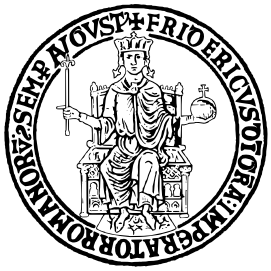


UNIVERSITY OF NAPLES "FEDERICO II"

DEPARTMENT OF ECONOMICS AND STATISTICS

DOCTORAL THESIS

Essays in statistical methods for asset allocation

PhD Candidate:
Raffaele MATTERA

Supervisor:
Giovanni Walter PUOPOLO
Germana SCEPI
PhD Director:
Marco PAGANO

March 3, 2022

UNIVERSITY OF NAPLES "FEDERICO II"

Abstract

Department of Economics and Statistics

Doctor of Philosophy

Essays in statistical methods for asset allocation

by Raffaele MATTERA

The present work, divided into three main chapters, discusses the development and the application of novel statistical techniques for portfolio selection problems. The first chapter is devoted to the *estimation theory*, and a new estimator for the precision matrix, called precision shrinkage, is developed to reduce the estimation error. The analysis provided in the chapter show that the use of precision shrinkage lead to the construction of more desirable portfolios in terms of return/risk trade-off with respect to well established alternatives. The second chapter studies the ability of *forecasting* techniques in constructing more attractive portfolios than strategies based on static estimation. Classical model-based econometric methods are compared with data-driven machine learning ones. We find that, for both low and large dimensions, the use of forecasts improves the out-of-sample portfolio performances if model-based approaches are employed. The last chapter discusses the usefulness of *clustering* in portfolio selection. Clustering can be used to reduce the asset allocation dimensionality. Several algorithms are compared in terms of out-of-sample profitability. As a main result, we show that clustering-based portfolios dominate the classical approaches.

Contents

List of Figures

List of Tables

Thesis overview

Asset allocation involves the decision about how many and which kind of assets to include in a portfolio for investment purposes. As argued by Markowitz (1952), the portfolio selection process can be divided into two stages. In the first stage, the investor observes the historical assets' returns and, in the second one, he/she estimate or predict their current or future characteristics. According to the mean-variance framework of Markowitz (1952), two main characteristics are used by the investor in order to make optimal choices: the expected value of asset's returns, which are contained in a vector μ, and the inverse of the assets' covariance matrix Σ, that is called *precision matrix* and denoted by Σ^{-1}. Note that optimal choices are referred to those that maximize the investor's utility function that is supposed to be mean-variance in Markowitz (1952).

In general, asset allocations can be based on either historical data or forecasts. An asset allocation based only on historical data involves the estimation of the *current* vector of expected returns and the current covariance structure Σ. However, the investor can also chose to adopt a forward looking approach, by anticipating future market conditions. This can be done by predicting what will be the *future* expected returns and covariances. When the investor uses some forecasts instead of static estimates, we say that he/she is implementing a *timing* asset allocation strategy.

The main problem related to Mean-Variance (MV, Markowitz, 1952) allocations[1] is that both μ and Σ are unknown quantities and need to be estimated or predicted. How close are the investor's estimates, called $\hat{\mu}$ and $\hat{\Sigma}$, to the true mean vector μ and covariance matrix Σ? The fact that economic agents provide estimates and forecasts about unknown quantities rise the so-called *estimation error* problem. Every time something is estimated, there is a certain probability of making mistakes. Hence, if the investors do not accurately estimate all the required quantities, asset allocation becomes sub-optimal and performs poorly in out-of-sample. Moreover, estimation error further increases in a so-called *large dimensional setting*, where the number of

[1] Note that actually, the same problem applies to any asset allocation strategy.

assets N is larger than the time series observations T, i.e. $N > T$. In this case, the covariance matrix cannot be inverted, and the *precision matrix* cannot be computed. Estimation error makes optimal portfolios less attractive than a naive asset allocation strategy where all the assets are equally weighted (e.g. see Frost and Savarino, 1986a; Michaud, 1989; Chopra and Ziemba, 1993; De Miguel, Garlappi, and Uppal, 2007). Among the others, De Miguel, Garlappi, and Uppal (2007) explicitly show that the naive (also called the $1/N$) strategy leads to the construction of a more desirable portfolio in terms of return/risk trade-off compared to more complex and advanced techniques. This result holds because an equally weighted strategy has zero estimation error because nothing is estimated. From this empirical fact, an important question arises: how can the estimation error be reduced in an asset allocation strategy? Can means and covariances be estimated or predicted so that the out-of-sample performance is maximized? How can investors deal with highly dimensional settings? These are the main research questions underlying this thesis.

In what follows, we provide some insights into such questions. In particular, we decided to focus on the covariance matrix estimation for implementation of the static asset allocation. However, for the forecasts-based asset allocation we consider the problem of predicting both means and covariances. There is an important motivation that justifies a specific focus on the covariance matrix estimation for static asset allocation. Indeed, Kourtis, Dotsis, and Markellos (2012) demonstrated that a Global Minimum Variance (GMV) strategy, where only the covariance structure is used to build the optimal portfolios, contains a lower estimation error than a Mean-Variance (MV) allocation. Intuitively, this happens because GMV avoids the estimation error contained in the expected returns' vector $\hat{\mu}$. As a result, throughout the thesis, we will consider both MV and GMV asset allocation strategies.

The thesis is structured as follows.

In the **first Chapter**, following most of previous literature (Barry, 1974; Frost and Savarino, 1986b; Kan and Zhou, 2007; De Miguel, Garlappi, and Uppal, 2007), we focus on *static* asset allocation within a standard *low-dimensional* setting where there are more time observations than assets (i.e. $T > N$). More in detail, we study how different ways of estimating the covariance matrix affect the estimation error. Intending to reduce the estimation error, we propose a new estimator for the covariance inverse (the *precision matrix*) based on the shrinkage technique of Stein (1956) and Ledoit and Wolf (2003) and Ledoit and Wolf (2004a). In particular, in the case

of MV allocation, we are able to find a closed formula for the optimal shrinkage intensity derived to maximize the investor's expected utility. Unfortunately, the same does not apply in the case of the GMV setting. In this second setting, we still derive the optimal shrinkage intensity from maximizing investor preferences but only through simulations. For both MV and GMV strategies, we provide simulation studies to demonstrate the superiority of the proposed *precision shrinkage estimator* with respect to the most important alternative estimators. We conclude the first Chapter by considering an application to real data of the developed *precision shrinkage estimator* in the case of MV asset allocation, providing evidence of its superiority in out-of-sample.

In the **second Chapter**, we consider the case of forecasts-based asset allocation in order to deeper understand the usefulness of forecasting in portfolio selection problems. The idea is based on the evidence highlighted by some studies (e.g. Pesaran and Timmermann, 1995; Fleming, Kirby, and Ostdiek, 2001; Fleming, Kirby, and Ostdiek, 2003; Marquering and Verbeek, 2004) that forecasting improves the profitability of asset allocation strategies. As stated previously, the over performances of these approaches can be found in the concept of *timing*, i.e. anticipating the future market conditions. Two important novelties are introduced in Chapter 2. First of all, we assess if and how much a mean-variance investor gains from using predictions rather than static estimated quantities for both low and high-dimensional settings. Studying the usefulness of forecasting in high-dimensional asset allocation is still an unexplored topic and this Chapter provides the first attempt in this direction. Second, we provide the first comparison between econometric (model-based) approaches and novel machine learning (data-driven) ones for the implementation of timing strategies. Briefly, we demonstrate that forecasting either the mean or the covariance seems better than predicting both (i.e. full-timing) and that forecasting is helpful for portfolio selection problems in large dimensional settings. Then, we show that machine learning forecasts are not useful, especially in low dimension. However, this result could be driven by the adopted methodological framework (as in DeMiguel, Garlappi, and Uppal, 2009) based on rolling-window approach, where relatively few time observations are used for training the ML algorithm in each recursion. Perhaps, this number is not enough large to ensure an accurate training of the ML algorithm. Therefore, the results simply show that machine-learning based

timing strategies are not appropriate for long-run (low-frequency) portfolio analysis.

The **third Chapter** explores the usefulness of clustering for portfolio selection problems in both low and large dimensional frameworks. Clustering is an unsupervised learning technique used to alleviate the course of dimensionality. Indeed, clustering can be used to build roughly diversified portfolios that than become the input of a low-dimensional asset allocation problem. Although clustering is recently being applied for portfolio selection, it is still unclear which technique provides more desirable portfolios in out-of-sample. Therefore, different clustering approaches are compared and a new clustering technique based on the time series distribution characteristics is also developed in the Chapter. The performances of investment strategies based on clustering are evaluated in detail. Overall, the empirical findings suggest that in the case of common stocks, clustering-based asset allocation is helpful to the extent to which roughly diversified funds are constructed. However, there is not a single clustering algorithm that consistently outperforms the others for all the experiments (i.e different datasets and portfolio rules), but the distribution-based approaches are the best in the majority of the cases. Therefore, the last Chapter shows that clustering is useful when dealing with common stocks, especially the approaches based on distributional characteristics developed therein.

Chapter 1

Improved asset allocation with precision shrinkage

1.1 Uncertainty on portfolio weights: the mean-variance investor

Asset allocation involves deciding how many and which kind of assets to include in a portfolio for investment purposes. The critical contribution of financial economics literature to this topic is given by the mean-variance approach of Markowitz (1952). The Mean-variance asset (MV) allocation can be summarized as follows.
Suppose to have a $N \times N$ matrix X of jointly normally distributed N asset returns $X \sim \mathcal{N}(\mu, \Sigma)$, observed for T times. To find the vector of optimal portfolio weights w, we suppose that the investor maximizes the following utility function:

$$U(w) = w'\mu - \frac{\gamma}{2} w'\Sigma w \tag{1.1}$$

where $w'\mu$ is the portfolio's expected return and $w'\Sigma w$ its portfolio variance. Hence the optimal portfolio weights vector is:

$$w^* = \frac{1}{\gamma}\Sigma^{-1}\mu \tag{1.2}$$

However, in practice, the optimal weights vector w^* is not observable because the mean and the covariance are unknown. Hence, to implement a mean-variance strategy, the portfolio weights are usually *estimated* by the plug-in of the sample estimates

$\hat{\mu}$ and $\hat{\Sigma}$. This is the source of the estimation error. A standard approach of measuring the estimation error is to compute the loss in utility the investors face because of estimation. An intuitive measure of this economic loss is the following difference in certainty equivalents:

$$E[L(w^*, \hat{w})] = U(w^*) - E[U(\hat{w})], \tag{1.3}$$

where both $U(w^*)$ and $U(\hat{w})$ are defined as (1.1). The (1.3) can be interpreted as an opportunity cost of using estimated quantities rather than the true unobserved ones. While for the mean, the sample counterpart $\hat{\mu} = T^{-1}\sum X_t$ is usually considered, for the covariance matrix Σ we study the estimation error obtained with different estimators. The standard alternatives that we use as a benchmark are:

- Maximum Likelihood (ML) Estimator: $\hat{\Sigma}_{ML} = T^{-1}D'D$,
- Sample Covariance (SC) Estimator: $\hat{\Sigma}_{SC} = (T-1)^{-1}D'D$,
- Unbiased Precision Matrix (PM) Estimator: $\hat{\Sigma}_{PM} = (T-N-2)^{-1}D'D$.

where $D = \sum_{t=1}^{T}(X_t - \hat{\mu})$, Obviously, with different estimators we get different allocations, namely $\hat{w}_{ML}$, $\hat{w}_{SC}$ and $\hat{w}_{PM}$ and, as we are going to show, different estimation errors.

1.1.1 Markowitz with standard sample estimates

The most simple plug-in strategy is the one based on maximum likelihood estimation. As briefly mentioned above, it involves the estimator $\hat{\Sigma}_{ML}$ for covariance matrix assuming joint i.i.d Gaussian distribution of asset returns. The maximum likelihood estimator is biased for the actual covariance matrix Σ as well as its asset allocation since:

$$E[\hat{w}_{ML}] = \frac{T}{(T-N-2)}w^* \neq w^*.$$

In particular, being $T/(T-N-2) > 1$, this asset allocation is more aggressive than the optimal one since the investor does not recognize estimation error and risky

assets are in a certain sense considered less risky. Then Kan and Zhou (2007) analytically derived closed form of the estimation error for the maximum likelihood strategy by considering the following expected loss function:

$$E\left[L(w^*, \hat{w})\right] = U(w^*) - E\left[U(\hat{w})\right] = \\ w^{*'}\mu - \frac{\gamma}{2} w^{*'}\Sigma w^* - E\left[\hat{w}'\mu - \frac{\gamma}{2}\hat{w}'\Sigma\hat{w}\right].$$

Since we know that $w^* = \gamma^{-1}\Sigma^{-1}\mu$ and that $\hat{w} = \gamma^{-1}\hat{\Sigma}_{ML}^{-1}\hat{\mu}$, we have to replace these quantities inside (1.3). As showed by Kan and Zhou (2007), estimation error with a maximum likelihood strategy is:

$$E[L(w^*, \hat{w}_{ML})] = (1 - k_1)\frac{\theta^2}{2\gamma} + \frac{1}{2\gamma}\frac{T(T-2)N}{(T-N-1)(T-N-2)(T-N-4)}.$$

Given:

$$k_1 = \frac{T}{(T-N-2)}\left[2 - \frac{T(T-2)}{(T-N-1)(T-N-4)}\right],$$

As noted by Frost and Savarino (1986b), the estimation error is inversely related to the precision of the estimates. The precision is calculated as the difference between the number of observations T and the number of assets N. In other words, increasing N or decreasing T leads to a higher estimation error. An explanation could be the following. In the extreme case where $T \to \infty$, the actual parameters are learned, so the loss is zero. On the other hand, the greater the number of assets N, the greater the number of elements of $\hat{\mu}$ that have to be estimated, the more the errors and the greater is the loss.

Therefore, we could reduce estimation error by selecting an "appropriate" sample. Nevertheless, as it will be shown later on, employing a different estimator for the covariance matrix has an essential role in reducing parameter uncertainty.

The Maximum likelihood is a biased estimator for the covariance matrix. A more

efficient alternative is to invest in an allocation based on sample covariance estimator $\hat{\Sigma}_{SC}$ which is unbiased for the actual covariance matrix Σ. It has the following relationship with the maximum likelihood estimator $\hat{\Sigma}_{ML}$:

$$\hat{\Sigma}_{SC} = \frac{T}{T-1}\hat{\Sigma}_{ML}$$

Nevertheless, also this strategy $\hat{w}_{SC}$ is biased since the inverse of sample the covariance is a biased estimator for the true inverse Σ^{-1}:

$$E[\hat{w}_{SC}] = \frac{(T-1)}{(T-N-2)}w^*.$$

Also, this strategy is more aggressive than the optimal w^* but is a bit more conservative than the maximum likelihood allocation $\hat{w}_{ML}$. It is possible to show that the estimation error associated with sample covariance strategy is:

$$E[L(w^*, \hat{w}_{SC})] = (1-k_2)\frac{\theta^2}{2\gamma} + \frac{1}{2\gamma}\frac{(T-1)^2(T-2)N}{T(T-N-1)(T-N-2)(T-N-4)},$$

given k_2:

$$k_2 = \frac{T-1}{(T-N-2)}\left[2 - \frac{(T-1)(T-2)}{(T-N-1)(T-N-4)}\right].$$

As proved by Kan and Zhou (2007), $E[L(\hat{w}_{SC}, \hat{w}_{ML})] > 0$. Therefore, estimation error could be reduced by estimating differently the covariance matrix.
Starting from the fact that the inverse of the sample covariance estimator is biased, an alternative estimator is the following $\hat{\Sigma}_{PM}$:

$$\hat{\Sigma}_{PM} = \frac{T}{T-N-2}\hat{\Sigma}_{ML},$$

This estimator is unbiased for the actual covariance matrix inverse (i.e. the *precision matrix*, PM), such that it leads to an unbiased asset allocation. In other words, the

investor who uses $\hat{w}_{PM}$ will, on average, invest the same amount of money in the risky asset as would be in the optimal portfolio w^*. The loss function associated with the scaled strategy $\tilde{w}$ is:

$$E[L(w^*, \hat{w}_{PM})] = (1 - k_3)\frac{\theta^2}{2\gamma} + \frac{1}{2\gamma}\frac{(T-N-2)(T-2)N}{T(T-N-1)(T-N-4)},$$

where k_3 is:

$$k_3 = 2 - \frac{(T-N-2)(T-2)}{(T-N-1)(T-N-4)}.$$

It is possible to verify that $E[U(\hat{w}_{PM})] > E[U(\hat{w}_{SC})]$, so we can further reduce estimation error by employing this plug-in strategy. One can think that the overperformance of this approach is due to unbiasedness, but it is not true. Indeed, Kan and Zhou (2007) consider a general asset allocation as follows:

$$\hat{w}_c = \frac{c}{\gamma}\hat{\Sigma}_{ML}^{-1}\hat{\mu}$$

where c is a scaling constant. It is easy to recognize that if $c = 1$ we get $\hat{w}_{ML}$, while with $c = \frac{T-1}{T}$ we obtain $\hat{w}_{SC}$ and for $c = \frac{T-N-2}{T}$ we end up with $\hat{w}_{PM}$. Kan and Zhou (2007) demonstrated the existence of an optimal scalar c^* that maximize expected utility:

$$c^* = \left[\frac{(T-N-1)(T-N-4)}{T(T-2)}\right]\left(\frac{\theta^2}{\theta^2 + \frac{N}{T}}\right)$$

These allocation strategies can be viewed as a plug-in approach that estimates covariance matrix with an estimator of the form $\hat{\Sigma}_c = \hat{\Sigma}/c$. Moreover, although it is optimal in terms of utility maximization, it leads to a biased asset allocation since $E[\hat{w}_c] \neq w^*$. Not only $\hat{w}_{c^*}$ rule is biased respect to w^* but also the associated estimator $\hat{\Sigma}_{c^*}$ is biased.

However, this *optimal scaling* strategy is not directly observable because of θ^2. A way to overcome this issue is to consider a quasi-optimal, always biased, alternative where the scaling $c = c_3 = (T-N-1)(T-N-4)/T(T-2)$ represents the first term in the square bracket. It could be easily obtained supposing that $\theta^2 \to \infty$.

In conclusion, the discussion made so far highlight some relevant points. First, it is possible to reduce the estimation error by estimating the covariance matrix differently by taking advantage of some estimators that reduce investors' opportunity costs. Secondly, such estimators do not need to be unbiased concerning the actual unobserved covariance matrix or lead unbiased asset allocation rule.
These two facts are essential since shrinkage estimators usually miss these two characteristics, as we will show in the next section. They do not ensure unbiasedness but are able to reduce estimation error through diversification further.

1.1.2 Precision matrix shrinkage estimator

Stein (1956) noted that, for $N > 2$ independent normal random variables, the sample mean estimator $\hat{\mu}$ is dominated in terms of mean-squared error by a convex combination of the sample means and a common constant μ_0:

$$\hat{\mu}_s = \alpha\hat{\mu} + (1-\alpha)\mu_0, \tag{1.4}$$

The estimator (1.4) is called *shrinkage* with $0 < \alpha < 1$ the *shrinkage intensity*, that represents the optimal trade-off between bias and variance. Jorion (1986) and Jorion (1991) propose the use of shrinkage in finance for better estimating the expected returns vector. Shrinkage can also be applied to covariance matrix estimation. In the portfolio selection context, Ledoit and Wolf (2003) and Ledoit and Wolf (2004a) propose a covariance matrix estimator equal to a convex combinations of the usual sample covariance matrix $\bar{\Sigma}$ and a shrinkage target $\hat{\Omega}$:

$$\hat{\Sigma}_s = \alpha\hat{\Sigma}_{SC} + (1-\alpha)\hat{\Omega} \tag{1.5}$$

It is straightforward to verify that shrinkage estimators as (1.5) are unbiased only if both $\hat{\Sigma}_{SC}$ and $\hat{\Omega}$ are both unbiased.
The selection of the shrinkage target $\hat{\Omega}$ as well as the optimal shrinkage intensity α^* is crucial for constructing a shrinkage estimator. What Ledoit and Wolf (2003) proposed is an operation called "shrinkage towards the market". It consists in shrinking the sample covariance estimator towards the covariance matrix of a single factor model of Sharpe (1963) $\hat{\Sigma}_F$. Ledoit and Wolf (2003) justify this choice by assessing

that, even if there is no consensus about which and how many factors to consider, the market return is the most intuitive and accepted factor for explaining the cross-section of asset returns. Moreover, they derived optimal shrinkage intensity by minimizing the following Fornebious norm:

$$\min_{\alpha} E[||\alpha\hat{\Sigma}_{SC} + (1-\alpha)\hat{\Sigma}_F - \Sigma||^2] \tag{1.6}$$

The resulting α^* is not observable, since it involves several unknown quantities. Therefore Ledoit and Wolf (2003) derived a consistent estimator for $\hat{\alpha}^*$. By plugging-in the Ledoit and Wolf (2003) linear shrinkage estimator in (1.2) we obtain the allocation $\hat{w}_s$.

Ideally, the shrinkage estimators allow the diversification of estimation risk by the averaging approach. As we have seen, shrinkage is a trade-off between bias and variance. Shrinking towards a constant (Stein, 1956) means shrinking towards a highly biased target with zero variance. Ledoit and Wolf (2003) derived the optimal shrinkage intensity claiming that α^* depends on the correlation between the estimation error in the sample covariance and the estimation error in the target. According to the authors, if the two are positively (negatively) correlated, the benefit of combining the information they contain is smaller (larger).

As in the case of the other approaches mentioned above for covariance estimation, we were interested in computing the expected loss using this shrinkage estimator. Unfortunately, this is not an easy task because of the difficulties in computing the inverse of two matrices weighted sum.

Moreover, an additional limitation of the Ledoit and Wolf (2003) shrinkage approach lies in the way in which optimal shrinkage intensity is determined. Indeed, since it is based on statistical arguments, the resulting α^* is not necessarily consistent with the portfolio selection problem. Therefore in what follows, we overcome this limitation by proposing a utility-maximizer shrinkage estimator.

Estimation framework

In order to build a preferences-maximizer shrinkage estimator, first, we need to accurately specify the value of the expected utility when the covariance matrix is estimated with the shrinkage approach. As we have already mentioned, it is impossible

with a classical shrinkage approach because of difficulties in inverting the sum of matrices. To overcome the problem, following the idea of several authors in literature (e.g. see Efron and Morris, 1973; Haff, 1977; Haff, 1979; Dey, 1987; Kubokawa and Srivastava, 2008; Kourtis, Dotsis, and Markellos, 2012; Sun, Ma, and Liu, 2018), we propose to shrink what matters for portfolio selection: the inverse of covariance matrix Σ^{-1}, also called the *precision matrix*. More specifically, we propose the following shrinkage estimator:

$$\hat{\Sigma}_s^{-1} = \alpha\hat{\Omega}_1^{-1} + (1-\alpha)\hat{\Omega}_2^{-1} \tag{1.7}$$

where $\hat{\Omega}_1^{-1}$ and $\hat{\Omega}_2^{-1}$ are the inverse of two symmetric and positive definite $N \times N$ covariance matrices. We define the estimator (1.7) as *precision shrinkage estimator*.
A natural choice for $\hat{\Omega}_1^{-1}$ is $\hat{\Sigma}_{PM}^{-1}$, that is the inverse of the Unbiased Precision Matrix covariance. This choice, if fairly motivated by its unbiasedness, respect the true covariance matrix inverse (Marx and Hocking, 1977) and because of its superior efficiency in terms of economic loss compared to usual maximum likelihood or sample covariance estimators. Similar arguments have been used by Tu and Zhou (2011) for combining portfolio rules based on this estimator instead of maximum likelihood.
Before analysing the estimation error of the proposed estimator, we want to highlight that with the plug-in of the precision shrinkage estimator we obtain a three-fund rule.

Proposition 1. The portfolio rule constructed trough the precision matrix shrinkage $\hat{\Sigma}_s^{-1}$ could be interpreted as two fund rule, whatever $\hat{\Omega}_1$ and $\hat{\Omega}_2$ are.

Proof. To see that, we can define the proposed $\hat{w}_s$ strategy as:

$$\begin{aligned}
\hat{w}_s = \frac{1}{\gamma}\hat{\Sigma}_s^{-1}\hat{\mu} = \\
= \frac{1}{\gamma}(\alpha\hat{\Omega}_1^{-1} + (1-\alpha)\hat{\Omega}_2^{-1})\hat{\mu} = \\
= \frac{\alpha}{\gamma}\hat{\Omega}_1^{-1}\hat{\mu} + \frac{(1-\alpha)}{\gamma}\hat{\Omega}_2^{-1}\hat{\mu} = \\
\hat{w}_s = \alpha\hat{w}_1 + (1-\alpha)\hat{w}_2.
\end{aligned} \tag{1.8}$$

Therefore, by estimating covariance matrix with this shrinkage approach, we invest in both portfolios $\hat{w}_1$ and $\hat{w}_2$. Hence, the relationship (1.8) highlights that our asset allocation strategy can be interpreted as a combination of portfolio rules, where the combination coefficient α is equal to the optimal shrinkage intensity α^*.
Once the prior matrix $\hat{\Omega}_1$ has been defined, the shrinkage target $\hat{\Omega}_2$ has to be carried out. The main idea is to find a target which estimation error is weakly correlated to the one of the prior $\hat{\Omega}_1^{-1}$. A natural candidate is the Identity matrix I as proposed by several authors (e.g. Haff, 1979; Ledoit and Wolf, 2004a; Kourtis, Dotsis, and Markellos, 2012):

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

There is a simple explanation that justifies the usage of this target. Indeed, even if highly biased, I does not contain estimation error at all. Therefore the estimation error in $\hat{\Sigma}_{PM}^{-1}$ is orthogonal to that of I. Moreover, following the idea highlighted by Proposition 1, another interesting target matrix could be the implied covariance of the equally weighted portfolio:

$$\Sigma_{EW} = \begin{pmatrix} N\hat{\mu}_1 & 0 & \cdots & 0 \\ 0 & N\hat{\mu}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & N\hat{\mu}_N \end{pmatrix} \tag{1.9}$$

That represents the implicit covariance structure assumed by a mean-variance investor who decides to invest across all the N assets equally. In other words, by plug-in of the estimator (1.9) within mean-variance rule (1.2), we obtain the equally weighted portfolio if $\gamma = 1$. For $\gamma \neq 1$, we end up with a two-fund rule where the agent invests a proportion of his wealth in an equally weighted portfolio and the rest in the riskless asset. Interestingly, with the following shrinkage estimator:

$$\hat{\Sigma}_s^{-1} = \alpha\hat{\Sigma}_{PM}^{-1} + (1-\alpha)\Sigma_{EW}^{-1}$$

if $\gamma = 1$ we end up with Tu and Zhou (2011) combination rule between Markwoitz mean-variance portfolio and equally weighted one:

$$\tilde{w}_s = \alpha \hat{w}_{PM} + (1 - \alpha) w_{EW}$$

On the other hand, another crucial aspect for shrinkage estimation is how the optimal shrinkage intensity α^* is derived. In what follows, we find α^* to maximize the investor's expected utility.

Optimal shrinkage intensity

In what follows, we propose to find optimal shrinkage intensity that maximize investor's preferences. In other words, given a general precision shrinkage estimator $\hat{\Sigma}_s^{-1}$, the α^* is determined by:

$$\max_{\alpha} E[U(\hat{w}_s)]. \tag{1.10}$$

Expanding the expression:

$$E[U(\hat{w}_s)] = E\left[\frac{1}{\gamma}\hat{\mu}'\hat{\Sigma}_s^{-1}\mu\right] - \frac{\gamma}{2}E\left[\left(\frac{1}{\gamma}\hat{\Sigma}_s^{-1}\hat{\mu}\right)'\Sigma\left(\frac{1}{\gamma}\hat{\Sigma}_s^{-1}\hat{\mu}\right)\right] =$$
$$E\left[\frac{1}{\gamma}\left(\alpha\hat{\Omega}_1^{-1} + (1-\alpha)\hat{\Omega}_2^{-1}\right)'\hat{\mu}\right] - \frac{\gamma}{2}E\left[\left(\frac{1}{\gamma}(\alpha\hat{\Omega}_1^{-1} + (1-\alpha)\hat{\Omega}_2^{-1})\hat{\mu}\right)'\Sigma\left(\frac{1}{\gamma}(\alpha\hat{\Omega}_1^{-1} + (1-\alpha)\hat{\Omega}_2^{-1})\hat{\mu}\right)\right].$$

Theorem 1. The optimal solution to the problem (1.10) is:

$$\alpha^* = \frac{E[a'\mu] - E[b'\mu] + E[b'\Sigma b] - E[a'\Sigma b]}{E[a'\Sigma a] + E[b'\Sigma b] - 2E[a'\Sigma b]}.$$

Therefore, the *oracle* estimator of precision matrix from a mean-variance perspective is:

$$\hat{\Sigma}_s^{-1} = \frac{E[a'\mu] - E[b'\mu] + E[b'\Sigma b] - E[a'\Sigma b]}{E[a'\Sigma a] + E[b'\Sigma b] - 2E[a'\Sigma b]}\hat{\Omega}_s^{-1} + \frac{E[b'\mu] - E[a'\mu] + E[a'\Sigma a] - E[a'\Sigma b]}{E[a'\Sigma a] + E[b'\Sigma b] - 2E[a'\Sigma b]}\hat{\Omega}_2^{-1}$$

Proof. We start treating the two expressions separately for convenience. The first part of the expression:

$$\begin{aligned} E[\tilde{w}_s']\mu &= E\left[\frac{1}{\gamma}(\alpha\hat{\Omega}_1^{-1} + (1-\alpha)\hat{\Omega}_2^{-1})\hat{\mu}\right]'\mu = \\ &= E\left[\frac{\alpha}{\gamma}\hat{\mu}'\hat{\Omega}_1^{-1}\mu + \frac{(1-\alpha)}{\gamma}\hat{\mu}'\hat{\Omega}_2^{-1}\mu\right] = \\ &= \frac{\alpha}{\gamma}E[\hat{\mu}'\hat{\Omega}_1^{-1}\mu] + \frac{(1-\alpha)}{\gamma}E[\hat{\mu}'\hat{\Omega}_2^{-1}\mu]. \end{aligned}$$

where the first term is the squared Sharpe ratio associated to the first portfolio rule while the second is the squared Sharpe ratio of a portfolio with $\Sigma^{-1} = \hat{\Omega}_2^{-1}$. Then we can evaluate the second part of the utility function:

$$\begin{aligned} &E[\tilde{w}_s'\Sigma\tilde{w}_s] = E\left[\left(\frac{\alpha}{\gamma}\hat{\Omega}_1^{-1}\hat{\mu} + \frac{(1-\alpha)}{\gamma}\hat{\Omega}_2^{-1}\hat{\mu}\right)'\Sigma\left(\frac{\alpha}{\gamma}\hat{\Omega}_1^{-1}\hat{\mu} + \frac{(1-\alpha)}{\gamma}\hat{\Omega}_2^{-1}\hat{\mu}\right)\right] = \\ &= E\left[\left(\frac{\alpha}{\gamma}\hat{\mu}'\hat{\Omega}_1^{-1}\Sigma + \frac{(1-\alpha)}{\gamma}\hat{\mu}'\hat{\Omega}_2^{-1}\Sigma\right)\left(\frac{\alpha}{\gamma}\hat{\Omega}_1^{-1}\hat{\mu} + \frac{(1-\alpha)}{\gamma}\hat{\Omega}_2^{-1}\hat{\mu}\right)\right] = \\ &= E\left[\frac{\alpha^2}{\gamma^2}\hat{\mu}'\hat{\Omega}_1^{-1}\Sigma\hat{\Omega}_1^{-1}\hat{\mu} + \frac{(1-\alpha)^2}{\gamma}\hat{\mu}'\hat{\Omega}_2^{-1}\Sigma\hat{\Omega}_2^{-1}\hat{\mu} + \frac{2\alpha(1-\alpha)}{\gamma^2}\hat{\mu}'\hat{\Omega}_1^{-1}\Sigma\hat{\Omega}_2^{-1}\hat{\mu}\right]. \end{aligned}$$

Considering the overall equation:

$$\begin{aligned} &E[U(\hat{w}_s)] = E[\hat{w}_s']\mu - \frac{\gamma}{2}E[\tilde{w}_s'\Sigma\tilde{w}_s] = \\ &= \frac{\alpha}{\gamma}E[\hat{\mu}'\hat{\Omega}_1^{-1}\mu] + \frac{(1-\alpha)}{\gamma}E[\hat{\mu}'\hat{\Omega}_2^{-1}\mu] - \frac{\gamma}{2}E\left[\frac{\alpha^2}{\gamma^2}\hat{\mu}'\hat{\Omega}_1^{-1}\Sigma\hat{\Omega}_1^{-1}\hat{\mu} + \frac{(1-\alpha)^2}{\gamma}\hat{\mu}'\hat{\Omega}_2^{-1}\Sigma\hat{\Omega}_2^{-1}\hat{\mu} + \right. \\ &\left. + \frac{2\alpha(1-\alpha)}{\gamma^2}\hat{\mu}'\hat{\Omega}_1^{-1}\Sigma\hat{\Omega}_2^{-1}\hat{\mu}\right]. \end{aligned}$$

Defining $a = \hat{\Omega}_1^{-1}\hat{\mu}$ and $b = \hat{\Omega}_2^{-1}\hat{\mu}$:

$$
\begin{aligned}
&E[U(\hat{w}_s)] = E[\hat{w}_s']\mu - \frac{\gamma}{2}E[\tilde{w}_s'\Sigma\tilde{w}_s] = \\
&= \frac{\alpha}{\gamma}E[a'\mu] + \frac{(1-\alpha)}{\gamma}E[b'\mu] - \frac{\gamma}{2}E\left[\frac{\alpha^2}{\gamma^2}a'\Sigma a + \frac{(1-\alpha)^2}{\gamma}b'\Sigma b + \frac{2\alpha(1-\alpha)}{\gamma^2}a'\Sigma b\right] = \\
&= \frac{\alpha}{\gamma}E[a'\mu] + \frac{(1-\alpha)}{\gamma}E[b'\mu] - \frac{\alpha^2}{2\gamma}E[a'\Sigma a] - \frac{(1-\alpha)^2}{2\gamma}E[b'\Sigma b] - \frac{\alpha(1-\alpha)}{\gamma}E[a'\Sigma b].
\end{aligned}
\tag{1.11}
$$

The expression (1.11) is the expected utility associated to a generic precision shrinkage (1.7), whatever the objective and the target matrices are. In the end, we obtain the optimal shrinkage intensity α^* by deriving the (1.11) with respect to α:

$$
\begin{aligned}
&\frac{\partial E[U(\hat{w}_s)]}{\partial\alpha} = 0 \\
&\frac{1}{\gamma}E[a'\mu] - \frac{1}{\gamma}E[b'\mu] - \frac{\alpha}{\gamma}E[a'\Sigma a] - \frac{(\alpha-1)}{\gamma}E[b'\Sigma b] - \frac{(1-2\alpha)}{\gamma}E[a'\Sigma b] = 0 \\
&\frac{1}{\gamma}\left[E[a'\mu] - E[b'\mu] + E[b'\Sigma b] - E[a'\Sigma b]\right] = \frac{\alpha}{\gamma}\left[E[a'\Sigma a] + E[b'\Sigma b] - 2E[a'\Sigma b]\right] \\
&\alpha^* = \frac{E[a'\mu] - E[b'\mu] + E[b'\Sigma b] - E[a'\Sigma b]}{E[a'\Sigma a] + E[b'\Sigma b] - 2E[a'\Sigma b]}.
\end{aligned}
\tag{1.12}
$$

Hence $(1-\alpha^*)$ is equal to:

$$
\begin{aligned}
&(1-\alpha^*) = 1 - \left(\frac{E[a'\mu] - E[b'\mu] + E[b'\Sigma b] - E[a'\Sigma b]}{E[a'\Sigma a] + E[b'\Sigma b] - 2E[a'\Sigma b]}\right) = \\
&= \frac{E[a'\Sigma a] + E[b'\Sigma b] - 2E[a'\Sigma b] - E[a'\mu] + E[b'\mu] - E[b'\Sigma b] + E[a'\Sigma b]}{E[a'\Sigma a] + E[b'\Sigma b] - 2E[a'\Sigma b]} = \\
&(1-\alpha^*) = \frac{E[b'\mu] - E[a'\mu] + E[a'\Sigma a] - E[a'\Sigma b]}{E[a'\Sigma a] + E[b'\Sigma b] - 2E[a'\Sigma b]}.
\end{aligned}
$$

Interestingly, the optimal shrinkage intensity does not depend on the investor risk aversion coefficient γ. What matters, instead, is only the estimation error associated with the involved estimators and their correlation. More in details, α^* depends by

the squared Sharpe ratio of the proposed strategies ($E[a'\mu]$ and $E[b'\mu]$), the estimation error of asset allocation with $\hat{\Omega}_1^{-1}$ as plug-in ($E[a'\Sigma a]$), the estimation error of asset allocation with $\hat{\Omega}_2^{-1}$ ($E[b'\Sigma b]$) and the quantity $E[a'\Sigma b]$. *Ceteris paribus*, we assign a higher weight to the strategy with a higher squared Sharpe ratio.
In the end, as the shrinkage estimator of Ledoit and Wolf (2003), Ledoit and Wolf (2004b), and Ledoit and Wolf (2004a), the *precision shrinkage* belong to the class of rotation-equivariant estimators.

Proposition 2.
The proposed shrinkage estimator is equivalent of shrinking eigenvalues' reciprocal of prior matrix $\hat{\Omega}_1$ towards the target's ($\hat{\Omega}_2$) one such that:

$$\lambda_s = \alpha^* \lambda_1^{-1} + (1 - \alpha^*)\lambda_2^{-1}$$

with λ_1 be the prior matrix eigenvalues and λ_2 the one of target matrix. Clearly, we can write our new estimator as $\hat{\Sigma}_s = U\Lambda_s U'$ with $\Lambda_s = Diag(\lambda_s)$. The proof is provided in the Appendix A.

What rotation-equivariant estimation means is that rotating the original variables results in the same rotation being applied to the covariance estimator. Rotation equivariance is appropriate in the general case where we do not have a priori information about the orientation of the eigenvectors of the actual covariance matrix (Ledoit and Wolf (2012)).
The fact that we keep the sample eigenvectors does not mean that we assume they are close to the population one. Differently, we do not know how to improve upon them. If we believed that the sample eigenvectors were close to the population one, then the optimal covariance matrix estimator would have eigenvalues very close to the population eigenvalues. This is not necessarily optimal from the portfolio selection point of view. What we do, instead, is to find the optimal way of shrinking eigenvalues such that it maximizes investor's preferences.

Shrinkage rules

Here we develop different shrinkage rules with their optimal shrinkage intensities. All the proofs are provided in Appendix A. Suppose, first, that $\hat{\Omega}_1 = \hat{\Sigma}_{PM}$. This choice is natural since it is common to consider an unbiased prior estimator $\hat{\Omega}_1$. We shrink the Unbiased Precision estimator towards three different targets: the Identity matrix I, the equally weighted portfolio Σ_{EW} and the optimal scaling c^*. In the case of shrinkage, $\hat{\Sigma}_{PM}^{-1}$ towards the Identity I, the optimal shrinkage intensity is:

$$\alpha^*_{PM,I} = \frac{\theta^2 - 2\lambda^2 + Q}{c_1 \frac{N}{T} + c_1\theta^2 + Q - 2\lambda^2} \tag{1.13}$$

with $\lambda = tr(\Sigma) + \mu'\mu$ and $Q = tr(\Sigma\Sigma) + \mu'\Sigma\mu$. Both the results are showed in the Appendix A and are based on the trace-expectation relationship. It is evident that the (1.13) is not observable and, therefore, we need an estimate for it. In order to estimate θ^2 we take advantage of the unbiased estimator of Kan and Zhou (2007):

$$\hat{\theta}^2_u = \frac{(T - N - 2)\hat{\theta}^2 - N}{T} \tag{1.14}$$

where $\hat{\theta} = \hat{\mu}'\hat{\Sigma}^{-1}\hat{\mu}$ is the, biased, sample counterpart of θ^2. Then, an unbiased estimator for Q is $(T-1)/(T-N-2)\hat{Q}$, where $\hat{Q} = \hat{\mu}'\hat{\Sigma}_{SC}\hat{\mu}$. The corrective factor is necessary to ensure unbiasedeness of the estimator (see Appendix A). Then, for λ^2 its sample counterpart $\hat{\lambda}^2 = \hat{\mu}'\hat{\mu}$ is unbiased since:

$$E[\hat{\lambda}^2] = E[\hat{\mu}' I \hat{\mu}] = \mu'\mu + tr(\Sigma) = \lambda^2.$$

By expectation-trace relationship (Appendix A). Then, we evaluate the case where the shrinkage target is the implied equally weighted covariance matrix (1.9). It is equal to:

$$\alpha^*_{PM,EW} = \frac{\theta^2 - 2w_e'\mu + w_e'\Sigma w_e}{c_1 \frac{N}{T} + c_1\theta^2 + w_e'\Sigma w_e - 2w_e'\mu}. \tag{1.15}$$

with $w_e = (1/N, \ldots, 1/N)$ a constant vector with equal weights and $c_1 = (T -$

$2)(T-N-2)/(T-N-1)(T-N-4)$. In the end, we also define a shrinkage operation of the Unbiased Precision Matrix estimator towards another scaled estimator that, even if biased, is truly a utility maximizer. Since the estimator c^* is not observable because of θ^2, we consider the quasi-optimal version based on the assumption $\theta^2 \to \infty$. In other words, we study the performance of the following shrinkage estimator:

$$\hat{\Sigma}_s^{-1} = \alpha \hat{\Sigma}_{PM}^{-1} + (1-\alpha)\hat{\Sigma}_{c_3}^{-1}$$

Optimal shrinkage intensity is given by:

$$\alpha^*_{PM,c_3} = \frac{\left(\frac{1-c_1}{c_1}\right)\frac{N}{T}}{\left(\frac{(1-c_1)^2}{c_1}\right)\frac{N}{T} + \left(\frac{(1-c_1)^2}{c_1}\right)\theta^2}. \tag{1.16}$$

Suppose we believe in doing better by shrinking the Unbiased Precision Matrix estimator towards the optimal Σ_{c^*} without assuming $\theta^2 \to \infty$. In this case, the optimal shrinkage intensity is $\alpha^* = 0$ because, from the utility maximization point of view, c^* is already the best one.

Despite it is possible to prove this result analytically, for convenience we include the proof in the Appendix A. For the same reason, the same result applies if we shrink the two optimal c^* estimators, namely c_3 (c^* with $\theta^2 \to \infty$) and c^* (with estimated θ^2). Suppose, instead, $\hat{\Omega}_1 = \hat{\Sigma}_{c^*}$ and $\hat{\Omega}_2 = \hat{\Sigma}_{c_3}$. In this case $\alpha^* = 1$ (see Appendix A). The main conclusion is that, if we shrink any plug-in covariance (e.g. maximum likelihood, sample covariance) towards the $\hat{\Sigma}_{c^*}$ estimator, the resulting optimal α^* is always zero or one, depending on if the estimator with optimal c^* is considered as prior matrix $\hat{\Omega}_1$ or as target $\hat{\Omega}_2$.

Suppose, now, to consider $\hat{\Omega}_1 = \hat{\Sigma}_{c_3}$. We study the shrinkage of c_3 towards both Identity and equally weighted. In the case of shrinkage $\hat{\Sigma}_{c_3}^{-1}$ towards Identity optimal shrinkage intensity is equal to:

$$\alpha^*_{c_3,I} = \frac{\frac{1}{c_1}\theta^2 - \left(1+\frac{1}{c_1}\right)\lambda^2 + Q}{\frac{1}{c_1}\left(\theta^2 + \frac{N}{T} - 2\lambda^2\right) + Q}. \tag{1.17}$$

where c_1 is defined as before and we replace estimators for λ and Q. Then, suppose

to shrink the same estimator towards the implied equally weighted covariance:

$$\alpha^*_{c_3,EW} = \frac{\frac{1}{c_1}\theta^2 - \left(1+\frac{1}{c_1}\right) w_e'\mu + w_e'\Sigma w_e}{\frac{1}{c_1}\frac{N}{T} + \frac{1}{c_1}\theta^2 + w_e'\Sigma w_e - 2w_e'\mu}. \tag{1.18}$$

The last shrinkage rule that we develop is the case of $\hat{\Omega}_1 = \hat{\Sigma}_{c^*}$. As we proved in Appendix A, any sample estimator we use lead to an optimal shrinkage intensity equal to either zero or one. Therefore, we directly shrink $\hat{\Sigma}_{c^*}$ with the Identity I and the implied covariance Σ_{EW}. For the shrinkage towards the Identity, the optimal shrinkage intensity is:

$$\alpha^*_{c^*,I} = \frac{\frac{1}{c_1}\left(\frac{\theta^4}{\theta^2+N/T}\right) - \lambda^2 + Q - \frac{1}{c_1}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)\lambda^2}{\frac{1}{c_1}\left(\frac{N}{T}+\theta^2\right) + Q - 2\frac{1}{c_1}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)\lambda^2}. \tag{1.19}$$

The last exercise is to shrink c^* towards the equally weighted covariance matrix Σ_{EW}. In this case we get:

$$\alpha^*_{c^*,EW} = \frac{\frac{1}{c_1}\left(\frac{\theta^4}{\theta^2+N/T}\right) - w_e'\mu + w_e'\Sigma w_e - \frac{1}{c_1}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right) w_e'\mu}{\frac{1}{c_1}\left(\frac{N}{T}+\theta^2\right) + w_e'\Sigma w_e - 2\frac{1}{c_1}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right) w_e'\mu}. \tag{1.20}$$

All the presented strategies based on shrinkage are 3-fund rules, where the agent invests in riskless assets and two different risky portfolios with the same goal (mean-variance optimization).

1.1.3 A simulation study

In what follows, we evaluate the expected out-of-sample performance of the proposed *precision shrinkage estimator* through a simulation study as in Kan and Zhou (2007). In doing so, we compare its performance with the alternative plug-in strategies shown in the paper.

More in detail, we assume a mean-variance investor with different risk aversion coefficients, $\gamma = 1$ and $\gamma = 3$, where the simulation parameters are calibrated from real data under the assumption of multivariate normality for stock returns. At this aim, we develop different scenarios. First, we assume there are $N = 5$ risky assets of different lengths $T = 60, 120, 240, 480, 960$ with mean and covariances calibrated based on real data. For calibration, we develop the sample estimates of the Fama-French

five industry portfolio monthly excess returns[1] from July 1926 to October 2019. The expected out-of-sample performances are determined as the average of $M = 10000$ replications. Results are showed in Tab. 1.1.

