]$ is the elastic-net parameter and provides the mix between the lasso and the ridge penalty.

In geometric terms, these methods modify the constraint area of the problem leading to different

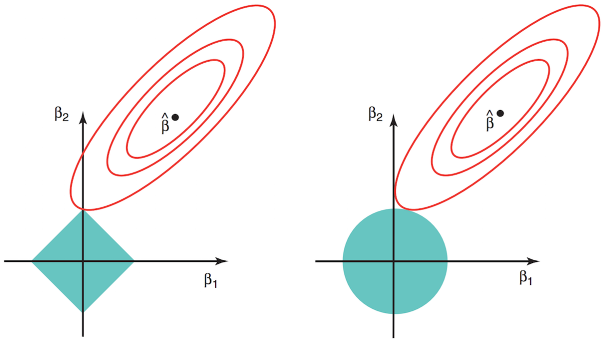


Fig. 1: Estimated graphical representation of equation (6) *for* $q = 1$ *(right) and* $q = 2$ *(left) in* $\mathbb{R}^2$*. The estimated graph of the lasso (left) and ridge regression (right). The red ellipses are the contours of the Residual Sum of Squares, RSS, and the solid blue areas correspond to the lasso constraint (diamond shape) and rigde constraint (ball shape)*

results. In figure 1, it is exhibited the constraint shape of both lasso and ridge regression. The fact that lasso, as formulated by (6), promotes sparsity can be easily observed in figure 1. Due to its diamond shape, the set of possible solutions are more likely to be tangent at one of the corners in comparison to the ridge regression, which its disk shape has no corners. Thus, a key attribute of the ℓ_1-penalization is the ability to obtain solutions which are sparse.

In many high dimensional settings, the ℓ_1-constraint is the desired technique because it is the smallest dimension of q that keeps the convexity property and, at the same time, causes some of the estimated coefficients to be exactly zero (i.e. performs variable selection). From higher dimensions of ℓ_q ($q > 1$), the set of solutions obtained are not sparse which might not be desirable in the context of high-dimensionality. On the other hand, for $q < 1$ the set of solutions turns out to be sparse but at the expense of non-convexity of the problem, which leads in most of the cases to a great increase in the computational complexity. These different properties can be observed in figure 2.

This project seeks to investigate the impact of these penalized approaches in finance, more concretely in the estimation of the inverse covariance matrix. This problem has been exhaustively studied in the recent years and it has seized the attention of many statisticians who have proposed

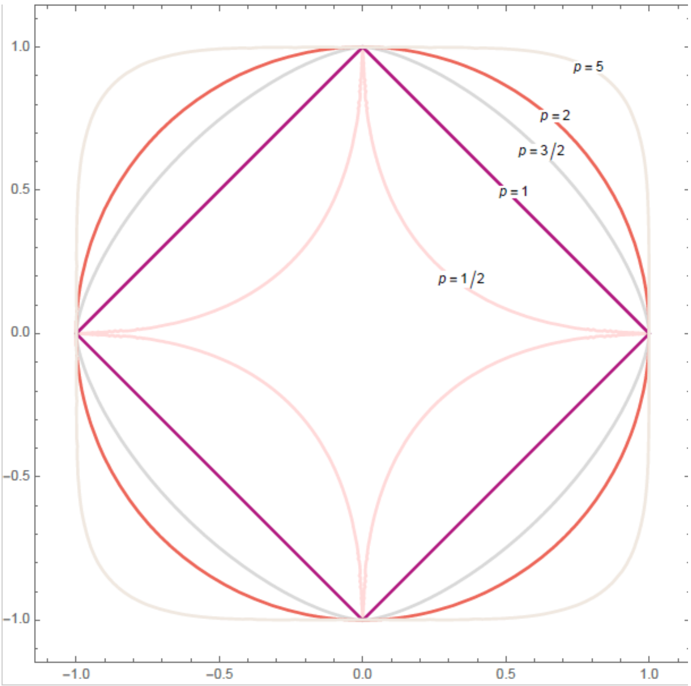


Fig. 2: Representation of the constraint areas according to the different p-norms