TABLE 1.1: Economic loss (3) with $N = 5$ risky assets

	T=60	T=120	T=240	T=480	T=960
Panel A: $\gamma = 1$					
$1/N$	12.45284	12.45284	12.45284	12.45284	12.45284
I	473.9204	394.6301	352.6923	334.92783	325.8351
EW	12.45284	12.45284	12.45284	12.45284	12.45284
Maximum Likelihood	0.053788	0.0242263	0.0114308	0.0056082	0.0027802
Sample Covariance	0.051930	0.0238023	0.0113317	0.0055838	0.0027742
Unbiased Precision Covariance	0.041751	0.0214083	0.0107638	0.0054431	0.0027393
Optimal $c^* = c_3$ (with $\theta^2 \to \infty$)	0.033863	0.0193509	0.0102518	0.0053132	0.0027068
Optimal c^* (with $\hat{\theta}_u^2$)	0.021993	0.0140119	0.0084696	0.0048443	0.0026099
Unbiased Precision towards I	0.042343	0.0214288	0.0107432	0.0054384	0.0027383
Unbiased Precision towards EW	0.042205	0.0214852	0.0107740	0.0054457	0.0027399
Tu and Zhou (2011)	0.042205	0.0214852	0.0107740	0.0054457	0.0027399
Unbiased Precision towards c_3	0.021993	0.0140119	0.0084696	0.0048443	0.0026099
Optimal $c^* = c_3$ towards I	0.036466	0.0198455	0.0103400	0.0053359	0.0027126
Optimal $c^* = c_3$ towards EW	0.032352	0.0189597	0.0101524	0.0052882	0.0027006
Optimal c^* towards I	0.026670	0.0144855	0.0080667	0.0045324	0.0024660
Optimal c^* towards EW	0.025316	0.0143474	0.0080626	0.0045196	0.0024587
Ledoit and Wolf (2004a)	0.045716	0.0222611	0.0110882	0.0055211	0.0027591
Ledoit and Wolf (2003)	0.063583	0.0265632	0.0121394	0.0057821	0.0028240
	T=60	T=120	T=240	T=480	T=960
Panel B: $\gamma = 3$					
$1/N$	39.26009	39.26009	39.260090	39.26009	39.26009
I	158.50539	131.11738	118.97522	112.542477	108.992236
EW	4.1509496	4.1509496	4.1509496	4.15094960	4.15094960
Maximum Likelihood	0.0177424	0.0079754	0.0038491	0.00186997	0.00091713
Sample Covariance	0.0171269	0.0078367	0.0038155	0.00186179	0.00091514
Unbiased Precision Covariance	0.0137572	0.0070541	0.0036226	0.00181457	0.00090357
Optimal $c^* = c_3$ (with $\theta^2 \to \infty$)	0.0111486	0.0063819	0.0034486	0.00177094	0.00089277
Optimal c^* (with $\hat{\theta}_u^2$)	0.0072304	0.0046407	0.0028367	0.00161301	0.00086097
Unbiased Precision towards I	0.0139577	0.0070592	0.0036167	0.00181315	0.00090327
Unbiased Precision towards EW	0.0138005	0.0070618	0.0036244	0.00181511	0.00090369
Tu and Zhou (2011)	0.0139079	0.0070789	0.0036279	0.00181617	0.00090392
Unbiased Precision towards c_3	0.0072304	0.0046407	0.0028367	0.00161301	0.00086097
Optimal $c^* = c_3$ towards I	0.0119930	0.0065443	0.0034815	0.00177958	0.00089492
Optimal $c^* = c_3$ towards EW	0.0106757	0.0062511	0.0034146	0.00176221	0.00089060
Optimal c^* towards I	0.0087343	0.0047937	0.0027224	0.00151719	0.00081510
Optimal c^* towards EW	0.0082680	0.0047518	0.0027233	0.00151457	0.00081317
Ledoit and Wolf (2004a)	0.0152734	0.0036993	0.0036993	0.00183161	0.00091867
Ledoit and Wolf (2003)	0.0212171	0.0040517	0.0040517	0.00191788	0.00094024

Naive approaches. The first three strategies are the most simple and, at the same time, the most biased. In particular, the first one (that we call $1/N$) suppose an investment based on just one fund, where the investor splits all his wealth equally across the risky assets. The strategy defined as EW is a scaled version of the former $1/N$, where we suppose a mean-variance investor that estimates the covariance matrix with the estimator (1.9). As it appears from the Panel A of Tab. 1.1, the resulting EW strategy is equivalent to $1/N$ when $\gamma = 1$. The strategy that we call I represents a mean-variance rule where the covariance matrix is estimated with the Identity.

[1]We get data from Kenneth French website https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

A first interesting aspect to highlight is that the Identity-based allocation is the worst one among all the alternatives considered in the Tab. 1.1, even asymptotically. The reason is that the bias contained in the Identity matrix is very high[2]. Moreover, the simulation results show that the equally weighted strategy is one of the worst, even if the empirical evidence suggests the opposite. With this respect, we have to stress that the simulated economy is risk factor free as in Kan and Zhou (2007). Surprisingly, Kan and Zhou (2007) did not show performance comparisons of such strategy with the others within their simulations. In this sense, the results presented on the Tab. 1.1 are the first in showing the performances of the equally weighted strategy in terms of the loss function. Nevertheless, in this ideal setting, the evidence of such poor performances of the naive rule can potentially explain why it is so effective in the real world. Indeed, risk factors that affect the fluctuations in stock prices make reality a complex system.

Moreover, some interesting results are also highlighted from Panel B of the Tab. 1.1 that shows the estimation error for an investor with a three times higher risk aversion than in Panel A. First of all, the loss in utility due to the $1/N$ strategy becomes larger, from a value of 12 to a value of 39. Moreover, the EW is different from the $1/N$ one because it represents a two fund rule where a proportion of wealth is also invested in the riskless asset. The resulting naive two fund rule implies a considerable reduction in the utility's loss from 39 to a value of 4. Moreover, the Identity matrix plug-in becomes more attractive than before, especially asymptotically but with a very high estimation error level.

Sample estimators. While the naive rules show inferior performances, the optimal mean-variance strategies show low estimation error. However, as discussed previously, different covariance estimators are associated with different levels of estimation error. What both Panel A and Panel B of the Tab. 1.1 show is that among the classical plug-in approaches, the maximum likelihood-based strategy contains the highest estimation error. In contrast, the Kan and Zhou (2007) optimal estimator $\hat{\Sigma}_{c^*}$ has the lowest one. However, the differences among different estimators vanish asymptotically. These results confirm those of Kan and Zhou (2007).

Precision shrinkage estimators. Now let us consider the performances of the

[2]On one hand, it is reasonable to assume a unitary main diagonal for Σ if we consider it as a correlation matrix. On the other hand, it is tough to believe that all the assets have zero correlation. Indeed, we have to note that the calibrated actual covariance matrix has not zero off-diagonal elements.

shrinkage estimators. In the simulation study, we develop different linear shrinkage estimators, the shrinkage towards the Identity of Ledoit and Wolf (2004a), the shrinkage towards the market of Ledoit and Wolf (2003) and all the proposed *precision shrinkage* estimators that we have developed in the Chapter.

First of all, we have to note that shrinkage of the unbiased precision matrix $\hat{\Sigma}_{PM}$ towards the implied covariance of the equally weighted strategy has the same performance of the Tu and Zhou (2011) combination rule. However, as shown in Panel B of the Tab. 1.1, the performance is higher for the proposed precision shrinkage than the former combination rule since we are genuinely diversifying the estimation error contained in the covariance matrix.

A second relevant aspect is that with the precision shrinkage approaches, we improve the performances concerning the classical plug-in, with overperformances that always become greater with an increasing sample. For example, if we shrink the unbiased precision estimator $\hat{\Sigma}_{PM}$ towards the Identity I we obtain, asymptotically, a loss reduction as well as the shrinkage of $\hat{\Sigma}_{c3}$ towards the implied equally weighted covariance matrix $\hat{\Sigma}_{EW}$.

Another interesting example of this result is the case of the shrinkage of the unbiased precision matrix estimator towards $\hat{\Sigma}_{c3}$. Indeed, in this case, we get the same performance of the optimal estimator of Kan and Zhou (2007) $\hat{\Sigma}_{c^*}$, so with the precision shrinkage, we improve upon both estimators. Moreover, it is not possible to perform better of $\hat{\Sigma}_{c^*}$ if we shrink two estimators that proportional to the maximum likelihood, since it results in an optimal shrinkage intensity equal to 1, as shown in Appendix A.

Moreover, we also found that the greatest improvements in the asset allocation performances are achieved using the information from equally weighting. Indeed, if we shrink the $\hat{\Sigma}_{c_*}$ estimator towards the implied equally weighted covariance (1.9), the overall loss is considerably lower asymptotically. Actually we need $T \geq 240$ to get this result, that is valid for both $\gamma = 1$ and $\gamma = 3$. In particular, the following shrinkage estimator:

$$\hat{\Sigma}_s^{-1} = \alpha \hat{\Sigma}_{c^*}^{-1} + (1 - \alpha)\hat{\Sigma}_{EW}^{-1}$$

Is the best plug-in in terms of the loss function is the optimal shrinkage intensity is

derived as in Theorem 1. However, also with the shrinkage of $\hat{\Sigma}_{c_*}$ towards Identity, we obtain excellent results since it is the second-best strategy.

This result can be explained by the fact that the estimation error of the involved estimators is weakly correlated. Indeed, if we shrink $\hat{\Sigma}_{c^*}$ towards Identity I and/or Σ_{EW} the loss is lower than $\hat{\Sigma}_{c^*}$ alone because the estimation errors associated to both I and Σ_{EW} are orthogonal to the estimation error contained in$\hat{\Sigma}_{c^*}$. Therefore, their combination improves the out of sample performance.

In the end, another interesting comparison is between the proposed precision shrinkage approaches and the linear shrinkage estimators of Ledoit and Wolf (2003) and Ledoit and Wolf (2004a). The comparison in Tab. 1.1 shows that the shrinkage towards the Identity (Ledoit and Wolf (2004a)) always performs better than the shrinkage towards the market (Ledoit and Wolf (2003)). Moreover, the averaging approach reduces the estimation error concerning the sample covariance matrix, since the Ledoit and Wolf (2004a) has a lower estimation error than $\hat{\Sigma}_{SC}$ also for a small sample size. However, it is worst than the unbiased precision matrix $\hat{\Sigma}_{PM}$. On the other hand, despite the empirical evidence of previous papers showing superior performances of this shrinkage approach, the shrinkage towards the market performs poorer than both the sample covariance and the shrinkage towards the Identity.

A possible explanation can be found in the economy under consideration. Indeed, as in Kan and Zhou (2007), the economy of the simulation study is i.i.d. Gaussian without risk factors. If we assume the presence of risk factors (e.g. the market), the difference in the two linear shrinkage approaches results is likely to be overthrown. Nevertheless, it is also interesting to highlight that both the linear shrinkage of Ledoit and Wolf (2003) and Ledoit and Wolf (2004a) have much poorer performances than the proposed precision shrinkage estimators.

The reason for this result lies in how the optimal shrinkage intensity is derived. Indeed, while the linear shrinkage approach of Ledoit and Wolf (2003), Ledoit and Wolf (2003), and Ledoit and Wolf (2004b) is based on statistical arguments, the proposed precision shrinkage derive α^* from maximizing investor's preferences. In other words, despite the Ledoit & Wolf approach being more general and can also be used in other sciences (e.g. biology, chemometrics), the proposed approach based on precision shrinkage is more appropriate for the portfolio selection problem. Therefore, it returns much higher performances than the alternatives. However, we have to mention that the Ledoit and Wolf (2003) and Ledoit and Wolf (2004a) approaches

are taught to be applied in large dimensional settings, where the number of assets N is greater than the observations T, such that the resulting sample estimators are all ill-conditioned. Despite our precision estimators cannot be applied in this setting, in a standard scenario where T is large enough if compared with N (e.g. daily returns' data), it represents the best estimator possible.

These conclusions are valid for the case of $N = 5$ assets. What if the number of assets increases? In the second scenario, the number of assets is six times larger, so we have $N = 30$ assets. The time series, with different lengths $T = 60, 120, 240, 480, 960$, are simulated from a multivariate normal distribution with a mean vector and covariance matrix calibrated from the sample estimates of the monthly excess returns the 30-industry portfolio of Fama-French.

The results are shown in Tab. 1.2. Also, in this case, the ranking of the classical sample estimators is the same of Kan and Zhou (2007), confirming that the maximum likelihood covariance estimator is the one with the highest level of estimation. In contrast, the optimal scaling covariance $\hat{\Sigma}_{c^*}$ is the estimator with the lowest estimation error.

TABLE 1.2: Economic loss (3) with $N = 30$ risky assets

	T=60	T=120	T=240	T=480	T=960
Panel A: $\gamma = 1$					
$1/N$	14.55155	14.55155	14.55155	14.55155	14.55155
I	22115.24	18358.98	15926.51	14883.06	14469.55
EW	14.55155	14.55155	14.55155	14.55155	14.55155
Maximum Likelihood	0.869859	0.205889	0.080465	0.036273	0.0173473
Sample Covariance	0.839873	0.202262	0.079755	0.036113	0.0173091
Unbiased PrecisionCovariance	0.183397	0.108915	0.060083	0.031518	0.0161919
Optimal $c^* = c_3$ (with $\theta^2 \to \infty$)	0.055685	0.063321	0.046595	0.027862	0.0152383
Optimal c^* (with $\hat{\theta}_u^2$)	0.041339	0.040276	0.031052	0.020525	0.0125020
Unbiased Precisiontowards I	0.185097	0.109077	0.060080	0.031514	0.0161912
Unbiased Precision towards EW	0.229433	0.110878	0.060078	0.031483	0.0161825
Tu and Zhou (2011)	0.229433	0.110878	0.060078	0.031483	0.0161825
Unbiased Precision towards c_3	0.041339	0.040276	0.031052	0.027862	0.0125020
Optimal $c^* = c_3$ towards I	0.062772	0.065278	0.047019	0.027965	0.0152679
Optimal $c^* = c_3$ towards EW	0.063049	0.063122	0.046311	0.027765	0.0152097
Optimal c^* towards I	0.047650	0.040944	0.030282	0.019774	0.0120458
Optimal c^* towards EW	0.044228	0.040856	0.030194	0.019664	0.0119894
Ledoit and Wolf (2004a)	0.6679396	0.2064624	0.06149322	0.02113893	0.01270083
Ledoit and Wolf (2003)	0.4201915	0.1560763	0.05248413	0.0199346	0.01200532
	T=60	T=120	T=240	T=480	T=960
Panel B: $\gamma = 3$					
$1/N$	45.53245	45.53245	45.53245	45.53245	45.53245
I	7431.359	5921.821	5313.502	4988.945	4793.515
EW	4.850518	4.850518	4.850518	4.850518	4.850518
Maximum Likelihood	0.291843	0.068181	0.0268983	0.0121100	0.0057678
Sample Covariance	0.281787	0.066984	0.0266604	0.0120565	0.0057551
Unbiased Precision Covariance	0.061572	0.036174	0.0200691	0.0105234	0.0053849
Optimal $c^* = c_3$ (with $\theta^2 \to \infty$)	0.018669	0.021116	0.0155491	0.0093034	0.0050690
Optimal c^* (with $\hat{\theta}_u^2$)	0.013833	0.013476	0.0103367	0.0068515	0.0041652
Unbiased Precision towards I	0.062133	0.036228	0.0200673	0.0105223	0.0053846
Unbiased Precision towards EW	0.076904	0.036675	0.0200670	0.0105124	0.0053801
Tu and Zhou (2011)	0.063793	0.036257	0.0200752	0.0105232	0.0053843
Unbiased Precision towards c_3	0.013833	0.013476	0.0103367	0.0068515	0.0041652
Optimal $c^* = c_3$ towards I	0.021045	0.021712	0.0156861	0.0093385	0.0050773
Optimal $c^* = c_3$ towards EW	0.021201	0.021041	0.0154574	0.0092714	0.0050597
Optimal c^* towards I	0.015963	0.013624	0.0100701	0.0066037	0.0040044
Optimal c^* towards EW	0.014866	0.013496	0.0100453	0.0065670	0.0039832
Ledoit and Wolf (2004a)	0.232274	0.078929	0.0305392	0.01322661	0.006154851
Ledoit and Wolf (2003)	0.149943	0.062082	0.0275302	0.01269284	0.006053937

About the naive strategies $1/N$, EW, and the Identity I plug-in, the same conclusion of Tab. 1.1 can be highlighted. Indeed, with $\gamma = 1$, we have that $1/N$ and EW are equivalent, while with $\gamma = 3$, the proposed two fund rule EW is less affected by the estimation error. Moreover, it is interesting to note how the estimation error contained in the Identity matrix dramatically increases with respect to the case $N = 5$. This is another evidence of the fact that the number of assets N is a vital quantity to account for to explain the estimation error problem.

Moreover, in this second scenario, the benefit of estimating the covariance matrix differently becomes more prominent than before. Indeed, in this case, estimating the covariance matrix with the $\hat{\Sigma}_{c^*}$ estimator reduces almost ten times the estimation error concerning the maximum likelihood or the sample covariance estimators.

The precision shrinkage estimators also provide, in this case, the best response in reducing the estimation error problem. While with $\gamma = 1$ the unbiased precision

estimator returns the same performance of Tu and Zhou (2011) combination rule if it is shrunk towards the implied covariance matrix $\hat{\Sigma}_{EW}$, with $\gamma = 3$ the two approaches differs. Moreover, the precision shrinkage $\hat{\Sigma}_s^{-1} = \alpha^* \hat{\Sigma}_{PM}^{-1} + (1 - \alpha^*)\hat{\Sigma}_{EW}^{-1}$ overperforms asymptotically. Further, if we shrink any prior matrix towards a target according to Theorem 1, we considerably reduce the estimation error in all the cases. Asymptotically the benefit is always increasing.

The best shrinkage approach is the shrinkage of the optimal scaling estimation $\hat{\Sigma}_{c^*}$ towards the implied equally weighted portfolio's covariance $\hat{\Sigma}_{EW}$, precisely as in the previous case of $N = 5$. Even if a tiny sample size (e.g. $T = 60$) is the second-best, the precision shrinkage is the best estimator in reducing the estimation error for $T \geq 240$. With monthly data, $T = 240$ means 20 years of observation usually available in empirical applications.

In the end, the comparison between the proposed approach and the linear shrinkage estimators of Ledoit and Wolf (2003) and Ledoit and Wolf (2004a) has to be carried out. Briefly, the results in Tab. 1.2 are consistent with those of Tab. 1.1, where the proposed precision matrix shrinkage has been revealed to perform better.

1.1.4 Full parameter uncertainty

What Kan and Zhou (2007) showed is, however, still an ideal setting. Even if we admit the possibility of making mistakes in weights estimation because of uncertainty on parameters μ and Σ, we assume the actual parameters in determining the expectations. This is logical if we want to understand the utility loss due to the under/overestimation of portfolio weights. Indeed, what changes between $U(w^*)$ and $E[U(\hat{w})]$ is just the presence of some estimation within the weights in the second one, but the rest of the utility is the same. Therefore, by computing the difference as (1.3), we are determining the loss in the utility only due to the weights' uncertainty. However, what if the investors do not know **anything** about the actual parameters? It is reasonable to assume that they make a mistake in estimating the portfolio's weights and estimating the utility that they get from the implementation of a given asset allocation strategy.

Let suppose an investor with mean-variance preferences. There is an effortless way to analyse this "miss perception" that replaces the actual parameters within the expectation with the investor's estimates (his/her beliefs about the actual parameters).

Hence, in this case, the expected utility associated with a given asset allocation strategy $\hat{w}$ is:

$$E[U(\hat{w})|\hat{\Sigma}, \hat{\mu}] = E[\hat{w}'\hat{\mu}] - \frac{\gamma}{2}E[\hat{w}'\hat{\Sigma}\hat{w}]. \tag{1.21}$$

We can evaluate the "full" impact of parameter uncertainty on investors' expected utility within this setting. Suppose, first, that the investor knows the actual covariance matrix but does not know the mean and suppose that she estimates it by the maximum likelihood estimator. The investor's perceived utility is equal to (the proof is provided in Appendix A):

$$E[U(\hat{w})|\Sigma, \hat{\mu}] = \frac{1}{2\gamma}\left(\frac{N}{T} + \theta^2\right). \tag{1.22}$$

Which reduces with increasing risk aversion and the number of assets while reducing with increasing observations T. Suppose, instead, that the investor does not know the true covariance, which is estimated via maximum likelihood, while the actual mean vector is known. In this case, the perceived utility is equal to:

$$E[U(\hat{w})|\hat{\Sigma}, \mu] = \frac{1}{2\gamma}\frac{T-N-2}{T}\theta^2. \tag{1.23}$$

Kan and Zhou (2007) showed that when the ratio N/T is small, the estimation error in the mean is larger than the estimation error in the covariance and that when it is large, the reverse applies.

In the following Table 1.3 we analyse differences in estimation error severity due to mean and covariance within this setting. Some important conclusions can be derived in the case of $N = 5$ assets. First, the perceived utility is decreasing with sample size T. This evidence suggests that once more information is available, the investors understand that their utility is lower than what they believed. Second, the utility with the estimated mean is greater than the one associated with the covariance estimation. Therefore, mean estimation increase miss perception.

Moreover, the perceived utility with the estimated expected returns vector is not reducing quickly with increasing T. On the other hand, the perceived utility with estimated covariances reduces significantly with an increasing sample size T. Hence,

we can conclude that overall utility's miss perception is mainly due to the errors in estimating the mean vector. The same conclusions almost apply when $N = 30$. Nevertheless, with increasing N and fixed T, the perceived utility is greater and, therefore, the miss perception increases with a higher N/T ratio.

The way the miss perception is affected by N and T is similar to estimation error. Therefore in a setting with no information, the utility's miss perception represents an additional source of deviation from the first best.

N	T	$\Sigma, \hat{\mu}$	$\hat{\Sigma}, \mu$
$\gamma = 0.3$			
5	60	0.01979254	0.08255701
	120	0.01354254	0.007744286
	240	0.01041754	0.007511625
	480	0.008855036	0.007400459
30	60	0.02606017	0.01535114
	120	0.01981017	0.01440018
	240	0.01668517	0.01396756
	480	0.01512267	0.01376085
$\gamma = 1$			
5	60	0.06597512	0.027519
	120	0.04514179	0.02581429
	240	0.03472512	0.02503875
	480	0.02951679	0.0246682
30	60	0.08686724	0.05117046
	120	0.06603391	0.04800061
	240	0.05561724	0.04655853
	480	0.05040891	0.0458695
$\gamma = 3$			
5	60	0.1979254	0.08255701
	120	0.1354254	0.07744286
	240	0.1041754	0.07511625
	480	0.08855036	0.07400459
30	60	0.2606017	0.1535114
	120	0.1981017	0.1440018
	240	0.1668517	0.1396756
	480	0.1512267	0.1376085

TABLE 1.3: Percieved Expected Utilities (1)

Now, suppose that the investor does not know the covariance and the expected returns vector. Assuming that the investor estimates these quantities by maximum likelihood, the perceived utility is:

$$E[U(\hat{w}_{ML})|\hat{\Sigma}, \hat{\mu}] = \frac{1}{2\gamma} \frac{1}{T - N - 2} \left(N + T\theta^2\right) \tag{1.24}$$

Suppose she believes that the sample covariance estimator's estimates are the same as the true covariances. In this case, the perceived utility is equal to:

$$E[U(\hat{w}_{SC})|\hat{\Sigma},\hat{\mu}] = \frac{1}{2\gamma}\frac{T-1}{T-N-2}\left(\frac{N}{T}+\theta^2\right). \tag{1.25}$$

Then, if we develop the Unbiased Precision (PM) estimator:

$$E[U(\hat{w}_{PM})|\hat{\Sigma},\hat{\mu}] = \frac{1}{2\gamma}\left(\frac{N}{T}+\theta^2\right). \tag{1.26}$$

As in Kan and Zhou (2007), all the covariance estimators differ each other by a scaling factor c, such that exists a general estimator of the form $\hat{\Sigma}_c = \hat{\Sigma}_{ML}/c$ as suggested also by Haff (1979). Indeed, in the case of MLE $c = 1$, for the sample covariance $c = (T-1)/T$ while for the unbiased precision matrix $c = (T-N-2)/T$.

An interesting question to answer is whether there is an optimal constant c^* within this framework. Surprisingly, within a perceived utility framework, the optimal scaling is $c^* = 1$ and, therefore, the maximum likelihood estimator is the one that maximizes the investor's utility.

This result is significant since, within a perceived utility framework, the optimal two-fund rule is a maximum likelihood-based strategy, which is the worst one within Kan and Zhou (2007) framework. This means, in other words, that despite all the investors making their choices rationally (maximizing their utilities), they will choose the worst allocation strategy in the case of full uncertainty.

One can think that there is a straightforward way to reduce errors due to uncertainty: avoid estimation at all. Hence, the solution could be employing an equally weighted strategy. This does not seem right. Indeed, equally-weighted is a way of reducing errors since no weight is estimated. However, the investor assumes a certain mean and covariance for computing her utility, not only the weights. In the case of equally-weighted strategy, where the investor estimate means and covariance via MLE, perceived utility is:

$$E[U(\hat{w}_{ew})|\hat{\Sigma},\hat{\mu}] = w_e'\mu - \frac{\gamma}{2}\frac{T-N-2}{T}w_e'\Sigma w_e. \tag{1.27}$$

where $w_e = (1/N,\ldots,1/N)$ is a constant vector. Table 1.4 reports comparisons in terms of perceived utilities between strategies.

N	T	ML	SC	US	EW
$\gamma = 0.3$					
5	60	0.2489627	0.2448133	0.2199171	-2.570922
	120	0.1597939	0.1584623	0.1504726	-2.8056269
	240	0.1192279	0.1187311	0.1157504	-2.9229795
	480	0.09984537	0.09963736	0.09838929	-2.98165580
30	60	2.1085755	2.0734326	0.9840019	-1.1716088
	120	0.7736390	0.7671920	0.5673352	-2.4120503
	240	0.4142330	0.4125070	0.3590019	-3.0322710
	480	0.2730378	0.2724689	0.2548352	-3.3423814
$\gamma = 1$					
5	60	0.07468882	0.07344400	0.06597512	-10.86383898
	120	0.04793818	0.04753870	0.04514179	-11.64618967
	240	0.03576836	0.03561933	0.03472512	-12.03736501
	480	0.02995361	0.02989121	0.02951679	-12.23295268
30	60	0.6325727	0.6220298	0.2952006	-6.2367448
	120	0.2320917	0.2301576	0.1702006	-10.3715496
	240	0.1242699	0.1237521	0.1077006	-12.4389521
	480	0.08191133	0.08174068	0.07645057	-13.47265328
$\gamma = 3$					
5	60	0.02489627	0.02448133	0.02199171	-34.55788836
	120	0.01597939	0.01584623	0.01504726	-36.90494042
	240	0.01192279	0.01187311	0.01157504	-38.07846646
	480	0.009984537	0.009963736	0.009838929	-38.66522947
30	60	0.21085755	0.20734326	0.09840019	-20.70856177
	120	0.07736390	0.07671920	0.05673352	-33.11298
	240	0.04142330	0.04125070	0.03590019	-39.31518
	480	0.02730378	0.02724689	0.02548352	-42.41629

TABLE 1.4: Percieved Expected Utilities (2)

The results show that, first, the differences in expected utilities between alternative strategies become lower with increasing sample size. Bigger sample size means higher information about the actual data distribution. Asymptotically, we have similar results to Kan and Zhou (2007).

Second, we have to note that, with $N = 30$, the expected utility differences seem to be much more significant from the investor's point of view. For example, with $T = 60$, a mean-variance strategy with maximum likelihood estimation has a perceived utility two times greater than an unbiased precision matrix estimator. This does not seem right. Nevertheless, a higher number of assets implies a greater estimation

error and, therefore, we also have a greater level of miss perception. Also, with a greater number of assets we have, asymptotically, the differences become lower, and the perceived utility becomes almost ten times lower. Third, the equally weighted strategy has still a negatively perceived utility and, therefore, it is not possible to reduce the estimation error without estimation. Again, the bigger is the sample size, the lower is the perceived utility.

One can ask if we would also be able to construct an optimal precision shrinkage estimator within this setting. Unfortunately, here this is not possible. To see why it is sufficient to write the perceived utility for a general precision shrinkage estimator:

$$
\begin{aligned}
E\left[U(\hat{w}_s)|\hat{\Sigma},\hat{\mu}\right] &= E\left[\hat{w}_s'\hat{\mu}\right] - \frac{\gamma}{2}E\left[\hat{w}_s'\hat{\Sigma}_s\hat{w}_s\right] = \\
&= E\left[\frac{1}{\gamma}\hat{\mu}'\hat{\Sigma}_s^{-1}\hat{\mu}\right] - \frac{\gamma}{2}E\left[\frac{1}{\gamma^2}\hat{\mu}'\hat{\Sigma}_s^{-1}\hat{\Sigma}_s\hat{\Sigma}_s^{-1}\hat{\mu}\right] = \\
&= \frac{1}{\gamma}E\left[\hat{\mu}'\hat{\Sigma}_s^{-1}\hat{\mu}\right] - \frac{1}{2\gamma}E\left[\hat{\mu}'\hat{\Sigma}_s^{-1}\hat{\mu}\right] = \\
&= \frac{1}{2\gamma}E\left[\hat{\mu}'(\alpha\hat{\Omega}_1^{-1} + (1-\alpha)\hat{\Omega}_2^{-1})\hat{\mu}\right] = \\
&= \frac{1}{2\gamma}\alpha E\left[\hat{\mu}'\hat{\Omega}_1^{-1}\hat{\mu}\right] + \frac{1}{2\gamma}(1-\alpha)E\left[\hat{\mu}'\hat{\Omega}_2^{-1}\hat{\mu}\right].
\end{aligned} \tag{1.28}
$$

Clearly, there is no α that maximizes perceived utility, since first derivative of (1.28) does not depend by α.

1.2 Uncertainty on portfolio weights: the minimum variance investor

The mean-variance setting of Markowitz (1952) involves the estimation of both means and covariance. However, it is known that estimating expected returns is much more challenging than estimating covariances (Merton, 1980). Therefore, several scholars focused on the following portfolio problem for reducing estimation error: find an asset allocation that minimizes portfolio variance instead of maximizing the investor's expected utility. This is called the "global minimum variance" (GMV) strategy and involves only the estimation of the covariance matrix inverse. Kourtis, Dotsis, and Markellos (2012) demonstrated that the estimation error contained in a Global Minimum Variance (GMV) strategy, where only the covariance structure is used to build

the optimal portfolios, contains a lower estimation error than a Mean-Variance (MV) allocation. Intuitively, this happens because GMV avoids the estimation error contained in the expected returns' vector $\hat{\mu}$. In what follows, we study the estimation error for this asset allocation strategy.

1.2.1 Minimum variance with standard sample estimates

Unlike classical mean-variance allocations, such a strategy involves just risky assets: no proportion of an investor's wealth has to be used to buy a riskless asset. Let us assume to have N risky assets in the investment universe. The portfolio problem can be writen as:

$$\min_{w} w'\Sigma w$$

$$s.t. \sum_{i=1}^{N} w_i = 1$$

The optimal global minimum variance weight w^* vector, as solution of the minimization problem above, is given by:

$$w^* = \frac{\Sigma^{-1}\mathbb{1}_N}{\mathbb{1}'_N\Sigma^{-1}\mathbb{1}_N} \tag{1.29}$$

where $\mathbb{1}_N = (1, 1, \ldots, 1)$ and $\mathbb{1}'_N\Sigma^{-1}\mathbb{1}_N$ is the sum of all elements within the vector $\Sigma^{-1}\mathbb{1}_N$. By replacing Σ^{-1} with $\hat{\Sigma}^{-1}$ we get the optimal *estimated* GMV portfolio weights that we call $\hat{w}$. In what follows we provide some insights about the estimation error for this type of strategy.

As proved by Okhrin and Schmid (2006), the sample counterpart of weights (2.3) follows an elliptical t-distribution with $T - N - 1$ degrees of freedom. The first two moments are equal to:

$$E[\hat{w}] = w \tag{1.30}$$

and:

$$Var[\hat{w}] = \frac{1}{T-N-1}\frac{\mathbf{R}}{\mathbb{1}_N'\hat{\Sigma}_{ML}^{-1}\mathbb{1}_N} \tag{1.31}$$

where:

$$\mathbf{R} = \frac{\hat{\Sigma}_{ML}^{-1}\mathbb{1}_N\mathbb{1}_N'\hat{\Sigma}_{ML}^{-1}}{\mathbb{1}_N'\hat{\Sigma}_{ML}^{-1}\mathbb{1}_N}.$$

About the expected return and variance of a GMV portfolio, Kourtis, Dotsis, and Markellos (2012) provided some useful results. In particular, the authors demonstrate that the following relationships hold:

$$E[\hat{w}'\mu] = w'\mu \tag{1.32}$$

and:

$$E[\hat{w}'\Sigma\hat{w}] = \frac{T-2}{T-N-1}w'\Sigma w. \tag{1.33}$$

where (1.32) is the portfolio expected returns and (1.33) its expected variance.
Using these results makes it possible to derive a closed formula for estimation error for such a strategy. In this case, however, we compute the estimation error as the deviation of portfolio variance under actual weights from the variance under the estimated portfolio weights. Intuitively, we follow an approach similar to the one of Kan and Zhou (2007) but for a GMV setting.
The estimation error for a maximum likelihood-based GMV strategy is:

$$\begin{aligned} \ell(w,\hat{w}) &= w'\Sigma w - E[\hat{w}'\Sigma\hat{w}] = \\ &= w'\Sigma w - \frac{T-2}{T-N-1}w'\Sigma w = \\ \ell(w,\hat{w}) &= \left(1 - \frac{T-2}{T-N-1}\right)w'\Sigma w. \end{aligned} \tag{1.34}$$

where the loss is a negative number. Indeed, the variance of the estimated portfolio

weights is greater since $T-2 > T-N-1$ so $T-2/T-N-1 > 1$.

Can we reduce estimation error by estimating differently precision matrices? The answer is no, at least with usual sample estimators. The reason is that all of them are proportional to maximum likelihood covariance. Suppose to consider the case of the Unbiased Precision estimator that is unbiased with respect to the true covariance inverse:

$$\hat{\Sigma}_{US} = \frac{1}{T-N-2}\sum_{t=1}^{T}(X_t - \hat{\mu})'(X_t - \hat{\mu}) = \frac{T-N-2}{T}\hat{\Sigma}_{ML}$$

where $\hat{\Sigma}_{ML} = T^{-1}\sum_{t=1}^{T}(X_t - \hat{\mu})'(X_t - \hat{\mu})$ be the usual maximum likelihood estimator for covariance matrix. Expected portfolio variance by plug-in of $\hat{\Sigma}_{US}^{-1}$ in (2.3) is:

$$E[\hat{w}'\Sigma\hat{w}] = E\left[\frac{\mathbb{1}_N'\hat{\Sigma}_{US}^{-1}}{\mathbb{1}_N'\hat{\Sigma}_{US}^{-1}\mathbb{1}_N}\Sigma\frac{\hat{\Sigma}_{US}^{-1}\mathbb{1}_N}{\mathbb{1}_N'\hat{\Sigma}_{US}^{-1}\mathbb{1}_N}\right] =$$
$$E\left[\frac{\frac{T-N-2}{T}\mathbb{1}_N'\hat{\Sigma}_{ML}^{-1}}{\frac{T-N-2}{T}\mathbb{1}_N'\hat{\Sigma}_{ML}^{-1}\mathbb{1}_N}\Sigma\frac{\frac{T-N-2}{T}\hat{\Sigma}_{ML}^{-1}\mathbb{1}_N}{\frac{T-N-2}{T}\mathbb{1}_N'\hat{\Sigma}_{ML}^{-1}\mathbb{1}_N}\right] =$$
$$E[\hat{w}'\Sigma\hat{w}] = E\left[\frac{\mathbb{1}_N'\hat{\Sigma}_{ML}^{-1}}{\mathbb{1}_N'\hat{\Sigma}_{ML}^{-1}\mathbb{1}_N}\Sigma\frac{\hat{\Sigma}_{ML}^{-1}\mathbb{1}_N}{\mathbb{1}_N'\hat{\Sigma}_{ML}^{-1}\mathbb{1}_N}\right] = \frac{T-2}{T-N-1}w'\Sigma w.$$

Hence, the portfolio variance is equivalent in the case of Unbiased Precision with respect maximum likelihood case. Obviously, the same applies to portfolio expected return since:

$$E[\hat{w}'\mu] = E[\hat{w}']\mu = E\left[\frac{\mathbb{1}_N'\hat{\Sigma}_{US}^{-1}}{\mathbb{1}_N'\hat{\Sigma}_{US}^{-1}\mathbb{1}_N}\right]\mu =$$
$$= E\left[\frac{\frac{T-N-2}{T}\mathbb{1}_N'\hat{\Sigma}_{ML}^{-1}}{\frac{T-N-2}{T}\mathbb{1}_N'\hat{\Sigma}_{ML}^{-1}\mathbb{1}_N}\right]\mu = E\left[\frac{\mathbb{1}_N'\hat{\Sigma}_{ML}^{-1}}{\mathbb{1}_N'\hat{\Sigma}_{ML}^{-1}\mathbb{1}_N}\right]\mu = w'\mu.$$

Therefore, by the plug-in of whatever estimator is proportional to maximum likelihood in this framework, we cannot reduce estimation error.

As further evidence, we provide simulation results (see Tab. 1.5) as well. We assume there are $N = 5$ and $N = 30$ risky assets with mean and covariances are chosen

based on the sample estimates from the monthly excess returns on the 5 and 30 industry portfolios of Fama-French[3] from July 1926 to October 2019. Expected out of sample performances are determined for all the cases by $M = 10000$ simulations.

N	T	ML	SC	US
5	60	-1.670225	-1.670225	-1.670225
	120	-0.7964584	-0.7964584	-0.7964584
	240	-0.3896277	-0.3896277	-0.3896277
	480	-0.1948903	-0.1948903	-0.1948903
30	60	-11.59008	-11.59008	-11.59008
	120	-3.801216	-3.801216	-3.801216
	240	-1.619598	-1.619598	-1.619598
	480	-0.7515538	-0.7515538	-0.7515538

TABLE 1.5: Estimation Error for alternative GMV strategies

The estimation error dynamics do not change with respect to the mean-variance case. In other words, it increases with N as T is fixed while it decreases with increasing sample size since more information is available. Nevertheless, the estimation error can be reduced by the plug-in of not proportional estimators. In what follows, we discuss if and how the *precision shrinkage estimator* can be helpful to this aim.

1.2.2 The precision shrinkage estimator

In the whole chapter, we claim that a limitation of the existing shrinkage estimators relies on how optimal shrinkage intensity is determined. Indeed, in all the cases, it is based on statistical arguments, and it is not necessarily consistent with the portfolio selection problem: finding portfolio composition that maximizes investors' preferences.

Therefore in what follows, we aim to overcome this limitation, proposing a portfolio volatility-minimizer shrinkage estimator of the precision matrix.

Estimation framework and optimal shrinkage intensity

From Ledoit and Wolf (2017) we know that minimizing the following loss function:

[3] Also, in this case, we get data from Kenneth French website https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

$$\ell(\hat{\Sigma}^{-1}, \Sigma, \mathbb{1}_N) = \hat{w}'\Sigma\hat{w} = \frac{\mathbb{1}_N'\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\mathbb{1}_N}{\left(\mathbb{1}_N'\hat{\Sigma}^{-1}\mathbb{1}_N\right)^2} \tag{1.35}$$

essentially means that we are minimizing the portfolio variance. Moreover, minimizing the loss (1.35) is also equivalent of maximizing out of sample Sharpe ratio (Ledoit and Wolf, 2017). Consider now the following general shrinkage estimator of precision matrix:

$$\hat{\Sigma}_s^{-1} = \alpha\hat{\Omega}_1^{-1} + (1-\alpha)\hat{\Omega}_2^{-1} \tag{1.36}$$

where, as previously, $\hat{\Omega}_1^{-1}$ is called *prior* and $\hat{\Omega}_2^{-1}$ is the *target*. The quantity α is called *optimal shrinkage intensity*. By substituting (1.36) in (1.35) we get:

$$\begin{aligned} E\left[\ell(\hat{\Sigma}_s^{-1}, \Sigma, \mathbb{1}_N)\right] &= E\left[\hat{w}'\Sigma\hat{w}\right] = \\ &= E\left[\frac{\mathbb{1}_N'\left[\alpha\hat{\Omega}_1^{-1} + (1-\alpha)\hat{\Omega}_2^{-1}\right]\Sigma\left[\alpha\hat{\Omega}_1^{-1} + (1-\alpha)\hat{\Omega}_2^{-1}\right]\mathbb{1}_N}{\left[\mathbb{1}_N'\left[\alpha\hat{\Omega}_1^{-1} + (1-\alpha)\hat{\Omega}_2^{-1}\right]\mathbb{1}_N\right]^2}\right]. \end{aligned} \tag{1.37}$$

Given two general estimators $\hat{\Omega}_1^{-1}$ and $\hat{\Omega}_2^{-1}$, we can find optimal shrinkage intensity α^* that minimize portfolio variance or, equivalently, the loss (1.37):

$$\min_{\alpha} E\left[\ell(\hat{\Sigma}_s^{-1}, \Sigma, \mathbb{1}_N)\right]. \tag{1.38}$$

Unfortunately, a closed formula in this case cannot be obtained. To overcome this issue, we develop the following algorithm able to optimally choose α^*:

1. Consider a sequence of possible intensities ranging from 0 to 1 with a sequence increment of k, so you have a vector $\underline{\alpha}$ of length k;
2. Given a value of α_k compute the shrinkage estimator $\hat{\Sigma}_s^{-1}$;
3. Given $\hat{\Sigma}_s^{-1}$ determine optimal portfolio weights $\hat{w}_s$;
4. Compute portfolio variance $\sigma_k = \hat{w}_s'\Sigma\hat{w}_s$;

5. Repeat points 3 and 4 M times for each value of α_k and store results in a matrix Φ of dimension $M \times k$:

$$\Phi = \begin{pmatrix} \sigma_{1,\alpha_1} & \sigma_{1,\alpha_2} & \cdots & \sigma_{1,\alpha_k} \\ \sigma_{2,\alpha_1} & \sigma_{2,\alpha_2} & \cdots & \sigma_{2,\alpha_k} \\ \vdots & \ddots & \vdots & \\ \sigma_{M,\alpha_1} & \sigma_{M,\alpha_2} & \cdots & \sigma_{M,\alpha_k} \end{pmatrix}$$

6. Compute column means of Φ to get expected out of sample variances $E[\sigma_k]$ for each α_k.

7. Choose the optimal α^* is the value with the minimum expected out of sample variance.

An important choice is related to the prior and the target that has to be optimally combined. The obvious choice for the prior is the Unbiased Precision covariance matrix. Its estimation error is equivalent to all the other classical plug-ins, but, differently from the others, it is unbiased with respect to the actual precision matrix. On the other side, a natural candidate as the target is the Identity matrix I (Haff (1979), Ledoit and Wolf (2004a), and Kourtis, Dotsis, and Markellos (2012)).
Two simple explanations justify the usage of such a target. First of all, the Identity matrix I does not contain estimation error.
However, another interesting argument arises in the minimum variance setting. Indeed, within this setting, we have additional motivation: if we plug in the Identity matrix in (2.3), we exactly get the equally weighted portfolio. Showing this result is straightforward. Just remember that $I\mathbb{1}_N = \mathbb{1}_N$ and $\mathbb{1}_N' I \mathbb{1}_N$ is the sum of all elements within the vector $I\mathbb{1}_N$ that is $\mathbb{1}_N'\mathbb{1}_N = N$. Therefore, GMV weights with Identity plug-in is:

$$\hat{w} = \frac{\hat{\Sigma}^{-1}\mathbb{1}_N}{\mathbb{1}_N'\hat{\Sigma}^{-1}\mathbb{1}_N} = \frac{I\mathbb{1}_N}{\mathbb{1}_N' I \mathbb{1}_N} = \frac{\mathbb{1}_N}{\mathbb{1}_N'\mathbb{1}_N} = w_e.$$

with $w_e = (1/N, \ldots, 1/N)$. In other words, the Identity matrix is the implied covariance assumed by a GMV investor that invests in an equally weighted portfolio. As showed by De Miguel, Garlappi, and Uppal (2007), surprisingly, this strategy is in

the empirical application the one with the highest out of sample Sharpe ratio. Hence we propose the following shrinkage estimator of the precision matrix:

$$\hat{\Sigma}_{sI}^{-1} = \alpha\hat{\Sigma}_{US}^{-1} + (1-\alpha)I \tag{1.39}$$

where α^* is chosen by the recursive algorithm presented above. If we apply (1.39) to estimate the GMV weights (2.3), another interpretation can be provided. Indeed, plug-in of the precision matrix shrinkage leads to a portfolio that is the result of the combination of the other two portfolios (Kourtis, Dotsis, and Markellos (2012)). In the specific case of (1.39), we obtain the following portfolio:

$$\hat{w}_s = \frac{\alpha\hat{\Sigma}^{-1}\mathbb{1}_N + (1-\alpha)I\mathbb{1}_N}{\alpha\mathbb{1}_N'\hat{\Sigma}^{-1}\mathbb{1}_N + (1-\alpha)\mathbb{1}_N'I\mathbb{1}_N} = \delta\hat{w} + (1-\delta)w_e \tag{1.40}$$

where:

$$\delta = \frac{\alpha\hat{\Sigma}^{-1}\mathbb{1}_N}{\alpha\mathbb{1}_N'\hat{\Sigma}^{-1}\mathbb{1}_N + (1-\alpha)\mathbb{1}_N'I\mathbb{1}_N}$$

In other words, by plug-in of precision matrix estimator (1.39) we invest a wealth's proportion δ in the sample global minimum variance portfolio and $(1-\delta)$ in the equally weighted portfolio. Clearly, if we substitute α with α^* in (1.40), we get the optimal δ^*.

Shrinkage targets

Since the shrinkage intensities are derived through simulations instead of closed formulas, we can easily consider the performances of more complex targets for which closed solutions for the expectations are not available. Among them, we have the Sharpe (1963) single index implied covariance matrix that has also been used by Ledoit and Wolf (2003) and a general factor model covariance matrix $\hat{\Sigma}_F$. In this case, the two competitive estimators are:

$$\hat{\Sigma}_{sF}^{-1} = \alpha\hat{\Sigma}_{US}^{-1} + (1-\alpha)\hat{\Sigma}_F^{-1} \tag{1.41}$$

$$\hat{\Sigma}_{sM}^{-1} = \alpha\hat{\Sigma}_{US}^{-1} + (1-\alpha)\hat{\Sigma}_{M}^{-1} \tag{1.42}$$

For a review on the possible targets that can be considered, see Bai and Shi (2011). In what follows we provide more details about the employed targets.

The first target, considered in 1.41 is the covariance matrix implied by a factor model. The factor models have been widely applied in both theoretical and empirical finance. Derived by Ross (1976) and Chamberlain (1983), according to the multi-factor model the excess return of any asset $r_{i,t}$ over the risk free rate satisfies:

$$r_{i,t} = \alpha_i + b_{i1}f_{1,t} + \cdots + b_{i,K}f_{K,t} + \epsilon_{i,t} \tag{1.43}$$

Where K are the number of factors, f the factors themselves and b_{ij} are the parameters associated with the factors with ϵ_i the idiosyncratic error associated with the return i, orthogonal to the K factors. The model presented in (2.15) is a general one, allowing for multiple factors. Nevertheless, a factor model can also be based on a single factor. A widespread single-factor model is the single-index of Sharpe (1963). Moreover, factor models can be either static or dynamic. First of all, in a static factor model, the intercept α_i and the parameters b_{ij} are time-invariant. Second, the conditional covariance matrix of the factors and the errors are also time-invariant.

Then, another critical distinction has to be made between observed and latent factor models. Observed factors are known (e.g. macroeconomic variables or firm characteristics) and based on outside information, either from economic theory or empirical facts. The leading example is the Fama and French (1993) three-factor model, but a vast literature exists on the so-called "return predictive signals" (RPS). Green, Hand, and Zhang (2013) found more than 330 RPS from previous papers. From a macroeconomic factors point o view, an example is Chen, Roll, and Ross (1986). On the other side, latent factors are unknown and need to be estimated. The most common approach in estimating static latent factors is through the principal components (e.g. Bai and Ng (2013)). Observed factor models are the simplest and the most convenient from the parsimony point of view. Indeed, instead of having a covariance matrix with $N(N+1)/2$ parameters, with a three-factor model, we have just $4N$

of them to estimate. Nevertheless, with observed factor models, the probability of omitting some relevant ones is high. Moreover, there is no consensus regarding the factors to be included. This is the main reason why latent factor models are usually preferred.

In a static factor model[4], once the factor parameters are stored in a matrix $\mathbf{B}$, given Σ_f the covariance matrix of factors and Σ_ϵ the (diagonal) covariance matrix of the errors, the returns' actual covariance matrix is (Bai, 2003):

$$\Sigma = \mathbf{B}\Sigma_F\mathbf{B}' + \Sigma_\epsilon \tag{1.44}$$

What we need is an estimator for both Σ_F and Σ_ϵ, with also the estimated factor parameters $\hat{\mathbf{B}}$.

What we can do is the following. Starting from the (2.15), we estimate the intercept and the parameters by OLS in order to obtain $\hat{\mathbf{B}}$ and, then, we use the "textbook estimator to get the factors' covariance matrix $\hat{\Sigma}_F$. In the end, if Σ_ϵ is diagonal, we take the residuals from the OLS regression and compute $\hat{\Sigma}_e$. In this way the estimated covariance matrix of returns in a factor model is given by:

$$\hat{\Sigma} = \hat{\mathbf{B}}\hat{\Sigma}_F\hat{\mathbf{B}}' + \hat{\Sigma}_e \tag{1.45}$$

An estimate of precision matrix could be obtained just inverting (1.45). We could do that by using the Sherman–Morrison–Woodbury formula:

$$\hat{\Sigma}^{-1} = \hat{\Sigma}_e^{-1} - \hat{\Sigma}_e^{-1}\hat{\mathbf{B}}\left[\hat{\Sigma}_F^{-1} + \hat{\mathbf{B}}'\hat{\Sigma}_e^{-1}\hat{\mathbf{B}}\right]^{-1}\hat{\mathbf{B}}'\hat{\Sigma}_e^{-1} \tag{1.46}$$

The theoretical properties of the estimators (1.45) and (1.46) are studied by Fan, Fan, and Lv (2008).

The main problem with the observed factor model is that, as we have mentioned, there is not a common consensus about which factor to include. For this reason, latent factor models are popular. Moreover, Ledoit and Wolf (2003) claim that factors working well in a dataset may not work for another dataset (Ledoit and Wolf, 2003).

[4] De Nard, Ledoit, and Wolf (2019) shown that dynamic factor models do not provide better results than static ones, that are easier to estimate.

An alternative way to get a reasonable estimate of the covariance matrix is based on the so-called shrinkage method due to Stein (1956), which we have already discussed. This statistical technique consists of taking a weighted average between two estimators to build a new one with some desirable properties.

The shrinkage method has been applied for the covariance matrix estimation by Ledoit and Wolf (2004a), shrinking the sample covariance matrix with the identity I. The most important element is the associated weights to the two matrices called "shrinkage intensity".

The optimal shrinkage intensity can be computed according to a loss function to be minimized. Ledoit and Wolf (2004a) proposed the Frobenius norm between the shrinkage estimator Σ^* and the true covariance matrix as follows:

$$\min_{\rho_1,\rho_2} E[||\Sigma^* - \Sigma||^2]$$

given that:

$$\Sigma^* = \rho_1 I + \rho_2 \hat{\Sigma} \tag{1.47}$$

With the key assumption that returns have finite fourth moments, the optima shrinkage intensity as solution of the previous problem is given by:

$$\rho_1 = \frac{\beta^2}{\delta^2}$$

$$\rho_2 = \frac{\alpha^2}{\delta^2}$$

where $\alpha^2 = E[||\Sigma - \mu I||^2]$, $\beta^2 = E[||\hat{\Sigma} - \Sigma||^2]$ and $\delta^2 = E[||\hat{\Sigma} - \mu I||^2]$.

The main drawback of this shrinkage estimator is that it needs the knowledge of four scalar functions of the true (and unobserved) covariance matrix Σ, namely μ, α^2, β^2 and δ^2. The authors overcome this problem by providing asymptotically consistent estimators for all of them under certain assumptions in a general asymptotic framework, where both $N \to \infty$ and $T \to \infty$.

The former is just one of the possible shrinkages for the covariance matrix, and it is useful when no obvious other shrinkage targets are available.

Indeed, shrinkage is also related to the factor models previously described. For example, Ledoit and Wolf (2003) proposed a shrinkage where the target matrix was the covariance from a factor model as (2.15). However, in their specification, the factor was just one, the market returns, according to the single-index model of Sharpe. They called this operation "shrinkage towards the market". They justify this choice by assessing that, even if there is no consensus about which factor to include in (2.15), the market returns is the most intuitive and accepted factor for the case of portfolio selection.

Also, in this case, they derived a formula for the optimal shrinkage intensity based on a consistent estimator for unknown quantities. Given the single-index factor model:

$$r_{it} = \alpha_i + \beta r_{Mt} + \epsilon_{it} \tag{1.48}$$

The associated covariance matrix is:

$$\Sigma_M = \sigma_M^2 \beta\beta' + \Sigma_\epsilon \tag{1.49}$$

where σ_M^2 is the market returns variance, β the parameter from (2.15) and Σ_ϵ the covariance matrix of the error. Replacing all the quantities with the estimated values, the shrinkage estimator becomes:

$$\hat{\Sigma}_s = \alpha\hat{\Sigma}_M + (1-\alpha)\hat{\Sigma} \tag{1.50}$$

Then, in order to determine the α^*, Ledoit and Wolf (2003) proposed to minimize the following Frobenius norm:

$$\min_\alpha E[||\alpha\hat{\Sigma}_M + (1-\alpha)\hat{\Sigma} - \Sigma||^2]$$

which leads to the following solution:

$$\alpha^* = \frac{\sum_{i=1}^N \sum_{j=1}^N Var(s_{ij}) - Cov(f_{ij}, s_{ij})}{\sum_{i=1}^N \sum_{j=1}^N Var(f_{ij} - s_{ij}) + (\phi_{ij}^2 - \sigma_{ij}^2)}$$

where f_{ij} are the entries of the covariance matrix $\hat{\Sigma}_M$, s_{ij} are the entries of $\hat{\Sigma}$, ϕ_{ij} the value associated to the unobserved true covariance matrix of the single-index $\mathbf{\Phi}$ and σ_{ij} the entries of the unobserved true covariance matrix Σ. Also here the optimal shrinkage intensity depends by unobserved quantities. Therefore, the authors provide some consistent estimator for them.

1.2.3 Simulation study

Case I: i.i.d. Gaussian economy

In what follows, we evaluate the expected out-of-sample performance of the proposed shrinkage estimator through a simulation study[5] that adapt the Kan and Zhou (2007) scheme within the minimum variance framework. Then we compare the performances with several alternative plug-in strategies. Mainly, we compare the proposed estimation approach with the usual sample estimator and the shrinkage of Ledoit and Wolf (2003) and Ledoit and Wolf (2004a).

More in detail, we assume an investor that aims to find a portfolio with minimum variance. The problem parameters (the expected returns mean and covariances) are calibrated from real time seriedata under the assumption of multivariate normality for stock returns. At this aim, we assume there are $N = 5$ risky assets with their mean and covariance matrix chosen based on the sample estimates from the monthly excess returns on the five industry portfolios of Fama-French[6] from July 1926 to October 2019. Expected out of sample performances in terms of out of sample portfolio variance and Sharpe ratios are determined for all the cases by 1000 simulations.

The portfolio variances respect to different value of α are shown in Figures 1.1-1.3. Points of the minimum are the selected optimal α^*.

[5] Unfortunately, we are not able to provide intense simulations because, despite we take advantage of parallel computing, the proposed algorithm for α^* becomes too challenging and the computer takes several days to accomplish the task.

[6] We get data from Kenneth French website https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

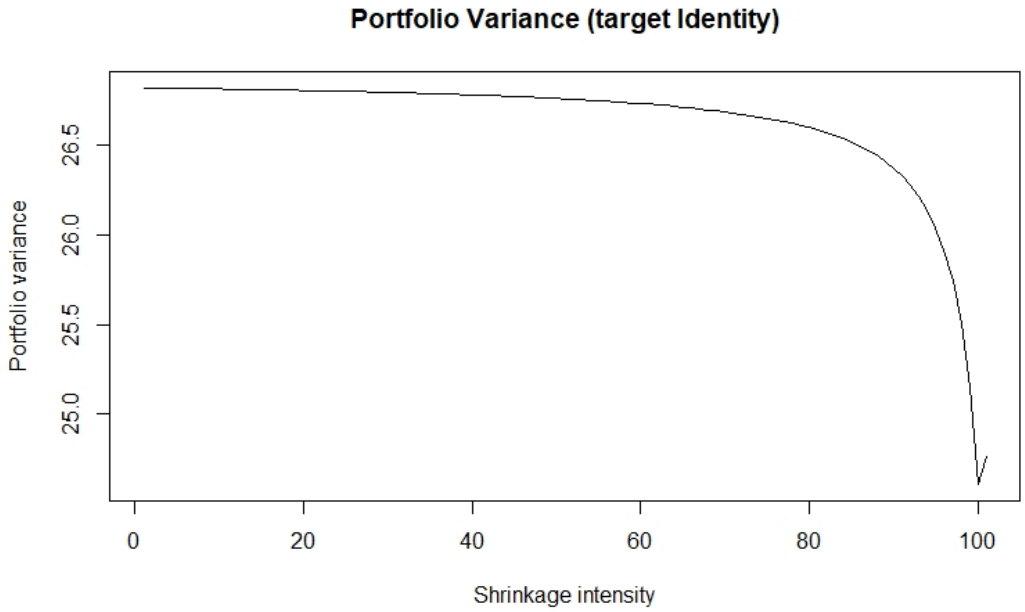


FIGURE 1.1: Portfolio variance of shrinkage towards I

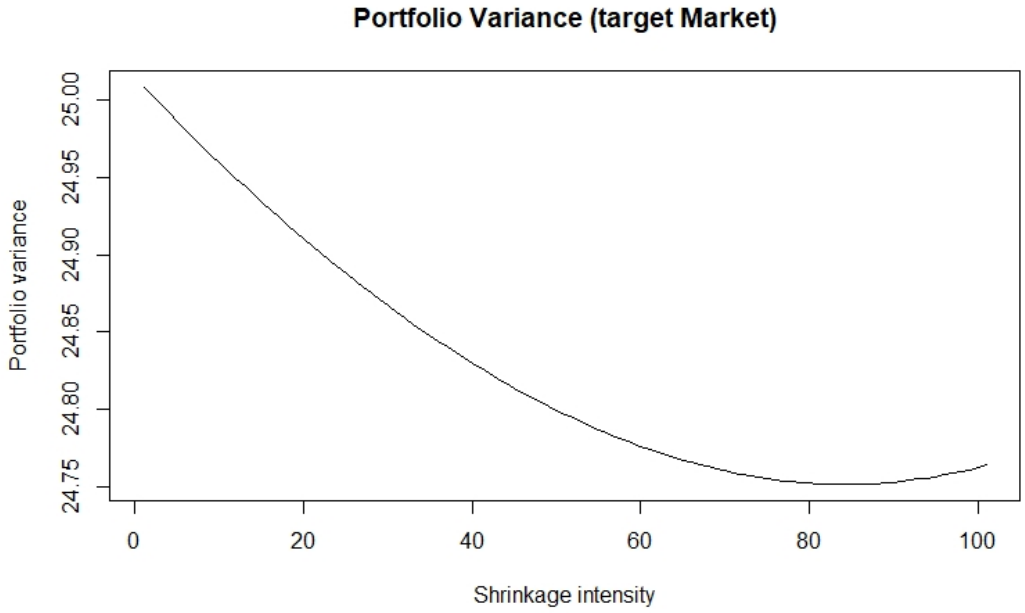


FIGURE 1.2: Portfolio variance of shrinkage towards $\hat{\Sigma}_M^{-1}$

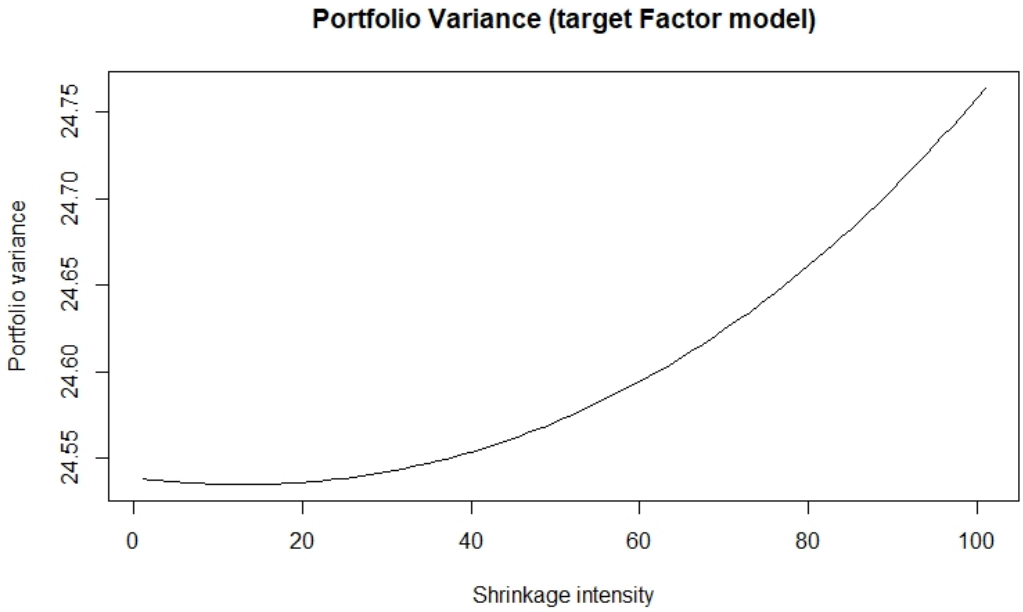


FIGURE 1.3: Portfolio variance of shrinkage towards $\hat{\Sigma}_F^{-1}$

7

In Tab. 1.6 the expected out of sample portfolio variances for all alternative strategies is shown. With $N = 5$ risky assets and $T = 60$ observations, the worst portfolio in terms of variance is the equally weighted one. Its variance is incredibly high. On the other hand, the precision shrinkage estimators best reduce sample portfolio variance. Notably, the shrinkage towards a factor model returns the best results in terms of estimation error (the distance from the "true" portfolio variance is the lowest).

N	T	True	$1/N$	Sample	LW_{2004}	LW_{2003}	$\hat{\Sigma}_{sI}^{-1}$	$\hat{\Sigma}_{sM}^{-1}$	$\hat{\Sigma}_{sF}^{-1}$
5	60	23.1290	670.58	24.7639	24.7489	24.8117	24.6070	24.7515	24.5347
5	120	23.1290	670.58	23.5282	23.5743	23.5297	23.5282	23.5282	23.5079
5	240	23.1290	670.58	23.9698	23.9662	23.9726	23.9698	23.9693	23.8966

TABLE 1.6: Portfolios out of sample variance with $k = 0.01$ and $M = 1000$

Clearly, in evaluating asset allocation performances, portfolio variance is essential but not the only important measure. Indeed, most studies analyse the Sharpe ratio. For a strategy p, its Sharpe ratio is given by the ratio between portfolio return and variance:

$$\hat{SR}_p = \frac{\hat{\mu}_p}{\hat{\sigma}_p}$$

Results are shown in Tab. 1.7 below. Analysing Sharpe ratios, we get the same results: the $1/N$ strategy is the worst one and the shrinkage towards factor model results in the highest (and the closest to the true one) Sharpe ratio. Moreover, also all the other shrinkage approaches return improved performances with respect to the shrinkage estimator of Ledoit and Wolf (2003) and Ledoit and Wolf (2004a).

N	T	True	$1/N$	Sample	LW_{2004}	LW_{2003}	$\hat{\Sigma}_{sI}^{-1}$	$\hat{\Sigma}_{sM}^{-1}$	$\hat{\Sigma}_{sF}^{-1}$
5	60	4.5072%	0.7331%	4.2161%	4.2179%	4.2261%	4.1077%	4.2251%	4.2796%
5	120	4.5072%	0.7331%	4.4292%	4.4292%	4.4330%	4.4292%	4.4292%	4.4472%
5	240	4.5072%	0.7331%	4.3507%	4.3509%	4.3582%	4.3505%	4.3507%	4.3851%

TABLE 1.7: Portfolios Sharpe ratio (%) with $k = 0.01$ and $M = 1000$

Case II: i.i.d. economy with skenwess

All the assumptions we made about stock returns do not fit with real data. For example, it is well known that stock returns are neither independent nor normally distributed. These results are also called "stylized fact" of finance (Cont (2001)). Indeed, the unconditional distribution of returns is usually asymmetric and heavy-tailed. Even though with increasing time scale (e.g. annual returns), returns' distribution looks closer to a Gaussian, this still represents an unreliable assumption. In what follows, we consider an economy with independent returns, but we allow for non-Gaussian distribution.

As before, we assume an investor that aims to find a portfolio with minimum variance, with problem parameters calibrated from real data. Nevertheless, now the assumption we make about returns' distribution is different. Again, we develop $N = 5$ risky assets with their mean and covariance matrix chosen based on the sample estimates from the monthly excess returns on the five industry portfolios of Fama-French[7] from July 1926 to October 2019. First, we generate data from a multivariate Skew Normal distribution with skew parameters vector estimated from actual data. This represents the effect of skewness in returns' distribution, that was not considered before.

Within this framework, we aim to evaluate an effect due to skewness. In the following Table 1.8 we compare out of sample variances. Also, in this case, the precision shrinkage estimators allow for a considerable reduction in portfolio variance and, therefore, estimation error. However, the best strategy is based on the plug-in of shrinkage towards the market under this scenario. Differently from Ledoit and Wolf (2003), both the idea of shrinking precision matrix and the way in which optimal shrinkage intensity is determined to allow to overperform the benchmark even if the target matrix is the same.

[7]We get data from Kenneth French website https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

N	T	True	$1/N$	Sample	LW_{2004}	LW_{2003}	$\hat{\Sigma}_{sI}^{-1}$	$\hat{\Sigma}_{sM}^{-1}$	$\hat{\Sigma}_{sF}^{-1}$
5	60	23.1290	670.58	25.4016	25.3950	25.7347	25.4016	25.4016	25.0911
5	120	23.1290	670.58	24.6799	24.6813	24.8854	24.6799	24.6799	24.4729
5	240	23.1290	670.58	24.3709	24.3725	24.3989	24.3709	24.3709	24.1803

TABLE 1.8: Portfolios out of sample variance with $k = 0.01$ and $M = 1000$: **multivariate skew normal data**

Then, the following Table 1.9 reports results in terms of Sharpe ratios. Most of the strategies are almost equivalent under this scenario. However, despite it being a portfolio with higher variance, the shrinkage towards the factor model portfolio returns the highest Sharpe ratio out of the sample. Once again, we overperform the benchmark shrinkage operations Ledoit and Wolf (2003) and Ledoit and Wolf (2004a) also in a not Gaussian economy.

N	T	True	$1/N$	Sample	LW_{2004}	LW_{2003}	$\hat{\Sigma}_{sI}^{-1}$	$\hat{\Sigma}_{sM}^{-1}$	$\hat{\Sigma}_{sF}^{-1}$
5	60	4.5072%	0.7331%	4.0098%	4.0101%	3.9658%	4.0098%	4.0098%	4.0737%
5	120	4.5072%	0.7331%	4.1134%	4.1129%	4.08782%	4.1134%	4.1134%	4.1674%
5	240	4.5072%	0.7331%	4.1632%	4.16283%	4.1630%	4.1632%	4.1632%	4.2171%

TABLE 1.9: Portfolios Sharpe ratio (%) with $k = 0.01$ and $M = 1000$: **multivariate skew normal data**

1.2.4 Limitations and future developments

The absence of a closed formula for the shrinkage intensity and the high computational burden required for the simulations dramatically limit the empirical applicability of the proposed variance-minimizer shrinkage estimators.
With this respect, further studies have to be devoted to overcoming these issues. Indeed, the simulations show that to some extent, this class of estimator is undoubtedly helpful in reducing estimation error also in the case of a minimum variance setting. These preliminary insights should serve as a good motivation for future researches on the topic.

1.3 Empirical Analysis

1.3.1 Methodology and datasets

In what follows, we study the empirical performance of the proposed precision shrinkage estimator with respect to several alternative estimators in the case of a mean-variance investor. The case of a minimum variance strategy has been excluded because of the computationally challenging properties of the precision shrinkage in this setting and the absence of a closed formula for shrinkage intensity that limits its applicability.
The validity of the proposal is evaluated across a variety of datasets by implementing the strategy of De Miguel, Garlappi, and Uppal (2007). The selected datasets are the industry-based portfolios of Fama-French[8], namely the five industry portfolio in the case $N = 5$, the ten industry portfolio for $N = 10$ and so forth. To evaluate also the performances in a large dimensional setting, we conducted an experiment with the French's 100 portfolios based on size. In the large dimensional setting, all covariance inverses, with the only exception of Ledoit and Wolf (2003) and Ledoit and Wolf (2004a) strategies, are defined as Moore-Penrose pseudo-inverse.
Based on a "rolling-sample" approach, the empirical strategy works as follows. Given a T month-long dataset of asset returns, following De Miguel, Garlappi, and Uppal (2007) we choose an estimation window of length $M = 60, 120$. Therefore, we have a high dimensional setting when $N = 100$ and $M = 60$. Then, in each month t, starting from $t = M + 1$, we use the M observations to estimate the parameters (the

[8] We get data from Kenneth French website https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

expected returns vector μ and the covariance matrix Σ) needed for implementing the mean-variance asset allocation strategy. These estimated parameters are then used to determine the relative portfolio weights. M represents the number of observations we use in estimating covariances and means. In this sense, M is the equivalent of T in the simulation study. Therefore, the greater is the ratio N/M, the more ill-conditioned is the covariance matrix and its inversion become challenging.
This process is repeated $T - M$ times by adding the return for the next period in the dataset and dropping the earliest one until the end of the dataset is reached. The outcome is, for each strategy, a time series of $T - M$ monthly out-of-sample portfolio returns.
Given the time series of monthly out-of-sample returns generated by each strategy, we compute the out-of-sample Sharpe ratio of strategy, defined as the sample means of out-of-sample portfolio returns divided by their standard deviation:

$$SR_p = \frac{\hat{\mu}_p}{\hat{\sigma}_p}$$

where $\hat{\mu}_p$ is the average of the $t - M$ out of sample returns for the p-th strategy and $\hat{\sigma}_p$ the standard deviation.

1.3.2 Out of sample Sharpe ratio: results

Consider first the case where $M = 60$, with an estimation window of 5 years of monthly observations. Results are reported in Tab. 1.10. Let us analyze first the scenario with $N = 5$ assets. In this case, the ratio N/M is enough higher to make covariance estimation with sample estimators not so reliable. In this sense, the estimation error is high, and the equally weighted strategy much overperforms all the alternatives. Ledoit and Wolf (2003) and Ledoit and Wolf (2004a) linear shrinkages, which work better more N/M is higher, provides better out-of-sample performance with respect to most alternatives. Nevertheless, our shrinkage estimator based on the shrinkage of c^* estimator of Kan and Zhou (2007) towards the Identity I improved upon both Ledoit & Wolf estimators. In a large dimensional setting, instead, shrinkage towards the market of Ledoit and Wolf (2003) is the best strategy. Nevertheless, precision shrinkage estimators that shrink (Moore-Penrose) inverse sample

estimators towards the implied equally weighted covariance matrix improve upon Ledoit and Wolf (2004a).

TABLE 1.10: Out of sample Sharpe ratios ($M = 60$)

Estimators	N=5	N=10	N=30	N=100
$1/N$	19,499%	19.788%	18.742%	15.867%
I	18.803%	10.548%	9.823%	6.755%
Sample Covariance	3.158%	-0.178%	6.119%	-8.223%*
Unbiased Precision Covariance	3.158%	-0.178%	6.119%	-8.223%*
Optimal $c^* = c_3$ (with $\theta^2 \to \infty$)	3.158%	-0.178%	6.119%	-8.223%*
Optimal c^* (with $\hat{\theta}_u^2$)	3.158%	-0.178%	6.119%	-8.223%*
Unbiased Precision towards I	3.158%	-0.178%	-0.641%	-8.223% *
Unbiased Precision towards EW	1.344%	10.384%	6.673%	1.330% *
Optimal $c^* = c_3$ towards I	3.158%	-0.178%	-6.641%	-8.223% *
Optimal $c^* = c_3$ towards EW	2.137%	8.936%	4.394%	1.330% *
Optimal c^* towards I	4.161%	-0.067%	2.814%	-1.634% *
Optimal c^* towards EW	-3.103%	0.335%	4.653%	-1.634% *
Ledoit and Wolf (2003)	-2.413%	4.511%	5.007%	3.991%
Ledoit and Wolf (2004a)	3.405%	-1.953%	-2.479%	1.291%

Note: Ledoit and Wolf (2003) is called "shrinkage towards the market" while Ledoit and Wolf (2004a) is the "shrinkage towards Identity". *Moore-Penrose inverses are used. Differently from other columns, the last one with $M = 60$ and $N = 100$ is defined as large dimensional setting.

Consider now the case with $N = 10$ assets. The N/M ratio is lower than the case analyzed before. In this setting, strategies based on sample estimators have negative out-of-sample Sharpe ratios, and the $1/N$ strategy still overperforms all the alternatives. Once again, linear shrinkage performs better than sample estimators. Nevertheless, precision shrinkage estimators return the highest out of sample performances. In particular, the "Unbiased Precision towards implied equally weighted covariance" can be identified as the best strategy. Therefore, also in this setting, the proposed shrinkage approach over-performs Ledoit and Wolf (2003).

When $N = 30$, we still have such ahigh estimation error that strategies without estimation givegive the highest sample performance. Despite that, our shrinkage approach is still better than the benchmark shrinkage of Ledoit & wolf. Particularly, Ledoit and Wolf (2003) over-performs several of the proposed shrinkage, but, again, the "Unbiased Precision towards implied equally weighted covariance" performs better.

Hence, from an empirical point of view, the proposed shrinkage improves upon Ledoit & Wolf, where the optimal shrinkage intensity is defined according to statistical criteria rather than economic. Determining shrinkage intensity from a mean-variance perspective allows our estimators of covariance inverses to overperform in a mean-variance asset allocation.

What happens if we increase the estimation window M? Results with $M = 120$ are reported in Tab. 1.11.

ł

TABLE 1.11: Out of sample Sharpe ratios ($M = 120$)

Estimators	N=5	N=10	N=30	N=100
$1/N$	22.612%	23.130%	21.321%	19.049%
I	22.667%	23.601%	23.783%	6.928%
Sample Covariance	26.305%	31.467%	4.852%	2.334%
Unbiased Precision Covariance	26.305%	31.467%	4.852%	2.334%
Optimal $c^* = c_3$ (with $\theta^2 \to \infty$)	26.305%	31.467%	4.852%	2.334%
Optimal c^* (with $\hat{\theta}_u^2$)	26.305%	31.467%	4.852%	2.334%
Unbiased Precision towards I	26.305%	31.467%	4.852%	13.138%
Unbiased Precision towards EW	26.619%	32.401%	19.166%	2.333%
Optimal $c^* = c_3$ towards I	26.305%	31.467%	4.852%	13.138%
Optimal $c^* = c_3$ towards EW	26.634%	32.481%	23.184%	2.322%
Optimal c^* towards I	23.927%	31.467%	4.852%	47.984%
Optimal c^* towards EW	28.296%	35.911%	51.409%	15.186%
Ledoit and Wolf (2003)	26.293%	30.421%	35.215%	4.455%
Ledoit and Wolf (2004a)	26.304%	31.535%	4.907%	2.132%

Note: Ledoit and Wolf (2003) is called "shrinkage towards the market" while Ledoit and Wolf (2004a) is the "shrinkage towards Identity".

Now estimation window is double than before. The first important result is that now $1/N$ is not always the best one. Particularly, for certain small N/M ratios it is not especially for $N = 5$ and $N = 10$. Instead, with an increasing number of assets, the ratio increases (with $N = 30$ and $N = 100$) and fixed window M again estimation error become so relevant that this biased strategy overperforms. The same conclusions can be made for plug-in of Identity matrix that does not require any estimation.

Another significant result confirms that our shrinkage estimators can perform better than Ledoit & Wolf shrinkages. With $N = 5$ and $N = 10$, Ledoit and Wolf (2003) estimator is even worst than sample estimators, despite being better than equally

weighted. Within these more ideal settings, we get the same result of simulations: shrinkage of c^* estimator towards equally weighted covariance matrix is overall the best strategy possible. This is because it reduces estimation error more than all the alternatives.

If we look at what happened with $N = 30$ or $N = 100$, it is easy to recognize that now sample estimators are again the worst one and, as we already said, strategies without estimation are better. Both Ledoit and Wolf (2003) and Ledoit and Wolf (2004a) improve upon sample estimators but are not able to reach the performances of our shrinkages.

1.4 Conclusions

The contributions of this chapter can be summarized as follows. First of all, we show that the application of shrinkage techniques reduces estimation error also in a low dimensional setting (when $T > N$) rather than just in the high dimensional one (when $N > T$). With this respect, we developed a new class of shrinkage estimators explicitly taught to deal with this issue, which we call *precision shrinkage*. The proposed estimators perform much better than the previous literature ones (e.g. the linear covariance shrinkage estimators of Ledoit and Wolf (2003), Ledoit and Wolf (2004a), and Ledoit and Wolf (2004b)).

The main advantage of the proposed precision shrinkage lies in how the optimal shrinkage intensity is determined. Indeed, while previous literature on shrinkage estimators derived the optimal shrinkage intensity through statistical arguments, we determine it consistently with the portfolio selection problem. Indeed, the proposed shrinkage intensity is obtained by maximizing the investors' preferences. Moreover, we have derived a general optimal shrinkage intensity formula valid for any kind of shrinkage. What we evaluated in this paper are just some possible specifications.

Another relevant contribution consists of deriving a closed formula for the estimation error in the presence of shrinkage estimation. Computing the estimation error formula is not easy for classical shrinkage estimators because of the difficulties in computing the inverse of two matrices weighted sum. we overcome this problem by shrinking the covariance matrix inverse (the *precision matrix*) directly; that is what matters for portfolio selection problems.

Further, we proposed an investment strategy that is a scaled version of the naive

rule. This strategy involves a two fund-rule where the mean-variance investor estimate the covariance matrix through the (1.9). This kind of mean-variance portfolio rule is more effective in reducing the estimation error within the proposed simulated economy with respect to the standard $1/N$ and, if optimally combined with the optimal scaling estimator of Kan and Zhou (2007), it results in the best plug-in possible. Moreover, we explicitly showed that within a simulated Gaussian i.i.d. economy, the equally weighted strategy is the best one. This result raises a possible explanation of empirical over performances of the naive strategy: the real world is complex, and within an economy, with several risk factors such easy diversification strategy works well, while it should not work under the theoretical assumptions underlying the Markowitz (1952) mean-variance framework. If investors can correctly identify all risk factors, they could over-perform the $1/N$ strategy. Further investigations in this direction are needed in future works.

Then, we also demonstrated that the investors could improve their utility within a simple i.i.d. economy by using the information from the equally weighted strategy. Indeed, shrinking the optimal scaling covariance estimators towards the implied equally weighted covariance is the best plug-in estimator among all the alternatives. In the end, we am the first in analyzing a completely new scenario under which no information at all is available about the actual parameters. we called the expected utility that investors get within this setting *perceived utility*. Interestingly, the worst strategy under an ideal world (i.e. maximum likelihood) in this setting is the best one from the investor's point of view. This raises a miss perception problem: any investor behaving rationally chooses the worst allocation strategy in the absence of information. The main conclusion we get is that the consequences of parameter uncertainty are even more severe than what we believed.

The last point to highlight regards the applicability of the precision shrinkage estimators in the minimum variance setting. Although some simulations show the usefulness of this proposal also under this alternative asset allocation rule, several computational problems limit its applicability in practice. Therefore, future studies are needed to overcome these issues.

Chapter 2

Timing asset allocation: model-based or data-driven forecasts?

2.1 Introduction

In the first Chapter we focused on the static implementation of portfolio rules, specifically both the Markowitz (1952) mean-variance (MV) and the Global Minimum Variance (GMV). Static implementation of these portfolio rules involve the estimation of current covariance structure among the assets and, in the case of mean-variance diversification, current expected returns' vector.
However, from asset allocation perspective, investors can improve the performances of their portfolios by statistically guessing the future covariance structure and expected returns. Indeed, as argued by DeMiguel, Garlappi, and Uppal (2009), if the first and second moments of returns vary over time and can be predicted, classical standard models may perform poorer than forecasts-based ones. However, because of the larger number of parameters that need to be estimated for the implementation of forecasting models, it is not clear if gains can be achieved in out-of-sample. Therefore, aim of this Chapter is to deeper understand the usefulness of forecasting in asset allocation.
The use of predicted rather than current quantities can be motivated by the concept of *market timing*, i.e. anticipating future market conditions. More in detail, market timing refers to the practice of predicting whether the market prices and volatility will rise or fall, and investing appropriately (Grant, 1978). It has been demonstrated that the anticipation of future market conditions generates higher performances in

out-of-sample (e.g. see Fleming, Kirby, and Ostdiek, 2001; Marquering and Verbeek, 2004; Kong et al., 2011; Almadi, Rapach, and Suri, 2014). Indeed, as demonstrated by Engle and Colacito (2006), the portfolio variance is minimised when the correct forecast is used to build the portfolio. Moreover, if investors can predict when the market will go up and down, they can make trades to turn that market move into a profit. Then, market timing enables traders to curtail the effects of market volatility and to reap the benefits of short-term price movements.

Since forecasts are used, returns and covariance predictability is an important issue for the empirical implementation of a timing strategy. The academic literature provide mixing findings regarding stock market predictability. In particular, while the prevalent literature of the 1970s argued that stock prices are appropriately described by a random walk, more recent empirical papers report the evidence that stock returns are to some extent predictable, either from their own past or by the use of other publicly available information (Breen, Glosten, and Jagannathan, 1989; Fama and Schwert, 1977; Fama and French, 1993; Fama and French, 2015). Recently Green, Hand, and Zhang (2013) identify more than 500 predictors, called *return predictive signals*. On the other hand, from the seminal papers of Engle (1982) and Schwert (1989), the evidence of predictability for volatility is generally consistent across a broad range of assets and econometric specifications, so literature agree on the idea that volatility is to some extent more predictable than returns.

Overall, because of financial markets' complexity, it is natural to see timing as a very complicated task. This is the main reason why static asset allocation is more often considered. In what follows we aim to show that using predicted quantities, obtained with *appropriate statistical models*, allows important improvements in portfolio out-of-sample performances for both MV and GMV portfolio rules. Hence, since the selection of an *appropriate statistical model* is crucial, the fundamental aim of this Chapter is to provide deeper insights about which kind of statistical models have to be preferred by investors for the implementation of timing strategies.

A first contribution of this Chapter is to assess if and how much timing is useful also in a large dimensional setting, since previous literature only focused on the standard low-dimensional one [1].

[1]The economic significance in a large dimensional setting is studies by considering an economy where the number of available assets N is higher than the time-span T. In the large dimensional setting it is well known that estimation error is more severe (Michaud, 1989; Ledoit and Wolf, 2003; Ledoit and Wolf, 2004b; Ledoit and Wolf, 2017)

Another, and perhaps the most important, novelty introduced in this chapter is to study if and to what extent machine learning, also defined as data-driven, approaches can be successfully used for the implementation of a timing strategy. With this respect, we compare, in terms of financial performances, the use of standard model-based approaches and novel machine learning tools for predicting returns and covariances. A recent literature (e.g. see Gu, Kelly, and Xiu, 2020; Götze, Gürtler, and Witowski, 2020; Bianchi, Büchner, and Tamoni, 2021) demonstrate the usefulness of machine learning (ML) techniques in many aspects of finance but it is still unclear if the predictions obtained with these techniques allow generating economic value to the investors.

By considering monthly returns, we show that, for both low and large-dimensional settings, model-based approaches dominate with respect those based on machine learning. Even if this result could seem surprising, it is not because the time series dimension with monthly frequencies is not enough to ensure a correct estimation of any ML-based technique. Therefore, we demonstrate that for this reason the investors that employ a timing and long-run asset allocation should use standard model-based (econometric) approaches rather than data-driven (ML) ones.

The portfolio performances are computed with the Sharpe ratio and the Certainty Equivalent (CEQ) return. However, following other authoritative studies (Fleming, Kirby, and Ostdiek, 2001; Fleming, Kirby, and Ostdiek, 2003; Marquering and Verbeek, 2004; Pesaran and Timmermann, 1995), questioning about the relevance of market timing means understanding whether predictability has an economic value. In the context of portfolio selection, the economic value is usually calculated in terms of gains in utility function. Hence, according to both MV and GMV diversification rules, we determine the economic value of machine learning by considering an investor with a quadratic utility function.

Considering an economy with many assets, we compare four optimal strategies: the first, a *static* investment strategy that does not employ timing; the second, a *return timing*, that uses forecasts from the returns only; the third, a *volatility timing*, that forecasts only the volatility; and the last one, a *full timing*, that uses both return and covariance forecasts to estimate portfolio weights. Note that in the case of GMV diversification the full-timing coincides with the volatility timing, while the return-timing does not exist.

The remainder of the chapter is organized as follows. Paragraph 2.2 introduces the

framework underlying the Chapter. Paragraph 2 describes the forecasting models used to build timing portfolios, distinguishing between classical model-based procedure and novel machine learning ones. Then, in the paragraph 2.4 are discussed data used for the analysis, while the results related to the economic significance of the alternative trading strategies are showed in paragraph 2.5 and 2.6. Finally, paragraph 2.8 concludes with a discussion of the results and some final remarks.

2.2 The underlying framework

2.2.1 Portfolio rules

As already stated in the Introduction, in what follows we consider both Markowitz (1952) mean-variance and Global Minimum Variance (GMV) diversification rules.
In the canonical mean-variance framework of Markowitz (1952) the investor faces an asset universe composed by N risky assets and a riskless asset. Particularly, he/she chooses the amount of wealth invested in the $i = 1, \dots, N$ risky assets that maximize his utility. At this aim, two main quantities are of interest: the mean of the asset's returns and their covariance matrix. Supposing to have a matrix $\mathbf{R}_t$ of N asset returns observed for T times, the static portfolio choice can be formalized as follows:

$$\max_{w} U(w) = w'\mu - \frac{\gamma}{2} w'\Sigma w \tag{2.1}$$

where w is the portfolio weights of the N risky assets, $w'\mu$ is the portfolio expected return and $w'\Sigma w$ the portfolio variance. The solution to problem (1) returns the optimal portfolio weight vector w^*:

$$w^* = \frac{1}{\gamma}\Sigma^{-1}\mu \tag{2.2}$$

Clearly, both Σ and μ are unknown and need to be estimated properly. The estimation phase rises the problem of estimation error, i.e. the estimated quantities are not close to the population ones. Moreover, as argued by Merton (1973), estimation error in the mean is more difficult to handle that the one due to the covariance estimation. Therefore, differently from classical mean-variance allocations, many scholars proposed to use the so-called Global Minimum Variance (GMV) diversification strategy,

that does not involve mean estimation. Assuming to have N risky assets in the investment universe, the GMV portfolio problem can we written as:

$$\min_{w} w'\Sigma w$$

$$s.t. \sum_{i=1}^{N} w_i = 1$$

The optimal global minimum variance weight w^* vector, as solution of the minimization problem above, is given by:

$$w^* = \frac{\Sigma^{-1}\mathbb{1}_N}{\mathbb{1}'_N\Sigma^{-1}\mathbb{1}_N} \tag{2.3}$$

where $\mathbb{1}_N = (1, 1, \ldots, 1)$ and $\mathbb{1}'_N\Sigma^{-1}\mathbb{1}_N$ is the sum of all elements within the vector $\Sigma^{-1}\mathbb{1}_N$. By replacing Σ^{-1} with $\hat{\Sigma}^{-1}$ we get the optimal *estimated* GMV portfolio weights that we call $\hat{w}$. As showed by Kourtis, Dotsis, and Markellos (2012), the estimation error associated to the GMV strategy is lower than those based on mean-variance because the estimation of mean vector is not considered.

2.2.2 Static estimation

Clearly, although it is optimal, the allocation (2.2) is not feasible because μ and Σ are unknown quantities and need to be estimated. A feasible static mean-variance portfolio in a given point in time t can be obtained by plug-in of some estimates, $\hat{\mu}$ and $\hat{\Sigma}$, such that:

$$\hat{w}_t = \frac{1}{\gamma}\hat{\Sigma}^{-1}\hat{\mu} \tag{2.4}$$

An easy way to obtain the feasible portfolio (2.4) is to replace $\hat{\mu}$ with the sample averages of asset returns and $\hat{\Sigma}^{-1}$ with some sample covariance estimators. The most common covariance estimator is its sample counterpart:

$$\hat{\Sigma}_{SC} = (T-1)^{-1}\sum_{t=1}^{T}(\mathbf{R}_t - \hat{\mu})(\mathbf{R}_t - \hat{\mu})' \tag{2.5}$$

However, the estimator (2.5) is highly affected by the estimation error (Kan and Zhou, 2007). Therefore, in order to account for estimation error explicitly, we propose to use the optimal *precision shrinkage* estimator proposed in the Chapter 1 of the thesis:

$$\hat{\Sigma}_{sP}^{-1} = \alpha^* \hat{\Sigma}_{c^*}^{-1} + (1 - \alpha^*)\hat{\Sigma}_{EW}^{-1} \tag{2.6}$$

that optimally shrinks the inverse of the covariance estimator $\hat{\Sigma}_{c^*}^{-1}$ developed by Kan and Zhou (2007) with the inverse of the implied covariance matrix of the equally weighted portfolio[2].

However, it is well known that strategies based on the plug-in of the sample covariance estimator[3] only works in low-dimensional settings, where $N < T$. As claimed before, an interesting aspect is to evaluate the usefulness of timing portfolios also in a large dimensional setting where $N > T$ or the *concentration ratio* N/T is such that usual estimators result in ill-conditioning covariance matrices. In other words, the matrix cannot be inverted and a feasible asset allocation cannot be obtained.

To obtain feasible static mean-variance and GMV portfolios when the dimension is large, we plug-in the shrinkage estimator of Ledoit and Wolf (2003) that shrinks the sample covariance towards the market. This estimation technique considerably reduce estimation error in this setting and lead to an invertible covariance estimator. Another useful estimator for large-dimensional setting is the POET (Fan, Liao, and Mincheva, 2013) obtained by applying thresholding in an approximate factor model.

2.2.3 Timing strategies

Clearly the aforementioned strategies are static, i.e. not based on timing. How can we obtain a *timing* portfolio?

In general, a market timing portfolio can be formed in a very simple way, e.g. investing the 100% of the wealth on stocks or bonds depending by some predictions (as in Pesaran and Timmermann, 1995) or trough more sophisticated approaches, based

[2]The estimator $\hat{\Sigma}_{sP}^{-1}$ is optimal because the shrinkage intensity α^* is chosen such that it maximize investor's expected utility and minimize the estimation error. The performances of the estimator $\hat{\Sigma}_{c^*}^{-1}$ are empirically equivalent to those of the sample covariance. However, if optimally combined, it results in superior performances.

[3]Note that the same applies also to the precision shrinkage estimator.

on Markowitz (1952) mean-variance or GMV optimization.

Pesaran and Timmermann (1995) assess the economic value of predictions considering the out of sample performances of a simple trading strategy based on the forecasts on market portfolio. This strategy, that employs the concept of market timing defined before, involves the decision of investing entirely in bonds if the return of the market portfolio is predicted to be negative in the next time period or investing entirely in the market vice versa. They find that the simple *returns timing* trading strategy generates an higher profit than a buy-and-hold strategy in the market portfolio. The two main limitations of Pesaran and Timmermann (1995) is that the authors consider an economy with only two assets and that they only predict the market returns, such that volatility is not considered.

By overcome this limitation, Fleming, Kirby, and Ostdiek (2001) and Fleming, Kirby, and Ostdiek (2003) provide evidence on the usefulness of volatility predictions in asset allocation, considering the out-of-sample performances of a *volatility timing* trading strategy. Moreover, they consider the Global Minimum Variance (GMV) portfolio problem with N risky assets. In order to measure the economic value of volatility predictions, they compare the performances of a timing portfolio where the covariance matrix is predicted following the approach of Foster and Nelson (1996) with a static strategy that employs the static sample covariance estimator.

Another important contribution on the topic is due to Marquering and Verbeek (2004) that study the combined effect of returns and volatility predictions. Nevertheless, as in Pesaran and Timmermann (1995), they consider an economy where the market portfolio and a bond are the only assets available.

In what follows, as in Fleming, Kirby, and Ostdiek (2001), we study the usefulness of timing in an economy where N risky assets are available. However, since both Markowitz (1952) MV and GMV diversification rules are involved, we are able to disentangle, as in Marquering and Verbeek (2004), the economic value of forecasting return, volatility or both.