different techniques to solve it. In 2006, the first attempt to solve this problem was proposed by Meinshausen and Bühlmann. They applied a lasso regression on each variable of the precision matrix, holding the others as predictors, to cause sparsity. Asymptotically, this approach was shown to consistently estimate the active elements of the inverse covariance matrix [15]. On the other hand, other recent studies Yuan & Lin (2007), Banerjee et al. (2007) and Friedman et al. (2007) propose new solutions based on the maximization of the Maximum Likelihood Estimator with a lasso penalty on the precision matrix, which are commonly known as graphical lasso. These sparse estimation approaches establish zeros in the inverse covariance matrix which brings interest since they correspond to conditional independences between the variables, as we describe below.

We focus on the second approach which is based on the log-likelihood maximization with a lasso penalty. Let $X \sim N_p\left(\mu, \mathbf{\Sigma}^0\right)$ be a Gaussian distribution in p-dimensions, with mean vector $\mu \in \mathbb{R}^p$

and covariance matrix $\mathbf{\Sigma}^0 \in \mathbb{R}^{p \times p}$:

$$\mathbb{P}_{\mu,\mathbf{\Sigma}^0}(x) = \frac{1}{(2\pi)^{\frac{p}{2}}(\det[\mathbf{\Sigma}^0])^{\frac{1}{2}}} e^{(-\frac{1}{2}(x-\mu)^T(\mathbf{\Sigma}^0)^{-1}(x-\mu))}. \tag{9}$$

When X follow multivariate normal distribution with mean zero and a precision matrix $\mathbf{\Theta} \in \mathbb{R}^{p \times p}$, the rescaled log-likelihood form, $\mathcal{L}(\mathbf{\Theta}; X)$, of the expression above (9) can be rewritten as:

$$\mathcal{L}(\mathbf{\Theta}; X) = \frac{1}{N}\sum_{i=1}^{N} \log \mathbb{P}_{\mathbf{\Theta}}(x_i) = \frac{1}{2}\log|\mathbf{\Theta}| - \frac{1}{2}\mathrm{tr}\big(\mathbf{S}\mathbf{\Theta}\big), \tag{10}$$

where $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T$ is the sample covariance matrix. The log-likelihood function in equation (10) is strictly concave which means that the maximum, when it is achieved, is guaranteed to be unique. However, in high-dimensional settings, the $\mathbf{S}$ is rank deficient, thus the MLE fails to exist. Therefore, a regularization parameter has to be included in order to induce some sparsity on the precision matrix and make the computation feasible.

Our estimator will be of the form

$$\hat{\mathbf{\Theta}} \in \arg\max_{\Theta > 0}\big\{ \log|\mathbf{\Theta}| - \mathrm{tr}\big(\mathbf{S}\mathbf{\Theta}\big) - \lambda_{\rho_1}(\mathbf{\Theta})\big\}, \tag{11}$$

where $\rho_1(\mathbf{\Theta}) = \sum_{i \neq j} |\mathbf{\Theta}_{ij}|$ is the ℓ_1-norm of the off-diagonal entries of the precision matrix. From the maximization of the penalized MLE problem stated above, a consistent estimator of the inverse covariance matrix is obtained [4]. The precision matrix contains the information about the partial variance and partial covariance between the different random variables. In the case of multivariate Gaussian distribution, this is extremely useful since zero partial covariance corresponds to conditional independence. More precisely by [10, Proposition 5.2], we have:

$$\mathbf{X}_a \coprod \mathbf{X}_b \mid \mathbf{X}_{V\setminus\{a,b\}} \iff \mathbf{\Theta}_{a,b} = 0. \tag{12}$$