In particular, we define a MV *full timing* portfolio as the one obtained by plug-in of forecasts for both covariances and means, such that the today's weights $\hat{w}_t$ are equal to:

$$\hat{w}_t\left(\mu_{t+1}, \Sigma_{t+1}\right) = \frac{1}{\gamma}\hat{\Sigma}_{t+1}^{-1}\hat{\mu}_{t+1} \tag{2.7}$$

where $\hat{\Sigma}_{t+1}$ and $\hat{\mu}_{t+1}$ are the forecasts for the covariance and expected returns vector, respectively. Then, a simple MV *return timing* trading rule can be obtained by employing static covariance estimation and forecasts only for the returns vector:

$$\hat{w}_t\left(\mu_{t+1}, \Sigma\right) = \frac{1}{\gamma}\hat{\Sigma}^{-1}\hat{\mu}_{t+1} \tag{2.8}$$

where $\hat{\mu}_{t+1}$ represents a forecast of the return vector at times $t+1$ and $\hat{\Sigma}^{-1}$ be a static sample covariance estimator . Similarly, a MV *volatility timing* trading rule can be obtained by means of:

$$\hat{w}_t\left(\mu, \Sigma_{t+1}\right) = \frac{1}{\gamma}\hat{\Sigma}_{t+1}^{-1}\hat{\mu} \tag{2.9}$$

where $\hat{\Sigma}_{t+1}^{-1}$ is the predicted covariance matrix at time $t+1$ and $\hat{\mu}$ is computed with the sample averages.

The timing strategies (2.7), (2.9) and (2.8) represent a novelty introduced by this thesis. Then, the Fleming, Kirby, and Ostdiek (2001) timing approach can be defined as:

$$w_t\left(\hat{\Sigma}_{t+1}\right) = \frac{\hat{\Sigma}_{t+1}^{-1}\mathbb{1}_N}{\mathbb{1}_N'\hat{\Sigma}_{t+1}^{-1}\mathbb{1}_N} \tag{2.10}$$

where $\hat{\Sigma}_{t+1}^{-1}$ is the inverse of the predicted covariance matrix at time $t+1$. In this case, the volatility-timing coincides with the full-timing and the return-timing approach does not exist.

Clearly, we would get favorable evidence of economic significance for timing portfolios if the performances associated to these strategies are superior than those of static portfolios. Therefore it is crucial to discuss how the portfolio's economic performance is evaluated.

2.2.4 Economic evaluation

A first approach is to consider traditional measures of financial performance. To evaluate whether a trading rule outperforms another one, the investors usually compute the Sharpe ratios (SR). The Sharpe ratio of a given portfolio p is defined as the ratio between the mean excess return on the portfolio p, μ_p and its standard deviation σ_p:

$$SR_p = \frac{\mu_p}{\sigma_p}$$

If the Sharpe ratio of a given portfolio $p = a$ exceeds the one of an alternative portfolio $p = b$, we say that portfolio a is more attractive than b. Since Sharpe ratio increases with increasing portfolio returns and decreases with increasing portfolio variance it consistent with a mean-variance analysis. However, the risk of the timing strategies is typically overestimated by the sample standard deviation, because the ex post unconditional standard deviation is an inappropriate measure for the conditional risk that an investor faces at each point in time (Marquering and Verbeek, 2004). Therefore, it is clear that the Sharpe ratio is not always the most appropriate measure if the aim is to compare static strategies with forecasts-based ones based on timing.

An alternative is to account for the preferences of the investor in evaluating strategy's performance. Assuming a mean-variance investor, it is possible to calculate the certainty-equivalent (CEQ) return, defined as the riskfree rate that an investor is willing to accept rather than adopting a particular risky portfolio strategy. The CEQ return can be computed as:

$$CEQ_p = \mu_p - \frac{\gamma}{2}\sigma_p \tag{2.11}$$

Another approach has been suggested by Fleming, Kirby, and Ostdiek (2001) and Fleming, Kirby, and Ostdiek (2003) and Marquering and Verbeek (2004). Indeed, we can determine for any given trading rule the economic value of timing portfolios by calculating the maximum fee Δ, in percent per month, that the investor should be willing to pay for holding the timing portfolio rather than a static one. It could done by equating the average utilities of the compared strategies. In the case of minimum

variance portfolio choice, Fleming, Kirby, and Ostdiek (2001) and Fleming, Kirby, and Ostdiek (2003) proposed to equate quadratic utilities:

$$\frac{1}{T}\sum_{t=1}^{T}(r_{a,t}-\Delta)-\frac{\gamma}{2(1+\gamma)}(r_{a,t}-\Delta)^2=\frac{1}{T}\sum_{t=1}^{T}r_{b,t}-\frac{\gamma}{2(1+\gamma)}r_{b,t}^2 \tag{2.12}$$

where the indices a,t and b,t refers to two different strategies estimated in a given point in time t and $r_{a,t}$ is the return of the portfolio a. Such approximation is possible considering that a quadratic utility function is a second-order approximation to the investor's true utility. However, since we are assuming a mean-variance investor, to compute Δ we consider as in Marquering and Verbeek (2004) a slightly different version of the (2.12):

$$\frac{1}{T}\sum_{t=1}^{T}(r_{a,t}-\Delta)-\frac{\gamma}{2}(r_{a,t}-\Delta)^2=\frac{1}{T}\sum_{t=1}^{T}r_{b,t}-\frac{\gamma}{2}r_{b,t}^2 \tag{2.13}$$

Clearly this formulation allows also the comparison between both alternative static and alternative timing strategies. Note that we use the same (2.13) also in the case which the mean-variance investor uses a GMV diversification rule. In doing so, we assume that the investor implements a GMV only to reduce the effect of the estimation error contained in the return vector.

In the end, it is well known that naive strategy usually overperforms any kind of optimal one (De Miguel, Garlappi, and Uppal, 2007) because of the estimation error. Hence, we also compute, for each strategy, the return-loss with respect to the 1/N strategy. The return-loss is defined as the additional return needed for strategy p to perform as well as the 1/N strategy in terms of the Sharpe ratio. To compute the return-loss per month, we consider μ_{ew} and σ_{ew}, the monthly out-of-sample mean and volatility of the net returns from the 1/N strategy, and μ_p and σ_p, i.e. the corresponding quantities for strategy p. Then, the return-loss from strategy p is:

$$RL_p=\frac{\mu_{ew}}{\sigma_{ew}}\times\sigma_p-\mu_p \tag{2.14}$$

Empirically, in order to compute the performances measures, we use the following

rolling-window approach. We first consider an initial estimation window equal to M. Such M observations are used to estimate parameters $\hat{\mu}$ and $\hat{\Sigma}^{-1}$ in the case of static portfolio choices at $t = M$ or are considered as the sample size of the information set used to make forecasts in the case of timing strategies. In the next time period $t = M + 1$, before to rebalance the portfolio, we can assess its return and variance at time $t = M$:

$$r_{p,t} = \hat{w}_{p,t} r_{p,t+1}$$

At the end of this rolling-window approach we obtain, as in as De Miguel, Garlappi, and Uppal (2007), a time series of $T - M$ observations of portfolio returns $r_{p,t}$ for each portfolio strategy p that we use in computing the Sharpe ratio, the CEQ return and the Δ.

2.3 Data-driven timing with machine learning

As previously stated, Fleming, Kirby, and Ostdiek (2001) consider the Foster and Nelson (1996) approach in forecasting volatility, based on rolling variance estimation. Then, Marquering and Verbeek (2004) use a similarly simple approach, where the forecasts on the single-asset volatility are obtained by regressing the adjusted rolling estimator of variability on a set of macroeconomic variables. Then, in Marquering and Verbeek (2004) the forecasts for the expected market returns are obtained by adopting standard model-based approaches suggested by the market premium forecasting literature (for an overview see Rapach and Zhou, 2013).
Nevertheless, forecasting the entire covariance structure requires multivariate statistical models. A very common model-based approach taught with the aim of predicting the future covariance structure is the Dynamic Conditional Correlation developed by Engle (2002). Fleming, Kirby, and Ostdiek (2001) do not consider DCC because not yet popularized, while Marquering and Verbeek (2004) explicitly state that "*these techniques were certainly not available to investors in the major part of the sample*". Nevertheless, the DCC-type of models, that are multivariate extensions of the GARCH processes, are commonly used by investors worldwide because of their high performances in terms of forecasting accuracy, computational speed and simplicity. Indeed, the great advantage of the model-based procedures is due to their

simplicity, since these models are based on clear and specific economic intuitions and can be explained by the financial theory. For this reason they also have a great degree of interpretability (for a deeper discussion about the motivation and usefulness of GARCH-type processes see the Appendix B).

Nevertheless, all the model-based procedures have to deal with possible misspecification. The perverse effects induced by the use of misspecified models in forecasting are well known (e.g. see Davies and Newbold, 1980 and Chatfield, 1996). Indeed, if the forecasts are obtained on the basis of a misspecified model, that is subject to estimation error, the prediction error dramatically increases (Patton, 2020). Further, in the case of financial time series, the presence of features like non linearities and long memory in stock returns and volatilities makes the forecasts obtained by simple model-based procedures often unreliable. Choosing the right statistical model is important as well as the selection of the predictors to be used. Indeed, choosing the wrong predictors can be seen as an additional source of misspecification (Geweke and Amisano, 2012).

Nowadays, the economists face a data-rich environment where many variables can be used in forecasting returns and volatilities. However, not all these variables are necessarily relevant. As a result, the econometricians have to choose the most important predictors by looking at the literature or by following established economic theories. Therefore, it is evident that the implementation of a timing strategy would potentially benefit from the employment of data-driven procedures.

A very recent literature (e.g. see Gu, Kelly, and Xiu, 2020; Götze, Gürtler, and Witowski, 2020; Bianchi, Büchner, and Tamoni, 2021) demonstrated the suitability of machine learning (ML) techniques in overcoming the aforementioned issues. Actually machine learning and deep learning techniques have a long history in the statistics' community. As argued by Mullainathan and Spiess (2017), an explanation of this slow adoption of ML techniques by the economics community lies on the fact that these methods are not suitable for structural analysis while they are explicitly taught for prediction.

In what follows, we first present the classical model-based approaches and, then, we discuss in detail the machine learning approaches that can be used to forecast returns and volatilities in the financial market with the aim of obtaining a timing strategy.

2.3.1 Classical model-based approaches

Traditionally, returns and volatility forecasts are obtained by means of model-based approaches. Clear examples are the CAPM for predicting excess returns and the GARCH (or the DCC in the multivariate case) for volatilities.

The *return timing* trading rule can be obtained by expected returns forecasting. With this respect, we consider two simple alternatives. The first one is only based on statistical arguments, where the one step ahead forecast is obtained by a generic AR(1) process applied to each i-th stock in the portfolio. On the other hand, we call the alternative *economic model* that is based on excess return forecasting literature[4]. Factor models are commonly used at this aim. According to the multi-factor model, the excess return of any asset $r_{i,t}$ over the risk free rate satisfies:

$$r_{i,t} = \alpha_i + b_{i1} f_{1,t} + \cdots + b_{i,K} f_{K,t} + \epsilon_{i,t} \tag{2.15}$$

Where K are the number of factors, f the factors themselves and b_{ij} are the parameters associated with the factors with ϵ_i the idiosyncratic error associated with the return i, orthogonal to the K factors. The model presented in (2.15) is a general one, allowing for multiple, possibly observed, factors. Nevertheless, a factor model can also be based on a single factor. A widespread single-factor model is the CAPM, where it is supposed that the stock's excess return $r_{i,t}$ depends by the market return m_t:

$$r_{i,t} = \beta_i m_t + \epsilon_{i,t} \tag{2.16}$$

Then, according to the market premium forecasting literature (e.g. see Stambaugh, 1999; Ang and Bekaert, 2007; Campbell and Thompson, 2008), we forecast the market premium as follows:

$$m_t = \delta \mathbf{X}_t + \eta_t \tag{2.17}$$

where $\mathbf{X}_t$ is a collection of covariates useful in forecasting m_t. Then we obtain a

[4] for an overview see (Rapach and Zhou, 2013)

forecast for $r_{i,t}$ with the following three step procedure:

1. Estimate $\hat{\beta}_i$ from (2.16)
2. Forecast market premium $\hat{m}_{t+1}$ with (2.17)
3. Forecast i-th return as: $\hat{r}_{i,t+1} = \hat{\beta}_i \hat{m}_{t+1}$ (2.18)

On the other side, the *volatility timing* trading rule is based on some forecasts of the covariance matrix. The easiest way to forecast a covariance matrix is to assume that all of the off-diagonal elements in Σ_{t+1}, the covariances, are restricted to be zero and to model each diagonal entries, the volatilities, with a GARCH(1,1) process.
However, it is natural to expect that true covariances are different from zero. The conditional correlation models are nowadays the most used to forecast covariance matrices. The Dynamic Conditional Correlation model of Engle (2002) separates the estimation of the volatility and the correlation, by estimating several univariate GARCH models for volatilities and many GARCH models for correlations. In this way the model can be applied to a large set of time series with a small computational effort. The DCC takes adavantage of the decomposition of the conditional covariance matrix Σ_t into conditional standard deviations $\mathbf{D}_t$ and conditional correlations Γ_t:

$$\Sigma_t = \mathbf{D}_t \Gamma_t \mathbf{D}_t \tag{2.19}$$

Let define $\mathbf{E}_t = \hat{\Sigma}^{-1/2} \mathbf{R}_t$ the matrix of devolatilized returns, such that covariance of $\mathbf{E}_t$ is equal to the identity matrix $\mathbf{I}_N$. Then, the DCC(1,1), model of Engle (2002) can be specified as follows:

$$\mathbf{Q}_t = (\mathbf{I}_N - \mathbf{A} - \mathbf{B}) \circ \bar{\mathbf{Q}} + \mathbf{A} \circ (\mathbf{E}_{t-1} \mathbf{E}'_{t-1}) + \mathbf{B} \circ \mathbf{Q}_{t-1} \tag{2.20}$$

$$\Gamma_t = \text{Diag}\,(\mathbf{Q})_t^{-0.5}\, \mathbf{Q}_t \text{Diag}\,(\mathbf{Q})_t^{-0.5} \tag{2.21}$$

where $\mathbf{Q}_t$ is the *pseudo*-conditional correlation matrix, $\bar{\mathbf{Q}}$ is the long-run correlation

matrix usually computed with sample estimator, the symbol $\circ$ represents the Hadarmand product, and Γ_t is the conditional correlation matrix. The matrix $\mathbf{Q}_t$ is called *pseudo*-correlation because the element on its main diagonal $q_{ii,t}$ are close but not equal to 1 since devolatilized returns are used as input. Therefore, the correction in (2.21) ensures that the diagonal element of Γ_t are exactly equal to 1. In the end, the off-diagonal elements of $\mathbf{Q}_t$, called $q_{ij,t}$, are modeled as GARCH-type processes:

$$q_{ij,t} = (1 - \alpha - \beta)\bar{\rho}_{ij} + \alpha e_{i,t-1}e_{j,t-1} + \beta q_{ij,t-1} \tag{2.22}$$

with $e_{i,t}$ the i-th column of $\mathbf{E}_t$ and $\bar{\rho}_{ij}$ the unconditional correlation between $e_{i,t}$ and $e_{j,t}$. Parameter estimation could be done by maximum likelihood. The the DCC model achieves parsimony in the dynamics of conditional correlations and maintains enough simple the estimation process.

Alternatively, we can follow the idea that each time series $r_{i,t}$ is explained by some K uncorrelated unobserved factors $y_{k,t}$, normalized to have unit variance. In other words, we are considering an *approximate* factor structure for stock returns such that:

$$\mathbf{R}_t = \mathbf{Z}\mathbf{Y}_t \quad \mathbf{Y}_t \sim \mathcal{N}(0, \mathbf{H}_t) \tag{2.23}$$

where $\mathbf{Z}$, the linear map that links the unobserved components with the observed variables, is constant over time and invertible and $\mathbf{H}_t$, the covariance matrix of the unobserved component, is diagonal such that each element $h_{k,t}$ can be written as a GARCH(1,1) process:

$$h_{k,t} = (1 - \alpha_k - \beta_k) + \alpha_k y_{k,t-1}^2 + \beta_k h_{k,t-1} \tag{2.24}$$

The conditional covariance of $\mathbf{R}_t$ is then computed as follows:

$$\Sigma_t = \mathbf{Z}\mathbf{H}_t\mathbf{Z}' \tag{2.25}$$

This is the idea behind the Generalized Orthogonal GARCH (GO-GARCH) of Weide

(2002), where the parameter estimation is done by maximum likelihood. A faster estimation can be reached by method of moments as showed in Boswijk and Weide (2011).

In a *large dimensional* setting it is not possible to compute parameters in the Dynamic Conditional Correlation since in the likelihood function appears the inverse of the long-run covariance. If we shall to use the Dynamic Conditional Correlation model, we have to find the way to make its estimation feasible in large dimension.

Hafner and Reznikova (2012) proposed to use the Ledoit and Wolf (2004a) shrinkage estimator for $\bar{\mathbf{Q}}$. As alternative, Pakel et al. (2017) proposed to use a *composite likelihood* approach to reduce the course of dimensionality. In the composite likelihood approach pairs of assets are used to estimate parameters of DCC. Therefore, rather than using the full N-dimensional Gaussian likelihood, Pakel et al. (2017) use all pairwise likelihoods to construct a composite likelihood. In the end, Engle, Ledoit, and Wolf (2019) proposed to combine both aforementioned approaches, using the non-linear shrinkage estimator of Ledoit and Wolf (2017) for $\bar{\mathbf{Q}}$ together with the composite likelihood approach. Because of its relative simplicity, in what follows we use the DCC approach of Hafner and Reznikova (2012) to make forecasts in large dimension.

2.3.2 Neural networks: generalities

The ML algorithms automatically find patterns in a large amount of data with the aim of predicting future evolution of the phenomenon under consideration. Among the many ML algorithms available, the Artificial Neural Networks (ANN) have been recognized as the most effective in predicting financial returns by several studies[5].

There are many reasons why the ANN are useful in forecasting (Zhang, Patuwo, and Hu, 1998). First of all, the ANN are based on few a priori assumptions about the model underlying the time series to be predicted. Differently from the traditional model-based approaches, the ANN learn from the experience and capture suitable functional relationships among the data even if the underlying relationship is unknown. Therefore, the ANNs are well suited for problems whose solutions require knowledge that is difficult to specify but for which there are enough data to train the network. Second, ANNs can generalize. After learning from the data, the ANNs

[5]For example see Hill, O'Connor, and Remus, 1996; Kaastra and Boyd, 1996; Enke and Thawornwong, 2005; Ahmed et al., 2010; Wang et al., 2011; Rather, Agarwal, and Sastry, 2015; Hsu et al., 2016 just to mention few examples

can often correctly infer the unseen part of a population even if the sample data contain noisy information. Third, and perhaps one of the most important, the ANNs are universal functional approximators (Hornik, Stinchcombe, and White, 1989; Hornik, 1991). It has been shown that a network can approximate any continuous function to any desired accuracy. Finally, the ANNs are nonlinear. This fact makes them very well suited for predicting stock market quantities such as returns and volatilities where strong non-linear relationships are present.

Typically, any ANN consists of three layers: an input layer, which contains the input variables, one or more hidden layers, and an output layer, with one or more output variables. The number of neurons in each layer and the number of layers are the hyperparameters of the network. In general, each neuron in the hidden layers has input, weight and bias terms. In addition, each neuron has a nonlinear activation function, which produces a cumulative output of the preceding neurons. Commonly used activation functions are the a logistic or hyperbolic tangent function, and allow to introduce nonlinearity. The weights are trained from the data by minimizing the mean squared error, usually trough a backpropagation algorithm (BP). Once the input and the output vectors are read by the BP algorithm, the training starts with random weights. After calculating the mean squared error between the observed and the predicted output, the network adjusts the parameters with the aim of reducing the error until there is no further improvement.

In econometrics, a significant part of the model specification consists in identifying the explanatory variables and the number of lags leading the most accurate forecasts. When constructing a neural network, the overall task is much longer, because the process does not only involve the choice of inputs, but also the identification of the network architecture. Firstly, a researcher that aims to construct a NN should choose the type of network to implement: feed-forward or recurrent. In feed-forward neural network (FNN), the information moves forward from a layer to the next one. Assuming a single hidden layer architecture, a FNN with multiple outputs can be depicted as in Figure 2.1.

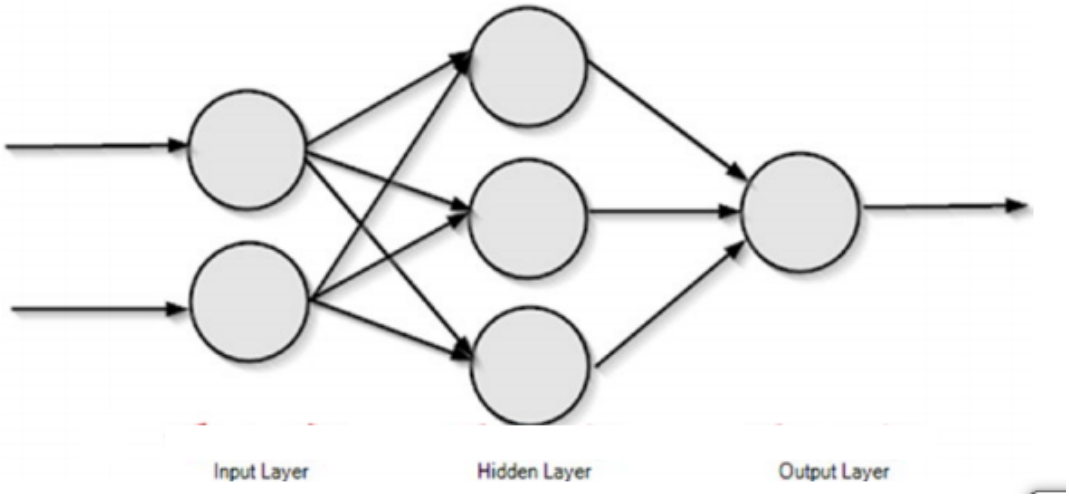


FIGURE 2.1: Example of FNN with $K = 1$ target variable, 1 hidden layer with 3 nodes and 2 input layers.

The output function of the k-th target variable can be specified as follows:

$$\hat{y}_{k,t} = F\left(\beta_{0,k} + \sum_{j=1}^{q} G\left(\delta_j \mathbf{X}_t\right)\beta_{j,k}\right) \tag{2.26}$$

where $G(\cdot)$ is the hidden layer activation function, $\beta_{j,k}$ is the weight from the j-th hidden unit to the output unit, $\mathbf{X}_t = \{1, x_{1,t}, \ldots, x_{m,t}\}$ is the $t \times m$ vector of input variables at time t, $\beta_{0,k}$ is the bias of the k-th output unit, $\delta_j = \{\delta_{1,j}, \ldots, \delta_{m,j}\}$ is the vector of weights of dimension $1 \times m$ connecting the input variables and the j-th hidden neuron and q is the number of hidden units. Usually, $F(\cdot)$ is an identity activation function such that $F(a) = a$ but also the logistic or other functions can be choosen. In the case of Identity activation the equation (2.26) takes the following form:

$$\hat{y}_{k,t} = \beta_{0,k} + \sum_{j=1}^{q} G\left(\delta_j \mathbf{X}_t\right)\beta_{j,k} \tag{2.27}$$

In order to replicate the activation state of a biological neuron, all neural networks use nonlinear activation functions at some point, usually through $G(\cdot)$. An ideal activation function should be continuous and differentiable to implement the back-propagation algorithm. The most common choice is the logistic function, but again this choice depends by the kind of the problem to be faced. In the case of logistic, we have a function bounded between 0 and 1, where a value closer to 0 (1) equals to

a low (high) activation level, and can be specified as follows:

$$G(x) = \frac{1}{1 + e^{-x}} \tag{2.28}$$

However, in some cases can be more appropriate to consider an activation function bounded to be bounded between -1 and 1. In this case we have the tangent hyperbolic (tanh):

$$G(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.29}$$

Lastly, another interesting activation function is the Rectified Linear Unit (ReLU):

$$G(x) = \{x = 0 \quad \text{if} \quad x \leq 0\} \quad \text{or} \quad \{x = x \quad \text{if} \quad x > 0\} \tag{2.30}$$

if we want an activation function that returns non-negative values.
The FNN can be seen as a static neural network without any autoregressive or lagged effect. However, if one aims to introduce a lagged effect he/she has to use a recursive neural network (RNN). Indeed, the prediction obtained by a RNN is dependent on the values of the previous time periods (Eliasy and Przychodzen, 2020). Therefore, in practical applications RNNs are more appropriate than FNNs in forecasting nonlinear time series. The difference between a FNN and a RNN is showed in Fig. 2.2

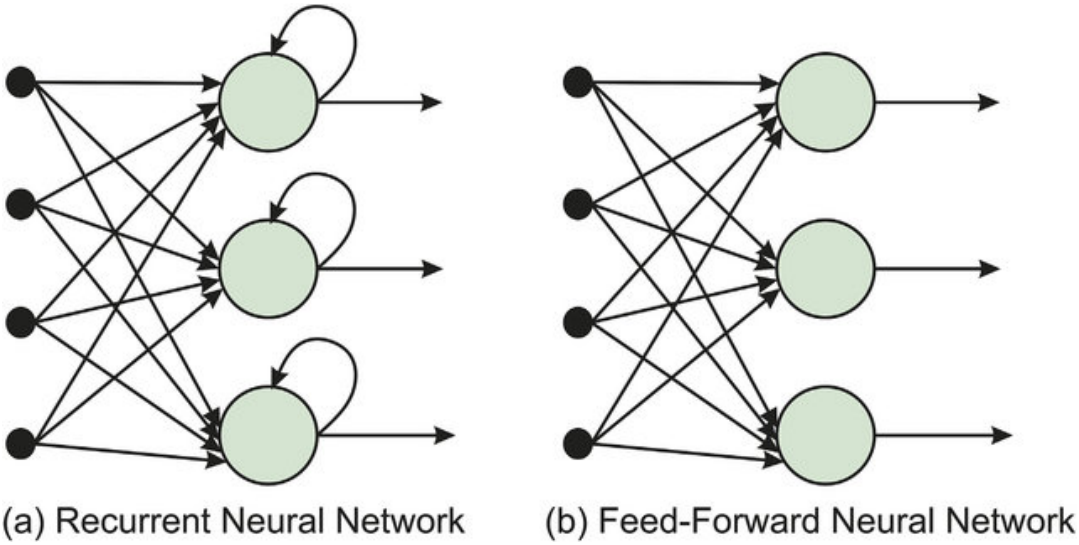


FIGURE 2.2: Comparison between RNN and FNN (Eliasy and Przychodzen, 2020)

There are several ways of implementing a RNN. For example, in the Elman (1990) network the hidden nodes with a time delay are used as additional input neurons. Assuming an identity output function $F(\cdot)$ and a single lag for the hidden nodes, the output of the Elman (1990) network with multiple outputs can be written as follows:

$$\hat{y}_{k,t} = \beta_{0,k} + \sum_{j=1}^{q} h_{t,j}\beta_{j,k} \tag{2.31}$$

with:

$$h_{t,j} = G\left(\delta_j \mathbf{X}_t + \gamma_j h_{t-1}\right) \quad j = 1, \ldots, q \tag{2.32}$$

where h_{t-1} is the vector of lagged hidden units and γ_j is the vector of connection weights between the j-th hidden node and the lagged hidden nodes. Another possibility is the Jordan (1997) network. The Jordan (1997) network exhibits a feedback from the output to the input layer. The lagged output units are then used as additional neurons such that:

$$\hat{y}_{k,t} = \beta_{0,k} + \sum_{j=1}^{q} G\left(\delta_j \mathbf{X}_t + \psi_{j,k}\hat{y}_{k,t-1}\right)\beta_{j,k} \tag{2.33}$$

where $\psi_{j,k}$ is the weight between the lagged output and the j-th hidden unit for the

target variable k.

2.3.3 NN-based return timing

Essentially the implementation of an NN-based timing strategy involves forecasts on the returns' vector obtained with a NN. Usually in econometrics the most important step lies on the selection of the variables to be used in forecasting a given quantity. If we aim to obtain forecasts within a neural network framework, things get more complicated. Indeed, in order to implement a NN is important to choose not only the output and the variables (i.e. the input layers) but also the entire architecture of the network. First of all, the kind of NN has to be chosen. Moreover, the network architecture essentially regards also the choice of the number of hidden nodes and layers. About the first aspect, it is well known that an architecture with a single-hidden layer is sufficient to approximate a wide range of nonlinear functions (Xiang, Ding, and Lee, 2005). The second aspect, i.e. the choice of the number of nodes, is perhaps the most tricky since there is no a theoretical basis to determine the appropriate number of hidden layers or nodes in a network. On one side, considering a small number of hidden units can be not enough in detecting complex nonlinear patterns in the data. On the other side, considering a too large number of hidden units may lead to overfitted out-of-sample forecast and dramatically increases the computational complexity of the network.

In what follows, we propose a network architecture based on economic intuitions. First of all, we consider a feed-forward structure of the neural network. This choice can be justified by the fact that the FNN has similarities with the CAPM models, showed in (2.16). Indeed, let us consider the i-th stock return as the single output layer of the network. The similarity can appear evident if we consider the predictive signals $\mathbf{X}_t$ as the input layers. Remember that $\mathbf{X}_t$ can be used to predict m_t that forecasts r_t. In other words, a transformation of the input variables generate forecasts of the output such that $f(\mathbf{X}_t) \rightarrow r_{t+1}$. Therefore, we can consider an hidden layer with a single node, where nonlinearities in $f(\cdot)$ are accounted by the activation function $G(\cdot)$. Then, previous values of the transformed predictive signals $\mathbf{X}_t$ are used as additional predictors. The FNN structure is displayed in Fig. 2.3 in the case of single stock return.

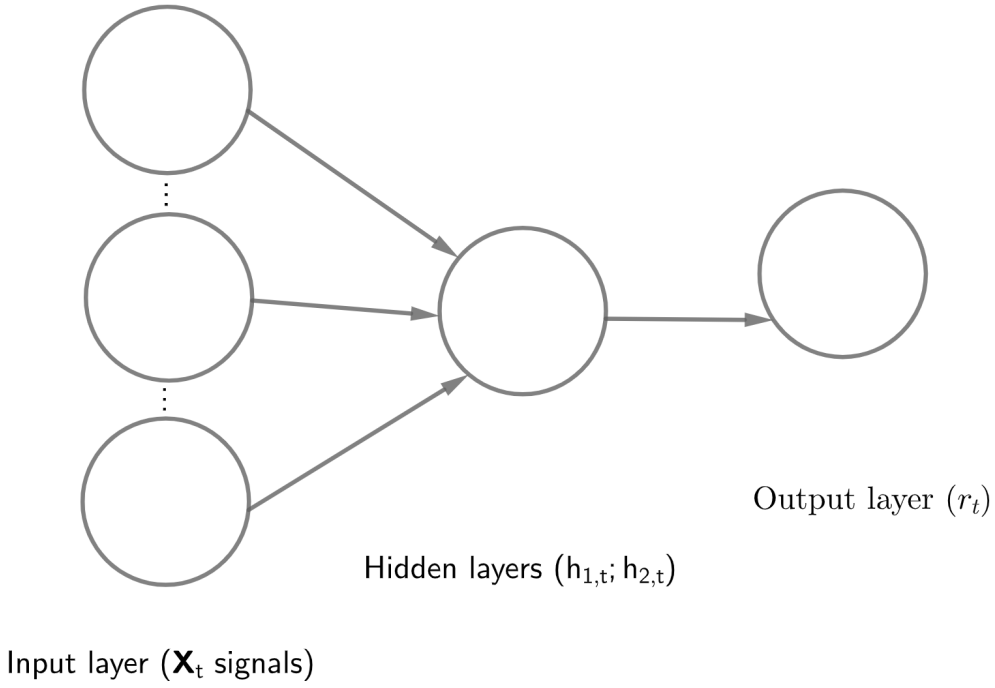


FIGURE 2.3: FNN in the case of single stock returns forecasting.

To better understand the parallelism between the architecture in Fig. 2.5 and the equation (2.16) we can start considering the following relationship (2.17):

$$m_t = \delta \mathbf{X}_t + \eta_t$$

that states that the market premium can be predicted by a set of $\mathbf{X}_t$ variables, defined *predictive signals*. By replacing it within (2.16):

$$r_t = \beta \left(\delta \mathbf{X}_t + \eta_t \right) + \epsilon_t$$

Let us define $h_t = \delta \mathbf{X}_t + \eta_t$. Hence we can write:

$$r_t = \beta h_t + \epsilon_t$$

where $\delta \mathbf{X}_t$ is equal to the predicted market risk premium $\hat{m}_t$. With the introduction of the bias term and the usage of an activation function such that $h_t = G(h_t)$, we

obtain the following prediction:

$$\hat{r}_t = b_0 + \beta G\left(\delta \mathbf{X}_t\right) \tag{2.34}$$

that looks like the one obtained with the FNN (2.27) in the case of a single hidden layer with one node. As activation function, since returns can assume any value between $(-\infty, \infty)$ we adopt an Identity activation. However, there are several ways of improving the presented network architecture in order to fully exploit the potentialities of ANNs. First of all we aim to forecast the returns within the same network hence, within a multivariate setting. Second, it is possible to specify more than 1 hidden nodes. The example of 2 hidden nodes is showed in Fig. 2.4.

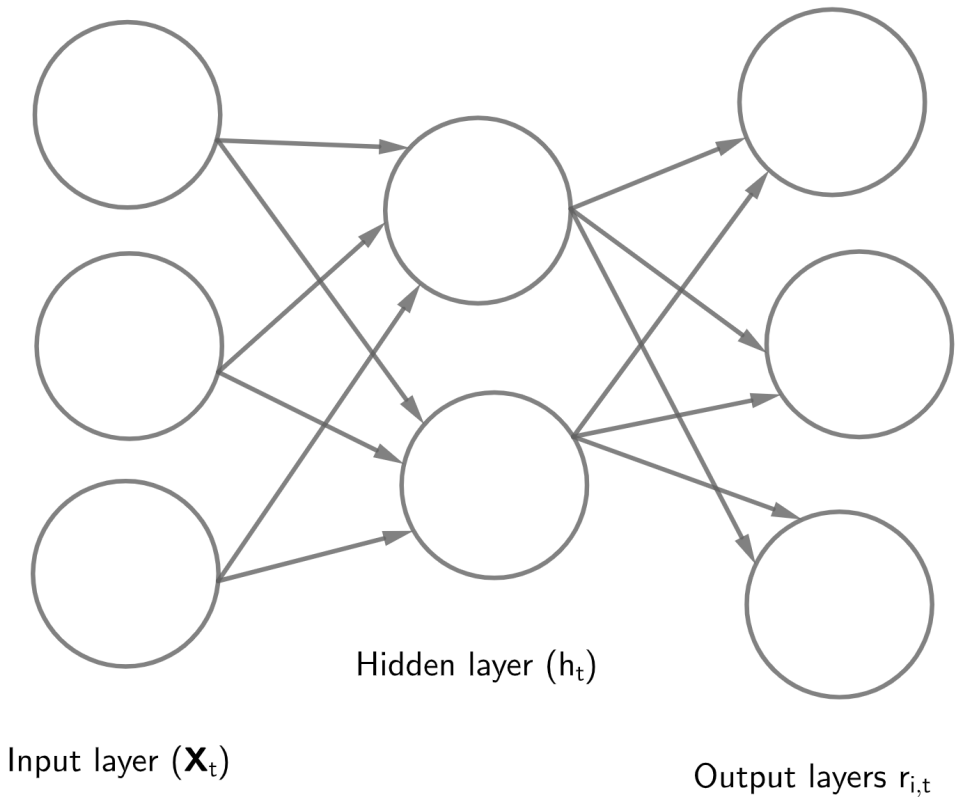


FIGURE 2.4: Implemented FNN for returns' forecasting with 2 hidden nodes.

Another improvement can be the consideration of an Elman (1990) recurrent neural network (RNN) structure. Indeed a FNN is not well suited for forecasting time series since it ignores the temporal order within the inputs and every new input is

considered in isolation (Hewamalage, Bergmeir, and Bandara, 2021). On the other side, the RNN incorporates time patterns in the network (see Fig. 2.5).

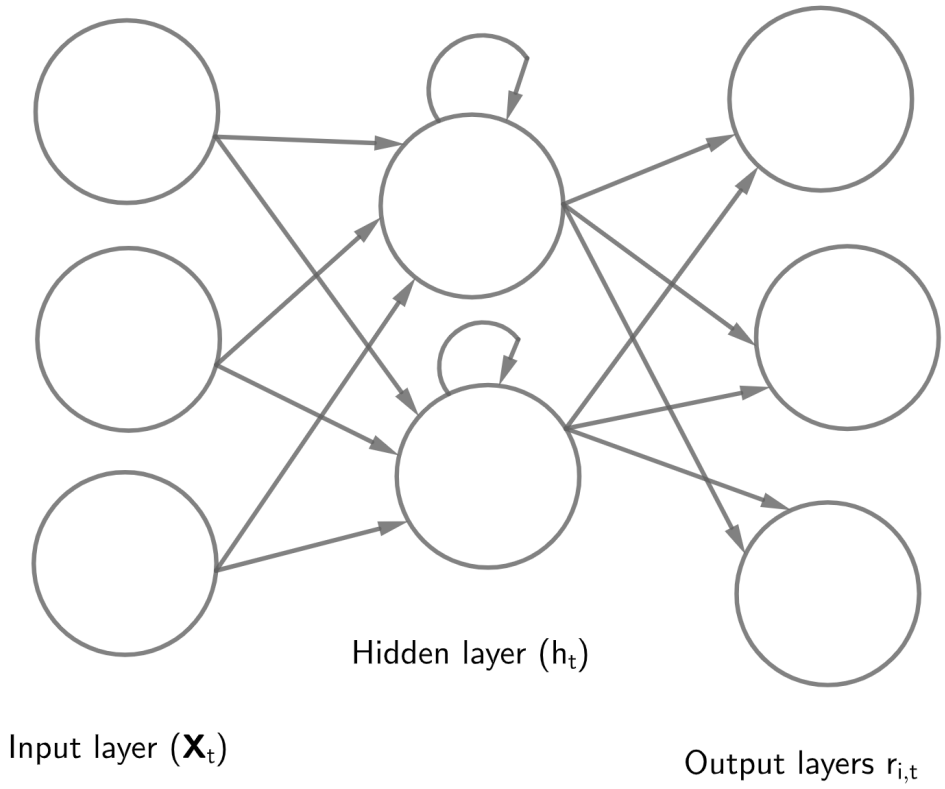


FIGURE 2.5: RNN (Elman, 1990) for returns' forecasting: example with 2 hidden nodes.

The main difference with the previous architecture lies on the presence of the additional lagged term h_{t-1} such that $h_t = h_t + h_{t-1}$. In this case of a single hidden node we obtain:

$$\hat{r}_t = b_0 + \beta G\left(\delta \mathbf{X}_t + h_{t-1}\right)$$

The introduction of h_{t-1} means that previous values of the inputs are able to explain future values of the returns. Clearly, as shown in Fig. 2.5, also in this case more than one node in the hidden layer can be considered.

2.3.4 NN-based volatility timing

Nowadays it is well known that stock market volatility shows heteroskedasticity and high nonlinearities. Moreover, since Schwert (1989) it is known that there are

many macroeconomic factors that are able to explain returns' volatility. These facts motivated a great debate about the usage of ANNs in forecasting volatility. One of the first contribution in this direction is due to Donaldson and Kamstra (1997a). Since this paper many other authors have written about the topic[6].

However, while forecasting volatilities has been widely discussed, the forecast of covariance matrices with ANNs is a still poorly explored topic. Recently, in this direction, Bucci (2020a) proposed a Cholesky-based ANNs approach for forecasting realized covariances. Nevertheless, in what follows we dot not make use of realized quantities and a different approach is used. In particular, following many previous studies (e.g. Wang, 2009; Kristjanpoller, Fadic, and Minutolo, 2014; Kim and Won, 2018), we consider an hybrid model. However, instead of considering a simple hybrid NN-GARCH model, we construct a DCC-NN model that is based on the Dynamic Conditional Correlation of Engle (2002).

First of all, we consider the usual decomposition:

$$\Sigma_t = \mathbf{D}_t \Gamma_t \mathbf{D}_t$$

where $\mathbf{D}_t = \text{Diag}\left(\sigma_{1,t}, \ldots, \sigma_{N,t}\right)$ is a diagonal matrix with conditional standard deviation and Γ_t be the conditional correlation matrix. Following Engle (2002), the conditional standard deviation can be modeled as GARCH-type processes and, by using the devolatilized returns, the same applies to the conditional correlations. Instead of considering GARCH-type processess, we take advantage of ANNs.

Therefore, we use a 2-step approach to forecast the conditional covariance matrix, where in the first step $\mathbf{D}_{t+1}$ and Γ_{t+1} are predicted and, then, they are aggregated to obtain Σ_{t+1} in the second step. Also in this case the most important step lies on the choice of the network architecture. As noted by Bucci (2020a), the mechanism underlying the Jordan (1997) network has similarities with GARCH models and this makes this network structure the most suitable for forecasting volatility and correlations. Since the aim is to forecast the covariance matrix, we necessarily have to structure a neural network with many output layers. In this case the output correspond to volatilities and correlations. The network architecture, in the case of 1 hidden node, is shown in Fig. 2.6.

[6]For example see Wang, 2009; Kristjanpoller, Fadic, and Minutolo, 2014; Kristjanpoller and Minutolo, 2015; Kristjanpoller and Minutolo, 2016; Kim and Won, 2018; Bucci, 2020b

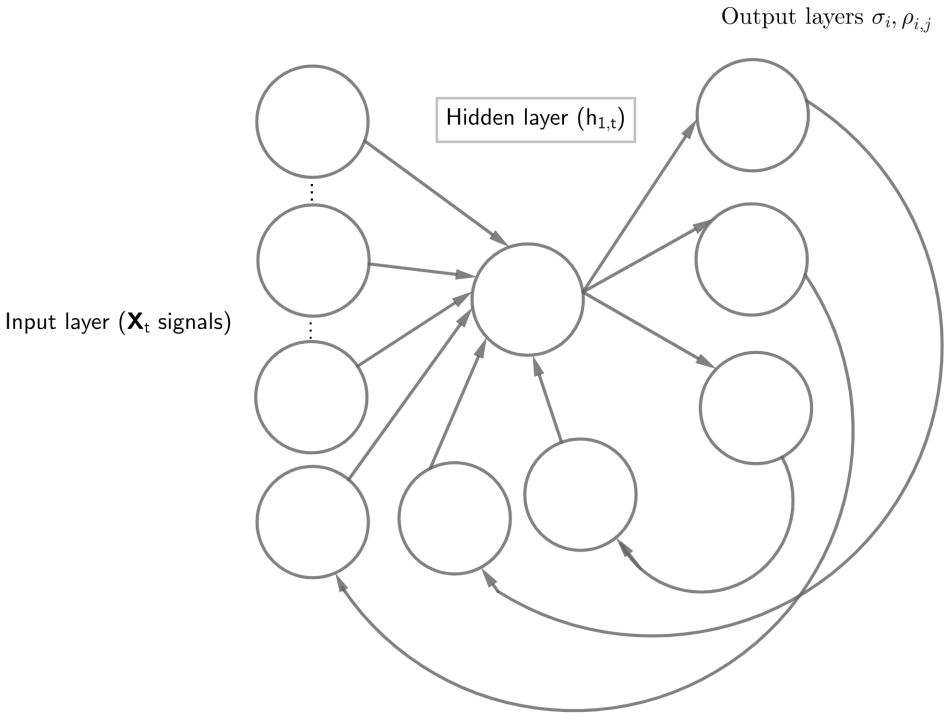


FIGURE 2.6: RNN (Jordan, 1997) for covariance forecasting: example with 1 hidden nodes.

First of all, we consider a Jordan RNN for predicting conditional volatilities in $\mathbf{D}_t$ where some macroeconomic and financial factors are considered as inputs as well as GARCH forecasts. Hence, volatilities are in fact predicted by means of an hybrid NN-GARCH model as in Donaldson and Kamstra (1997b), Kristjanpoller, Fadic, and Minutolo (2014), and Kim and Won (2018). In this case, since volatility cannot assume negative values, a ReLU activation function is considered in the network.

Then, the predictions for the conditional correlation matrix Γ_t are obtained by means of pseudo-conditional correlation $\mathbf{Q}_t$ forecasts. In doing so, we again consider as inputs of the network the set of macroeconomic and financial variables that are used also in forecasting volatilities. However, in this case a Jordan RNN is applied to forecasts of devolatilized cross product of returns $E_t E_t'$ that are the outputs of the network. In this case a tangent hyperbolic activation function is considered in order to ensure that the pseudo-conditional correlation forecasts are bounded between -1 and 1.

Clearly, in both settings it is possible to include multiple hidden nodes. The network architecture with two hidden nodes is shown in Fig. 2.7.

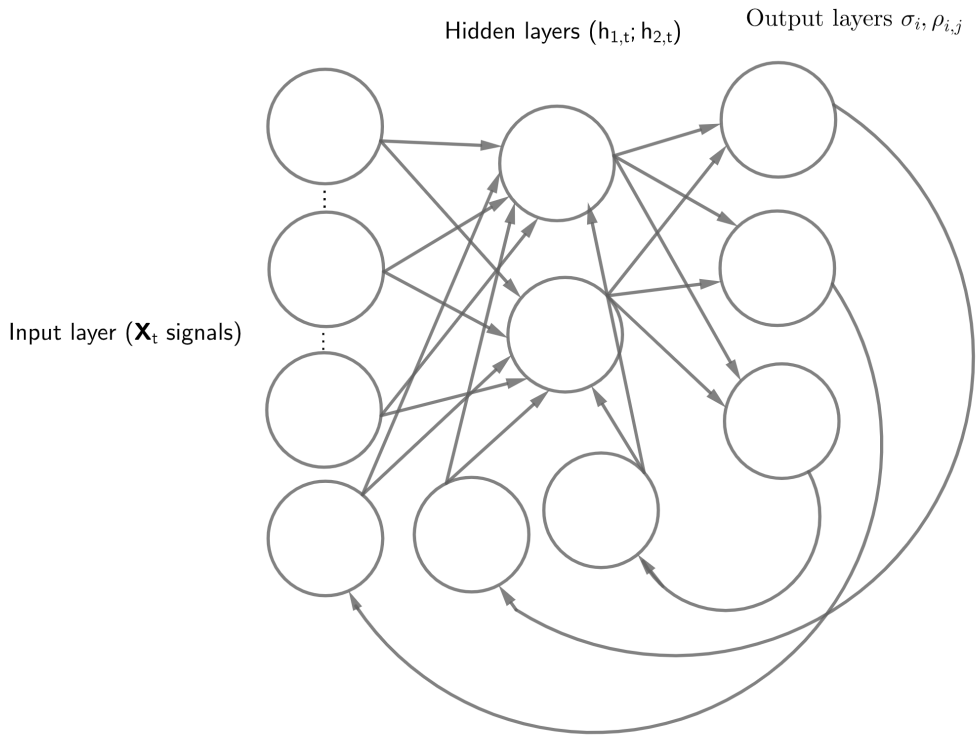


FIGURE 2.7: RNN (Jordan, 1997) for covariance forecasting: example with 2 hidden nodes.