Additionally to the computation of the precision matrix, in this project, we are also interested in the interpretation behind the precision matrix. For this reason, we use the undirected graphical models, which are direct representations of the random variables in the precision matrix. On a high level graphical models can be defined as follows [12]: Let $\mathbf{X} = (X_1, \dots, X_p)$ be a random vector which follows a distribution $\mathbf{P} = N(\mu, \mathbf{\Sigma})$. Denote $\mathbf{G} = (V, E)$ to be the undirected graph corresponding to $\mathbf{P}$ which comprises a vertex set V and a set E of edges. Each random variable X_i corresponds to an element of the set V. That is, V has p elements. Edges indicate relationship between the elements of V. The set E of edges can be regarded as a $p \times p$ adjacency matrix where $E(i,j) = 1$ if there is an edge between X_i and X_j and 0 otherwise. Thus, the set E excludes the edge (i,j) if and only if X_i and X_j are conditional independent given the other variables $O_{\setminus\{i,j\}} \equiv (X_s : 1 \leq s \leq p, s \neq i, j)$, written

$$X_i \coprod X_j | O_{\setminus\{i,j\}}. \tag{13}$$

Under multivariate Gaussian distribution, equation (13) holds if and only if $\mathbf{\Theta}_{i,j} = 0$ which follows by (12). The undirected graphs are extremely useful because they give an easy graphical interpretation about the joint distribution of the variables.

4 Sparse Estimation of the Precision Matrix

In this section, we put into practice the regularized maximum log-likelihood estimator of the precision matrix and the undirected graph models with a real dataset. We retrieve from the Yahoo database monthly returns of each NASDAQ component from January 2016 until May 2017. Our data set is composed of $n = 17$ observations for each of the $p = 107$ components. That is, we have 1819 observations but we need to estimate $\binom{p+1}{2} = 5778$ parameters. The goal is to estimate the optimal inverse covariance matrix. Thus high dimensional methods are required to impose some sparse structure on the precision matrix otherwise the covariance matrix would be very unstable. All the computations in this section are done in R [1].

Prior to estimating the precision matrix, we run a Shapiro-Wilk multivariate normality test to check whether our data follow a Gaussian distribution. The p-value of the test is $7.556e - 14$, trivially zero, therefore there is strong evidence that our data are not normally distributed. As Guassianity is violated an adaptation of the methods is required. All the techniques we have studied so far assume that the data follow a multivariate Gaussian distribution. While Gaussian graphical models can be useful, relying on exact normality is limiting. In fact, in many cases, the data violate the normality assumption. Therefore a more flexible model that relaxes such assumption is needed in order to increase the estimator's accuracy. In our case, as the matrix of returns is not normally distributed, we will apply a nonparametrical extension of the undirected graphical models based on multivariate Gaussian distribution to relax the normality assumption, as proposed by Liu, Lafferty and Wasserman (2009) [12].

It is known that any multivariate joint distribution can be modelled by separately estimating the univariate marginals distribution function and copulae, which links the marginal distributions together. We will assume that the joint distribution of our model follows a semiparametric Gaussian distribution called *nonparanormal* distribution. It combines nonparametric marginals, $\{f_j\}$, with a high dimensional Gaussian copula, $C(\cdot)$.

[1] The code is available upon request.

By Sklars Theorem [17]:

$$H\big(x_1, .., x_p\big) = C\big(f_1(X_1), ..., f(X_p)\big), \tag{14}$$

where $H(\cdot)$ is the joint distribution, $C(\cdot)$ is the copula and $F(\cdot)$ are the marginal distributions. The nonparanormal distribution depends on the nonparametric marginal functions, $\{f_j\}$, and the mean μ and correlation matrix $\mathbf{\Sigma}^0$ of the Gaussian copula. Hence, our p-dimensional stock data $\mathbf{X} = \big(X_1, ..., X_p\big)^T$ has a nonparanormal distribution $X \sim NPN_p\big(f, \mathbf{\Sigma}^0\big)$ if $C(f_1(X_1), ..., f_p(X_p))^T \sim N_p\big(0, \mathbf{\Sigma}^0\big)$.

Furthermore, it is recommended that the empirical distribution of the marginals is truncated (for details see section IV [12]) since using the entire empirical distribution leads to inaccurate inference, as stated by Liu, Lafferty and Wasserman (2009) [12]. Rearranging the formula (11) to correct for nonparanormal distribution and truncated marginals distribution, we obtain the following minimization problem:

$$\hat{\mathbf{\Theta}}_n = \arg\max_{\Theta > 0}\Big\{\log|\mathbf{\Theta}| - \operatorname{tr}\big(\mathbf{\Theta}\mathbf{S}_n(\tilde{f})\big) - \lambda_n\|\mathbf{\Theta}\|\Big\}, \tag{15}$$

where $\mathbf{S}_n(\tilde{f})$ is the Winsorized sample covariance matrix [2], λ is the regularization parameter and $\|\,\mathbf{\Theta}\,\| = \sum_{j \neq k}|\mathbf{\Theta}_{jk}|$. The choice of λ is critical since it determines the sparsity level of the estimated precision matrix. As we see in our undirected graph models, figure 3, higher values of λ lead to sparser but very stable estimates and smaller values of λ yield denser and less stable estimations.

There exist different selection procedures for λ such as *K-fold Cross Validation* (K-CV), *Akaike information criterion* (AIC), *Bayesian information criterion* (BIC) or *Stability Approach to Regularization Selection* (StARS). We will use the latter one as it is proven to outperform the other procedures and be the suitable choice in high dimensional settings, as shown by Liu, Roeder and Wasserman (2010) [13]. The StARS approach chooses λ based on stability. Obviously, there is a

[2] The Winsorized covariance corrects the heavy influence extreme values have by setting the tail values equal to a certain percentile value, (see [12] for details).

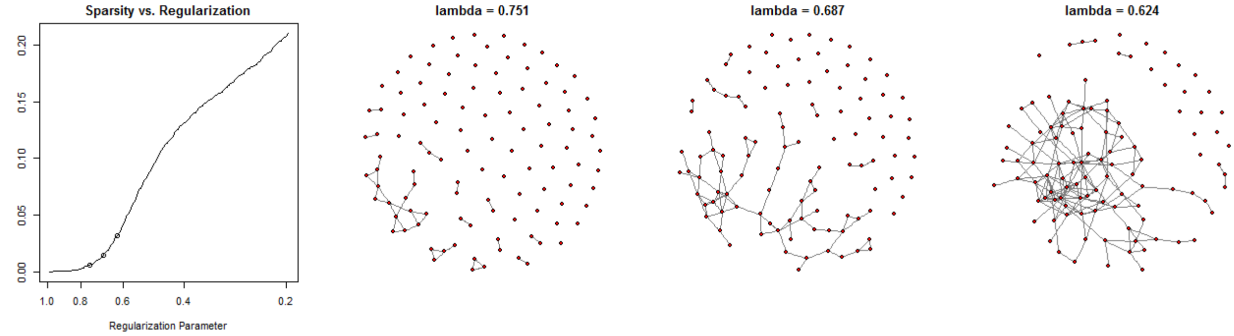


Fig. 3: *On the left, the graph corresponds to the different values of lambdas, as a log scale of the max to min λ, and its associated sparsity level (here, 0 sparsity corresponds to no edges). Thus, we observe that as the ratio decreases (i.e. smaller values of λ), the graph becomes denser, as shown in the 3 graphical models. (Own source)*

trade-off between sparser and denser graphs. Denser graphs will be more likely to contain the true graph but at the expense of less sparsity and an increase of false positive cases. What the approach does is to find the right trade-off with a tendency to overselect rather than underselect, this is, it is preferred to have some false positives rather than false negatives cases.

To estimate our precision matrix we use the huge package [18] in R which contains implementations of all these different models, transformations and selection criteria. We first apply a nonparanormal transformation to the returns matrix. We then estimate high-dimensional undirected graphs for 200 different values of λ. The chosen method to solve the minimization problem (15) is the graphical lasso algorithm, Friedman (2007) [4]. Afterwards, we select the optimal regularization parameter using the StARS criteria. The corresponding undirected graphical model of the optimal estimated sparse precision matrix is shown in figure 4.

4.1 Interpretation of the Results

In this subsection, we present the main findings from the exhaustive analysis of our optimal undirected graph. From figure 4 we can assess the conditional independences among the NASDAQ components during the period studied.