Then, in order to reconstruct the conditional correlation Γ_t we use:

$$\hat{\Gamma}_t = \text{Diag}\left(\hat{\mathbf{Q}}_t\right)^{-0.5} \hat{\mathbf{Q}}_t \text{Diag}\left(\hat{\mathbf{Q}}_t\right)^{-0.5}$$

where $\hat{\mathbf{Q}}_t$ is the forecast of the pseudo-correlation matrix obtained with the Jordan RNN. At the end the forecast for the covariance matrix Σ_t by means of the hybrid DCC-NN model is obtained by means of:

$$\hat{\Sigma}_t = \hat{\mathbf{D}}_t \hat{\Gamma}_t \hat{\mathbf{D}}_t$$

2.4 Data and strategies

For the empirical assessment, we consider both data in a *low dimension* ($N < T$) and in a *large dimension* ($N > T$). Therefore, we'll have different datasets about stock returns. The market excess return m_t is the one obtained by the Kenneth French website from the 10/1926 up to the 10/2019. The covariates used to forecast market

premium are some selected according to Ang and Bekaert (2007) and Campbell and Thompson (2008) suggestions. At this aim we selected the Industrial Production Index (code: INDPRO), the inflation rate (code: CUUR0000SA0R) and the short-term interest rates (code: M1329AUSM193NNBR from 10/1926 to 03/1934 and code: TB3MS from 03/1934 to 10/2019) obtained from the FRED website[7]. we use these variables as inputs $\mathbf{X}_t$ of the neural networks as well.

Before to start, a brief description of the implemented strategy is provided. The $\hat{\mu}, \hat{\Sigma}$ indicates strategies based on plug-in, where some estimator (models) are used to estimate (forecast) the covariance matrix or the expected return vector. Implemented choices in the case of expected returns vector are: $\hat{\mu}$ for the sample average (static); $\hat{\mu}_{AR}$ for the AR based forecasts (return timing) and $\hat{\mu}_{CAPM}$ for the economic forecasting approach (return timing) based on the three step procedure specified in (2.16). Implemented choices for the covariance matrix are, instead: $\hat{\Sigma}_{SC}$ for the sample covariance (static); $\hat{\Sigma}_{sP}$ for the optimal precision shrinkage developed in the first chapter (static); $\hat{\Sigma}_{LW}$ for the Ledoit and Wolf (2003) shrinkage estimator of the sample covariance towards the market (static); $\hat{\Sigma}_{POET}$ the estimator of Fan, Liao, and Mincheva (2013) (static); $\hat{\Sigma}_{DCC}$ for the forecasted covariance froma Dynamic Conditional Correlation model (volatility timing); $\hat{\Sigma}_{GO}$ the forecasted covariance matrix from a GO-GARCH model (volatility timing). Moreover, we also implement timing strategies based on machine learning approaches, namely $\hat{\mu}_{ANN}$ for the ANN-based returns forecasts and $\hat{\Sigma}_{ANN}$ for the ANN-based covariance forecasts.

In the case of excess return, there is only one static estimator, the sample average, that we use as a standard choice for implementing the *voltatility timing* strategies. Similarly, in order to implement a *return timing* strategy, we use the sample covariance as the static estimator. In the end, in the case of the *full-timing strategy*, we take the best forecasting models for both the returns vector and the covariance matrix by comparing standard versus machine learning approaches.

As stated previously, in what follows we conduct empirical experiments with both a low-dimensional setting (with $N = 5$ and $N = 30$) where $N < T$ and an high-dimensional setting where $N > T$. The economic significance of timing versus static strategies is studied but also a comparison between classical econometric models and machine learning one is conducted.

[7] We get data from Kenneth French website `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html` and FRED Database at `https://fred.stlouisfed.org/`

2.5 Results: mean-variance diversification

2.5.1 Low-dimension

N=5

First of all, we consider the case in which few assets are available in the universe. As the $N = 5$ assets we consider the *5 Industry portfolios* of Fama-French. The monthly time series are sampled from the 10/1926 up to the 10/2019. Therefore, $N = 5$, $T = 1120$, $M = 120$ and $T - M = 1000$. Being M the estimation window's length, we are in a *low-dimensional setting* because $M > N$.

Static strategies. Let's analyse first the static strategies' performances. The first result to note is that large dimensional techniques are not useful in improving financial performances in a low-dimensional setting. First recommendation for practitioners: *never be fancy*. Indeed, the POET estimator of Fan, Liao, and Mincheva (2013) and the shrinkage estimator of Ledoit and Wolf (2003) provide the lowest performances in terms of Sharpe ratio, even if the shrinkage estimator of Ledoit and Wolf (2003) shows very close performances to sample covariance estimator (see the first column of the Table 2.1). On the other side, instead, the precision shrinkage estimator allows a considerably improvement in the performance. This is due to the fact that, with the plug-in of $\hat{\Sigma}_{sP}$ we are explicitly taking into account and *optimally reducing* the estimation error in the sample covariance matrix estimation.

TABLE 2.1: Results for $N = 5$ assets

Strategies	SR	CEQ	RL
$1/N$	22.492%	0.888%	0.000
Static strategies:			
$\hat{\mu}, \hat{\Sigma}_{SC}$	12.781%	0.646%	0.585%
$\hat{\mu}, \hat{\Sigma}_{sP}$	19.254%	0.866%	0.102%
$\hat{\mu}, \hat{\Sigma}_{LW}$	12.763%	0.648%	0.591%
$\hat{\mu}, \hat{\Sigma}_{POET}$	11.836%	0.632%	0.763%
Return-timing:			
$\hat{\mu}_{AR}, \hat{\Sigma}_{SC}$	0.095%	-446.595%	63.118%
$\hat{\mu}_{CAPM}, \hat{\Sigma}_{SC}$	21.342%	0.855%	-0.006%
$\hat{\mu}_{ANN}, \hat{\Sigma}_{SC}$	0.257%	-412.143%	60.204%
Volatility-timing:			
$\hat{\mu}, \hat{\Sigma}_{DCC}$	15.789%	0.894%	0.401%
$\hat{\mu}, \hat{\Sigma}_{GO}$	3.959%	-0.937%	3.141%
$\hat{\mu}, \hat{\Sigma}_{ANN-DCC}$	-3.422%	-30.424%	18.388%
Full timing:			
$\hat{\mu}_{CAPM}, \hat{\Sigma}_{DCC}$	21.747%	0.896%	-0.025%
$\hat{\mu}_{ANN}, \hat{\Sigma}_{ANN-DCC}$	-3.749%	-389.103%	68.689%

One can ask whether such Sharpe ratios statistically differs each other. At this aim, we performed the Ledoit and Wolf (2008) test, an improved and robust version of the usual Jobson and Korkie (1980) and Jobson and Korkie (1981) test. What we conclude is that, while the Sharpe ratio associated to the precision matrix shrinkage statistically differs and outperform the others, the remaining static approaches do not statistically differ in their performance. In other words, using the sample covariance estimator, the Ledoit and Wolf (2003) or the POET of Fan, Liao, and Mincheva (2013) lead to exactly the same results. Almost the same results apply if we look at the CEQ returns. Indeed the plug-in of the Ledoit and Wolf (2003) shrinkage estimator does not improve the performance with respect the sample covariance, and still the static strategy based on the precision shrinkage provides the best results with a superior performances.

In the end, confirming the results of De Miguel, Garlappi, and Uppal (2007), none of

the proposed investment strategies is able to overperform the naive one. However, the Ledoit and Wolf (2008) test results highlight that, while the other approaches have statistically different and lower performances with respect the naive strategy, the strategy based on the precision shrinkage plug-in performs as well as the naive one. These two strategies are statistically indistinguishable. In other words, the precision matrix shrinkage is as good as the naive asset allocation.

Timing strategies. For the *return-timing* strategies we analysed the case of statistical forecasting with AR models, the 3-step economic forecasting approach and the neural network-based technique.

Surprisingly both the AR and NN-based timing strategies are not able to outperform any of the static strategies. On the contrary, the model-based economic forecasting approaches show Sharpe ratios that are also greater than the precision shrinkage plug-in. In particular, the return timing with CAPM-based forecasts provides a Sharpe ratio equal to 21.3%, while the volatility timing with DCC forecasts reaches a Sharpe ratio of 15.7%. Hence, timing strategies based on the mean overperforms those based on the covariance forecasts. However, the Ledoit and Wolf (2008) test on the difference between the return-timing strategy with CAPM forecasts and the static with precision shrinkage shows a p-value equal to 0.53, so the two approaches lead to statistically equal Sharpe ratios. This is still an interesting result: *return-timing does not improve upon static strategies*, if the static strategy correctly accounts for the estimation error in the covariance matrix. Besides, these results provide an interesting evidence: forecasting can be useful *only if* the model we use to make the forecasts is accurate. Therefore, practitioners should first test their forecasting ability of the model before to use it in portfolio analysis. On the side of CEQ return, we can get exactly the same conclusions.

As highlighted before, an interesting result is that the volatility timing strategies are even less useful than the return-timing one. Indeed, the best timing strategy for volatility, obtained with the Dynamic Conditional Correlation approach, has a statistically different and lower Sharpe ratio than the best static approach. It seems, once again, that a correct static estimation of the covariance matrix is better than a forecasting approach. Indeed, the portfolio constructed with the precision shrinkage lead to superior out-of-sample Sharpe ratio with respect to DCC forecasts, even if this difference is not statistically significant.

From these evidences a natural question rise: is the timing usefull? Actually yes.

Indeed, a full-timing strategy, where both returns and covariances are predicted, is able to outperforms with respect the static strategy based on the shrinkage precision matrix estimator. However, again, the Sharpe ratio of the two strategies are not statistically significant (the p-value of the test is equal to 0.19). Therefore we can conclude that, from the financial performance point of view, in this low-asset world forecasting-based asset allocation have almost the same performance of static approaches that correctly accounts for the estimation error.

ML-based strategies. Overall, Tab. 2.1 shows that machine learning-based timing strategies are not useful at all. This could seem a surprising evidence. However, we should have in mind that the rolling-window approach that we used requires training the networks with only $M = 120$ time observations. This number is not enough large to ensure an accurate training of the network. As argued by Zhang, Patuwo, and Hu (1998), neural network-based forecasting models have been for long time considered as unnecessary tools by professional forecasters. The reason lies on their poor performances within environments with few-data. Nowadays both economists and investors face data-rich environment, where neural networks perform very well. Nevertheless, canonical approaches for economic significance evaluation require the analysis of monthly data, that are known to be less noisy than the daily or infra-daily data. Therefore, the results of Tab. 2.1 show that machine-learning based timing strategy are not appropriate for long-run (low-frequency) portfolio selection problems. This is not a so much unexpected result.

Economic evaluation. By economic evaluation we refer to the approach of Fleming, Kirby, and Ostdiek (2001) and Fleming, Kirby, and Ostdiek (2003) and Marquering and Verbeek (2004) in assessing economic performances of alternative trading strategies. As previously discussed, we compute the *annualized percentage* maximum fee Δ that a mean-variance investor is willing to pay to switch from a strategy to another one.

Table 2.6 shows the economic comparison between naive strategy and all the alternatives. With this respect, a mean-variance investor would prefer always a naive strategy. However, the investor requires the highest fees Δ to switch from a naive to a data-driven allocation. Indeed, in the case of data-driven approaches as well as those based on GO-GARCH and AR-based forecasts the mean-variance investor requires $\Delta = 468231$ basis points fee for the implementation of a full-timing strategy

based on machine learning. On the contrary, in the other cases where static estimation is used or where model-based forecasts are employed, the investor requires a much lower fee. The minimum fee that the investor is willing to accept is associated to the full-timing strategy where model-based forecasts are employed (in this case $\Delta = 112$ basis points).

TABLE 2.2: Economic fee Δ for $N = 5$ assets: optimal vs naive

	$\hat{\mu}, \hat{\Sigma}_{SC}$	$\hat{\mu}, \hat{\Sigma}_{LW}$	$\hat{\mu}, \hat{\Sigma}_{POET}$	$\hat{\mu}, \hat{\Sigma}_{sP}$
$\hat{\mu}_{AR}, \hat{\Sigma}_{SC}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{SC}$	$\hat{\mu}_{ANN}, \hat{\Sigma}_{SC}$	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{GO}$
$\hat{\mu}, \hat{\Sigma}_{ANN}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{DCC}$	$\hat{\mu}_{ANN}, \hat{\Sigma}_{ANN}$		
	0.0409	0.0407	0.0427	0.0148
53.6559	0.0159	49.5259	0.0116	0.2306
3.7696	0.0112	46.8288		

Table 2.3, instead, is useful for answering the following question: *it is useful using statistical methods explicitly taught to work in large dimensional setting also in a low dimension one?*

It depends. Indeed, the mean-variance investor asks positive fees Δ for renouncing to the precision shrinkage estimators in this low-dimensional setting. However, the Ledoit and Wolf (2003) estimator is preferred respect to the sample covariance because the investor is willing to pay a low fee of $\Delta = 2.32$ basis points to use the first instead of the second. However, the highest fee is required for using the precision shrinkage instead of the Fan, Fan, and Lv (2008) covariance estimator, with $\Delta = 279$ basis points. Therefore, in general large-dimensional estimators are useless in low-dimension if we use appropriate estimators such as the precision shrinkage.

TABLE 2.3: Economic fee Δ for $N = 5$ assets: low vs large dimension

	$\hat{\mu}, \hat{\Sigma}_{LW}$	$\hat{\mu}, \hat{\Sigma}_{POET}$
$\hat{\mu}, \hat{\Sigma}_{SC}$	-0.0002320396	0.0018032558
$\hat{\mu}, \hat{\Sigma}_{sP}$	0.02595272	0.02798802

In the next Table 2.5.1 we have a comparison in the spirit of previous papers between static and timing strategies. Results are useful to answer to the question: *it more convenient forecasting the excess returns' mean vector, the covariance matrix or both?*

Table 2.5.1 suggests that forecasting has economic utility if model-based procedures

are employed. Indeed, the static strategies make the investor better of in utility terms with respect neural network-based approaches. More in details, the precision shrinkage plug-in is the static strategy for which the investor requires the highest fee Δ.

However, on the side of model-based timing approaches, in most of the cases the mean-variance investor is willing to pay a fee Δ to switch from a static to a timing asset allocation. This is true for all the static strategies with only exception of the precision shrinkage plug-in. Indeed, while the investor always prefer timing (either return or volatility or both), the static allocation with the precision shrinkage plug-in is preferred to a return-timing strategy. However, both volatility timing and full timing strategy make the investor better of.

	$\hat{\mu}_{AR}, \hat{\Sigma}_{SC}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{SC}$	$\hat{\mu}_{ANN}, \hat{\Sigma}_{SC}$	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{GO}$	$\hat{\mu}, \hat{\Sigma}_{ANN}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{DCC}$	$\hat{\mu}_{ANN}, \hat{\Sigma}_{ANN}$
$\hat{\mu}, \hat{\Sigma}_{SC}$	53.6149	-0.0250	49.4850	-0.0294	0.1897	3.7286	-0.0298	46.7879
$\hat{\mu}, \hat{\Sigma}_{LW}$	53.6152	-0.0248	49.4852	-0.0291	0.1899	3.7289	-0.0296	46.7881
$\hat{\mu}, \hat{\Sigma}_{POET}$	53.6131	-0.0268	49.4832	-0.0312	0.1879	3.7268	-0.0316	46.7861
$\hat{\mu}, \hat{\Sigma}_{sP}$	53.6411	0.0012	49.5112	-0.0032	0.2159	3.7548	-0.0036	46.8140

In the last Table 2.4 are reported the comparisons between model-based and data-driven (machine learning) timing strategies. Briefly, switching from any machine learning-based timing strategy is costly and fees are very high. In particular, the highest fee that the investor is willing to pay is the one required to switch from a neural network-based return timing to a full-timing strategy with model-based predictions. The most competitive machine-learning strategy in terms of utility is the volatility-timing one because the fees are overall lower. Overall, the conclusions do not differ from those of Tab. 2.1.

TABLE 2.4: Economic fee Δ for $N = 5$ assets: model-based vs data-driven timing

	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{SC}$	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{GO}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{DCC}$
$\hat{\mu}_{AR}, \hat{\Sigma}_{SC}$	-53.6399	-53.6443	-53.4253	-53.6447
$\hat{\mu}_{ANN}, \hat{\Sigma}_{SC}$	-49.5100	-49.5144	-49.2953	-49.5148
$\hat{\mu}, \hat{\Sigma}_{ANN}$	-3.7536	-3.7580	-3.5389	-3.7584
$\hat{\mu}_{ANN}, \hat{\Sigma}_{ANN}$	-46.8129	-46.8172	-46.5982	-46.8176

In the next section we are moving further to the large dimensional case but still in a low-dimensional one, analysing an economy in which the number of asset is larger than 5.

N=30

Here we consider an asset universe of $N = 30$ risky assets represented by the *30 Industry portfolios* of Fama-French. As before, the monthly time series are sampled from the 10/1926 up to the 10/2019. Therefore, now we have $N = 30$, $T = 1120$, $M = 120$ and $T - M = 1000$. Being M the estimation window's length, we are still in a *low-dimensional setting* because $M > N$.

Static strategies. As the previous Section, let's have a look, first, to the static strategies' performances. From Table 2.5 it is evident that results differ from the other we have seen in the Table 2.1. Now most of the static strategies have extremely bad performances expect one: the plug-in of the optimal precision matrix estimator $\hat{\Sigma}_{sP}$. This last strategy, based on the plug-in of the estimator developed in the first Chapter of the thesis, ensures a Sharpe ratio equal to 14%. Statistical test of Ledoit and Wolf (2008) provides further evidence in this direction, because Sharpe ratios are all statistically different each other. Therefore, the out-performance of this plug-in

estimator is not sample-driven. The CEQ returns confirm that, while for the static strategies are negative, a mean-variance investor realizes a gain in her utility by investing in the static strategy with precision shrinkage plug-in.

Timing strategies. In this case, the return-timing strategy based only on statistical arguments (i.e. AR and neural network) perform very poor in out-of-sample and the static strategy based on the precision shrinkage guarantees much higher performances. However, the return timing strategy based on the CAPM forecasts provides a Sharpe ratio equal to 21%. Moreover, with a p-value of 0.018, the Ledoit and Wolf (2008) test highlights that the performances of the two portfolio are statistically different. Note that the Sharpe ratio of the return-timing with CAPM forecasts is greater than the one associated to the naive strategy. However, with a p-value of 0.68, the Ledoit and Wolf (2008) test suggests to accept the null hypothesis of equality between Sharpe ratios. Therefore, there is not enough statistical evidence to conclude that the two strategies perform differently in out-of-sample.

The main conclusions do not change by looking at volatility timing strategies. More in details, using a covariance matrix predicted with neural networks lead to a negative Sharpe ratio (-2.7%). By using the GO-GARCH model the out-of-sample Sharpe ratio is much lower than the one given by the precision matrix shrinkage, that is a static asset allocation (3.9% versus 14.2%). Then, using a DCC model provides almost the same out-of-sample performance, in terms of Sharpe ratio, if compared to the one based on the precision shrinkage (15.8% versus 14.2%). This result is confirmed by the Ledoit and Wolf (2008) test with a p-value of 0.63. In the end, we observe that the full-timing strategy provides exactly the same out-of-sample performance of the return-timing based on CAPM forecasts.

From the CEQ return point of view, we observe that the precision shrinkage approach provides the highest return in the utility for a mean-variance investor and that the the model-based full timing improves the investor utility more than the simple CAPM-based return-timing approach. Hence we can conclude that, once again, timing strategies improve the performances with respect a static strategy and that machine-learning is not useful at this aim. Moreover, we also confirm that a static strategy with the precision shrinkage estimator is very competitive also with respect to the timing strategies.

ML-based strategies. In terms of utility of data-driven procedures, also in this case we observe that neural network predictions are not competitive with respect

to model-based ones. Therefore, these results suggest that investors should avoid the usage of these techniques for the implementation of a static strategy. However, also in this case this result can be explained by the small size of the observations used for estimation (i.e. $M = 120$). In unreported Tables we also studied the effects of changes in the estimation window (with $M = 240$) but results were the same. This happen because also by doubling the estimation window size, it remains not enough large for accurate training of the networks. Therefore, in presence of monthly data the model-based approaches are very recommended.

TABLE 2.5: Results for $N = 30$ assets

Strategies	SR	CEQ	RL
$1/N$	21.227%	0.890%	0.000
Static strategies:			
$\hat{\mu}, \hat{\Sigma}_{SC}$	1.443%	-11.677%	9.165%
$\hat{\mu}, \hat{\Sigma}_{sP}$	14.248%	0.738%	0.381%
$\hat{\mu}, \hat{\Sigma}_{LW}$	5.040%	-0.118%	1.782%
$\hat{\mu}, \hat{\Sigma}_{POET}$	0.868%	-3.366%	5.093%
Return-timing:			
$\hat{\mu}_{AR}, \hat{\Sigma}_{SC}$	3.741%	-904%	69%
$\hat{\mu}_{CAPM}, \hat{\Sigma}_{SC}$	21.904%	0.879%	-0.090%
$\hat{\mu}_{ANN}, \hat{\Sigma}_{SC}$	-2.148%	-255%	49%
Volatility-timing:			
$\hat{\mu}, \hat{\Sigma}_{DCC}$	15.865%	0.898%	0.296%
$\hat{\mu}, \hat{\Sigma}_{GO}$	3.997%	-0.930%	2.886%
$\hat{\mu}, \hat{\Sigma}_{ANN}$	-2.755%	-1312%	115%
Full timing:			
$\hat{\mu}_{CAPM}, \hat{\Sigma}_{DCC}$	21.905%	0.902%	-0.094%
$\hat{\mu}_{ANN}, \hat{\Sigma}_{ANN}$	-3.741%	-388%	64%

Economic evaluation. As the previous case, we compute the *annualized percentage* maximum fee Δ that a mean-variance investor is willing to pay (or to receive) to switch from a strategy to another one.

Table 2.6 shows the economic comparison between naive strategy and all the alternatives. In this case, despite the naive strategy is not the most performing in terms

of both Sharpe ratio and CEQ return, the investor realized the highest utility by its implementation. Indeed, the investor requires fees equal to Δ for switching from the naive to any other alternative strategy. Among the static strategies, the one based on precision shrinkage plug-in is the "less costly", i.e. the investor requires the lowest fee $\Delta = 322$ basis points. However, overall the model-based full timing strategy is the most competitive among the considered alternatives with $\Delta = 129$ basis points. Evidently, the machine learning based timing strategy are very costly to implement, because of their poor performance in out-of-sample, so the investor requires very high annualized fees.

TABLE 2.6: Economic fee Δ for $N = 5$ assets: optimal vs naive

	$\hat{\mu}, \hat{\Sigma}_{SC}$	$\hat{\mu}, \hat{\Sigma}_{LW}$	$\hat{\mu}, \hat{\Sigma}_{POET}$	$\hat{\mu}, \hat{\Sigma}_{sP}$
$\hat{\mu}_{AR}, \hat{\Sigma}_{SC}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{SC}$	$\hat{\mu}_{ANN}, \hat{\Sigma}_{SC}$	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{GO}$
$\hat{\mu}, \hat{\Sigma}_{ANN}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{DCC}$	$\hat{\mu}_{ANN}, \hat{\Sigma}_{ANN}$		
	1.5203	0.1345	0.5238	0.0322
108.7084	0.0162	30.6743	0.0129	0.2314
157.0716	0.0126	46.6447		

Interestingly, in the Table 2.7 are reported the costs of using large-dimensional estimators. It is known that with increasing N the estimation error increases and, therefore, the sample covariance estimator performs always poorer. In this case using estimators taught for large-dimensional setting such as the on of Ledoit and Wolf (2003) makes the investor better of. Therefore, she is willing to pay an high fee of $\Delta = 13858$ basis points to continue using the Ledoit and Wolf (2003) estimator instead of the sample covariance. However, the same investor requires a positive fee to use the precision shrinkage estimator instead of both the Ledoit-Wolf and the POET estimators. In other words, if the investor uses an estimator especially taught for reducing estimation error in low-dimensional setting, it is very expensive in utility terms using other tools that are an inappropriate context. As in the previous case, this lead to a useful indication for practitioners: in low-dimensional setting is better to use the precision shrinkage estimator.

TABLE 2.7: Economic fee Δ for $N = 30$ assets: low vs large dimension

Strategies	$\hat{\mu}, \hat{\Sigma}_{LW}$	$\hat{\mu}, \hat{\Sigma}_{POET}$
$\hat{\mu}, \hat{\Sigma}_{SC}$	-1.3858368	-0.9965103
$\hat{\mu}, \hat{\Sigma}_{sP}$	0.1022415	0.4915680

Table 2.5.1 provides useful insights about the economic performance of static strategies compared to the timing one. Overall the timing strategy are preferred to their static alternatives if model-based procedures are implemented for forecasting. On the contrary, if NN-based forecasts are used the static strategy will be preferred also in utility terms. However, we have to note that the precision shrinkage plug-in remains the most competitive static approach.

	$\hat{\mu}_{AR}, \hat{\Sigma}_{SC}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{SC}$	$\hat{\mu}_{ANN}, \hat{\Sigma}_{SC}$	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{GO}$	$\hat{\mu}, \hat{\Sigma}_{ANN}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{DCC}$	$\hat{\mu}_{ANN}, \hat{\Sigma}_{ANN}$
$\hat{\mu}, \hat{\Sigma}_{SC}$	107.1881	-1.5050	29.1540	-1.5074	-1.2889	155.5513	-1.5077	45.1244
$\hat{\mu}, \hat{\Sigma}_{LW}$	108.5739	-0.1191	30.5398	-0.1216	0.0969	156.9371	-0.1219	46.5102
$\hat{\mu}, \hat{\Sigma}_{POET}$	108.1846	-0.5085	30.1505	-0.5109	-0.2924	156.5478	-0.5112	46.1209
$\hat{\mu}, \hat{\Sigma}_{sP}$	108.6762	-0.0169	30.6421	-0.0194	0.1991	157.0394	-0.0197	46.6124

Finally we have a comparison between timing strategies, model-based vs data-driven, in the Table 2.8. What emerges is that the investor is willing to pay to switch from ML-based to a model-based timing strategy, for all the cases (i.e. return-timing, volatility-timing, full-timing). Among the alternative timing implementations, the highest fees are required for the adoption of the model-based full-timing strategy.

TABLE 2.8: Economic fee Δ for $N = 30$ assets: model-based vs data-driven timing

	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{SC}$	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{GO}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{DCC}$
$\hat{\mu}_{AR}, \hat{\Sigma}_{SC}$	-108.6931	-108.6955	-108.4770	-108.6958
$\hat{\mu}_{ANN}, \hat{\Sigma}_{SC}$	-30.6590	-30.6614	-30.4429	-30.6617
$\hat{\mu}, \hat{\Sigma}_{ANN}$	-157.0563	-157.0587	-156.8402	-157.0590
$\hat{\mu}_{ANN}, \hat{\Sigma}_{ANN}$	-46.6293	-46.6318	-46.4133	-46.6321

Essentially, we find that model-based strategy are better suited than data-driven ones and that full-timing is better than partial timing. Therefore, we get the same conclusion for both $N = 5$ and $N = 30$ cases. The static strategy based on the precision shrinkage estimator reveals again to be very competitive and the best solution within a low-dimensional setting. However, timing ensures higher performances, even if not always statistically different from the best static approaches.
In the next Section we am going to investigate the usefulness of forecasting in large dimensional setting, with a huge increase in the concentration ratio N/T

2.5.2 Large dimension

In this case we consider all the 500 constitutes of the S&P500 Index from the 10/1999 up to 10/2019 in order to consider a similar time span of the previous two examples with $N = 5$ and $N = 30$. In doing so we assume that the asset universe is represented by only the stocks for which we have a complete time series, so those with missing values have been excluded. Particularly, we get $N = 286$ time series with $T = 240$. Considering, again, an estimation window of $M = 120$, we have now that $T - M = 120$. In this framework, being M our sample size for the estimation, we are in a *large dimensional setting* because $M < N$. We do not have any problem in estimating and predicting the excess returns but we know that standard sample estimators are ill-conditioned, resulting in singular matrices. Therefore, here we take advantages of ad-hoc statistical methods used by previous literature for the estimation and the

forecasting of *large dimensional covariances*.

Static strategies. Exactly as the other cases in low dimension, the equally weighted strategy returns one of the highest performances (see Tab. 2.9). As static estimators in large dimension we consider the shrinkage of Ledoit and Wolf (2003) and the POET of Fan, Liao, and Mincheva (2013). Both of them do not guarantee any over-performance with respect the naive allocation. However, Table 2.9 shows that by estimating the covariance matrix with the POET estimator the investor gains almost a 5% greater Sharpe ratio in out-of-sample. Moreover, also in terms of the certainty equivalent (CEQ) the POET-based allocation guarantees an higher return. However, despite one can wonder that the POET out-performance with respect to the Ledoit and Wolf (2003) covariance estimator can be sample driven, we want to highlight that these results confirm those of Ledoit and Wolf (2017). Indeed, the authors show that the Sharpe ratio associated to the POET plug-in is higher in all the considered cases[8] where $N = 30, 50, 100, 250, 500$.

TABLE 2.9: Results for $N = 286$ assets

Strategies	SR	CEQ	RL
$1/N$	30.125%	1.182%	0.000
Static strategies:			
$\hat{\mu}, \hat{\Sigma}_{LW}$	22.330%	0.833%	0.320%
$\hat{\mu}, \hat{\Sigma}_{POET}$	27.607%	0.946%	0.092%
Return-timing:			
$\hat{\mu}_{AR}, \hat{\Sigma}_{LW}$	11.893%	-46.609%	19.905%
$\hat{\mu}_{CAPM}, \hat{\Sigma}_{LW}$	31.667%	1.386%	-0.073%
$\hat{\mu}_{ANN}, \hat{\Sigma}_{LW}$	15.357%	-0.861%	5.251%
Volatility-timing:			
$\hat{\mu}, \hat{\Sigma}_{DCC}$	28.533%	1.316%	0.081%
$\hat{\mu}, \hat{\Sigma}_{ANN}$	6.872%	-40.236%	22.519%
Full timing:			
$\hat{\mu}_{CAPM}, \hat{\Sigma}_{DCC}$	16.664%	0.759%	0.733%
$\hat{\mu}_{ANN}, \hat{\Sigma}_{ANN}$	20.746%	-21.276%	8.366%

[8] Note that the Ledoit and Wolf (2017) assumes daily returns of a Global Minimum Variance (GMV) diversification strategy while in this paper we construct portfolios with monthly returns according to the more standard Markowitz mean-variance rule. However the POET over-performance is confirmed here, especially where the scenario of a large asset universe is considered.

Timing strategies. On the side of the timing strategies, we analyse first those based on *return-timing*. The Tab. 2.9 shows that model-based return-timing strategy is the one with the highest out-of-sample Sharpe ratio, also higher than the naive. Nevertheless, this difference is very small and also the Ledoit and Wolf (2008) test do not reject the null with a p-value of 0.89. In other words, the two approaches provide the same out-of-sample performance. This evidence documents the fact that naive strategy is very hard to defeat in out-of-sample, also if we use timing approaches.
Volatility-timing strategies are overall less competitive with respect to the return-timing in a large dimensional framework. Even if this result holds also for the other two analysed low-dimensional settings, it is reasonable in a large-dimensional one. Indeed, the analysis of covariance matrices is more problematic in large dimension respect to the mean vector and this fact has a consequence also in the forecasting activity.
Surprisingly, the data-driven approaches provides quite well in large dimension. Indeed, a return-timing strategy with AR forecasts guarantees a Sharpe ratio of 12%, while a neural-network one returns a Sharpe ratio of 15%. These values are much higher than those (negative) that we found in both low-dimensional settings. Nevertheless, their performances remain lower than the the model-based return-timing (Sharpe ratio close to 32%).
In the end, full-timing strategies seem to be less competitive than those based on either return or volatility timing. Indeed, while in previous cases the performances of return-timing strategies were very close to full-timing, in this setting we observe the opposite. Moreover, machine learning based full timing seems to perform better than the model-based approach. In terms of Sharpe ratio, however, the Ledoit and Wolf (2008) test does not allow to reject the null hypothesis. Hence, the wo approaches are not statistically different in out-of-sample even if the ML-based approach shows a 4% higher Sharpe ratio.
Economic evaluation. As in the low-dimensional setting, in what follows we report the annualized percentage maximum fee Δ, that a mean-variance investor is willing to pay to switch from a strategy to another one, for the large dimensional case.
Table 2.5.2 shows the economic comparison between naive strategy and all the optimal alternatives. In this case all the considered static strategies are less appealing to the mean-variance investor than the naive. Therefore, the investor requires

the payment of fees Δ to implement those strategies instead of the naive. However, among the timing strategies, there are some that are preferred by the investor. Indeed, the model-based volatility-timing with covariance matrix predicted by the Hafner and Reznikova (2012) approach and the model-based CAPM return timing provide utility gains to the investor that is implementing a naive asset allocation. Indeed, the investor is willing to pay $\Delta = 126$ basis points to switch the naive for the CAPM-based return timing and $\Delta = 43$ basis points to buy the Hafner and Reznikova (2012) volatility-timing. In the case of full-timing strategies, instead, the investor prefers the naive allocation. However, even if the machine learning full-timing returns higher out-of-sample Sharpe ratio than the model-based full-timing strategy, the investor asks a lower fee ($\Delta = 618$ basis points for model-based versus $\Delta = 28740$ basis points for the data-driven) for implementing the second one.

$\hat{\mu}, \hat{\Sigma}_{LW}$	$\hat{\mu}, \hat{\Sigma}_{POET}$	$\hat{\mu}_{AR}, \hat{\Sigma}_{LW}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{SC}$	$\hat{\mu}_{ANN}, \hat{\Sigma}_{LW}$	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{ANN}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{DCC}$	$\hat{\mu}_{ANN}, \hat{\Sigma}_{ANN}$
0.0530	0.0396	5.7913	-0.0126	0.2679	-0.0043	4.9637	0.0618	2.8740

Table 2.5.2 reports the economic comparison between static and timing strategies. The most important result to highlight is the negative fee associated to the timing strategies. This means that the investor in large dimension is willing to pay to switch from a static strategy to a timing one. However, the mean-variance investor prefers the static strategy to full-timing and, among the partial model-based timing strategies, she prefers the return timing. Then, static strategies are preferred to data-driven timing approaches. Therefore, forecasting is at some extent useful but only if model-based approaches are used.

	$\hat{\mu}_{AR}, \hat{\Sigma}_{LW}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{LW}$	$\hat{\mu}_{ANN}, \hat{\Sigma}_{LW}$	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{ANN}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{DCC}$	$\hat{\mu}_{ANN}, \hat{\Sigma}_{ANN}$
$\hat{\mu}, \hat{\Sigma}_{LW}$	5.7383	-0.0656	0.2148	-0.0573	4.9107	0.0088	2.8210
$\hat{\mu}, \hat{\Sigma}_{POET}$	5.7517	-0.0522	0.2283	-0.0439	4.9242	0.0222	2.8344

The last Table 2.10 shows a comparison across different timing strategies. As previously, the investor is always willing to pay a fee Δ to avoid the usage of data-driven timing strategies. This result confirm the lack of economic usefulness for the machine learning predictions. Note that, differently from the previous low-dimensional cases, in large dimension the data-driven approach perform quite well in out-of-sample. However, despite their good performances, data-driven approaches provide considerable losses in utility terms.

TABLE 2.10: Economic fee Δ for $N = 286$ assets: model-based vs data-driven timing

	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{LW}$	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}_{CAPM}, \hat{\Sigma}_{DCC}$
$\hat{\mu}_{AR}, \hat{\Sigma}_{LW}$	-5.8039	-5.7956	-5.7295
$\hat{\mu}_{ANN}, \hat{\Sigma}_{LW}$	-0.2805	-0.2721	-0.2061
$\hat{\mu}, \hat{\Sigma}_{ANN}$	-4.9763	-4.9680	-4.9019
$\hat{\mu}_{ANN}, \hat{\Sigma}_{ANN}$	-2.8866	-2.8783	-2.8122

These results highlight that forecasting is much more useful in large dimensional setting than in low dimension. However, model-based approaches are better suited than data-driven ones. In the end, the strategy based on return-timing seem to be the better suited for increasing investor utility and performances than full-timing.

2.6 Results: minimum variance diversification

As stated above, Kourtis, Dotsis, and Markellos (2012) showed that estimation error can be reduced by avoiding mean estimation. Following this intuition, in what follows we consider the case of a mean-variance investor that constructs her portfolios according to a minimum variance strategy.

2.6.1 Low-dimension

N=5

As previously, first of all we consider the case in which few assets are available. So, we consider $N = 5$ assets given by the *5 Industry portfolios* of Fama-French. The monthly time series are sampled from the 10/1926 up to the 10/2019. Therefore, $N = 5$, $T = 1120$, $M = 120$ and $T - M = 1000$. Being M the estimation window's

length, we are in a *low-dimensional setting* because $M > N$. The strategies under investigation are not the same of those showed in previous paragraphs. Indeed, now there are not return-timing strategies and full-timing because mean vector is not considered in GMV asset allocation. Moreover, we excluded GO-GARCH because its poor performances compared to the DCC model. Estimating less models allows for a reduction in the computational time of the procedure.

Static strategies. Let's analyse, first, the static strategies' performances (see Tab. 2.11). First of all, we note that the static strategies perform better than the naive. This fact highlights the role of estimation error in the mean: estimating the mean vector lead static strategies to perform poorer than the naive. On the contrary, avoiding mean estimation makes the static approaches more competitive than the $1/N$ diversification rule.

Interestingly, the sample covariance plug-in provides the highest out-of-sample Sharpe ratio. The Ledoit and Wolf (2003) shrinkage estimator performs very closely to sample covariance, while POET ranks as third.

Moreover, another interesting aspect to highlight is that now, differently from the mean-variance scenario, the GMV asset allocation with the precision shrinkage plug-in results in the wort performances among the static approaches. This happen because the precision shrinkage estimator is optimal under mean-variance allocation but not under GMV setting.

TABLE 2.11: Results for $N = 5$ assets - GMV approach

Strategies	SR	CEQ	RL
$1/N$	22.492%	0.888%	0.000
Static strategies:			
$\hat{\mu}, \hat{\Sigma}_{SC}$	23.968%	0.876%	-0.059%
$\hat{\mu}, \hat{\Sigma}_{sP}$	18.594%	0.759%	0.182%
$\hat{\mu}, \hat{\Sigma}_{LW}$	23.875%	0.872%	-0.055%
$\hat{\mu}, \hat{\Sigma}_{POET}$	22.814%	0.858%	-0.013%
Volatility-timing:			
$\hat{\mu}, \hat{\Sigma}_{DCC}$	24.528%	0.896%	-0.081%
$\hat{\mu}, \hat{\Sigma}_{ANN-DCC}$	5.114%	-432%	52%

One can ask whether such Sharpe ratios statistically differs each other. At this

aim, we performed the Ledoit and Wolf (2008) test. With this respect, as found by DeMiguel, Garlappi, and Uppal (2009), the sample covariance and naive approaches are statistically the same, since the test p-value is equal to 0.76 meaning that the Sharpe ratio are equal. Surprisingly, also the static GMV allocation with precision shrinkage plug-in provides the same Sharpe ratio of the naive strategy, because also in this case we do not reject the null hypothesis with a p-value of 0.47. In terms of CEQ return, instead, the naive strategy seems to lead to the highest performances.

Timing strategies. As stated before, in this setting there is not other timing strategy than the one based on volatility. In particular, we are mainly interested in understanding the differences between model-based (DCC forecasts) and data-driven (NN-DCC forecasts) approaches in implementing the volatility-timing within GMV setting.

Also in this case we get evidence of timing strategies over-performance. Indeed, the DCC-based approach ensures an out-of-sample Sharpe ratio equal to 24.5% versus the 23.9% of the static sample covariance. In other words, in this example with $N = 5$ assets we have that timing strategy based on econometric methods provides the highest performances in out-of-sample. On the contrary the timing strategy based on neural network forecasts is the worst to implement with a Sharpe ratio of 5%. Clearly, we should once again have in mind that the estimation window is not enough large to ensure accurate training of the networks. This fact can in principle explain the so poor performances of NN-based timing strategy. Then, is the performance of volatility timing statistically different than the naive approach? The answer is not. Indeed, the Ledoit and Wolf (2008) test does not reject the null with a p-value of 0.67. Hence, once again, the over performance that we get with timing strategy could be sample-driven and does not extend to population.

Economic evaluation. As previously, the economic evaluation is conducted by fallowing the approach of Fleming, Kirby, and Ostdiek (2001) and Fleming, Kirby, and Ostdiek (2003) and Marquering and Verbeek (2004). So, we compute the *annualized percentage* maximum fee Δ that a mean-variance investor is willing to pay to switch from a kind of GMV strategy to another one.

Table 2.12 shows the economic comparison between naive strategy and all the alternatives. With this respect, none of them is preferred to the naive. The minimum fee the investor would accept is associated to the *volatility-timing* strategy, where the covariance matrix is predicted with a model-based approach (i.e. the DCC model).

TABLE 2.12: Economic fee Δ for $N = 5$ assets: optimal vs naive

	$\hat{\mu}, \hat{\Sigma}_{SC}$	$\hat{\mu}, \hat{\Sigma}_{LW}$	$\hat{\mu}, \hat{\Sigma}_{POET}$	$\hat{\mu}, \hat{\Sigma}_{sP}$	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{ANN}$
$1/N$	0.0135	0.0140	0.0156	0.0274	0.0111	52.0933

Table 2.13, instead, is useful for answering the following question: *it is useful using statistical methods explicitly taught to work in large dimensional setting also in a low dimension one?*

The answer is no. Indeed, the investor requires a fee $\Delta = 5.21$ basis points to switch from sample covariance to the Ledoit and Wolf (2003) and another higher $\Delta = 21.37$ basis points for the POET estimator of Fan, Fan, and Lv (2008). However, large dimensional estimators are preferred with respect to the precision shrinkage in this case since the investor is willing to pay $\Delta = 133$ basis points for using the Ledoit and Wolf (2003) estimator. This happen because the precision shrinkage is not optimal for minimum variance setting.

TABLE 2.13: Economic fee Δ for $N = 5$ assets: low vs large dimension

	$\hat{\mu}, \hat{\Sigma}_{LW}$	$\hat{\mu}, \hat{\Sigma}_{POET}$
$\hat{\mu}, \hat{\Sigma}_{SC}$	0.0005205708	0.0021375964
$\hat{\mu}, \hat{\Sigma}_{sP}$	-0.01338207	-0.01176505

In the next Table 2.14 we have a comparison in the spirit of previous papers between static and timing strategies. From Table 2.14 we understand that, from the point of view of the mean-variance investor, the static strategies are worst than timing. However, this result holds only if timing is adopted by employing model-based approaches rather than data driven. Indeed, the investor is willing to pay a fee Δ for implementing a model-based volatility timing strategy, while she requires the payment of a positive fee Δ for implementing any of the static strategies instead of a data-driven timing.

TABLE 2.14: Economic fee Δ for $N = 5$ assets: static vs timing

	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{ANN}$
$\hat{\mu}, \hat{\Sigma}_{SC}$	-0.0024	52.0798
$\hat{\mu}, \hat{\Sigma}_{LW}$	-0.0029	52.0793
$\hat{\mu}, \hat{\Sigma}_{POET}$	-0.0045	52.0776
$\hat{\mu}, \hat{\Sigma}_{sP}$	-0.0163	52.0659

This means, in other words, that forecasting is useful. In the end, we have to make a comparisons between the two different timing strategies, we can compute a fee equal to $\Delta = 520800$, meaning that the investor requires the payment of the annualized fee $\Delta = 520800$ basis points to switch from the model-based timing to the data-driven one. Hence, model-based procedure makes the investor better off.

N=30

Here we consider an asset universe of $N = 30$ risky assets represented by the *30 Industry portfolios* of Fama-French. As before, the monthly time series are sampled from the 10/1926 up to the 10/2019. Therefore, now we have $N = 30$, $T = 1120$, $M = 120$ and $T - M = 1000$. Being M the estimation window's length, we are still in a *low-dimensional setting* because $M > N$.

Static strategies. Also in this case with $N = 30$ assets we have that the naive strategy is no longer the best one in terms of out-of-sample Sharpe ratio (see Tab. 2.15). This happen because we get rid of mean estimation, such that the estimation error reduces and the performances of alternatives. With increasing number of assets we have that the Ledoit and Wolf (2003) estimator improves with respect the sample covariance. Indeed, now the Ledoit and Wolf (2003) provides an out-of-sample Sharpe ratio of 24.9% versus the 22.4% of the sample covariance. With a p-value of 0.0012, the Ledoit and Wolf (2008) test suggests that the two strategies are statistically different, hence the shrinkage towards the market operation is definitively better. However, the Ledoit and Wolf (2003) static approach is not statistically different than the naive strategy, since the test on the equality between Sharpe ratios does not reject the null with a p-value of 0.43. Hence, the static and the naive approaches are indistinguishable in the population. This result confirm those of DeMiguel, Garlappi, and Uppal (2009). The precision shrinkage provides very bad performances

with a negative Sharpe ratio. This result confirm those showed in the case of $N = 5$, meaning that the precision shrinkage has to be used only for the implementation of the mean-variance allocation.

Timing strategies. Also in this case timing strategies provide quite good performances in out-of-sample only if a model-based approach is used to obtain the predictions. On the contrary, the machine learning (data-driven) approach performs very poorly with a negative Sharpe ratio equal to -2.6%. It performs even poorer than the static strategy based on the precision shrinkage plug-in (-1.4%). Moreover, in terms of CEQ the under performances is dramatically huge with respect all the alternatives.

However, the volatility-timing is not the best investment strategy because in terms of Sharpe ratio of the static strategy involving the Ledoit and Wolf (2003) plug-in returns higher performances. However, in this case timing allows for a greater Sharpe ration than the naive strategy.