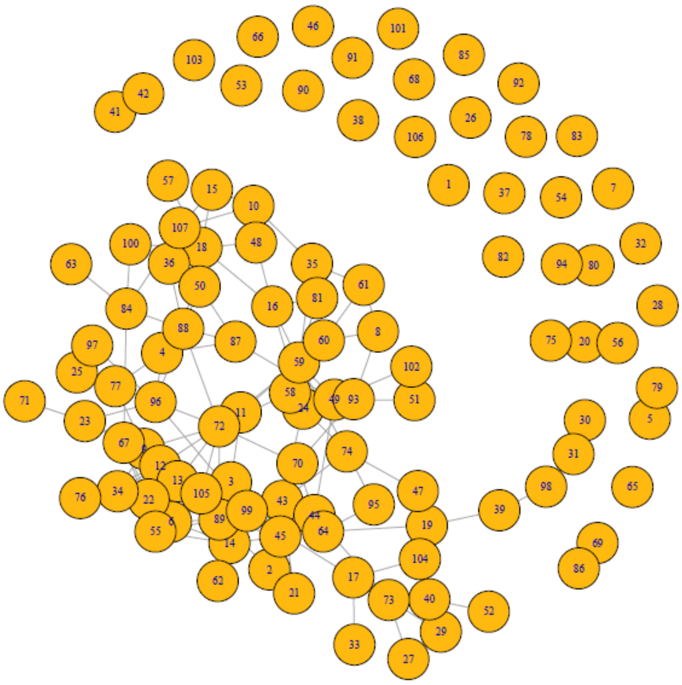
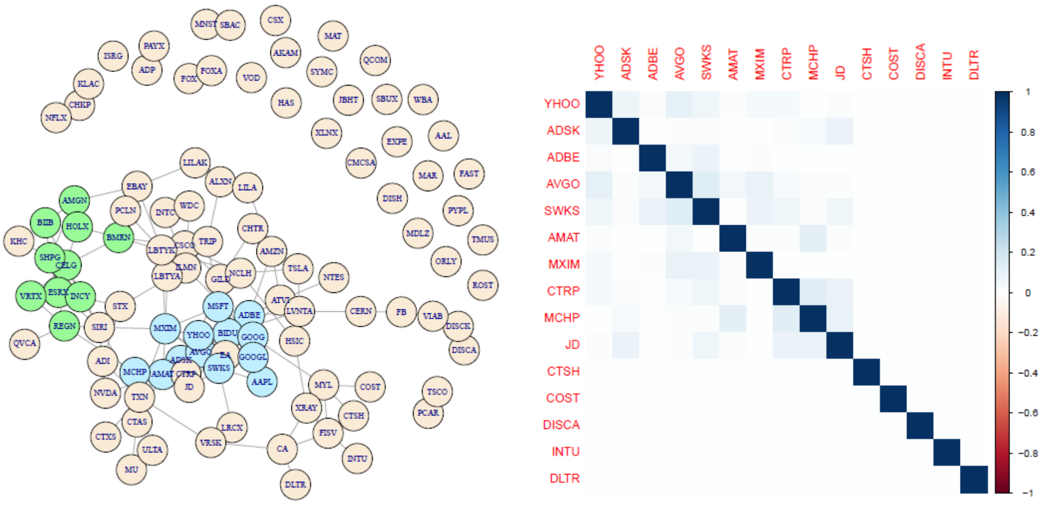


Fig. 4: Optimal Undirected Graph(Own source)

First, we can observe that there are 23 isolated nodes. That is, 23 components are marginally independent of all the other variables. Among these stocks we have firms such as American Airlines Group, Comcast Corporation, Expedia Inc., Marriot, Starbucks, T-Mobile US or Vodafone. The existence of such companies in this group can have a twofold explanation. On one hand, they are typically very diversified firms and this may explain their returns can be unrelated to other components return. And, on the other hand, most of them operate in many different countries and a big portion of their revenue is obtained overseas. This fact may reveal that they are less dependent on American market fluctuations.