TABLE 2.15: Results for $N = 30$ assets - GMV approach

Strategies	SR	CEQ	RL
$1/N$	21.227%	0.891%	0.000
Static strategies:			
$\hat{\mu}, \hat{\Sigma}_{SC}$	22.483%	0.738%	-0.045%
$\hat{\mu}, \hat{\Sigma}_{sP}$	-1.410%	-9.446%	9.525%
$\hat{\mu}, \hat{\Sigma}_{LW}$	24.961%	0.775%	-0.124%
$\hat{\mu}, \hat{\Sigma}_{POET}$	21.872%	0.726%	-0.023%
Volatility-timing:			
$\hat{\mu}, \hat{\Sigma}_{DCC}$	22.088%	0.743%	-0.032%
$\hat{\mu}, \hat{\Sigma}_{ANN-DCC}$	-2.606%	-28590%	569%

Economic evaluation. As the previous case, we compute the *annualized percentage* maximum fee Δ that a mean-variance investor is willing to pay to switch from a strategy to another one. Table 2.16 shows the economic comparison between naive strategy and the alternatives.

TABLE 2.16: Economic fee Δ for $N = 30$ assets: optimal vs naive

	$\hat{\mu}, \hat{\Sigma}_{SC}$	$\hat{\mu}, \hat{\Sigma}_{LW}$	$\hat{\mu}, \hat{\Sigma}_{POET}$	$\hat{\mu}, \hat{\Sigma}_{sP}$
$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{ANN}$			
	0.0320	0.0276	0.0334	1.2529
0.0314	3429.9109			

Table 2.16 shows that the naive allocation is preferred respect to all the alternative strategies, either static or based on timing. In fact the investor requires the payment of a fee Δ to change the naive strategy with another one. The strategy associated with the lowest fee Δ to pay is the static one based on Ledoit and Wolf (2003) estimation of the covariance matrix. Hence, volatility-timing in this framework does not guarantees over performances with respect the static approaches.

Interestingly, in the Table 2.17 are reported the costs of using large-dimensional estimator instead of low-dimensional one within the GMV setting.

TABLE 2.17: Economic fee Δ for $N = 30$ assets: low vs large dimension

Strategies	$\hat{\mu}, \hat{\Sigma}_{LW}$	$\hat{\mu}, \hat{\Sigma}_{POET}$
$\hat{\mu}, \hat{\Sigma}_{SC}$	-0.004337182	0.001452158
$\hat{\mu}, \hat{\Sigma}_{sP}$	-1.225293	-1.219504

Overall, Tab. 2.17 shows that in this case using large dimensional tool within this low dimensional environment is beneficial. Indeed, the investor is willing to pay for using the large dimensional estimators instead of the low dimensional ones. This result is in contrast with respect to what we found with the other experiments. Nevertheless, sample covariance is still more attractive than the POET estimstor, meaning that not all the tools build for large dimension are useful.

About the economic performance of static strategies compared to the timing one is reported in Table 2.18. The data-driven volatility timing approach is the worst strategy and, therefore, the investor requires huge fees Δ to change a static for a data-driven. Then, the model-based timing is preferred to the static approaches in most of the cases. For example, the investor is willing to pay a small fee of $\Delta = 6$ basis points to use model-based timing instead of sample covariance, $\Delta = 20$ basis

points to change the POET and a very huge fee $\Delta = 12215$ basis points to use timing instead of precision shrinkage. Nevertheless, the investor prefers the Ledoit and Wolf (2003) estimator, such that she required the payment of $\Delta = 38$ basis points to change this static strategy with the model-based timing.

TABLE 2.18: Economic fee Δ for $N = 30$ assets: static vs timing

	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{ANN}$
$\hat{\mu}, \hat{\Sigma}_{SC}$	-0.0006	3429.8790
$\hat{\mu}, \hat{\Sigma}_{LW}$	0.0038	3429.8833
$\hat{\mu}, \hat{\Sigma}_{POET}$	-0.0020	3429.8775
$\hat{\mu}, \hat{\Sigma}_{sP}$	-1.2215	3428.6580

Finally by comparing timing strategies, it is possible to note that the investor requires a fee of $\Delta = 3428$ to switch from the model-based to data-driven approach. Hence, we can conclude that while the model-based timing has economic utility, data-driven timing does not.

In conclusion, in this experiment we find that forecasting is useful at some extent but is not able to defeat particular static strategies. In what follows, we analyse what happen in a large dimensional setting.

2.6.2 Large dimension

In this case we consider all the 500 constitutes of the S&P500 Index from 10/1999 to 10/2019, assuming that the asset universe N is represented by only the stocks for which we have a complete time series. Particularly, we get $N = 286$ time series with $T = 240$. Considering, again, an estimation window of $M = 120$, we have now that $T - M = 120$. In this framework, being M our sample size for the estimation, we are in a *large dimensional setting* because $M < N$.

Static strategies. In this case static approaches provide very good performances if compared to the naive $1/N$ strategy. Indeed, the naive strategy guarantees an out-of-sample Sharpe ratio of 30%, while the Ledoit and Wolf (2003) and POET estimators provide Sharpe ratios equal to 41.3% and 37.2% respectively. However, also in this case the Ledoit and Wolf (2008) test does not reject the null hypothesis of Sharpe ratios equality with p-values equal to 0.37 and 0.58 for both shrinkage and POET estimators. This means that, as for almost all experiments conducted in this Chapter,

the naive and the best strategy are statistically indistinguishable in the population. In other words, the over performance of optimal strategies could be sample driven.

TABLE 2.19: Results for $N = 286$ assets - GMV approach

Strategies	SR	CEQ	RL
$1/N$	30.125%	1.182%	0.000
Static strategies:			
$\hat{\mu}, \hat{\Sigma}_{LW}$	41.357%	1.072%	-0.301%
$\hat{\mu}, \hat{\Sigma}_{POET}$	37.252%	0.997%	-0.198%
Volatility-timing:			
$\hat{\mu}, \hat{\Sigma}_{DCC}$	41.824%	1.157%	-0.335%
$\hat{\mu}, \hat{\Sigma}_{ANN-DCC}$	-4.407%	-10.340%	14.256%

Timing strategies. Volatility timing with model-based predictions performs even better than the static allocation with Ledoit and Wolf (2003) plug-in (Sharpe ratio equal to 41.8%), confirming the utility of forecasting. However, the Ledoit and Wolf (2008) test of Sharpe ratio equality suggest that the two approaches (DCC timing and Ledoit & Wolf static) are statistically the same (p-value equal to 0.38). Hence, timing is useful but is not able to provide statistically over performances with respect the best static approach. Data-driven timing is useless and leads to negative Sharpe ratio (-4.4%).

Economic evaluation. As in the low-dimensional setting, in what follows we report the annualized percentage maximum fee Δ, that the investor is willing to pay to switch from a strategy to another one. Table 2.20 shows the economic comparison between naive strategy and all the optimal alternatives. Surprisingly, despite it is not the best strategy in terms of Sharpe ratio, the naive provides the highest economic benefit compared to the alternatives. Indeed, the investor requires the payment of a fee Δ for renouncing to this strategy. The less costly is the volatility timing. Hence, in economic terms the volatility timing represents the best alternative even if it is not able to perform better than the naive strategy. Clearly, as volatility timing we refer to the model-based one, because the data-driven timing approach is the worst among all the considered investment strategies.

TABLE 2.20: Economic fee Δ for $N = 286$ assets: optimal vs naive

$\hat{\mu}, \hat{\Sigma}_{LW}$	$\hat{\mu}, \hat{\Sigma}_{POET}$	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{ANN}$
0.0246	0.0334	0.0144	1.3874

Table 2.21 reports the economic comparison between static and timing strategies. As expected, the investor is willing to pay Δ to use the model-based timing strategy instead of both the static ones. In particular, she is willin to pay $\Delta = 102$ basis points for changing the static with Ledoit and Wolf (2003) estimation and $\Delta = 190$ basis points for the POET. On the contrary, static approaches are much better than the data-driven timing.

TABLE 2.21: Economic fee Δ for $N = 286$ assets: static vs timing

	$\hat{\mu}, \hat{\Sigma}_{DCC}$	$\hat{\mu}, \hat{\Sigma}_{ANN}$
$\hat{\mu}, \hat{\Sigma}_{LW}$	-0.0102	1.3628
$\hat{\mu}, \hat{\Sigma}_{POET}$	-0.0190	1.3540

Finally, we can conclude that overall forecasting is useful in economic terms even if the naive allocation remains the best one for the mean-variance investor. This is due to estimation error perhaps. However, we note that overall timing should be preferred to static estimation. Moreover, we also find that data-driven approaches are very scars in this setting if compared to standard (econometric) model-based approaches.

2.7 Conclusions

In this second Chapter we studied the usefulness of forecasting in portfolio selection problem. What we found is that, in a low dimensional setting, forecasting is useful as documented by previous studies. However if a static estimation strategy that properly accounts for estimation error is used, instead, the investment strategies based on forecasts becomes less competitive. This does not mean that forecasting becomes useless, but the difference in terms of performances between specific static strategies and timing become thinner. Indeed, the case of precision shrinkage estimator developed in the first Chapter is a perfect example of this result. However, the results for

global minimum variance asset allocation show that the precision shrinkage performances are very poor. This happen because the precision shrinkage is taught to be optimal within a mean-variance rather than minimum variance allocation.

An important contribution of this chapter lies on the assessment of the forecasting usefulness in the case of large dimensional setting. This second case has never been considered by previous studied. In such a case, the implementation of forecasting becomes even more important from the point of view of return/risk trade-off. From the side of CEQ returns the results are not very different. Overall, a perhaps surprising result is that full-timing strategies are not generally as good as partial-timing ones, where only mean or covariances are predicted. In general, for mean-variance asset allocation (where both mean and covariance are involved) it seems that mean predictions provide much better results than covariance forecasts. Therefore, forecasting can be seen as a useful tool for reducing estimation error in the mean vector, especially in large dimension.

In the end, another important novelty introduced in this chapter lies on the comparison between machine learning (data-driven) versus econometric (model-based) approaches in forecasting for timing strategies' implementation. Almost everywhere in this chapter we showed that machine learning forecasts are not useful, especially in low dimension. As explained above, this result could be driven by the adopted rolling-window approach (DeMiguel, Garlappi, and Uppal, 2009). Indeed, we used only $M = 120$ time observations for training the networks. Perhaps, this number is not enough large to ensure an accurate training. As argued by Zhang, Patuwo, and Hu (1998), neural network-based forecasting models have been for long time considered as unnecessary tools by professional forecasters. The reason lies on their poor performances within environments with few-data. Nowadays both economists and investors face data-rich environment, where neural networks perform very well. Nevertheless, canonical approaches for economic significance evaluation require the analysis of monthly data, that are known to be less noisy than the daily or infra-daily data. Therefore, the results simply show that machine-learning based timing strategy are not appropriate for long-run (low-frequency) portfolio selection problems. Further studies should be devoted to the analysis of high-frequency portfolios, where much more observations are available for training the machine learning algorithm. In this case it is more reasonable to expect an economic usefulness of ML.

Chapter 3

On the performance of clustered portfolios

3.1 Introduction

Clustering is one of the most important data mining algorithms, usually implemented for exploratory purposes and more complex tasks like anomaly detection or classification. Once a dissimilarity matrix is computed, several algorithms can be used get a final classification. When time series are involved, the clustering task becomes more complicated because it is difficult to define a proper distance among the time series: what does it mean that two time series are similar?
Moreover, the task further complicates when time series of a peculiar type, such as stock prices and returns, are considered. If we aim to reach an accurate classification, the empirical regularities of financial time series, defined *stylized facts* (Cont, 2001), must be considered in calculating the dissimilarities. First of all, stock prices are generally integrated time series. Therefore, it is more common to consider returns rather than prices in financial time series clustering. Moreover, we have to consider that the squared returns (as well as the absolute values) are viewed as proxies for volatility (Forsberg and Ghysels, 2007). With this respect, a second crucial stylized fact lies in the evidence that squared (or absolute) returns are highly auto-correlated. This phenomenon is known as *volatility clustering*, meaning that volatility tends to be clustered in groups of low/high values over time. Hence, time variation in the volatility is usually considered for clustering financial time series. In the end, we should also take into account the fact that the empirical densities of the returns' time series are usually non-Gaussian, asymmetric and heavy-tailed.
When applied to stock returns, an immediate application of time series clustering

can be found in the portfolio construction (Mantegna, 1999; Caiado and Crato, 2010; Iorio et al., 2018; Raffinot, 2017), also defined as the *asset allocation* task. The asset allocation task involves deciding how many and which kind of assets to include in a portfolio for investment purposes.

Nowadays, investors usually face the problem of having many available assets N greater than the number of time observations T. Because of dimensionality, there are difficulties in estimating the inverse of asset returns' covariance matrix, which is singular. The impossibility of inverting the covariance matrix, which is a crucial operation needed to implement most optimization strategies, increases the estimation error. As a consequence, in a high-dimensional setting (i.e. when $N > T$) the estimation error cause a reduction of portfolio performances in out-of-sample (Michaud, 1989; Ledoit and Wolf, 2003; Jagannathan and Ma, 2003).

Following this evidence, clustering can be a very powerful tool for portfolio selection. Indeed, it allows to alleviate the course of dimensionality by finding smaller sets of stocks that can be used to build roughly diversified funds. Then, these funds become the input of a portfolio optimization strategy.

For example, DeMiguel, Garlappi, and Uppal (2009) grouped stocks on the basis of industry sectors rather than defining the clusters with an unsupervised learning approach. Caiado and Crato (2010), using an hierarchical unsupervised learning algorithm, proposed to define clusters of stocks with similar conditional volatilities by employing a GARCH-based distance across time series. D'Urso et al. (2013) followed a similar approach by considering a partitioning clustering algorithm rather than hierarchical. Lahmiri (2016) studied the idea of using time series' self-similarity (i.e. the Hurst exponent) to this aim. Raffinot (2017) considered a correlation-based distance to build portfolio of stocks. Iorio et al. (2018) proposed the adoption of a p-spline based distance in the definition of portfolios. More recently, Mattera, Giacalone, and Gibert (2021) exploited the role of stock returns' distributional characteristics in defining clustered portfolios. The aforementioned papers are just few examples that document the use of unsupervised learning in asset allocation.

These studies show the validity of using clustering in portfolio selection, by considering alternative approaches of measuring time series distances. However, it is still unclear which clustering approach is better than others from the portfolio selection task perspective, under what conditions and if the performances are robust for different datasets. In other words, a comparative study of the different clustering

approaches from an asset allocation perspective is still missing. This chapter aims to fill this gap, understanding not only the merits of clustering in finance but also under what conditions the portfolios constructed with different clustering approaches can be expected to perform better than others consistently.

Hence, this question turns back to the problem of measuring similarity. The most common approach employed for measuring distances in clustering is the standard Euclidean distance. However, it is well known that in the context of time series, the simple Euclidean distance between observations is not appropriate because it does not take into account important aspects of the time series, such as the auto-correlation structures trends and so on.

The general approaches for time series classification can be divided into three well-known classes: observation-based clustering, feature-based and model-based (for a detailed discussion, see Liao (2005)). Briefly, while observation-based approaches (Coppi and D'Urso, 2006; D'Urso, De Giovanni, and Massari, 2018) are useful for clustering short time series, they do not consider the data features that are very important, especially in clustering of financial time series (Bastos and Caiado, 2021). The features-based approaches, on the other hand, aim to cluster time series with similar characteristics like the auto-correlation or the partial auto-correlation function (D'Urso and Maharaj, 2009), conditional moments (Cerqueti, Giacalone, and Mattera, 2021), periodogram ordinates (Caiado, Crato, and Peña, 2006) or cepstral coefficients (Maharaj and D'Urso, 2011; D'Urso et al., 2020). In the end, the model-based approaches aim to cluster time series according to parameters estimated by statistical models (D'Urso, De Giovanni, and Massari, 2016; Piccolo, 1990; Otranto, 2008; Iorio et al., 2016; D'Urso, Maharaj, and Alonso, 2017; Mattera, Giacalone, and Gibert, 2021). Several ways of defining distances between financial time series will be reviewed in paragraph 3.2 of this chapter.

Moreover, this chapter also introduces a new approach for clustering financial time series based on the "deviation from Gaussianity" stylized fact. The deviation from Gaussianity relates to the assumption about the returns' probability distribution. Indeed, the Gaussian distribution is not a reliable choice for stock return modelling purposes, and more sophisticated probabilistic assumptions which account for normality deviation are needed. A vast literature considered an extension for non-Gaussian distribution of standard statistical models (e.g. GARCH) for forecasting

purposes, finding that alternative skewed and fat-tailed distribution allow for improvements in forecasting accuracy of both returns and volatilities (Bollerslev, 1987; Harvey and Zhou, 1993; Wilhelmsson, 2006; Curto, Pinto, and Tavares, 2009; Cerqueti, Giacalone, and Mattera, 2020). More in detail, we develop a clustering procedure of model-based type, assuming that the time series are generated by the same underlying probability distribution but with different parameters. Therefore, stock clusters can be formed by inspecting the differences across distribution parameters. Clearly, by specifying a very general underlying distribution, we can account for a wide range of possible exceptional cases. Different stock returns can be generated by different distributions that are exceptional cases of a more general family of probability distribution. Since the parameters are not allowed to change over time, we define this clustering model as *distribution-based clustering with static parameters*. On the contrary, in the time series context, the assumption of static parameters is not reliable because parameters are likely to be time-varying. If we introduce time-varying parameters, the clustering procedure becomes much more complex because the time variation in the distribution parameters has to be adequately modelled. As an additional source of innovation, this chapter introduces the so-called *distribution-based clustering with time-varying parameters*.

The structure of the rest of the chapter is as follows. In paragraph 3.2, an overview of clustering models developed for financial time series is provided. Paragraph 3.2.1 discusses the most common approaches for computing distances between financial time series, while paragraph 3.2.2 explains the clustering algorithm adopted in the entire chapter. In particular, we consider a Partition Around Medodids (PAM) clustering algorithm. The motivation of this choice over well-known alternatives is discussed in detail therein. Then, in paragraph 3.3, the new clustering procedures (i.e. static and time-varying distribution-based clustering). Paragraph 3.4 provide two application of the discussed clustering approaches to portfolio selection.

3.2 Clustering of financial time series

As argued by De Miguel, Garlappi, and Uppal (2007), allocating the wealth across portfolios of stocks rather than individual stocks reduces estimation error because diversified portfolios have lower idiosyncratic volatility than individual assets. Therefore, investing in already diversified funds can be seen as a good tool for alleviating

the problem of estimation error. Moreover, clustering reduces the course of dimensionality, which causes problems in asset allocation in high-dimensional settings. A clear advantage of the 1/N rule is that it is straightforward to apply to a large number of assets, in contrast to optimizing models, which typically require additional parameters to be estimated as the number of assets increases.
In what follows, we describe the most common approaches for measuring distances among financial time series and the clustering algorithm that will be used in the empirical experiments.

3.2.1 Measuring time series similarity

Given a pair of stocks returns' time series $r_{n,t}$ and $r_{n\prime,t}$, a first approach for clustering the time series could be simply the Euclidean distance between the two raw time series:

$$d_{\text{EUCL}_{n,n\prime}} = \sqrt{\sum_{t=1}^{T}(r_{n,t} - r_{n\prime,t})^2} \tag{3.1}$$

There are at least two reasons why this measure is inadequate for clustering financial time series. First of all, it does not account for the serial correlation structure of the data. This makes the simple Euclidean distance useless for time series in general (Díaz and Vilar, 2010). Moreover, it ignores the correlation structure of the assets, a crucial aspect of portfolio selection.
Starting from this idea, Mantegna (1999) and Raffinot (2017) proposed to quantify the dissimilarity among different stocks according to their estimated correlation coefficient. The simple difference between estimated correlation coefficients cannot be used as a distance since it does not fulfil the axioms that define a metric. To overcome this issue, Mantegna (1999) proposed to use the following distance:

$$d_{\text{COR}_{n,n\prime}} = \sqrt{2(1 - \rho_{n,n\prime})} \tag{3.2}$$

that depends by the correlation $\rho_{n,j}$ between the n-th stock returns $r_{n,t}$ and the $n\prime$-th returns $r_{n\prime,t}$. However, the correlation coefficient is still a static measure that does not properly account for the dynamic structure of the time series. If we aim to cluster

similar stocks according to the correlation structure, an alternative is to consider a distance based on their auto-correlation functions:

$$d_{\text{ACF}_{n,n\prime}} = \sqrt{\sum_{l=1}^{L} (\rho_{n,l} - \rho_{n\prime,l})^2} \tag{3.3}$$

with $\rho_{n,l}$ and $\rho_{n\prime,l}$ be the estimated autocorrelations of the time series n and $n\prime$, respectively, for some lags $l (l : 1, \ldots, L)$.

As briefly mentioned above, other approaches are based on the frequency domain representation of the time series. Frequency domain approaches consider the spectral density of the time series. An unbiased estimator of the actual spectral density is the periodogram. Hence, let:

$$I_n(\lambda_l) = \frac{1}{2\pi T} \left| \sum_{t=1}^{T} r_{n,t} e^{-i\lambda t} \right|^2 \quad \lambda \in [-\pi, \pi]$$

be the periodogram of the n-th returns' time series $r_{n,t}$ at frequencies $\lambda_l = 2\pi l/T$, given $\{l = -L, \ldots, L\}$ with $L = (T-1)/2$. Given σ_n^2 be the sample variance of $r_{n,t}$, Caiado, Crato, and Peña (2006) proposed to consider the following distance between the log normalized periodograms:

$$d_{\text{NPER}_{n,n\prime}} = \sqrt{\sum_{l=1}^{L} \left(\log \frac{I_n(\lambda_l)}{\sigma_n^2} - \log \frac{I_{n\prime}(\lambda_l)}{\sigma_{n\prime}^2} \right)^2} \tag{3.4}$$

Another interesting approach for clustering financial time series can be found in the fact that the returns are usually characterized by *long memory*. In a stationary time series, the term "long memory" means that exists a significant dependence between the present and all points in the past (Lo, 1991). Sometimes, it is also defined as "long-range dependence" or "long term persistence". Lahmiri (2016) proposed to cluster time series according to their long-memory values. The study of the "long-memory" behaviour of a time series is usually done through the Hurst exponent that the R/S analysis can estimate, that is, the range of partial sums of deviations of a time series from its mean, re-scaled by its standard deviation. More in detail, in the R/S analysis, estimation of the Hurst exponent H can be summarized as follows. Let $(r_1, r_2, \ldots, r_T)$ be the time series vector and denote by $\bar{r}$ the sample mean of these

observations. Then the usual re-scaled range statistic, that we call $\tilde{Q}$, can be obtained as:

$$\tilde{Q} = \frac{1}{\sigma}\left[\max_{1\leq k\leq T}\sum_{t=1}^{k}(r_t - \bar{r}) - \min_{1\leq k\leq T}\sum_{t=1}^{k}(r_t - \bar{r})\right]$$

with σ is the standard deviation of r_t. The time series of length T is firstly divided into K blocks where each $k(k = 1, \ldots, K)$-th block is of size K/T. Then, we compute the R/S statistics $\tilde{Q}$ for each k. Finally, the Hurst exponent H is obtained as the coefficient of linear regression where the logarithm of $\tilde{Q}$ is the dependent variable and $\log(T)$ is the regressor. Then, we can compute a dissimilarity measure based on the H estimates as follows (Lahmiri, 2016):

$$d_{\text{HURST}_{n,n\prime}} = \sqrt{(H_n - H_{n\prime})^2} \tag{3.5}$$

Instead, a different approach is to consider that the returns' time series are generated by a specific statistical model, such that we can measure the proximity between the fitted models. With this respect, an important contribution is Piccolo (1990) that defined a metric in the class of invertible ARIMA processes as the Euclidean distance between the AR(∞) representation of a given stock returns series $r_{n,t}$. In practice, we compute an AR(K) representation where K is selected according to information criteria. Then, the AR-distance of Piccolo (1990) could be written as:

$$d_{\text{ARMA}_{n,n\prime}} = \sqrt{\sum_{k=1}^{K}(\pi_{n,k} - \pi_{n\prime,k})^2} \tag{3.6}$$

with $\pi_{n,k}$ and $\pi_{n\prime,k}$ the vector of the autoregressive coefficients for the n-th and $n\prime$-th stocks, respectively. If $K_1 \neq K_2$, we take $K = \max(K_1, K_2)$ and $\pi_{n,k} = 0$ for $K > K_1$ and, similarly, $\pi_{n\prime,k} = 0$ for $K > K_2$.

All these measures ignore a crucial stylized fact: the time-varying nature of volatility. As proposed by many authors (Otranto, 2008; Caiado and Crato, 2010; D'Urso et al., 2013), if we aim to cluster time series with similar volatility behaviour, we should consider a distance of model-based type between estimated parameters of GARCH processes. The standard GARCH(p,q) model of Bollerslev (1986) can be specified as

follows:

$$\begin{aligned} r_t - \mu_t &= \epsilon_t \\ \epsilon_t &= \sigma_t z_t \quad \text{with} \quad z_t \sim \mathcal{N}(0,1) \end{aligned}$$

where z_t is called *innovation* and it is a process with zero mean and unit variance, while σ_t is a univariate stochastic process independent from z_t of the following form:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-p}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-q}^2$$

with $\omega > 0$, $0 \leq \alpha_i < 1$, $0 \leq \beta_j < 1$ and $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$. Parameters' estimation can be easily done by maximum likelihood. According to Caiado and Crato, 2010; D'Urso et al., 2013, assuming a GARCH(1,1) process[1], the estimated parameters $\hat{\alpha}$ and $\hat{\beta}$ for each i-th time series can be stored into a matrix $\mathbf{T} = (\hat{\alpha}, \hat{\beta})$ with Ω the covariance matrix associated to the estimates contained in the $\mathbf{T}$'s. Therefore, we can consider the following Mahalanobis-like distance between two returns time series $r_{n,t}$ and $r_{n\prime,t}$:

$$d_{\text{GARCH}_{n,n\prime}} = \sqrt{(\mathbf{T}_n - \mathbf{T}_{n\prime})' \, \Omega^{-1} \, (\mathbf{T}_n - \mathbf{T}_{n\prime})} \tag{3.7}$$

where trough the weighting matrix inverse Ω^{-1} we also account for the uncertainty in the parameter estimation step. Nevertheless, it could also be set equal to the Identity matrix $\mathbf{I}$.

3.2.2 Clustering algorithms

Once a dissimilarity measure between the objects has been specified, a clustering algorithm must be chosen to obtain the partitions. Indeed, as stated by Liao, 2005, most clustering techniques for time series *"try to modify the existing algorithms for clustering static data in such a way that time series can be handled"*. This is usually done using proper time series distance matrices.

[1] The specification of a GARCH(1,1) model is a parsimonious representation of an ARCH(∞) model (for the proof see Bollerslev, 1986).

Most of the applications to financial time series clustering have been based on either the hierarchical Mantegna, 1999; Caiado and Crato, 2010; Raffinot, 2017 or the k-means algorithms Nanda, Mahanty, and Tiwari, 2010; D'Urso et al., 2013; D'Urso, De Giovanni, and Massari, 2016; D'Urso et al., 2020.

Hierarchical clustering methods work by grouping time series into a tree of clusters Nanda, Mahanty, and Tiwari, 2010. However, the performance of a hierarchical clustering method often suffers from its inability to adjust once a merge decision has been executed Liu et al., 2021. Moreover, the hierarchical algorithms typically are time-consuming Xie et al., 2020 with quadratic complexity, while one of k-means like algorithms is linear. In other words, despite it representing a widely used alternative for clustering financial time series, hierarchical clustering is not recommended when there are many assets with long time series. This is especially true when dealing with high-dimensional settings, when very large N is involved. In this case the hierarchical approaches are proved to perform quite poorly.

Differently, the k-means based approaches are computationally less challenging. Therefore, in this Chapter, we consider k-means like clustering algorithms. The k-Means algorithm is one of the most popular clustering approaches aiming to partition the time series into a predetermined number of clusters. It relies on an iterative scheme based on the minimization of an objective function, which is usually chosen to be the total distance between all patterns from their respective cluster centres:

$$\min : \sum_{i=1}^{N} \sum_{c=1}^{C} d_{i,c} \tag{3.8}$$

where N is the number of the time series to be grouped, C is the number of clusters (a priori fixed), c represents the centre such that $d_{i,c}$ is the distance between each time series i from the centroid of the c-th cluster.

However, instead of considering all time series, one can analyze prototypal time series, i.e., time series that retain the main features of similar time series classified in the same group. To this end, we adopt the so-called *Partitioning Around Medoid* (PAM) approach. To summarize, the PAM approach provides a robustification of the usual k-means clustering algorithm. To see why, with the PAM approach, the centroid object is randomly selected before calculating the distances of each time series data concerning the centroid itself, such that initial partition is made based on the

closeness of each object to the clustering centroid. The prototypes of each group, the *medoid time series*, are time series that belongs to the sample and are not virtual, as happen with the *k*-means algorithm Ushakov and Vasilyev, 2021.

The possibility of obtaining non-fictitious representative time series in the clusters is very appealing and helpful in many applications. Moreover, the fact that the centroids are real-time series also improves interpretability results. As a further advantage, as shown by Arora, Varshney, et al., 2016, the PAM approach is better in terms of execution time, less sensitive to outliers and reduces noise with respect to the *k*-means. Moreover, as argued by D'Urso, De Giovanni, and Massari, 2016, since the medoids are observed time series the constraints on the GARCH coefficients needed to compute the GARCH-based dissimilarity are always satisfied. This is not guaranteed when using the *k*-means approach. Therefore we claim that a PAM approach should be preferred to a classical *k*-means, especially if financial time series are involved.

The main drawback of *k*-means and *k*-medoids algorithms is that the number of clusters has to be identified in advance. Several approaches can be used to this aim. In this paper, following many other authoritative studies Arbelaitz et al., 2013; Batool and Hennig, 2021, we choose the number of clusters employing the Average Silhouette Width (ASW) criterion of Rousseeuw, 1987.

3.3 A unified framework for distribution-based clustering

There is pervasive evidence documenting the distributional characteristics of stock returns. Due to the relevance of returns' distribution in finance, clustering stocks on the basis of this information can be handy.

The idea of clustering time series based on their distributional characteristics is mainly due to Nanopoulos, Alcock, and Manolopoulos (2001) that considered both skewness and kurtosis in the clustering process. Successively, Wang, Smith, and Hyndman (2006), and Fulcher and Jones (2014) proposed approaches of clustering based on multiple features, including static mean, variance, skewness and kurtosis. The studies mentioned above do not assume any underlying probability distribution for the time series but use sample estimators for those features. Differently, following a model-based approach, D'Urso, Maharaj, and Alonso (2017) proposed to cluster time series using extremes, i.e. according to static parameters estimated

by a Generalized Extreme Value (GEV) distribution. Similarly, in a recent contribution (Mattera, Giacalone, and Gibert, 2021) we considered an approach of clustering based on (static) parameters estimated by a Skewed Generalized Error Distribution (SGED). In that paper, we applied a distribution-based clustering algorithm with static parameters to financial time series, demonstrating that portfolios can be formed based on the obtained clusters. Further, we demonstrated the overperformance of the asset allocation strategy based on this clustering approach concerning to alternative clustering methods.

In the next paragraph 3.3.1, we provide the details about the clustering approaches with static parameters. Then, the extension to time-varying parameters is discussed in paragraph 3.3.2.

3.3.1 Static parameters

The distribution-based clustering approach with static parameters can be formalized as follows. Let $\mathbf{Y}\{y_{n,t} : n = 1,\dots,N; t = 1,\dots,T\}$ be the matrix containing the N time series of length T:

$$\mathbf{Y} = \begin{bmatrix} y_{1,1} & \cdots & y_{n,1} & \cdots & y_{N,1} \\ \vdots & \dots & y_{n,t} & \dots & \vdots \\ y_{1,T} & \cdots & y_{n,T} & \cdots & y_{N,T} \end{bmatrix} \tag{3.9}$$

Let us suppose that each column of the (3.9) is generated by a probability density function $p(\cdot)$ that is characterized by the presence of J parameters. The number of J parameters depends on the underlying distributional assumption. For example, if we suppose a Gaussian density:

$$p(y;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2} \tag{3.10}$$

we have $J = 2$ because $p \sim N(\mu,\sigma^2)$, where μ is the mean and σ^2 is the variance. In presence of a general $p(\cdot)$ density, the distribution-based clustering considers the following $(N \times J)$ matrix F as the input of the algorithm:

$$\mathbf{F} = \begin{bmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,J} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,J} \\ \vdots & \vdots & \vdots & \vdots \\ f_{n,1} & f_{n,2} & \cdots & f_{n,J} \\ \vdots & \vdots & \vdots & \vdots \\ f_{N,1} & f_{N,2} & \cdots & f_{N,J} \end{bmatrix} \tag{3.11}$$

where on the J columns of $\mathbf{F}$ there are the $j = 1, \ldots, J$ parameters for the N assets that are indexed by the rows. In the case of Gaussian density (3.24) the matrix (3.11) becomes:

$$\mathbf{F}_{\text{norm}} = \begin{bmatrix} \mu_1 & \sigma_1^2 \\ \mu_2 & \sigma_2^2 \\ \vdots & \vdots \\ \mu_n & \sigma_n^2 \\ \vdots & \vdots \\ \mu_N & \sigma_N^2 \end{bmatrix} \tag{3.12}$$

Clearly, in specifying the density $p(\cdot)$, it would be advantageous to choose a very general distribution to account for a wide range of possible exceptional cases. Indeed, the observed characteristics of financial time series motivated the exploration of distributions that can accommodate properties such as fat-tailedness and skewness. A critical desired property of these classes is that maximum likelihood estimation of the parameters is possible.

In finance, a commonly employed distribution for relaxing the gaussianity assumption is the t-student with fatter tails than the Gaussian. The t-student distribution is characterized by $J = 3$ parameters with the following density function:

$$p(y; \mu, \phi, v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\phi\sqrt{\pi v}} \left(1 + \frac{(y-\mu)^2}{v\phi}\right)^{\frac{v+1}{2}} \tag{3.13}$$

where μ is the location, ϕ is the scale, and v is the shape parameter that controls the degree of tails' fatness. Supposing that the time series in (3.9) are distributed as t-student with different parameters lead to the construction of the following matrix:

$$\mathbf{F}_{\text{std}} = \begin{bmatrix} \mu_1 & \phi_1 & v_1 \\ \mu_2 & \phi_2 & v_2 \\ \vdots & \vdots & \vdots \\ \mu_n & \phi_n & v_n \\ \vdots & \vdots & \vdots \\ \mu_N & \phi_N & v_N \end{bmatrix} \tag{3.14}$$

that then becomes the input of the clustering algorithm. However, a class of asymmetric distributions that accommodate heavy tails and skewness is represented by the Skewed Exponential Power Distribution (SEPD) Fernandez, Osiewalski, and Steel, 1995; Fernández and Steel, 1998; Theodossiou, 2015; Komunjer, 2007, that generalizes the Exponential Power Distribution for skewness. The SEPD is characterized by 4 parameters, i.e. location μ, scale ϕ, skewness λ and shape v. A random variable Y is said to have a Skewed Exponential Power Distribution if its probability density function is the following (Ayebo and Kozubowski, 2003):

$$p(y;\mu,\sigma,v,\lambda) = \frac{v}{\sigma\Gamma\left(1+\frac{1}{v}\right)}\frac{\lambda}{1+\lambda^2}\exp\left(-\frac{\lambda^p}{\sigma^v}[(z-\mu)^+]^v - \frac{1}{\sigma^v\lambda^v}[(z-\mu)^-]^v\right) \tag{3.15}$$

where:

$$(z-\mu)^+ = \max(z-\mu;0) \quad \text{and} \quad (z-\mu)^- = \max(\mu-z;0)$$

Some papers (for example see Ayebo and Kozubowski, 2003; Komunjer, 2007; Zhu and Zinde-Walsh, 2009; Theodossiou, 2015) constructed seemingly different classes of SEPD distributions. However, as suggested by Zhu and Zinde-Walsh, 2009, all of them are reparametrizations of the SEPD proposed by Fernandez, Osiewalski, and Steel, 1995; Fernández and Steel, 1998. An essential feature of the EPD is that it includes many common distributions as special cases, depending on the value of shape v and skewness λ parameters (Fig. 3.1).

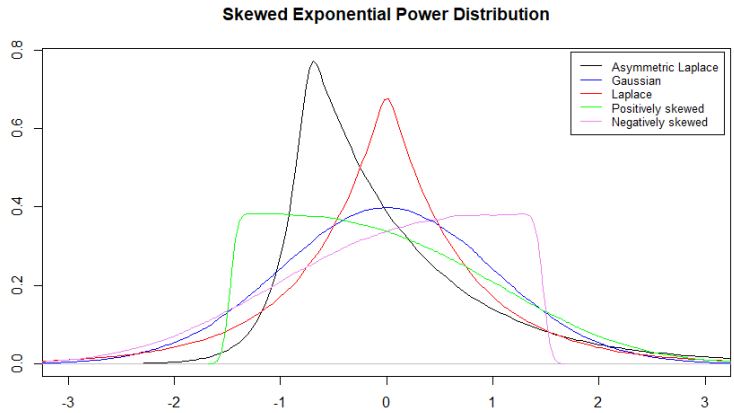


FIGURE 3.1: Skewed Exponential Power Distribution for different values of shape and skewness.

In particular, for $\lambda = 1$, the distribution is symmetric about μ so we obtain the symmetric exponential power distribution. If $\lambda = 1$ and $v = 2$ we obtain the Gaussian distribution. In the case $\lambda \neq 1$, by letting $v = 1$ we obtain the skewed Laplace distribution. Moreover, for $v = 2$ and $\lambda \neq 1$, we obtain the skewed normal distribution as defined in Mudholkar and Hutson, 2000. Many financial applications of the EPD as well as its skewed extensions have been considered[2].

The great flexibility of the SEPD can be successfully exploited in the clustering process if the aim is to form distribution-based clusters. In the case which the time series that are generated by a Skewed Exponential Power Distribution of parameters μ_n, σ_n, p_n and λ_n, the matrix (3.11) becomes:

$$\mathbf{F}_{\text{sepd}} = \begin{bmatrix} \mu_1 & \sigma_1 & p_1 & \lambda_1 \\ \mu_2 & \sigma_2 & p_2 & \lambda_2 \\ \vdots & \vdots & \vdots & \vdots \\ \mu_n & \sigma_n & p_n & \lambda_n \\ \vdots & \vdots & \vdots & \vdots \\ \mu_N & \sigma_N & p_N & \lambda_N \end{bmatrix} \tag{3.16}$$

As we have already said, regardless of the specified distribution, the main idea underlying the distribution-based approach with static parameters is to use the matrix (3.11) as the input of the clustering procedure. For the sake of consistency, in what

[2]For example see Hsieh, 1989; Nelson, 1991; Duan, 1999; Ayebo and Kozubowski, 2003; Komunjer, 2007; Christoffersen et al., 2010; Theodossiou, 2015; Cerqueti, Giacalone, and Panarello, 2019; Cerqueti, Giacalone, and Mattera, 2020

follows differently from Mattera, Giacalone, and Gibert (2021)[3] we use a Partition Around Medoids (PAM) as shown in (3.32) to generate clusters once the dissimilarity is computed. In the case of a generic distribution-based clustering with static parameters, a reasonable dissimilarity is given by the Euclidean distance between parameters such that:

$$d_{n,n\prime} = \sqrt{\sum_{j=1}^{J} \mathbf{F}_n - \mathbf{F}_{n\prime}} \tag{3.17}$$

where $\mathbf{F}_n$ represents the n-th row of the matrix (3.11). Therefore, the clustering problem can be written as follows:

$$\min : \sum_{i=1}^{N}\sum_{c=1}^{C} d_{i,c} = \sum_{i=1}^{N}\sum_{c=1}^{C} \sqrt{\sum_{j=1}^{J} \mathbf{F}_n - \mathbf{F}_c} \tag{3.18}$$

where c is the centroid time series of the c-th cluster. Clearly, in the case of Gaussian distribution the distance becomes:

$$d_{n,n\prime} = \sqrt{\sum_{j=1}^{2} \mathbf{F}_{\text{norm},n} - \mathbf{F}_{\text{norm},n\prime}} \tag{3.19}$$

because $J = 2$, while in the case of t-student it is equal to:

$$d_{n,n\prime} = \sqrt{\sum_{j=1}^{3} \mathbf{F}_{\text{std},n} - \mathbf{F}_{\text{std},n\prime}} \tag{3.20}$$

because $J = 3$ and so forth. In conclusion, we can define the solution to the problem (3.18) as the *distribution-based clustering with static parameters*. This clustering approach can potentially be applied to any time series with known distribution. For example, interesting applications of this clustering method can be devoted to classifying count time series that follow a Poisson distribution.

[3] In that paper, we used the Entropy-weighted k-means clustering algorithm of Jing, Ng, and Huang (2007) in order to assign specific weight to each static parameter. In what follows, instead, we implicitly suppose that all the parameters have the same importance. This assumption can be relaxed but not considered in the thesis and is the object of future research. Note that most of the previous studies also considered an implicit equal weighting scheme.

3.3.2 Time-varying parameters

The main drawback of the clustering approach discussed so far is that it considers static distribution features (i.e. static mean, static variance, etc.). In time series context, these parameters are likely time-varying. Despite there is a vast literature documenting this evidence with many statistical tools developed for modeling time variation in the parameters (Cox et al., 1981; León, Rubio, and Serna, 2005; Harvey, 2013; Creal, Koopman, and Lucas, 2013; Harvey and Sucarrat, 2014; Caivano and Harvey, 2014; Harvey and Lange, 2017), a clustering approach based on time-varying parameters has been only recently explored by Cerqueti, Giacalone, and Mattera (2021).

The idea underlying the clustering with time-varying parameters is, in principle, straightforward. Let us suppose to consider the matrix $\mathbf{Y}$ as in (3.9), that contains the N time series on the columns with their T observations in the rows. Let us now consider the simple case of Gaussian density $p(\cdot)$ with time-varying parameters, where $p \sim N\left(\mu_t, \sigma_t^2\right)$, i.e.:

$$p(y_t; \mu_t, \sigma_t^2) = \frac{1}{\sigma_t \sqrt{2\pi}} e^{-(y_t - \mu_t)^2 \big/ 2\sigma_t^2} \tag{3.21}$$

Also, in this case, the goal is to cluster time series with the same parameters μ_t and σ_t^2. However, differently from the static case, now the distribution parameters are time series themselves. Therefore, the similarity between time series is defined as the degree to which their parameters vary over time. In other words, two-time series n and $n\prime$ belong to the same cluster if they share similar time patterns in the time-varying mean μ_t and/or in the time-varying variance σ_t^2. For example, Cerqueti, Giacalone, and Mattera (2021) considered a distance similar to the (3.3), i.e. based on the auto-correlation structure of the time-varying parameters. Therefore, to some extent, the distribution-based clustering with time-varying parameters is even more related to the time series clustering literature with respect to its alternative with static parameters.

However, considering more than one time-varying parameter in the clustering process has a consequence in terms of the dataset structure. Indeed, for each n-th time series, there are $J \geq 1$ parameters that vary over time as well. In other words, in this

case, instead of dealing with a matrix $\mathbf{F}$ we have a three-dimensional *tensor*, specifically a time data array (D'urso, 2004; D'Urso and Maharaj, 2012; D'Urso et al., 2019), where three dimensions (N time series, J parameters, T time) are involved such that:

$$\mathbf{F} = \left\{ f_{n,j,t} : n = 1, \ldots, N; j = 1, \ldots, J; t = 1, \ldots, T \right\} \tag{3.22}$$

The graphical representation of the tensor is showed in Fig. 3.2.

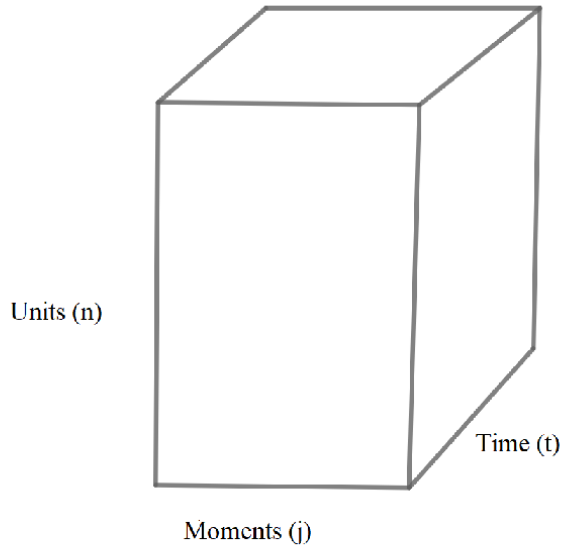


FIGURE 3.2: Graphical representation of $\mathbf{F}\{f_{n,j,t} : n = 1, \ldots, N; j = 1, \ldots, J; t = 1, \ldots, T\}$

Clustering time series within a tensor input is more difficult than the standard case where a matrix is involved. To overcome this issue, in Cerqueti, Giacalone, and Mattera (2021) we considered a clustering problem with a *target time-varying parameter*. In doing so, one can obtain J alternative classifications based on each j-th time-varying parameter. For example, in the case of Gaussian density, is it possible to get a mean-based or a variance-based classification.

Differently from Cerqueti, Giacalone, and Mattera (2021), in what follows, we propose a novel clustering approach that, instead of choosing a single target parameter, is based on multiple time-varying parameters. Therefore, it represents a direct extension for time-varying parameters of the model previously discussed in paragraph 3.3.1.

The proposed clustering procedure is based on three main steps. In the first one, the time-varying parameters are estimated and stored in a three-dimensional tensor **F**, i.e. the time data array. Then, the elements of the resulting time data array are clustered according to the two-step procedure of D'URSO (2004).

In order to model and estimate the time varying parameters we use the Generalized Autoregressive Score (GAS) model Creal, Koopman, and Lucas, 2013. Let $\mathbf{Y}\{y_{n,t} : n = 1, \dots, N; t = 1, \dots, T\}$ be the matrix containing N time series of length T generated by the following observation density $p(\cdot)$:

$$y_{n,t} \sim p(y_{n,t}|f_{n,t}, \mathcal{F}_{n,t}; \theta_n), \tag{3.23}$$

where θ_n is a vector of static parameters, $\mathcal{F}_{n,t}$ is the information set at time t, $f_{n,t}$ is a vector of length $J(j = 1, \dots, J)$ of time-varying parameters depending by the probability distribution. The model's information set at a given point in time t, $\mathcal{F}_{n,t}$, is obtained by the previous realizations of the time series $y_{n,t}$ and the time varying parameters $f_{n,t}$.

In this context, crucial is the role of the time-varying parameter vector $f_{n,t}$ since it represents the input of the proposed clustering procedure. In the paper we first suppose that all the time series are generated by a Gaussian density with different time-varying parameters. Therefore, for each n-th time series we have that the (3.23) is equal to:

$$p(y_t|f_t, \mathcal{F}_t; \theta) = \frac{1}{\sigma_t \sqrt{2\pi}} e^{-(y_t - \mu_t)^2 \big/ 2\sigma_t^2} \tag{3.24}$$

where the time-varying parameters are $f_{n,t} = (\mu_{n,t}, \sigma^2_{n,t})$ with $J = 2$, where $\mu_{n,t}$ and $\sigma^2_{n,t}$ represent respectively the conditional mean and the conditional variance for the n-th time series. By assuming different density functions, we get more (or less) J-th time varying parameters. An interesting example exploited in the paper is given by the assumption of t-student distribution for the density (3.23) for each n-th time series:

$$p(y_t|f_t, \mathcal{F}_t; \theta) = \frac{\Gamma\left(\frac{v_t+1}{2}\right)}{\Gamma\left(\frac{v_t}{2}\right) \phi_t \sqrt{\pi v_t}} \left(1 + \frac{(y_t - \mu_t)^2}{v_t \phi_t}\right)^{\frac{v_t+1}{2}} \tag{3.25}$$

where the time-varying parameters $f_t = (\mu_t, \phi_t, v_t)$ are the conditional location μ_t, the conditional scale ϕ_t and the conditional shape v_t, respectively. The assumption of t-student distribution is crucial for modelling time series with heavy-tails. Additional probability distributions can be considered in this framework, such as the Skew Normal, the Generalized Skew-t and many others.

In what follows, we propose to estimate the time-varying parameters (i.e. the conditional moments) by the Generalized Autoregressive Score model of order one, namely the $GAS(1,1)$. The $GAS(1,1)$, for any specification of the density in (3.23), can be written as:

$$f_{n,j,t} = \omega_{n,j} + \mathbf{A}_{n,j,1} s_{n,j,t-1} + \mathbf{B}_{n,j,1} f_{n,j,t-1} \tag{3.26}$$

where $\omega_{n,j}$ is a real vector and $\mathbf{A}_{n,j,1}$ and $\mathbf{B}_{n,j,1}$ are diagonal matrices. All the scalar parameters $\omega_{n,j}, \mathbf{A}_{n,j,1}, \mathbf{B}_{n,j,1}$ are collected in the vector θ_n. An appealing feature of the GAS model is that the vector of parameters θ_n is estimated by maximum likelihood (for the details see Creal, Koopman, and Lucas, 2013). Moreover, $s_{n,j,t}$ is the *scaled* score of the conditional density (3.23) in a time t with respect to a j-th parameter of the n-th time series.

In other words, in the GAS model we suppose that the evolution of the time-varying parameter vector $f_{n,t}$ depends both by a vector $s_{n,t}$, called *scaled score* since it is proportional to the score of the density, and by an autoregressive component.

Clearly, the choice of the underlying probability distribution in (3.23) is very important since it changes the kind of score considered in (3.26) and, therefore, the considered GAS model. The scaled score $s_{n,t}$ is given by:

$$s_{n,j,t} = S_{n,j,t} \cdot \nabla_{n,j,t} \tag{3.27}$$

where $\nabla_{n,j,t}$ is the conditional j-th score at time t for the n-th time series and it computed as:

$$\nabla_{n,j,t} = \frac{\partial \log p(y_{n,t} | f_{n,t}, \mathcal{F}_{n,t}; \theta_n)}{\partial f_{n,t}} \tag{3.28}$$

and $S_{n,j,t}$ is a *scaling matrix* of appropriate dimension that is usually given by the inverse of the Fisher information matrix:

$$S_{n,j,t} = \left(E \left[\nabla_{n,j,t} \nabla'_{n,j,t} \right] \right)^{-1} \tag{3.29}$$

The aforementioned approach is the standard GAS proposed by Creal, Koopman, and Lucas, 2013. However, different GAS model can be considered, for example assuming a different scaling matrix $S_{n,j,t}$. As previously mentioned, the assumption about the predictive density (3.23) is crucial also because it completely changes the kind of score considered in the model. For example, assuming a Gaussian density as in (3.24), the conditional score vector (3.28) is given by:

$$\begin{aligned} \nabla_{n,1,t} &= \frac{(y_t - \mu_t)}{\sigma_t^2} \\ \nabla_{n,2,t} &= \frac{(y_t - \mu_t)^2}{2\sigma_t^4} - \frac{T}{2\sigma_t^2} \end{aligned}$$

with $\nabla_{n,j,t} = (\nabla_{n,1,t}, \nabla_{n,2,t})$ where $\nabla_{n,1,t}$ is the conditional score for the first moment (i.e. the conditional mean) and $\nabla_{n,2,t}$ is the one for the second conditional moment (i.e. the conditional variance). Therefore the model's variables and parameters are given by:

$$f_t = \begin{pmatrix} \mu_t \\ \sigma_t^2 \end{pmatrix}, \quad \omega = \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} b_1 & 0 \\ 0 & b_2 \end{pmatrix}$$

Moreover, in the case of t-student density (3.13) the conditional score vector is given by:

$$\nabla_{n,1,t} = \frac{(v_t+1)(y_t-\mu_t)}{v_t\phi_t + (y_t-\mu_t)^2}$$
$$\nabla_{n,2,t} = \frac{1}{2\phi_t}\left[\frac{(v_t+1)(y_t-\mu_t)^2}{v_t\phi_t + (y_t-\mu_t)^2} - 1\right]$$
$$\nabla_{n,3,t} = \frac{1}{2}\left\{\psi\left(\frac{v_t+1}{2}\right) - \psi\left(\frac{v_t}{2}\right) - \frac{1}{v_t} - \log\left(1+\frac{(y_t-\mu_t)^2}{v_t\phi_t}\right) + \frac{(v_t+1)(y_t-\mu_t)^2}{v_t\left[v_t\phi_t + (y_t-\mu_t)^2\right]}\right\}$$

where $\psi(\cdot)$ is the Digamma function. In this case the conditional score vector $\nabla_t = (\nabla_{n,1,t}, \nabla_{n,2,t}, \nabla_{n,3,t})$ is composed by the conditional location score $\nabla_{n,1,t}$, the conditional scale score $\nabla_{n,2,t}$ and the conditional shape score $\nabla_{n,3,t}$. In compact form we have:

$$f_t = \begin{pmatrix}\mu_t\\ \phi_t\\ v_t\end{pmatrix}, \quad \omega = \begin{pmatrix}\omega_\mu\\ \omega_\phi\\ \omega_v\end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix}a_\mu & 0 & 0\\ 0 & a_\phi & 0\\ 0 & 0 & a_v\end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix}b_\mu & 0 & 0\\ 0 & b_\phi & 0\\ 0 & 0 & b_v\end{pmatrix}$$

This model has also been defined as Beta-t-EGARCH model by Harvey (2013) and Harvey and Sucarrat (2014).

Once the score are computed, the parameters in (3.26) are estimated by maximum likelihood. Then, the conditional moments can be obtained by the in-sample predictions $\hat{f}_{n,j,t}$ as in Cerqueti, Giacalone, and Mattera (2021):

$$\hat{f}_{n,j,t} = \hat{\omega}_{n,j} + \hat{\mathbf{A}}_{n,j,1} s_{n,j,t-1} + \hat{\mathbf{B}}_{n,j,1} f_{n,j,t-1} \tag{3.30}$$

In order to cluster the objects contained into a time data array, we take inspiration from the two-step procedure of D'urso (2004). In the first step, we compute, for each n-th time series, the dissimilarity matrix between each pair of the J conditional moments observed at T times. Thus, we obtain N distance matrices $\mathbf{D}_n$. Then, in the second step, we classify the N time series by considering a diversity measure between each pair of distances $\mathbf{D}_n$.

In more details, given the data time array $\mathbf{F}$, we compute the dissimilarity $\mathbf{D}_n = \left\{d_{n,j,j'} : j, j' = 1, J; j \neq j'\right\}$ for the n-th object between the moments j and j', observed

at T times. Because the dissimilarity matrices are squared and symmetric with a null diagonal, we can vectorize their lower triangular $\mathbf{L}_n$ obtaining vec $(\mathbf{L}_n)$. Then, we define the following pairwise Euclidean dissimilarity between the time series n and $n\prime$:

$$d_{n,n\prime} = \|\text{vec}\,(\mathbf{L}_n) - \text{vec}\,(\mathbf{L}_{n\prime})\| \tag{3.31}$$

For clustering, we adopt the fast k-medoids algorithm of Park and Jun, 2009. By using an object in the sample rather than a fictitious one as a medoid, the k-medoids algorithm leads to more interpretable results. Moreover, it is proved to be more robust to outliers than the k-means, and it is also faster in terms of computational time. The proposed clustering model can be formalized as follows:

$$\min : \sum_{n=1}^{N}\sum_{c=1}^{C} d_{n,c}^2 = \sum_{n=1}^{N}\sum_{c=1}^{C} \|\text{vec}\,(\mathbf{L}_n) - \text{vec}\,(\mathbf{L}_c)\|^2 \tag{3.32}$$

where subscript c represents the centroid time series.

3.3.3 A simulation study

To show the validity of the proposed distribution-based clustering procedure, we provide a simulation study. In particular, we compare the classification accuracy of the proposed clustering approaches, based on both static and time-varying parameters, with other classical alternatives such as the Euclidean distance on raw time series, based on correlations auto-correlations and periodogram ordinates.

Since in this case the ground truth is available, we measure the quality of classification of the different clustering approaches by means of the Rand index Rand (1971). Let Y be a set of N time series, a clustering $\mathbf{K}$ on Y allows to partition the set of time series into non-overlapping grpups $\{K_1, K_2, \ldots, K_C\}$, where $\cup_{i=1}^{C} K_i = Y$ and $K_i \cap K_j = \emptyset$ for $i \neq j$. Let define with N_{11} the number of objects that are in the same cluster in both $\mathbf{K}$ and $\tilde{\mathbf{K}}$, N_{00} those that are in different clusters in both $\mathbf{K}$ and $\tilde{\mathbf{K}}$ N_{01} the onse that are in the same cluster in $\mathbf{K}$ but in different clusters in $\tilde{\mathbf{K}}$, and N_{10} the number of time series that are in different clusters in $\mathbf{K}$ but in the same cluster in $\tilde{\mathbf{K}}$. As noted by Rand (1971), N_{11} and N_{00} can be used as measures about the degree of

agreement in the classification between $\mathbf{K}$ and $\tilde{\mathbf{K}}$ while, conversely, N_{01} and N_{10} can be seen as measures of disagreement. The Rand index is defined as:

$$\text{RI} = \frac{(N_{00} + N_{11})}{\binom{N}{2}} \tag{3.33}$$

However, since the RI often lies within the the range of $[0.5, 1]$, Hubert and Arabie (1985) proposed the following adjustment:

$$\text{ARI} = \frac{2\,(N_{00}N_{11} - N_{01}N_{10})}{(N_{00} + N_{01})\,(N_{01} + N_{11}) + (N_{00} + N_{10})\,(N_{10} + N_{11})} \tag{3.34}$$

that is always between the range $[0, 1]$. It is called the Adjusted Rand Index (ARI). An ARI value close to 0 indicates randomness in the partition, while a value close to 1 indicates a perfect classification. The comparison is made in terms of the average Adjusted Rand Index (ARI) over 300 trials as in Díaz and Vilar (2010).
A first simulation scenario is generated as follows. A first set of $N = 5$ time series is simulated considering a Gaussian distribution with a time-varying mean $\mu_{1,t}$ and variance $\sigma^2_{1,t}$ whose process are given by the Gaussian-GAS with parameters:

$$\omega_1 = (0.0490, 0.0154); \quad \mathbf{A}_1 = \begin{pmatrix} 0.0001 & 0 \\ 0 & 0.0534 \end{pmatrix}; \quad \mathbf{B}_1 = \begin{pmatrix} 0.0485 & 0 \\ 0 & 0.9891 \end{pmatrix}$$

Then, we simulate another set of $N = 5$ Gaussian time series with $\mu_{2,t}$ and $\sigma^2_{2,t}$ generated by a Gaussian-GAS process with the following parameters:

$$\omega_2 = (0.0840, 0.0456); \quad \mathbf{A}_2 = \begin{pmatrix} 0.00001 & 0 \\ 0 & 0.0139 \end{pmatrix}; \quad \mathbf{B}_2 = \begin{pmatrix} 0.0660 & 0 \\ 0 & 0.0968 \end{pmatrix}$$

These parameters are calibrated according to two different, randomly selected, real time series belonging to the Dow Jones 30 financial market index. we consider three different scenarios in terms of time series' length, namely $T = \{250, 1000, 2500, 5000\}$.

In the first scenario, we have the shortest time series, while the last three have an always increasing size. The results are shown in Tab. 3.1.

	$T = 250$	$T = 1000$	$T = 2500$	$T = 5000$
EUCL	0.0020	0.0007	0.0002	0.0010
COR	0.0019	0.0064	0.0195	0.0040
ACF	0.0123	0.0146	0.0017	0.0030
PER	0.0001	0.0002	0.0000	0.0010
DIST	0.0139	0.0408	0.1222	0.2530
DIST_t	0.1581	0.6001	0.9352	0.9879

TABLE 3.1: Results of the simulation study: average adjusted Rand index for $N = 10$ time series

Simulations in Tab. 3.1 clearly show that the distribution-based clustering with time-varying parameters is the best clustering approach among the considered alternatives. Indeed, its ARI is always the highest for all the simulated scenarios. Moreover, the approach based on static distribution parameters dominates with respect to the alternatives but performs worst than the approach with time-varying parameters. This is because the underlying DGP assumes dynamic rather than static distribution parameters. Overall, all the approaches improve their performances with an increasing time series length. Asymptotically, the distribution-based clustering with time-varying parameters approaches to the value of 1, where a perfect classification is achieved. On the other side, Euclidean and frequency domain distances provide very random partition.

Let now consider the case of an increasing number of time series. So, instead of considering $N = 5$ for both groups, we now consider $N = 15$ such that the total number of time series is three times greater. As in the previous experiment, three different scenarios in terms of time series' length have been simulated, namely $T = \{250, 1000, 2500, 5000\}$. The results are shown in Tab. 3.2.

	$T = 250$	$T = 1000$	$T = 2500$	$T = 5000$
EUCL	0.0015	0.0006	0.0004	0.0023
COR	0.0036	0.0027	0.0015	0.0031
ACF	0.0015	0.0037	0.0035	0.0008
PER	0.0003	0.0000	0.0000	0.0000
DIST	0.0223	0.0549	0.1258	0.2389
DIST_t	0.1499	0.6407	0.9352	0.9951

TABLE 3.2: Results of the simulation study: average adjusted Rand index for $N = 30$ time series

In general, the results highlighted by Tab. 3.2 are not very different concerning those of Tab. 3.1, meaning that the proposed clustering approaches based on distribution's parameters perform well regardless of the dimensionality of the dataset. Tab. 3.2 confirm that the distribution-based clustering with time-varying parameters provides the best classification. Moreover, asymptotically we reach an ARI value close to 1. Moreover, the static distribution-based approach remains the second-best clustering model, even if the performances are lower than those of Tab. 3.1. Therefore, the approach with static parameters seems to be more sensitive to the size of the dataset.

3.4 Methodology

In what follows, we study the empirical performances of clustering in portfolio selection, by evaluating if clustered portfolios return higher performances than standard ones. As briefly stated previously, clustered portfolios can be formed by the application of any diversification rule to the $C \geq 2$ subsets of stocks. In other words, in order to build clustered portfolios the first step requires the selection of the clustering approach (i.e. the employed distance and the kind of algorithm used). Once a partition into C groups is found, C clustered portfolios can be obtained on the basis of the stocks belonging to each of the $c \in C$ clusters. In the end, the optimal amount of wealth associated to each of the C funds can be defined according to any kind of optimization strategy. A simple approach consist in forming C roughly diversified portfolios by means of a naive diversification rule and, then, apply optimal

diversification across the C funds. For example, we can use the minimum-variance criterion:

$$\min_{w} w'\Sigma w, \quad \text{s.t.} \quad \sum_{c=1}^{C} w_c = 1 \rightarrow w^* = \frac{\Sigma^{-1}\mathbb{1}_N}{\mathbb{1}'_N\Sigma^{-1}\mathbb{1}_N} \tag{3.35}$$

where Σ^{-1} is the inverse of the C funds covariance matrix and:

$$w = (w_1, \ldots, w_c, \ldots, w_C)$$

is the vector of C weights associated to each c-th clustered portfolio. However, an investor can in principle construct C minimum variance portfolios and, then, uses the same diversification criterion to diversify across the C minimum-variance funds. Another alternative that we consider is to adopt a mean-variance diversification rule across clustered portfolios, i.e. by means of the following optimal weighting scheme:

$$\max_{w} w'\mu - \gamma w'\Sigma w \rightarrow w^* = \frac{1}{\gamma}\Sigma^{-1}\mu \tag{3.36}$$

with γ is the investor's risk aversion coefficient. In what follows, we assume $\gamma = 1$. Clustered portfolios can be formed on the basis of an either large or small asset universe N. In the first case, where $N > T$, we have a *large dimensional setting*, while in the case $N < T$ we have the standard *low dimensional setting*. It is well known that asset allocation in the first case is much more complex than in the second one because when $N > T$ the covariance matrix Σ is ill-conditioned and cannot be inverted. This fact leads to an asset allocation that is is unfeasible. By alleviating the course of dimensionality, clustering can be used as a tool to transform an high-dimensional problem into C low-dimensional ones. Nevertheless, using clustered portfolio can be in principle useful also in a low-dimensional setting.

In order to study the performances of clustered portfolios, we use the following rolling-window strategy (DeMiguel, Garlappi, and Uppal, 2009). Given a matrix $\mathbf{Y}$ with N time series of returns observed for T months, we choose an estimation widow equal to M, to form C clustered according to the different clustering procedures discussed so far in parapraph 3.2 and 3.3 that are used to form C naive clustered portfolios. Then, we estimate the covariance structure across the C funds that

is needed for the implementation of the minimum-variance strategy. To estimate the covariance structure we use the static sample covariance estimator.

This process is recursively repeated by adding the return for the next period in the dataset and dropping the earliest one until the end of the dataset is reached. The result is, therefore, a time series of length $(T - M)$ of returns.

Usually, a common choice is $M = 120$ for monthly data (see DeMiguel, Garlappi, and Uppal (2009)). However, the clustering algorithms perform differently for long and short time series (Díaz and Vilar, 2010). Indeed, it can be the case that clustering approaches that work well with short time series perform instead poorly with very long time series and vice versa. Since M represents the time series length within each iteration, we consider different values, i.e. short, medium and long M.

Given the time series of monthly out-of-sample portfolio returns, we compute the out-of-sample Sharpe ratio of the portfolio obtained using the k-th strategy, SR_k, defined as the sample mean of out-of-sample portfolio returns divided by its standard deviation:

$$\text{SR}_k = \frac{\hat{\mu}_k}{\hat{\sigma}_k} \tag{3.37}$$

where $\hat{\mu}_k$ is the average of the $(T - M)$ out of sample returns for the portfolio using the k-th clustering approach and $\hat{\sigma}_k$ its standard deviation. Another approach that can be used for assessing the out-of-sample performance lies, by assuming a mean-variance investor, on the calculation of the certainty-equivalent (CEQ) return. The CEQ return is defined as the risk-free rate that an investor is willing to accept rather than adopting a particular risky portfolio strategy. The CEQ return can be computed as follows:

$$\text{CEQ}_k = \hat{\mu}_k - \frac{\gamma}{2}\hat{\sigma}_k \tag{3.38}$$

Clearly, the approach with clustering is compared with the standard approaches where clustering is not involved. In the case of high-dimensional experiment, where $N < T$, the minimum-variance with Ledoit and Wolf (2003) covariance estimator is used as benchmark. In the low-dimensional setting standard minimum-variance with sample covariance is instead considered. Table 3.3 summarizes the investment

strategies that are implemented.

Symbol	Description
	Benchmark
w_{SC}	investment strategy with sample covariance plug-in
w_{LW}	investment strategy with Ledoit and Wolf (2003) plug-in
	Clustered portfolios
w_{EUCL}	investment strategy with Euclidean distance
w_{COR}	investment strategy with correlation-based distance
w_{ACF}	investment strategy with ACF-based distance
w_{NPER}	investment strategy with peridiogram-based distance
w_{HURST}	investment strategy with Hurst-based distance
w_{ARMA}	investment strategy with ARMA-based distance
w_{GARCH}	investment strategy with GARCH-based distance
w_{norm}	investment strategy with static Gaussian distance
w_{std}	investment strategy with static t-student distance
w_{norm_t}	investment strategy with time-varying Gaussian distance
w_{std_t}	investment strategy with time-varying t-student distance

TABLE 3.3: Implemented investment strategies

3.5 Experiments with common stocks

Common stocks have higher idiosyncratic risk if compared with already diversified funds. In what follows, we consider common stocks that belong to market indices as the input of asset allocation.

3.5.1 Low-dimension: Dow Jones constitutes

In what follows, we provide an application of clustered portfolios to Dow Jones Index constitutes. More in details, as in the other empirical applications within this thesis, we considered the monthly stock returns in the time period 10/1999-10/2019 and we excluded the stocks showing missing values. Hence, in this setting there are $N = 28$ stocks observed for $T = 240$. The stock returns are showed in Fig. 3.3.

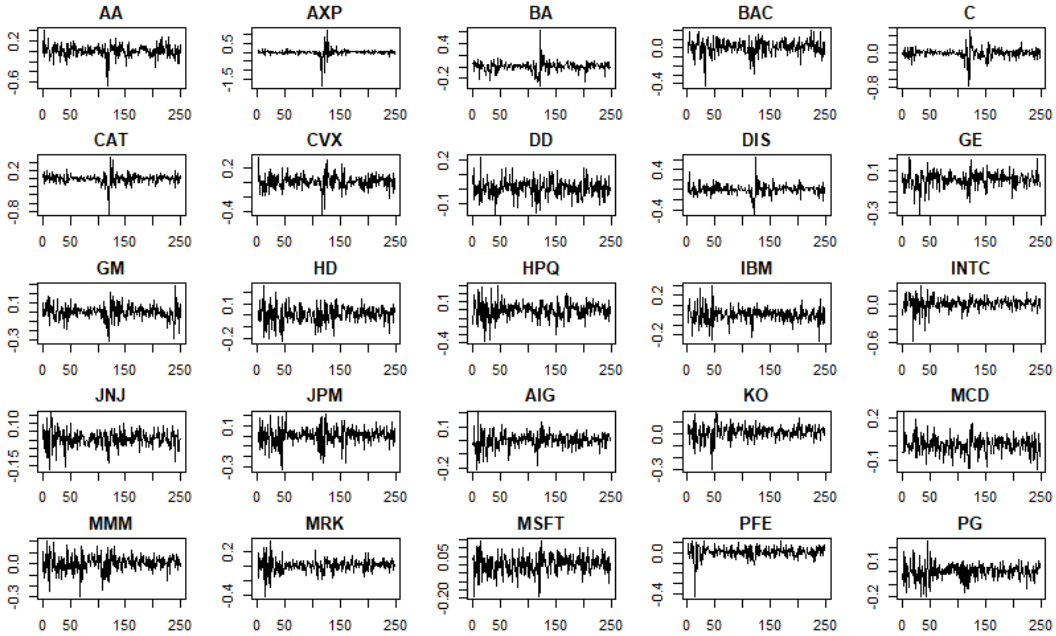


FIGURE 3.3: Dow Jones constitutes: returns

Main descriptive statistics are reported in Tab. 3.4. It is evident that stock returns show different mean and standard deviations. Moreover, most of them are negatively skewed and present excess of kurtosis larger than 0, meaning that most of them do not follow a Gaussian distribution. This evidence, that is in line with stylized facts, is also supported by the Jarque and Bera (1987) statistics that is very large for most of the stocks in the sample.

The descriptive analysis of distributional features is important in motivating the distribution-based clustering discussed so far. In particular, the fat tails justify the adoption of a non-Gaussian distribution for clustering, for example by means of the t-student density.

Stock	Mean	St. Dev.	Skewness	Kurtosis	JB
AA	-0.0024	0.1177	-0.8173	4.6474	239.0056
AIG	-0.0105	0.1975	-2.8406	41.5773	17459.1260
AXP	0.0066	0.0843	0.8403	12.9748	1692.6804
BA	0.0114	0.0826	-0.9460	3.0664	127.9152
BAC	0.0016	0.1177	-1.2942	10.2555	1106.2475
C	-0.0039	0.1258	-1.7257	12.7357	1723.4572
CAT	0.0093	0.0944	-0.4415	2.9940	95.3566
CVX	0.0076	0.0586	-0.0434	0.5928	3.2804
DD	0.0050	0.1016	0.4732	8.5129	724.0188
DIS	0.0066	0.0709	-0.5472	2.2548	61.4368
GE	-0.0029	0.0821	-0.4405	2.5041	68.8359
HD	0.0086	0.0741	-0.3331	0.9234	12.4918
HPQ	0.0027	0.1017	-0.3548	1.4433	25.0519
IBM	0.0034	0.0734	-0.1781	3.0076	89.5785
INTC	0.0033	0.1018	-1.2556	5.4498	355.6687
JNJ	0.0066	0.0490	-0.4096	1.4962	28.2570
JPM	0.0057	0.0890	-0.5694	1.5053	34.8375
KO	0.0042	0.0542	-0.5569	2.3146	64.5525
MCD	0.0088	0.0571	-0.8881	3.4394	147.5299
MMM	0.0079	0.0588	-0.0464	1.1873	13.5030
MRK	0.0036	0.0720	-0.5244	1.8045	42.5302
MSFT	0.0064	0.0857	-0.2992	3.4143	117.6315
PFE	0.0020	0.0580	-0.3336	0.2964	5.2234
PG	0.0062	0.0543	-2.4535	17.4012	3239.0990
T	0.0025	0.0666	-0.1888	1.3945	20.0616
VZ	0.0041	0.0654	0.3658	2.9530	90.4976
WMT	0.0056	0.0584	-0.3555	1.6771	32.1953
XOM	0.0050	0.0509	0.0017	1.1780	13.2006

TABLE 3.4: Descriptive statistics of the Dow Jones constitues. Jarque and Bera (1987) statistics is also reported.

Minimum-variance approach

As stated previously, the estimation window used for the construction of minimum-variance portfolios is initially set to be equal to $M = 120$. Hence, we start with a *medium sized* estimation window. Since $M > N$ we are in a low-dimensional setting. In this scenario the sample covariance can be used for the estimating covariances necessary for the implementation of the minimum-variance strategy. However, in what follows, we also study the possibility of adopting a minimum-variance strategy on the basis of clustered portfolios. In doing so, we recursively perform the cluster analysis in order to build C portfolios that then become the input of the diversification strategy.

As stated previously, in each recursive step we choose the optimal number of clusters according to the Average Silhouette Width (ASW) criterion. The maximum number of clusters is set to be equal to $N/2$, hence in this case we have that $C_{\max} = 14$. In other words, in the worst case we have that $C < M$, hence a low-dimensional setting is ensured.

The comparison in terms of out-of-sample performances is shown in Tab. 3.5.

Strategy	Sharpe ratio	CEQ
w_{SC}	34.33%	1.12%
w_{EUCL}	24.82%	0.83%
w_{COR}	27.93%	0.82%
w_{ACF}	33.77%	1.11%
w_{PER}	22.82%	0.94%
w_{HURST}	**39.21%**	1.23%
w_{ARMA}	36.22%	1.12%
w_{GARCH}	33.99%	1.15%
w_{norm}	32.92%	1.04%
w_{std}	31.28%	1.22%
w_{norm_t}	28.36%	0.98%
w_{std_t}	38.59%	**1.44%**

TABLE 3.5: Clustered portfolios: experiment with Dow Jones constitutes ($M = 120$)- GMV approach

First of all, we note that not all clustering procedures are able to outperform the standard miniumum variance strategy. Therefore, clustering is not always useful. Nevertheless, there are some strategies that perform very well compared to the benchmark. For example, the ACF-based clustered portfolios, as well as the GARCH-based and the static distribution based ones, perform almost the same. Most of them show also higher CEQ returns compared to the benchmark.

Moreover, there are also many strategies based on clustering that outperform the benchmark. This is the case of the Hurst-based clustered portfolios, ARMA-based and the student t-based with time varying parameters. In particular, the investment strategy with Hurst-based clustered portfolio has the highest Sharpe ratio (39.2%), followed by the student-t with time-varying parameters (38.6%). Note that in terms of CEQ return the student-t clustered portfolio with time-varying parameters is the best strategy. This means that the investor can improve the performances of her investment strategy by using clustering.

As stated above, the algorithms have different classification accuracy on the basis of the time series lengths. To exploit this evidence, we repeated the experiment by assuming an estimation window equal to $M = 60$, i.e. where the clustering of shorter time series is involved. The performances of the portfolios constructed in this way are reported in Tab. 3.6.

Strategy	Sharpe ratio	CEQ
w_{SC}	24.87%	0.92%
w_{EUCL}	17.35%	0.65%
w_{COR}	27.76%	0.81%
w_{ACF}	17.56%	0.55%
w_{PER}	12.68%	0.45%
w_{HURST}	17.33%	0.51%
w_{ARMA}	11.52%	0.38%
w_{GARCH}	19.69%	0.63%
w_{norm}	15.83%	0.50%
w_{std}	**39.93%**	**1.49%**
w_{norm_t}	20.88%	0.68%
w_{std_t}	12.52%	0.42%

TABLE 3.6: Clustered portfolios: experiment with Dow Jones constitutes ($M = 60$) - GMV approach

Tab. 3.6 highlights that the strategy based on the Hurst-based clustering portfolios is not anymore the best one. Moreover, only fewer clustering techniques perform better than the benchmark with respect to the results of Tab. 3.5 (with $M = 120$). Among the alternatives, however, the distribution-based clustering approach is confirmed to be one of best from the point of view of portfolio construction. Indeed, in the case of Tab. 3.6 this strategy leads to the highest performances in terms of both out-of-sample Sharpe ratio and CEQ return. More precisely, the t-student based clustered portfolios guarantees a Sharpe ratio close to 40%, versus the 25% of the benchmark. Therefore, we get again evidence in favor of clustering in portfolio selection.

Then, we also suppose a larger estimation window with $M = 180$, that is the case with the longest time series are considered for clustering. The results are showed in Tab. 3.7.

Strategy	Sharpe ratio	CEQ
w_{SC}	33.16%	1.06%
w_{EUCL}	22.50%	0.73%
w_{COR}	26.20%	0.74%
w_{ACF}	**39.10%**	**1.13%**
w_{PER}	22.50%	0.73%
w_{HURST}	30.61%	0.92%
w_{ARMA}	31.57%	0.86%
w_{GARCH}	27.23%	0.75%
w_{norm}	31.17%	0.85%
w_{std}	25.28%	0.74%
w_{norm_t}	16.36%	0.50%
w_{std_t}	16.34%	0.45%

TABLE 3.7: Clustered portfolios: experiment with Dow Jones constitutes ($M = 180$) - GMV approach

In this case the ACF-based clustering provides the highest performances in terms of out-of-sample Sharpe ratio and CEQ return. More in detail, the ACF-based clustered portfolios return a Sharpe ratio of 39% while the benchmark 33%. In this framework the distribution-based clustering approaches show enough robust performances, even if other approaches seem to work better with longer time series.
In conclusion, this first experiment within a low-dimensional setting with minimum-variance diversification across the C clustered funds reveal the usefulness of clustering in portfolio selection. The distribution-based approaches guarantee the most robust performances among the different scenarios, while the highest outcome in the case with short time series.

Mean-variance approach

Following the procedure explained in the methodology paragraph, we now consider the case of mean-variance diversification rule across clustered portfolios. Also in this case we consider a maximum number of clusters within each iteration of $C_{max} = 14$. The comparison in terms of out-of-sample performances is shown in Tab. 3.8.

Strategy	Sharpe ratio	CEQ
w_{SC}	19.77%	1.15%
w_{EUCL}	6.68%	0.15%
w_{COR}	7.05%	0.25%
w_{ACF}	14.30%	0.92%
w_{PER}	10.31%	-20.22%
w_{HURST}	9.40%	-1.03%
w_{ARMA}	18.93%	1.21%
w_{GARCH}	-6.68%	-10.56%
w_{norm}	10.22%	0.49%
w_{std}	**29.78%**	1.74%
w_{norm_t}	12.97%	-1.06%
w_{std_t}	28.85%	**2.52%**

TABLE 3.8: Clustered portfolios: experiment with Dow Jones constitutes ($M = 120$) - Mean-Variance approach

Tab. 3.8 highlights that the best investment strategy is once again one based on clustered funds. More in detail, the distribution-based clustering still returns the highest out-of-sample performances when the student-t distribution is considered. In particular, both the static and dynamic approaches provide almost the same performances in terms of Sharpe ratio (29.7% versus 28.8%) but there is a bigger difference in terms of CEQ return, where the dynamic approach outperforms the static one.

The results obtained by considering the experiment with $M = 60$ (short time series) are showed in Tab. 3.9.

Strategy	Sharpe ratio	CEQ
w_{SC}	22.33%	0.83%
w_{EUCL}	-12.39%	-3.88%
w_{COR}	11.10%	-3290%
w_{ACF}	40.75%	1.31%
w_{PER}	38.96%	1.28%
w_{HURST}	1.18%	-0.52%
w_{ARMA}	10.66%	0.48%
w_{GARCH}	23.87%	1.07%
w_{norm}	40.97%	1.23%
w_{std}	**46.52%**	**1.59%**
w_{norm_t}	16.40%	1.11%
w_{std_t}	8.21%	0.33%

TABLE 3.9: Clustered portfolios: experiment with Dow Jones constitutes ($M = 60$) - Mean-Variance approach

In this case many clustering-based investment strategies perform better than the benchmark (with a Sharpe ratio of 22.3%). More in detail, the most competitive approach is given again by a distribution-based clustering with student-t distribution (Sharpe ratio 46.5%). This result is confirmed also by looking at the CEQ return. Then, the second best strategy is represented by the Gaussian distribution-based clustering and, then, the ACF-based one. Note that in this case the approach with time-varying parameters perform very poor with respect to the static approaches. Therefore, while it can be an useful tool for clustering similar time series, it does not seem to help much with respect the static approach from the portfolio selection task. Then, we consider the case with $M = 180$, where longer time series are available for recursive clustering procedure. The results in terms of financial performance are reported in Tab. 3.10.

Strategy	Sharpe ratio	CEQ
w_{SC}	36.49%	**1.96%**
w_{EUCL}	22.89%	0.76%
w_{COR}	23.86%	0.75%
w_{ACF}	**41.27%**	1.37%
w_{PER}	22.89%	0.76%
w_{HURST}	32.47%	1.16%
w_{ARMA}	35.88%	1.00%
w_{GARCH}	21.89%	0.72%
w_{norm}	28.14%	0.83%
w_{std}	25.60%	0.83%
w_{norm_t}	8.67%	0.32%
w_{std_t}	17.95%	0.77%

TABLE 3.10: Clustered portfolios: experiment with Dow Jones constitutes ($M = 180$) - Mean-Variance approach

The results from Tab. 3.10 somehow confirm those of Tab. 3.7 because with longer time series the ACF-based clustering approach provides the highest out-of-sample performances. The Sharpe ratio associated to the strategy is equal to 41.2% versus the 36.4% of the benchmark. Therefore, we get again the same evidence: not all clustering algorithms are always useful but the ACF-based approach seems to lead to higher performances with long time series, while the distribution-based approaches are better suited if short time series (long-run portfolio analysis) are involved.

3.5.2 Large-dimension: S&P500 constitutes

In what follows we provide an experiment in a large-dimensional (or high-dimensional) setting, where the estimation window M is lower than the number of assets N. In this case, the covariance matrix is ill-conditioned and not invertible, so the asset allocation becomes unfeasible. To fix this issue, Ledoit and Wolf (2003) and Ledoit and Wolf (2004a) developed a covariance estimator with the shrinkage technique that is invertible within this setting. Hence, in such a way the standard minimum-variance or mean-variance allocation becomes feasible.

To study the usefulness of clustering in large dimensional setting we consider the S&P500 constitutes from the 10/1999 to 10/2019. From all the stocks, we exclude

those with missing values. Hence, we get a total asset universe of $N = 286$. Since $M = 60, 120, 180$, we always have that $M < N$.

Clustering is standard technique used to reduce the course of dimensionality. In this case, it is used to form $C < M < N$ clustered funds that are used as the input of a standard low-dimensional asset allocation. The diversification under Ledoit and Wolf (2003) shrinkage estimator is assumed as benchmark.

Providing descriptive statistics and plots of the time series is prohibitive in this setting and for this reason are skipped. Nevertheless, it is evident that there is a very high degree of heterogeneity within the $N = 286$ stocks included in the sample. According to unreported Jarque and Bera (1987) tests, most of the stock returns are non-Gaussian and with heavy tailed distribution. This evidence justifies the use of the student-t distribution for the implementation of the distribution-based clustering approach.

Minimum-variance approach

As usual, we first analyse the case with $M = 120$. The comparison of the different approaches in terms of out-of-sample performances is shown in Tab. 3.11.

Strategy	Sharpe ratio	CEQ
w_{LW}	41.36%	1.07%
w_{EUCL}	22.42%	1.01%
w_{COR}	-4.99%	-0.49%
w_{ACF}	39.05%	1.38%
w_{PER}	40.44%	1.32%
w_{HURST}	30.50%	1.21%
w_{ARMA}	24.63%	1.10%
w_{GARCH}	40.67%	1.43%
w_{norm}	42.26%	1.30%
w_{std}	**43.64%**	**1.54%**
w_{norm_t}	33.32%	1.35%
w_{std_t}	15.47%	0.58%

TABLE 3.11: Clustered portfolios: experiment with S&P500 constitutes ($M = 120$) - GMV approach

In this setting, the Ledoit and Wolf (2003) is very competitive with a 41.3% of Sharpe ratio but it does not provide the highest performances. Indeed, the distribution-based clustering approaches, both with Gaussian and t-student distributions, provide superior performances. For example, in terms of Shapre ratio the Gaussian-based clustering provides a value of 42.2% and the t-student one the 43.6% (it is the highest value). In terms of CEQ return, similarly, we have that the t-student based approach has a value of 1.54% that is again the highest return among all the alternatives. In other words, also in this setting clustering is useful. Despite the good performances of the benchmark, the distribution-based clustering is the most effective to build portfolios. However, also other clustering-based strategies are as good as the benchmark. With this respect, the GARCH clustering, the ACF-based and the peridiogram-based clustering approaches return Sharpe ratios close or equal to 40%. In terms of CEQ return, instead, they show superior performances respect the benchmark. Indeed, considering the CEQ return of 1.07% of the benchmark, the ACF-based clustering has a CEQ return of 1.38%, the peridiogram-based a CEQ of 1.32% and the GARCH-based approach a CEQ of 1.43%. Moreover, in terms of CEQ return the benchmark performs even poorer than the Hurst-based clustered portfolios (1.21%) and the ARMA (1.1%).

Let now consider the experiment with $M = 60$, such that the concentration ratio N/T is bigger than the previous case. Note that bigger is concentration ratio, higher is the estimation error. This is true for any kind of asset allocation problem. The out-of-sample comparisons are showed in Tab. 3.12.

Strategy	Sharpe ratio	CEQ
w_{LW}	32.65%	0.82%
w_{EUCL}	25.71%	0.8%7
w_{COR}	29.80%	0.85%
w_{ACF}	30.48%	1.32%
w_{PER}	20.60%	0.66%
w_{HURST}	8.89%	0.39%
w_{ARMA}	5.41%	0.13%
w_{GARCH}	26.28%	0.97%
w_{norm}	28.46%	0.83%
w_{std}	24.97%	0.88%
w_{norm_t}	13.10%	0.73%
w_{std_t}	**37.93%**	**2.41%**

TABLE 3.12: Clustered portfolios: experiment with S&P500 constitutes ($M = 60$) - GMV approach

In this case the considered algorithms take into account short time series for the definition of the clusters. The Ledoit & Wolf plug-in approach remains very competitive but, still we are able to find a clustered portfolios with higher performance. In particular, the distribution-based clustering provide again the highest financial performances. In particular, the clustering approach with time-varying parameters outperform all the alternatives in terms of both Sharpe ratio and CEQ return. The Sharpe ratio associated to this strategy is more than 5% higher respect to the benchmark, while the CEQ return is 1.5% higher as well. Moreover, we have to note that many clustering algorithms overpeform the benchmark in terms of CEQ return, thus they are able to generate greater economic benefits to the investor with mean-variance utility.

Let now consider the case with an expanded time dimension, i.e. $M = 180$. The results are reported in Tab. 3.13

Strategy	Sharpe ratio	CEQ
w_{LW}	41.05%	1.04%
w_{EUCL}	23.35%	0.73%
w_{COR}	24.97%	0.82%
w_{ACF}	24.35%	0.76%
w_{PER}	42.89%	1.11%
w_{HURST}	24.65%	0.79%
w_{ARMA}	38.84%	1.18%
w_{GARCH}	24.56%	0.73%
w_{norm}	41.62%	1.09%
w_{std}	37.02%	1.20%
w_{norm_t}	24.01%	0.78%
w_{std_t}	**45.53%**	**1.27%**

TABLE 3.13: Clustered portfolios: experiment with S&P500 constitutes ($M = 180$) - GMV approach

Also in this setting with relatively larger time series the distribution-based clustering approach with time-varying parameters provides the highest out-of-sample performances. In details, it guarantees a Sharpe ratio equal to 45.5%, versus the 41% of the benchmark. Moreover, also the CEQ return is higher for the clustered portfolio. However, also the approach with static parameters ensures relatively high performances if compared with the alternatives.
Overall, from this experiment is evident that the distribution-based clustering algorithms perform much better than the considered alternatives in out-of-sample, also within a large dimensional setting.

Mean-variance approach

Let analyse the case of mean-variance diversification within the large dimensional setting with common stocks. The out-of-sample comparisons for the experiment with estimation window $M = 120$ are reported in Tab. 3.14.

Strategy	Sharpe ratio	CEQ
w_{LW}	17.49%	**1.52%**
w_{EUCL}	11.13%	0.55%
w_{COR}	11.18%	0.60%
w_{ACF}	7.15%	-5115%
w_{PER}	8.13%	0.19%
w_{HURST}	1.95%	-0.52%
w_{ARMA}	-2.29%	-12%
w_{GARCH}	13.57%	0.80%
w_{norm}	19.34%	0.93%
w_{std}	**22.69%**	-13%
w_{norm_t}	7.97%	-332%
w_{std_t}	12.97%	0.38%

TABLE 3.14: Clustered portfolios: experiment with S&P500 constitutes ($M = 120$) - Mean-Variance approach

In this case most of the clustering-based strategies perform poorer than the benchmark. Nevertheless, the distribution-based clustering approaches with static parameters ensure higher out-of-sample return/risk trade-off. Indeed, the Gaussian-based approach has a Sharpe ratio equal to 19.3% versus the 17.4% of the benchmark, while the t-sudent approach has a Sharpe ratio equal to 22.6%. However, in terms of CEQ return the benchmark provides higher performances and the Gaussian-based clustered portfolios approach is the second best. Overall, the Tab. 3.14 again confirms the superiority of the distribution-based clustering in forming high performance portfolios.

The results associated to the experiment with $M = 60$ are showed in Tab. 3.15.

Strategy	Sharpe ratio	CEQ
w_{LW}	22.24%	**0.90 %**
w_{EUCL}	13.38%	0.48%
w_{COR}	17.12%	0.62 %
w_{ACF}	-6.55%	-7.37%
w_{PER}	15.48%	0.51%
w_{HURST}	-11.81%	-6.61%
w_{ARMA}	12.22%	-7.19%
w_{GARCH}	16.86%	0.71%
w_{norm}	**28.70%**	0.83%
w_{std}	18.17%	0.81%
w_{norm_t}	12.85%	-6.37%
w_{std_t}	7.17%	-1.49%

TABLE 3.15: Clustered portfolios: experiment with S&P500 constitutes ($M = 60$) - Mean-variance approach

The superiority of the distribution-based clustering is confirmed also for this scenario. Indeed, the Gaussian distribution-based approach ensures a 6.5% additional overperformance with respect the benchmark in terms of Sharpe ratio. According to the CEQ returns, instead, the two strategies share almost the same performance (0.07% of difference). Moreover, the distribution-based approach with t-student density is the second best strategy.

Let us consider the last case where $M = 180$. The performances are showed in Tab. 3.16.

Strategy	Sharpe ratio	CEQ
w_{LW}	26.01%	0.82%
w_{EUCL}	33.32%	1.39%
w_{COR}	22.47%	1.10%
w_{ACF}	22.85%	0.71%
w_{PER}	**43.18%**	**1.13 %**
w_{HURST}	9.48%	0.33%
w_{ARMA}	28.66%	1.03%
w_{GARCH}	28.26%	0.86%
w_{norm}	41.95 %	1.12%
w_{std}	32.56%	1.10%
w_{norm_t}	14.51%	0.60%
w_{std_t}	18.40%	0.81%

TABLE 3.16: Clustered portfolios: experiment with S&P500 constitutes ($M = 180$) - Mean-variance approach

In this last experiment, the frequency domain distance surprisingly performs very well. Indeed, it is the strategy with the highest performances, with a Sharpe ratio equal to 43.2% versus the 26% of the benchmark. Moreover, the distribution-based clustering represents the second best alternative with a Sharpe ratio much greater than the benchmark (it is equal to 42%). Moreover, it is interesting to note as most of the clustering-based investment strategies perform better than the benchmark.
From these results we can surely conclude that clustering is potentially useful and can improve the performances of an asset allocation strategy. In general, we do not find that the same clustering procedure ensures the highest performances, and the results changes on the basis of the considered sample. Nevertheless, it appear clearly that the clustering approaches based on distribution parameters are those with the most robust results, since they appear constantly to be the best approaches from asset allocation perspective.

3.6 Experiments with already diversified funds

While in the previous experiments we considered as starting points the stocks that virtually contain high idiosyncratic risk, in what follows we work with already diversified funds. In the previous experiments we showed the usefulness of clustering common stocks. In the following we study if those results hold also in this other setting.