From the graph we can also observe four pairs of nodes with only one edge. For instance, one of them is formed by FOX and FOXA which are the same company, 21st Century Fox, and seem to be unrelated to the other components return. And another pair corresponds to Paychex and Automatic Data Processing which are the only human resources companies listed in the NASDAQ stock market. On the other hand, we can also distinguish one group composed of three nodes and



(a) Optimal Undirected Graph *(b) Correlation Plot*

Fig. 5: Cluster analysis: The blue and green nodes in figure 5a correspond to the High-tech and Biotech clusters, respectively. 5b shows the correlations among several stocks, presenting positive correlation for stocks within the high-tech cluster (i.e. directly connected nodes) and zero correlation for indirectly connected nodes. (Own source)

two edges which is constituted by Netflix, Check Point Software Technologies Ltd. and Kla-Tencor Corporation.

Then, even though there exist interdependencies among all the sectors, through a more exhaustive analysis we can detect two big clusters with some exceptions. The first one corresponds to high-technology and semiconductors firms such as Google, Apple, Microsoft, Adobe, AutoDesk, Skyworks Solutions or Yahoo. The other cluster corresponds to Biotech and Pharmaceutical companies like Celgene, Shire PLC, Biogen, Incyte or Vertex Pharmaceuticals. This result is consistent since the return of the stocks within each cluster should be affected by the same external shocks and factors. These two clusters are visualized in figure 5a. However, as the sample length is relatively small we might have a significant randomness effect biasing our interpretation.

On the other hand, an important attribute of our estimated inverse covariance matrix is that almost all of the active entries are nonpositive. This is a very strong property: For $A \subseteq \{1, \ldots, p\}$ let X_A denote a subvector of $\mathbf{X} = (X_1, \ldots, X_p)$ with coordinates $X_{i:i \in A}$. If the inverse covariance matrix of $\mathbf{X}$ has nonpositive off-diagonal entries it implies that for any pair of variables X_i, X_j and any

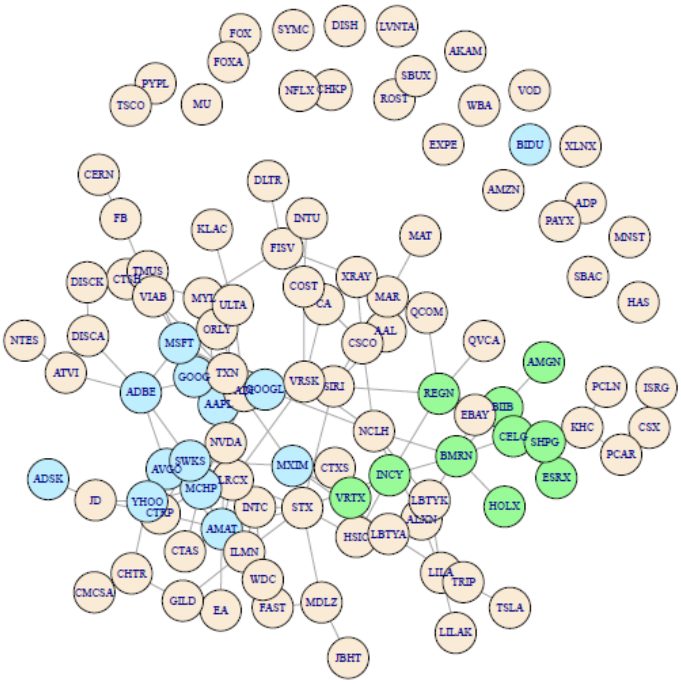


Fig. 6: Optimal Undirected Graph with stationary data(Own source)

subvector X_A of $\mathbf{X}$, the conditional correlation $corr(X_i, X_j|X_A)$ is nonnegative (see e.g. Karlin-Rinott 1983 [8]). Recently, this property has been linked to the estimation of graphical models also in the high-dimensional settings (see Fallat et al. [3], Lauritzen et al. [9]). This can be observed in figure 5b where all the correlations, if nonzero, are positive.

5 Correcting for difference stationary

In this section, we proceed to analyse the behaviour of our time series data and, after correcting for stationarity, we compare the outcome of the inverse covariance matrix and the undirected graph with the previous results.

We run an Augmented Dickey-Fuller test on the matrix of returns to test for stationarity and for some stocks the null hypothesis cannot be rejected at the 5% significance level. That is, there is evidence they have a unit root (i.e. non-stationary). To correct for stationarity we take the first differences of the matrix of returns and run again the Augmented Dickey-Fuller test. In this case, all the obtained p-values are smaller than 0.05. Thus, the transformed data are stationary. The same procedure used in the previous section is conducted to obtain the estimated sparse inverse covariance matrix and its optimal undirected graph.

The optimal undirected graph with stationary data, shown in figure 6, is denser than the non-stationarity graph, figure 5a, which implies that fewer stocks are marginally independent of all the other variables after correcting for difference stationarity. Nevertheless, notice that the graph preserves a similar structure compared to the previous scenario. For most of the cases, highly connected stocks in the figure 5a are still very connected with the transformed data, figure 6, and that is also true concerning the disconnected nodes. Therefore, the results in this section do not significantly differ from the results in the previous section.

Furthermore, in this instance, the transformed data set does not satisfy all the restrictive properties of the inverse covariance matrix having nonpositive entries. Consequently, all the estimated correlations are no longer guaranteed to be nonnegative. In fact, a substantial number of estimated correlations are negative for some specific stocks. That is, these stocks have weak, since values are close to zero, but negative co-movements with the other components.

6 Sparse Markowitz Optimization Problem

In this section we solve the Markowitz optimization problem with a lasso penalty on the vector of portfolio weights. We will show that adding such restriction has various advantageous results and it corresponds to penalizing short sales positions. We will use the untransformed matrix of stocks returns and the optimal sparse precision matrix obtained in section 4 to find the optimal sparse vector of portfolio weights. All the computations in this section are carried out in the software

Matlab using the CVX package [5] [3].

As originally stated by Markowitz, the portfolio selection problem is very unstable since small changes in the returns, variances or correlations can lead to dramatically different results in the optimal weighting factors. That is, the objective function is very sensitive to the parameter inputs and thus, it is said that the optimization problem is an ill-conditioned inverse problem [16]. Moreover, the unrestricted Markowitz framework has not been proven to significantly outperform the naive strategy [2], which consists on evenly dividing the contributions across the stocks (i.e. $w_i = \frac{1}{n}$). To overcome the aforementioned problems, two main approaches have been proposed and deeply studied: (1) Estimate the covariance matrix with the optimal convex combination of the identity matrix and the sample covariance matrix [11], or (2) Add a ℓ_q-norm penalty on the portfolio weights [1]. In line with our project, we implement the second approach which consists of solving the Markowitz optimization problem with a ℓ_1 norm penalty on the portfolio weights vector

$$\underset{w}{\text{minimize}} \quad (\omega^T \mathbf{\Sigma} \omega + \lambda \|\omega\|_1) \quad \text{subject to} \quad \omega^T \mathbf{1} = 1 \qquad \text{and} \quad \omega^T \mu = r^*, \tag{16}$$

where $\|\omega\|_1 = \sum_{j=1}^{p} |\omega_i|$.

Quite interestingly, unlike equation (1), this program will typically output a nonnegative optimal vector ω. This relies on three basic observations: (1) Adding a constant λ to the function being optimized does not change anything from the optimization point of view, (2) If the constraint that the capital has to be fully invested (i.e. $\omega^T \mathbf{1} = 1$)is satisfied, then $\lambda = \lambda \sum \omega_i$, and (3) $\sum |\omega_i| = \sum_{i \text{ with } w_i \geq 0} \omega_i - \sum_{i \text{ with } w_i < 0} \omega_i$. Thus, equation (16) can be rewritten as follows

$$\underset{w}{\text{minimize}} \quad (\omega^T \mathbf{\Sigma} \omega + 2\lambda \sum_{i \text{ with } w_i < 0} |\omega_i|) \quad \text{subject to} \quad \omega^T \mathbf{1} = 1 \qquad \text{and} \quad \omega^T \mu = r^* \tag{17}$$

[3] The code is available upon request.

which is equivalent to solving (1) with a penalization on short sales. As shown by Brodie, Daubechies, De Mol, Giannone and Loris (2009) [1], adding a ℓ_1 penalty on the objective function has several useful consequences:

- Stabilizes the optimization problem
- Promotes sparsity
- Accounts for transaction costs

The value of λ allows to adjust the importance of the penalty on the short sales. For large enough values of λ no-short-positions optimal portfolios are obtained. That is, there exists an extreme value of λ for which the penalized objective function (17) is equivalent to solving the unpenalized problem (1) with an additional constraint on short sales (i.e. $\omega_i > 0$). Notice that further increasing this value has no effect on the output of the optimization procedure since only short sales are penalized. We applied this method to our data set and, in our case, $\lambda = 0.1$ is the value which leads to the sparsest solution. Solving the equation (17) with such λ causes 26 stock weights to be exactly zero, which indeed correspond to all the components with negative weights in the unconstrained problem and some positively weighted stocks, which their values were very close to zero under the unpenalized setting. The fact that some sparsity for nonnegative positions is obtained derives from the construction of the problem. Penalizing short sales also implies shrinking the positive-active weights.

7 Conclusions

In this project we have studied the different shrinkage techniques to overcome the high-dimensionality problem. A particular focus has been placed on the graphical lasso which performs a penalized MLE with a ℓ_1-regularization penalty on the precision matrix. We have applied this technique to stock returns data to obtain the optimal sparse precision matrix. Multivariate Gaussian distribution has been relaxed and instead we assumed data follow a nonparanormal distribution. Through the undirected graphical models we have assessed the conditional independences among the NASDAQ stocks. Further research was needed to evaluate the time series nature of the dataset which turned out to be non-stationary. Thus, the data were transformed to correct for difference stationarity and the results of the optimal undirected graph with the transformed data did not significantly differ from the results with the non-treated data. Finally, we have used our estimated sparse precision matrix to solve the Markowitz optimization problem with an additional ℓ_1 penalty on the vector of portfolio weights. It has been showed that such penalization leads to sparse and stable optimal portfolios and it is equivalent to penalizing short-sales.

The covariance matrix can also be estimated using the third high-dimensional approach which is to compute the optimal projections of the predictors. This goes beyond the scope of the present thesis which was to study and implement the shrinkage techniques. In future work, we plan to extend this project by estimating the covariance matrix through dimension reduction techniques and compare it to the results obtained in the current project.

References

[1] Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

[2] Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *Review of Financial Studies*, 22(5):1915–1953, 2009.

[3] Shaun Fallat, Steffen Lauritzen, Kayvan Sadeghi, Caroline Uhler, Nanny Wermuth, and Piotr Zwiernik. Total positivity in Markov structures. *Ann. Statist.*, 45(3):1152–1184, 2017.

[4] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[5] Michael Grant, Stephen Boyd, and Yinyu Ye. Cvx: Matlab software for disciplined convex programming, 2008.

[6] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. 2015.

[7] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 6. Springer, 2013.

[8] Samuel Karlin and Yosef Rinott. M-matrices as covariance matrices of multinormal distributions. *Linear Algebra and its Applications*, 52:419–438, 1983.

[9] Steffen Lauritzen, Caroline Uhler, and Piotr Zwiernik. Maximum likelihood estimation in Gaussian models under total positivity. *arXiv preprint arXiv:1702.04031*, 2017.

[10] Steffen L Lauritzen. *Graphical Models*, volume 17. Clarendon Press, 1996.

[11] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[12] Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct):2295–2328, 2009.

[13] Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (StARS) for high dimensional graphical models. In *Advances in Neural Information Processing Systems*, pages 1432–1440, 2010.

[14] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[15] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

[16] Richard O Michaud. The Markowitz optimization enigma: Is optimized optimal? *ICFA Continuing Education Series*, 1989(4):43–54, 1989.

[17] M Sklar. *Fonctions de répartition à n dimensions et leurs marges.* Université Paris 8, 1959.

[18] Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, 13(Apr):1059–1062, 2012.