3.6.1 Low-dimension: 49 Industry Portfolios

First of all, we consider a low-dimensional setting where the number of assets is greater than the estimation window $N > M$. To this aim, we take the 49 Industry Portfolios from the Kenneth French website, from the time period between 10/1999-10/2019. Therefore we have that $N = 49$ and $T = 240$. Fig. 3.4

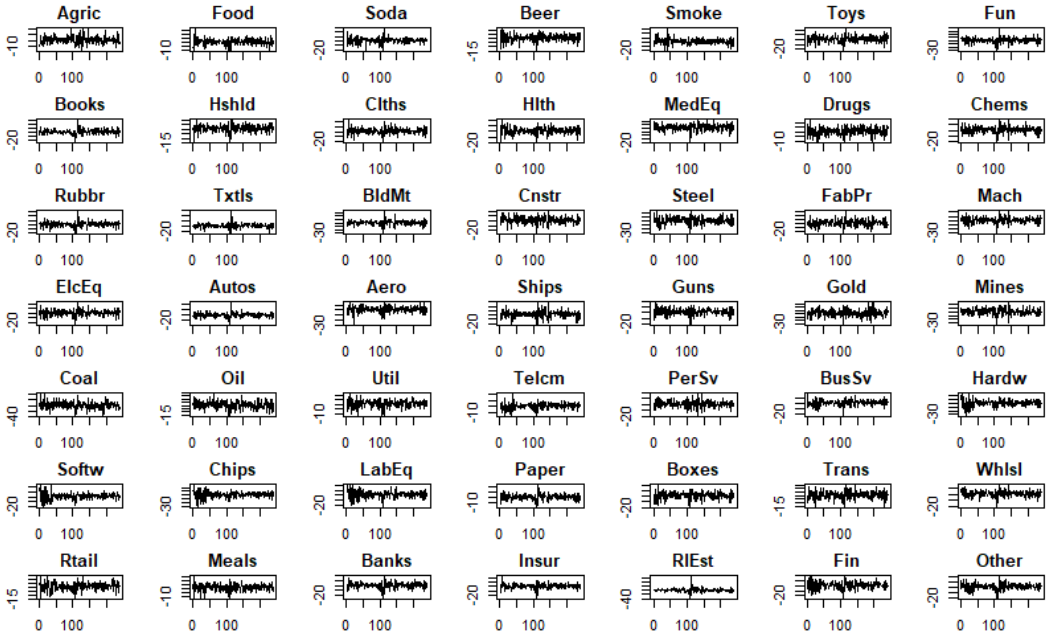


FIGURE 3.4: 49 Industry Portfolio: returns

As in the other experiments, we study the out-of-sample performances of the investment strategies based on both minimum-variance and mean-variance approaches. we start by presenting with the minimum-variance diversification rule.

Minimum-variance approach

Let us start with the analysis of this low dimensional setting where an investor build portfolios with a GMV diversification in the case of medium sized estimation window $M = 120$. Also in this case, in each recursive step we choose the optimal number of clusters according to the Average Silhouette Width (ASW) criterion. The maximum number of clusters is set to be equal to $N/2$, hence in this case we have that $C_{max} = 24$.

The comparison of the out-of-sample performances is reported in Tab. 3.17.

Strategy	Sharpe ratio	CEQ
w_{SC}	24.56%	**-531.45%**
w_{EUCL}	21.31%	-926.80%
w_{COR}	20.46%	-879.16%
w_{ACF}	**25.83%**	-813.52%
w_{PER}	-43.60%	-23284.84%
w_{HURST}	21.83%	-760.57%
w_{ARMA}	22.04%	-748.75%
w_{GARCH}	19.18%	-956.65%
w_{norm}	22.03%	-935.02%
w_{std}	22.27%	-941.25%
w_{norm_t}	21.04%	-882.23%
w_{std_t}	22.85%	-906.41%

TABLE 3.17: Clustered portfolios: experiment with 49 Industry Portfolio constitutes (M=120) - minimum variance approach

In this first case, most of the clustering-based portfolios show lower (but very similar) performances than the benchmark. The only strategy that allows a superior out-of-sample Sharpe ratio is represented by the clustering with ACF-based distance. In terms of CEQ returns, instead, all the strategies show negative values.

Tab. 3.24 shows the results in the case of a lower sample size for clustering, since $M = 60$.

Strategy	Sharpe ratio	CEQ
w_{SC}	**26.04%**	**-497.29%**
w_{EULC}	17.19%	-1057.17%
w_{COR}	24.84%	-871.54%
w_{ACF}	22.53%	-815.66%
w_{PER}	-51.26%	-30867.76%
w_{HURST}	23.32%	-768.09%
w_{ARMA}	21.44%	-835.57%
w_{GARCH}	21.19%	-1004.03%
w_{norm}	21.30%	-881.13%
w_{std}	16.53%	-1379.03%
w_{norm_t}	25.01%	-890.48%
w_{std_t}	18.79%	-1173.34%

TABLE 3.18: Clustered portfolios: experiment with 49 Industry Portfolio constitutes (M=60) - minimum variance approach

The results are in line with those of Tab. 3.17. The benchmark guarantees the highest performances in this case but many clustering-based portfolios show very close performances. In particular, the correlation-based and the distribution-based distances are those with the highest out-of-sample Sharpe ratios. Also in this case, all the investment strategies provide negative CEQ returns.

Let now consider the last experiment with an increasing estimation window $M = 180$. The results are showed in Tab. 3.25.

Strategy	Sharpe ratio	CEQ
w_{SC}	**26.71%**	**-492.49%**
w_{EUCL}	20.23%	-972.50%
w_{COR}	20.74%	-871.35%
w_{ACF}	21.42%	-898.18%
w_{PER}	-35.44%	-13545.61%
w_{HURST}	23.06%	-879.98%
w_{ARMA}	22.25%	-756.05%
w_{GARCH}	20.12%	-1006.42%
w_{norm}	21.94%	-947.54%
w_{std}	18.52%	-961.99%
w_{norm_t}	20.76%	-955.42%
w_{std_t}	24.03%	-822.61%

TABLE 3.19: Clustered portfolios: experiment with 49 Industry Portfolio constitutes (M=180) - minimum variance approach

As in the Tab. 3.18, the Tab. 3.25 shows that the benchmark outperform in out-of-sample the clustering-based alternatives. This happen, perhaps, because the stocks are already diversified funds.

Therefore, for the GMV setting, we observe that the clustering-based strategies do not improve the benchmark, when we consider already diversified portfolios. These results suggest that clustering is useful in presence of high idiosyncratic risk, as happen with common stock returns. Indeed, clustering-based portfolios are useful in the extent to which roughly diversified funds are constructed in the first step.

Mean-variance approach

Within GMV setting we get evidence against the usefulness of clustering-based portfolios for already diversified funds. In what follows, we repeat the same experiments by considering a mean-variance diversification rule. Let us start with the case of medium sized estimation window $M = 120$. The results in terms of out-of-sample Sharpe ration and CEQ returns are showed in Tab. 3.14.

Strategy	Sharpe ratio	CEQ
w_{SC}	-3.26%	-159129.03%
w_{EUCL}	1.68%	-447897.27%
w_{COR}	13.11%	**-1690.40%**
w_{ACF}	6.79%	-389463.72%
w_{PER}	-11.62%	-217010.49%
w_{HURST}	2.07%	-14611.38%
w_{ARMA}	4.55%	-4728.59%
w_{GARCH}	-0.80%	-20000.32%
w_{norm}	6.48%	-1862019.74%
w_{std}	-1.60%	-86709.07%
w_{norm_t}	-3.39%	-8392544.39%
w_{std_t}	**20.18%**	-3551.03%

TABLE 3.20: Clustered portfolios: experiment with 49 Industry Portfolio constitutes (M=120) - Mean-variance approach

In this case the benchmark has a negative Sharpe ratio and it is one of the worst strategies in terms of out-of-sample performances. Indeed, almost all clustering-based portfolios outperform the mean-variance allocation with sample covariance plug-in on the whole sample. The investment strategy with the highest performance is represented by the distribution-based clustering with time-varying parameters, with the t-student as the underlying probability distribution (Sharpe ratio of 20.1% versus the -3.2% of the benchmark).

The experiment with a shorter estimation window (M=60) is shown in Tab. 3.21.

Strategy	Sharpe ratio	CEQ
w_{SC}	4.43%	-125484.39%
w_{EUCL}	1.57%	-1980045.08%
w_{COR}	**12.02%**	**-14048.78%**
w_{ACF}	-0.74%	-59569.32%
w_{PER}	-16.79%	-207499.87%
w_{HURST}	3.52%	-901507.27%
w_{ARMA}	-3.84%	-110652188.30%
w_{GARCH}	3.90%	-605960.59%
w_{norm}	6.93%	-490497.18%
w_{std}	4.44%	-59738.79%
w_{norm_t}	8.49%	-22766.73%
w_{std_t}	-2.46%	-109084.91%

TABLE 3.21: Clustered portfolios: experiment with 49 Industry Portfolio constitutes (M=60) - Mean-variance approach

In this case the benchmark has a positive Sharpe ratio equal to 4.4% but still many clustering-based portfolios outperform it. In particular, the strategy constructed with correlation-based clustering ensures a Sharpe ratio equal to 12%. Moreover, we have that the distribution-based clustering is associated to the second best Sharpe ratio. Indeed, the approach with time varying parameters under Gaussian density ensures a Sharpe ratio of 8.5%.

In the end, we analyze the case with $M = 180$, hence with an extended estimation window (see Tab. 3.22).

Strategy	Sharpe ratio	CEQ
w_{SC}	2.91%	-40929.36%
w_{EUCL}	5.32%	-314768.48%
w_{COR}	**13.23%**	**-1225.61%**
w_{ACF}	7.52%	-48985.90%
w_{PER}	-8.06%	-558912.65%
w_{HURST}	13.12%	-1803.10%
w_{ARMA}	8.83%	-1552.75%
w_{GARCH}	0.77%	-27783.10%
w_{norm}	5.23%	-300801.94%
w_{std}	4.20%	-44890.47%
w_{norm_t}	2.43%	-42570.01%
$w_{\text{std}\,t}$	5.45%	-95376.47%

TABLE 3.22: Clustered portfolios: experiment with 49 Industry Portfolio constitutes (M=180) - Mean-variance approach

Also in this last case most of the clustering-based approaches outperform the benchmark. Hence, clustering is surely useful tool for asset allocation. In details, the approach based on the correlation provides the highest out-of-sample Sharpe ratio 13.2% versus the 2.9% of the benchmark.

Overall, in the case of already diversified funds, the mean-variance diversification rule the clustering-based approach generate important improvements in the out-of-sample financial performances.

3.6.2 Large-dimension: 100 Industry Portfolios

In the previous sub-section we have studied the usefulness of clustering in asset allocation involving already diversified funds within a low-dimensional setting. Overall, we find that clustering is useful only within a mean-variance diversification. This fact can be explained by the ability of clustering in performing better in situations where estimation error is higher. Therefore, following this intuition, in what follows we analyze the case of large-dimensional setting where $M < N$ and estimation error further increases.

To this aim, we consider the 100 industry portfolios, from the Kenneth French website, from the time period 10/1999-10/2019. Hence, in this case we have that $N =$

100 and $T = 240$. In order to force the analysis to be in a large dimensional setting, we consider two estimation windows $M = 50$ and $M = 100$. In both cases we have that $N \geq M$.

In particular, the case $M = 50$ represents a standard high dimensional setting, with a concentration ratio equal to $N/M = 2$. The second case with $M = 100$ is a limit one, where the concentration ratio takes value of 1. Within this framework, we study the benefit of implementing clustering-based asset allocation strategies by following both a Global Minimum Variance and a Mean-variance diversifications.

Minimum-variance approach

Let consider the GMV diversification strategy. The limit case with $M = 100$ is showed in Tab. 3.23.

Strategy	Sharpe ratio	CEQ
w_{LW}	**20.97%**	**-485.75%**
w_{EUCL}	18.40%	-981.78%
w_{COR}	17.82%	-962.08%
w_{ACF}	17.44%	-1018.16%
w_{PER}	20.24%	-1000.73%
w_{HURST}	17.32%	-940.86%
w_{ARMA}	17.69%	-900.97%
w_{GARCH}	17.71%	-1040.10%
w_{norm}	19.91%	-826.14%
w_{std}	15.72%	-1154.53%
w_{norm_t}	16.72%	-985.56%
w_{std_t}	16.79%	-1245.54%

TABLE 3.23: Clustered portfolios: experiment with 100 Industry Portfolio constitutes (M=100) - minimum variance approach

As we have seen for the low-dimensional setting, the benchmark with the Ledoit & Wolf estimator outperforms the alternatives with a Sharpe ratio equal to 20.9%. Nevertheless, the overperformance is not very high: both the peridiogram-based clustering portfolios and those based on distribution parameters (under Gaussian density) have Sharpe ratio equal to 20.2% and 19.9% respectively.

Let now consider the more standard case with $M = 50 < N$. The results are showed in Tab. 3.24.

Strategy	Sharpe ratio	CEQ
w_{LW}	14.21%	**-696.72%**
w_{EUCL}	17.27%	-947.70%
w_{COR}	15.24%	-999.64%
w_{ACF}	13.16%	-1171.05%
w_{PER}	14.76%	-910.51%
w_{HURST}	16.80%	-1216.09%
w_{ARMA}	12.03%	-1277.90%
w_{GARCH}	16.09%	-1159.04%
w_{norm}	**17.58%**	-858.81%
w_{std}	11.56%	-1325.09%
w_{norm_t}	11.51%	-1213.19%
w_{std_t}	15.52%	-1499.82%

TABLE 3.24: Clustered portfolios: experiment with 100 Industry Portfolio constitutes (M=50) - minimum variance approach

In this experiment we observe that many clustering-based investment strategies perform better than the benchmark. Among them, the one with the highest performance is the Gaussian distribution-based approach. Tab. 3.24 is the only exception where, under GMV diversification with already diversified funds, the clustering-based approaches guarantee higher performances than the benchmark.

Mean-variance approach

If clustering seems to be less usefull for allocating wealth across already diversified funds under GMV diversification, in this last section we study what happen if the ideal investor uses a mean-variance rule.
Tab. 3.25 shows the results in the limit case with $M = 100$.

Strategy	Sharpe ratio	CEQ
w_{LW}	7.63%	-68720.64%
w_{EUCL}	-3.83%	-3026235.27%
w_{COR}	-4.83%	-2091844.24%
w_{ACF}	0.96%	-1397942.58%
w_{PER}	**17.00%**	**-5368.76%**
w_{HURST}	2.66%	-671136.81%
w_{ARMA}	-5.01%	-2449516.68%
w_{GARCH}	-1.03%	-2548425.76%
w_{norm}	-10.14%	-30613.70%
w_{std}	4.81%	-522367.62%
w_{norm_t}	7.48%	-1041184.55%
w_{std_t}	0.74%	-7142.56%

TABLE 3.25: Clustered portfolios: experiment with 100 Industry Portfolio constitutes (M=100) - Mean-variance approach

These results confirm those of previous tables. Despite the benchmark performs quite well (Sharpe ratio 7.6%), we am still able to find at least a clustering-based allocation whose performances are higher. In this case, the peridiogram-based distance ensures a Sharpe ratio of 17% and also the CEQ return, even if negative, is the highest among the alternatives.

The case with shorter estimation window is reported in Tab. 3.26 below.

Strategy	Sharpe ratio	CEQ
w_{LW}	-4.18%	-415428.85%
w_{EUCL}	1.48%	-545694.03%
w_{COR}	3.04%	-23614126.54%
w_{ACF}	1.00%	-23773485.37%
w_{PER}	1.31%	-2074317.43%
w_{HURST}	0.09%	-1876150.70%
w_{ARMA}	**6.00%**	-3992195.84%
w_{GARCH}	3.08%	-1283954.12%
w_{norm}	2.09%	-545669.80%
w_{std}	1.33%	-539718.66%
w_{norm_t}	5.40%	-1112853.14%
w_{std_t}	-1.78%	**-79671.87%**

TABLE 3.26: Clustered portfolios: experiment with 100 Industry Portfolio constitutes (M=50) - Mean-variance approach

In this scenario all the clustering approaches are better than the benchmark, despite the fact that the algorithms deal with relatively short time series. The benchmark Sharpe ratio is negative and equal to -4.2%. Among the considered alternatives, the ARMA-based approach guarantees an out-of-sample Sharpe ratio equal to 6%. The second best approach is represented by the Gaussian distribution-based clustering with time-varying parameters.

3.7 Conclusions

Overall, the empirical findings suggest that in the case of common stocks (that contain high idiosyncratic risk), clustering-based asset allocation is useful to the extent to which, in the first step, they construct roughly diversified funds. With this respect, it is important to highlight that clustering allows to build already diversified funds on the basis of a better suited definition of similarity across time series.
In general, for common stocks we can always find clustering procedures that overperform the benchmark, regardless of the employed diversification rule (i.e. minimum-variance or mean-variance). In the case of already diversified funds (proxied by the

industry portfolios), with a minimum-variance diversification rule, the clustering-based approach seems not to generate a substantial improvement in the out-of-sample financial performances. In the case of mean-variance diversification, however, the opposite result hold. This evidence holds if both large and low dimensional settings are considered. The fact that clustering-based portfolios perform better with a mean-variance diversification suggests that clustering is useful to alleviate the estimation error problem that, as showed by Kourtis, Dotsis, and Markellos (2012), contains *ceteris paribus* more estimation error than minimum-variance.

Nevertheless, even if this evidence seems promising, we do not find a single clustering algorithm that performs better for all the datasets. Hence, the conclusions highlighted above cannot be considered general. Indeed, the findings of this study show that clustering is undoubtedly helpful if we deal with common stocks, but the best algorithm should be chosen case-by-case based on backtesting activities.

Among the considered alternatives, the distribution-based approaches show the most robust financial performances in out-of-sample. However, in this case, we cannot find a specific underlying distributional assumption that consistently outperforms the others. This may be due to the best fit obtained in each recursion needed to implement the asset allocation under the distribution-based clustering. Moreover, it is also clear that the hypothesis of time-varying parameters is beneficial and leads to better results, but not for all the considered experiments. This aspect deserves a deeper investigation.

The future development of the proposed work can be related to comparing different clustering algorithms rather than only dissimilarity measures. Indeed, we consider only PAM clustering because of its higher computational speed, interpretability and robustness to outliers than the k-means and the hierarchical clustering models. However, it would be interesting to evaluate the different algorithms' ability to generate high-performance portfolios.

Appendix A

Appendix for the Chapter 1

A.1 Proofs of Chapter 1

A.1.1 Expectation-trace relationship

Given $X \sim \mathcal{N}(\mu, \Sigma)$ the quadratic form:

$$X'AX = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}X_iX_j.$$

The expectation is:

$$\begin{aligned} E[X'AX] = E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}X_iX_j\right] = \\ = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}E\left[X_iX_j\right] = \\ = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}(\sigma_{i,j} + \mu_i\mu_j) = \\ = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}\sigma_{j,i} + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}\mu_i\mu_j = \\ = tr(A\Sigma) + \mu'A\mu. \end{aligned}$$

If we have an idempotent matrix A the quadratic form above is distributed as a chi-square with n degrees of freedom. (Muirhead (1982))
In the case $A = \Sigma^{-1}$ we get $E[X'\Sigma^{-1}X] = tr(I) + \mu\Sigma^{-1}\mu = N + \mu'\Sigma^{-1}\mu$ that follows a not centered chi-square distribution. In the paper we used this identity in the case $A = I$ and $A = \Sigma$ when we shrink the Unbiased Precision estimator towards the

Identity. Hence we get $E[X'IX] = tr(\Sigma) + \mu' I \mu = \lambda^2$ and $E[X'\Sigma X] = tr(\Sigma\Sigma) + \mu'\Sigma\mu = Q$.

A.1.2 Proposition 2: Proof

All covariances admit the classical spectral decomposition as $\Sigma = U\Lambda U'$, where U is an ortoghonal matrix such that $UU' = U'U = I$ and Λ is a diagonal matrix containing eigenvalues $\Lambda = Diag(\lambda)$ with $\lambda = (\lambda_1, \ldots, \lambda_N)'$. Therefore, also portfolio weights could be decomposed similarly. Following Ledoit and Wolf, 2017:

$$\hat{w} = \frac{1}{\gamma} U \hat{\Lambda}^{-1} U' \hat{\mu}$$

Hence, given two matrices $\hat{\Omega}_1$ and $\hat{\Omega}_2$ that both admit the spectral decomposition $\hat{\Omega}_1 = U\hat{\Lambda}_1^{-1}U'$ and $\hat{\Omega}_2 = U\hat{\Lambda}_2^{-1}U'$, we can write our maximization problem as follows:

$$E[U(\hat{w}_s)] = E\left[\frac{1}{\gamma}\hat{\mu}' U \hat{\Lambda}_s^{-1} U' \mu\right] - \frac{\gamma}{2} E\left[\left(\frac{1}{\gamma} U \hat{\Lambda}_s^{-1} U' \hat{\mu}\right)' U \Lambda U' \left(\frac{1}{\gamma} U \hat{\Lambda}_s^{-1} U' \hat{\mu}\right)\right]$$

where $\Lambda_s = \alpha^* \Lambda_1^{-1} + (1-\alpha^*)\Lambda_2^{-1}$. Hence:

$$\begin{aligned}
&E[U(\hat{w}_s)] = E\left[\frac{1}{\gamma}\hat{\mu}' U(\alpha\Lambda_1^{-1} + (1-\alpha)\Lambda_2^{-1})U'\mu\right] + \\
&- \frac{\gamma}{2} E\left[\frac{1}{\gamma^2}\hat{\mu}' U \left(\alpha\Lambda_1^{-1} + (1-\alpha)\Lambda_2^{-1}\right) U'U\Lambda U'U \left(\alpha\Lambda_1^{-1} + (1-\alpha)\Lambda_2^{-1}\right) U'\hat{\mu}\right] = \\
&= E\left[\frac{\alpha}{\gamma}\hat{\mu}' U \Lambda_1^{-1} U'\mu + \frac{(1-\alpha)}{\gamma}\hat{\mu}' U \Lambda_1^{-1} U'\mu\right] + \\
&- \frac{1}{2\gamma} E\left[\left(\alpha\hat{\mu}' U\Lambda_1^{-1}\Lambda + (1-\alpha)\hat{\mu}' U\Lambda_2^{-1}\Lambda\right)\left(\alpha\Lambda_1^{-1}U'\hat{\mu} + (1-\alpha)\Lambda_2^{-1}U'\hat{\mu}\right)\right].
\end{aligned}$$

Given orthogonality of U. The first term is exactly equivalent to $E\left[\frac{\alpha}{\gamma}\hat{\mu}'\hat{\Omega}_2^{-1}\mu + \frac{(1-\alpha)}{\gamma}\hat{\mu}'\hat{\Omega}_2^{-1}\mu\right]$. Let's analyse the second term:

$$\frac{1}{2\gamma}E\left[\left(\alpha\hat{\mu}'U\Lambda_1^{-1}\Lambda+(1-\alpha)\hat{\mu}'U\Lambda_2^{-1}\Lambda\right)\left(\alpha\Lambda_1^{-1}U'\hat{\mu}+(1-\alpha)\Lambda_2^{-1}U'\hat{\mu}\right)\right]=$$
$$=\frac{1}{2\gamma}E\left[\alpha^2\hat{\mu}'U\Lambda_1^{-1}\Lambda\Lambda_1^{-1}U'\hat{\mu}+(1-\alpha)^2\hat{\mu}'U\Lambda_2^{-1}\Lambda\Lambda_2^{-1}U'\hat{\mu}+2\alpha(1-\alpha)\hat{\mu}'U\Lambda_1^{-1}\Lambda\Lambda_2^{-1}U'\hat{\mu}\right]$$

Now, recognizing that $U\Lambda_1^{-1}\Lambda\Lambda_1^{-1}U'=\hat{\Omega}_1^{-1}\Sigma\hat{\Omega}_1^{-1}$, $U\Lambda_2^{-1}\Lambda\Lambda_2^{-1}U'=\hat{\Omega}_2^{-1}\Sigma\hat{\Omega}_2^{-1}$ and $U\Lambda_1^{-1}\Lambda\Lambda_2^{-1}U'=\hat{\Omega}_1^{-1}\Sigma\hat{\Omega}_2^{-1}$, the proof is completed.

A.1.3 Shrinkage rules

To prove that, we have to recall some properties. First of all, assuming that returns are jointly normally distributed and i.i.d., maximum likelihood estimators of mean and covariance are independent each other, where the covariance follows a Wishart distribution $\hat{\Sigma}\sim\mathcal{W}(T-1,\Sigma)$. Therefore we use the following relationships. First, $W=\Sigma^{-1/2}\hat{\Sigma}_{ML}\Sigma^{-1/2}$ that implies $\hat{\Sigma}_{ML}^{-1}=\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}$, where $E[W^{-1}]=T/(T-N-2)$. Second, $E[W^{-2}]=T^2(T-2)/(T-N-1)(T-N-2)(T-N-4)$. Third, the quadratic form $\hat{\mu}'\Sigma^{-1}\hat{\mu}$ follows a not centered chi-square distribution which expected value is $E[\hat{\mu}'\Sigma^{-1}\hat{\mu}]=(N+T\mu'\Sigma^{-1}\mu)/T$. (Muirhead (1982))
Suppose that $\hat{\Omega}_1=\hat{\Sigma}_{PM}$. In this case, $a=E[\hat{\Sigma}_{PM}^{-1}\hat{\mu}]$. Hence, we have to determine two quantities: $E[\hat{\mu}\hat{\Sigma}_{PM}^{-1}\mu]$ and $E[\hat{\mu}\hat{\Sigma}_{PM}^{-1}\Sigma\hat{\Sigma}_{PM}^{-1}\hat{\mu}]$. The first, namely the squared Sharpe ratio of the strategy, is equal to:

$$\begin{aligned}E[a'\mu]=E\left[\hat{\mu}'\hat{\Sigma}_{PM}^{-1}\mu\right]=\\=E\left[\frac{(T-N-2)}{T}\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\mu\right]=\\=\frac{(T-N-2)}{T}E\left[\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\mu\right]=\\=\frac{(T-N-2)}{T}\frac{T}{(T-N-2)}E\left[\hat{\mu}'\Sigma^{-1/2}\Sigma^{-1/2}\mu\right]=\mu'\Sigma^{-1}\mu=\\E[a'\mu]=\theta^2.\end{aligned}$$

Then the quantity $E[a'\Sigma a]$ is equal to:

$$
\begin{aligned}
E[a'\Sigma a] = E\left[\left(\hat{\Sigma}_{PM}^{-1}\hat{\mu}\right)'\Sigma\left(\hat{\Sigma}_{PM}^{-1}\hat{\mu}\right)\right] = \\
= \frac{(T-N-2)^2}{T^2}E\left[\left(\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\Sigma\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right)\right] = \\
= \frac{(T-N-2)^2}{T^2}E\left[\left(\hat{\mu}'\Sigma^{-1/2}W^{-2}\Sigma^{-1/2}\hat{\mu}\right)\right] = \\
= \frac{(T-N-2)^2}{T^2}\frac{T^2(T-2)}{(T-N-1)(T-N-2)(T-N-4)}\left(\frac{N+T\theta^2}{T}\right) = \\
= \frac{(T-2)(T-N-2)}{(T-N-1)(T-N-4)}\left(\frac{N+T\theta^2}{T}\right) = \\
E[a'\Sigma a] = c_1\frac{N}{T} + c_1\theta^2,
\end{aligned}
$$

where $c_1 = (T-2)(T-N-2)/(T-N-1)(T-N-4)$ as defined by Tu and Zhou, 2011. This two quantities will be always the same in our shrinkage operations. Now let us consider first the case where $\hat{\Omega}_2 = I$, hence $b = I\hat{\mu}$. In this case, the squared Sharpe ratio of this strategy is simply:

$$E[b'\mu] = E\left[\hat{\mu}'\hat{\mu}\right] = tr(\Sigma) + \mu'\mu = \lambda^2.$$

That could be derived by the expectation-trace identity showed in A.1. Then, with the same identity, we can find the value of the associated estimation error:

$$E[b'\Sigma b] = E\left[\hat{\mu}'\Sigma\hat{\mu}\right] = tr(\Sigma\Sigma) + \mu'\Sigma\mu = Q.$$

The last ingredient we need to determine optimal shrinkage intensity is the quantity $E[a'\Sigma b]$:

$$
\begin{aligned}
E[a'\Sigma b] = E\left[\hat{\mu}'\hat{\Sigma}_{PM}^{-1}\Sigma\hat{\mu}\right] = \\
= \frac{(T-N-2)}{T}E\left[\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\Sigma\hat{\mu}\right] = \\
= \frac{(T-N-2)}{T}E\left[\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\Sigma\hat{\mu}\right] = \\
= \frac{(T-N-2)}{T}\frac{T}{(T-N-2)}E\left[\hat{\mu}'\hat{\mu}\right] = tr(\Sigma) + \mu'\mu = \\
E[b'\mu] = \lambda^2.
\end{aligned}
$$

Hence, optimal shrinkage intensity is:

$$
\alpha^* = \frac{\theta^2 - 2\lambda^2 + Q}{c_1\frac{N}{T} + c_1\theta^2 + Q - 2\lambda^2}
$$

Consider now the case $\hat{\Omega}_2 = \Sigma_{EW}$. As the Identity case, we need to determine $E[b'\mu$ and $E[b'\Sigma b]$. Since $b = \Sigma_{EW}^{-1}\hat{\mu} = 1/N = w_e$, we get that $E[b'\mu] = w_e'\mu$ and $E[b'\Sigma b] = w_e'\Sigma w_e$. So we need to determine $E[a'\Sigma b]$ for this second shrinkage operation:

$$
\begin{aligned}
E[a'\Sigma b] = E\left[\hat{\mu}'\hat{\Sigma}_{PM}^{-1}\Sigma w_e\right] = \\
= \frac{(T-N-2)}{T}E\left[\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\Sigma w_e\right] = \\
= \frac{(T-N-2)}{T}E\left[\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\Sigma w_e\right] = \\
= \frac{(T-N-2)}{T}\frac{T}{(T-N-2)}E\left[\hat{\mu}'w_e\right] = \mu'w_e = \\
E[a'\Sigma b] = \mu'w_e.
\end{aligned}
$$

Since numerically $\mu'w_e = w_e'\mu$, we get the following optimal shrinkage intensity:

$$
\alpha^* = \frac{\theta^2 - 2w_e'\mu + w_e'\Sigma w_e}{c_1\frac{N}{T} + c_1\theta^2 + w_e'\Sigma w_e - 2w_e'\mu}.
$$

So we have easily proved that this equation is exactly the same of Tu and Zhou, 2011. Then, we study the performance of the following shrinkage estimator:

$$\hat{\Sigma}_s^{-1} = \alpha\hat{\Sigma}_{PM}^{-1} + (1-\alpha)\hat{\Sigma}_{c_3}^{-1}$$

Also here we have to determine the value of the three quantities. Starting from $E[b'\mu]$:

$$\begin{aligned}
E[b'\mu] = E\left[\hat{\mu}'\hat{\Sigma}_{c_3}^{-1}\mu\right] = \\
= E\left[\frac{(T-N-1)(T-N-4)}{T(T-2)}\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\mu\right] = \\
= \frac{(T-N-1)(T-N-4)}{T(T-2)}E\left[\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\mu\right] = \\
= \frac{(T-N-1)(T-N-4)}{T(T-2)}\frac{T}{(T-N-2)}E\left[\hat{\mu}'\Sigma^{-1/2}\Sigma^{-1/2}\mu\right] = \\
= \frac{(T-N-1)(T-N-4)}{(T-N-2)(T-2)}\mu'\Sigma^{-1}\mu = \\
E[b'\mu] = \frac{1}{c_1}\theta^2.
\end{aligned}$$

Then $E[b'\Sigma b]$ is:

$$\begin{aligned}
E[b'\Sigma b] = E\left[\left(\hat{\Sigma}_{c_3}^{-1}\hat{\mu}\right)'\Sigma\left(\hat{\Sigma}_{c_3}^{-1}\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)^2(T-N-4)^2}{T^2(T-2)^2}E\left[\left(\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\Sigma\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)^2(T-N-4)^2}{T^2(T-2)^2}E\left[\left(\hat{\mu}'\Sigma^{-1/2}W^{-2}\Sigma^{-1/2}\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)^2(T-N-4)^2}{T^2(T-2)^2}\frac{T^2(T-2)}{(T-N-1)(T-N-2)(T-N-4)}\left(\frac{N+T\theta^2}{T}\right) = \\
= \frac{(T-N-1)(T-N-4)}{(T-N-2)(T-2)}\left(\frac{N+T\theta^2}{T}\right) = \\
E[b'\Sigma b] = \frac{1}{c_1}\frac{N}{T} + \frac{1}{c_1}\theta^2.
\end{aligned}$$

In the end, we have to determine $E[a'\Sigma b]$:

$$E[a'\Sigma b] = E\left[\left(\hat{\Sigma}_{PM}^{-1}\hat{\mu}\right)'\Sigma\left(\hat{\Sigma}_{c_3}^{-1}\hat{\mu}\right)\right] =$$
$$= \frac{(T-N-1)(T-N-2)(T-N-4)}{T^2(T-2)}E\left[\left(\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\Sigma\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right)\right] =$$
$$= \frac{(T-N-1)(T-N-2)(T-N-4)}{T^2(T-2)}E\left[\left(\hat{\mu}'\Sigma^{-1/2}W^{-2}\Sigma^{-1/2}\hat{\mu}\right)\right] =$$
$$= \frac{(T-N-1)(T-N-2)(T-N-4)}{T^2(T-2)}\frac{T^2(T-2)}{(T-N-1)(T-N-2)(T-N-4)}\left(\frac{N+T\theta^2}{T}\right) =$$
$$E[a'\Sigma b] = \frac{N}{T}+\theta^2.$$

Hence, optimal shrinkage intensity is:

$$\alpha^* = \frac{\theta^2 - \frac{1}{c_1}\theta^2 + \frac{1}{c_1}\frac{N}{T} + \frac{1}{c_1}\theta^2 - \frac{N}{T} - \theta^2}{c_1\frac{N}{T} + c_1\theta^2 + \frac{1}{c_1}\frac{N}{T} + \frac{1}{c_1}\theta^2 - 2\frac{N}{T} - 2\theta^2} =$$
$$= \frac{(\frac{1}{c_1}-1)\frac{N}{T}}{(c_1+\frac{1}{c_1}-2)\frac{N}{T} + \left(c_1+\frac{1}{c_1}-2\right)\theta^2} =$$
$$= \frac{\left(\frac{1-c_1}{c_1}\right)\frac{N}{T}}{\left(\frac{1+c_1^2-2c_1}{c_1}\right)\frac{N}{T} + \left(\frac{1+c_1^2-2c_1}{c_1}\right)\theta^2} =$$
$$\alpha^* = \frac{\left(\frac{1-c_1}{c_1}\right)\frac{N}{T}}{\left(\frac{(1-c_1)^2}{c_1}\right)\frac{N}{T} + \left(\frac{(1-c_1)^2}{c_1}\right)\theta^2}.$$

Suppose, now, to consider $\hat{\Omega}_1 = \hat{\Sigma}_{c_3}$. We study the shrinkage of c_3 towards both Identity and equally weighted. Above we have already defined all the involved quantities, despite $E[a'\Sigma b]$ that in this case is:

$$
\begin{aligned}
E[a'\Sigma b] = E\left[\left(\hat{\Sigma}_{c_3}^{-1}\hat{\mu}\right)'\Sigma\left(I\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)(T-N-4)}{T(T-2)}E\left[\left(\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\Sigma\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)(T-N-4)}{T(T-2)}E\left[\left(\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\Sigma\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)(T-N-4)}{T(T-2)}\frac{T}{(T-N-2)}E\left[\left(\hat{\mu}'\Sigma^{-1/2}\Sigma^{-1/2}\Sigma\hat{\mu}\right)\right] = \\
E[a'\Sigma b] = \frac{1}{c_1}\lambda^2.
\end{aligned}
$$

Therefore, for shrinkage $\hat{\Sigma}_{c_3}^{-1}$ towards Identity I, the optimal shrinkage intensity is:

$$
\begin{aligned}
\alpha^* = \frac{\frac{1}{c_1}\theta^2 - \lambda^2 + Q - \frac{1}{c_1}\lambda^2}{\frac{1}{c_1}\frac{N}{T} + \frac{1}{c_1}\theta^2 + Q - 2\frac{1}{c_1}\lambda^2} = \\
\alpha^* = \frac{\frac{1}{c_1}\theta^2 - \left(1+\frac{1}{c_1}\right)\lambda^2 + Q}{\frac{1}{c_1}\left(\theta^2 + \frac{N}{T} - 2\lambda^2\right) + Q}.
\end{aligned}
$$

Then, we suppose to shrink $\hat{\Sigma}_{c_3}^{-1}$ towards the implied equally weighted covariace Σ_{EW}^{-1}. Now $E[a'\Sigma b]$ is equal to:

$$
\begin{aligned}
E[a'\Sigma b] = E\left[\left(\hat{\Sigma}_{c_3}^{-1}\hat{\mu}\right)'\Sigma\left(\Sigma_{EW}^{-1}\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)(T-N-4)}{T(T-2)}E\left[\left(\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\Sigma w_e\right)\right] = \\
= \frac{(T-N-1)(T-N-4)}{T(T-2)}E\left[\left(\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\Sigma w_e\right)\right] = \\
= \frac{(T-N-1)(T-N-4)}{T(T-2)}\frac{T}{(T-N-2)}E\left[\left(\hat{\mu}'w_e\right)\right] = \\
E[a'\Sigma b] = \frac{1}{c_1}\mu' w_e = \frac{1}{c_1}w_e'\mu.
\end{aligned}
$$

Hence optimal shrinkage intensity is:

$$\alpha^* = \frac{\frac{1}{c_1}\theta^2 - w_e'\mu + w_e'\Sigma w_e - \frac{1}{c_1}w_e'\mu}{\frac{1}{c_1}\frac{N}{T} + \frac{1}{c_1}\theta^2 + w_e'\Sigma w_e - 2w_e'\mu} =$$
$$\alpha^* = \frac{\frac{1}{c_1}\theta^2 - \left(1 + \frac{1}{c_1}\right)w_e'\mu + w_e'\Sigma w_e}{\frac{1}{c_1}\frac{N}{T} + \frac{1}{c_1}\theta^2 + w_e'\Sigma w_e - 2w_e'\mu}.$$

In the end, suppose $\hat{\Omega}_1 = \hat{\Sigma}_{c^*}$ and that we shrink c^* towards equally Identity matrix I. In this case $E[a'\Sigma b]$:

$$E[a'\Sigma b] = E\left[\left(\hat{\Sigma}_{c^*}^{-1}\hat{\mu}\right)'\Sigma\left(I\hat{\mu}\right)\right] =$$
$$= \frac{(T-N-1)(T-N-4)}{T(T-2)}\left(\frac{\theta^2}{\theta^2 + \frac{N}{T}}\right)E\left[\left(\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\Sigma\hat{\mu}\right)\right] =$$
$$= \frac{(T-N-1)(T-N-4)}{T(T-2)}\left(\frac{\theta^2}{\theta^2 + \frac{N}{T}}\right)E\left[\left(\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\Sigma\hat{\mu}\right)\right] =$$
$$= \frac{(T-N-1)(T-N-4)}{(T-N-2)(T-2)}\left(\frac{\theta^2}{\theta^2 + \frac{N}{T}}\right)\lambda^2 =$$
$$E[a'\Sigma b] = \frac{1}{c_1}\left(\frac{\theta^2}{\theta^2 + \frac{N}{T}}\right)\lambda^2.$$

Then optimal α^* is:

$$\alpha^* = \frac{\frac{1}{c_1}\left(\frac{\theta^4}{\theta^2 + N/T}\right) - \lambda^2 + Q - \frac{1}{c_1}\left(\frac{\theta^2}{\theta^2 + \frac{N}{T}}\right)\lambda^2}{\frac{1}{c_1}\left(\frac{N}{T} + \theta^2\right) + Q - 2\frac{1}{c_1}\left(\frac{\theta^2}{\theta^2 + \frac{N}{T}}\right)\lambda^2}.$$

The last exercise is to shrink c^* towards equally weighted covariance matrix Σ_{EW}. In this case $E[a'\Sigma b]$ is:

$$
\begin{aligned}
E[a'\Sigma b] = E\left[\left(\hat{\Sigma}_{c^*}^{-1}\hat{\mu}\right)'\Sigma\left(\hat{\Sigma}_{EW}^{-1}\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)(T-N-4)}{T(T-2)}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)E\left[\left(\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\Sigma w_e\right)\right] = \\
= \frac{(T-N-1)(T-N-4)}{T(T-2)}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)E\left[\left(\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\Sigma w_e\right)\right] = \\
= \frac{(T-N-1)(T-N-4)}{(T-N-2)(T-2)}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)\mu' w_e = \\
E[a'\Sigma b] = \frac{1}{c_1}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)w_e'\mu.
\end{aligned}
$$

Then optimal α^* is:

$$
\alpha^* = \frac{\frac{1}{c_1}\left(\frac{\theta^4}{\theta^2+N/T}\right) - w_e'\mu + w_e'\Sigma w_e - \frac{1}{c_1}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)w_e'\mu}{\frac{1}{c_1}\left(\frac{N}{T}+\theta^2\right) + w_e'\Sigma w_e - 2\frac{1}{c_1}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)w_e'\mu}.
$$

A.1.4 Unbiased estimator for Q

In what follows we derived the unbiased estimator for $Q = tr(\Sigma\Sigma) + \mu'\Sigma\mu$. Its sample counterpart $\hat{Q} = \hat{\mu}'\hat{\Sigma}_{SC}\hat{\mu}$ is biased since:

$$
\begin{aligned}
E[\hat{Q}] = E\left[\hat{\mu}'\hat{\Sigma}_{SC}\hat{\mu}\right] = \frac{T}{T-1}E\left[\hat{\mu}'\hat{\Sigma}_{ML}\hat{\mu}\right] = \\
= \frac{T}{T-1}E\left[\hat{\mu}'\Sigma^{1/2}W\Sigma^{1/2}\hat{\mu}\right] = \frac{T}{T-1}\frac{T-N-2}{T}E\left[\hat{\mu}'\Sigma\hat{\mu}\right] = \\
E[\hat{Q}] = \frac{T-N-2}{T-1}\left[tr(\Sigma\Sigma) + \mu'\Sigma\mu\right] \neq Q.
\end{aligned}
$$

Hence, with the corrective factor $(T-1)/(T-N-2)$ we get unibiasdeness.

A.1.5 Shrinkage of any sample estimate towards c^*

As mentioned in the main text, if we shrink any plug-in towards Biased Optimal Scaled with c^* as in (1.19) and in (1.20) lead to an $\alpha^* = 0$. This happen because, from an utility maximization point of view, the BOS is already an utility maximizer

estimator. To show that, consider first the case of shrinking Unbiased Precision estimator towards BOS. The value of $E[b'\mu]$ in this case:

$$\begin{aligned}
E[b'\mu] = E\left[\hat{\mu}'\hat{\Sigma}_{c^*}^{-1}\mu\right] = \\
= E\left[\frac{(T-N-1)(T-N-4)}{T(T-2)}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\mu\right] = \\
= \frac{(T-N-1)(T-N-4)}{T(T-2)}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)E\left[\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\mu\right] = \\
= \frac{(T-N-1)(T-N-4)}{T(T-2)}\frac{T}{(T-N-2)}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)E\left[\hat{\mu}'\Sigma^{-1/2}\Sigma^{-1/2}\mu\right] = \\
= \frac{(T-N-1)(T-N-4)}{(T-N-2)(T-2)}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)\theta^2 = \\
E[b'\mu] = \frac{1}{c_1}\left(\frac{\theta^4}{\theta^2+\frac{N}{T}}\right)
\end{aligned}$$

Then $E[b'\Sigma b]$ is:

$$\begin{aligned}
E[b'\Sigma b] = E\left[\left(\hat{\Sigma}_{c^*}^{-1}\hat{\mu}\right)'\Sigma\left(\hat{\Sigma}_{c^*}^{-1}\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)^2(T-N-4)^2}{T^2(T-2)^2}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)^2 E\left[\left(\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\Sigma\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)^2(T-N-4)^2}{T^2(T-2)^2}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)^2 E\left[\left(\hat{\mu}'\Sigma^{-1/2}W^{-2}\Sigma^{-1/2}\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)^2(T-N-4)^2}{T^2(T-2)^2}\frac{T^2(T-2)}{(T-N-1)(T-N-2)(T-N-4)}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)^2\left(\frac{N}{T}+\theta^2\right) = \\
= \frac{(T-N-1)(T-N-4)}{(T-N-2)(T-2)}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)^2\left(\frac{N}{T}+\theta^2\right) = \\
E[b'\Sigma b] = \frac{1}{c_1}\left(\frac{\theta^4}{\theta^2+\frac{N}{T}}\right).
\end{aligned}$$

In the end, we have to determine $E[a'\Sigma b]$:

$$E[a'\Sigma b] = E\left[\left(\hat{\Sigma}_{PM}^{-1}\hat{\mu}\right)'\Sigma\left(\hat{\Sigma}_{c^*}^{-1}\hat{\mu}\right)\right] =$$
$$= \frac{(T-N-1)(T-N-2)(T-N-4)}{T^2(T-2)}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)E\left[\left(\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\Sigma\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right)\right] =$$
$$= \frac{(T-N-1)(T-N-2)(T-N-4)}{T^2(T-2)}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)E\left[\left(\hat{\mu}'\Sigma^{-1/2}W^{-2}\Sigma^{-1/2}\hat{\mu}\right)\right] =$$
$$= \left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)\left(\frac{N}{T}+\theta^2\right) =$$
$$E[a'\Sigma b] = \theta^2.$$

Hence, optimal shrinkage intensity is:

$$\alpha^* = \frac{E[a'\mu] - E[b'\mu] + E[b'\Sigma b] - E[a'\Sigma b]}{E[a'\Sigma a] + E[b'\Sigma b] - 2E[a'\Sigma b]} =$$
$$\alpha^* = \frac{\theta^2 - \frac{1}{c_1}\left(\frac{\theta^4}{\theta^2+\frac{N}{T}}\right) + \frac{1}{c_1}\left(\frac{\theta^4}{\theta^2+\frac{N}{T}}\right) - \theta^2}{c_1\left(\frac{N}{T}+\theta^2\right) + \frac{1}{c_1}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right) - 2\theta^2}.$$

That is clearly equal to zero because of the numerator. Equally, $(1-\alpha^*) = 1$ since:

$$(1-\alpha^*) = \frac{E[b'\mu] - E[a'\mu] + E[a'\Sigma a] - E[a'\Sigma b]}{E[a'\Sigma a] + E[b'\Sigma b] - 2E[a'\Sigma b]} =$$
$$(1-\alpha^*) = \frac{\frac{1}{c_1}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right) - \theta^2 + c_1\left(\frac{N}{T}+\theta^2\right) - \theta^2}{c_1\left(\frac{N}{T}+\theta^2\right) + \frac{1}{c_1}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right) - 2\theta^2} = 1.$$

Then, the same apply if we shrink c_3 towards c^* estimators. Suppose now BOS is $\hat{\Omega}_1$. Optimal shrinkage intensity $\alpha^* = 1$. Remember the following quantities in this case, all that have already been derived in the Appendix A.1.3:

- $E[a'\mu] = \frac{1}{c_1}\left(\frac{\theta^4}{\theta^2+N/T}\right)$;
- $E[b'\mu] = \frac{1}{c_1}\theta^2$;
- $E[b'\Sigma b] = \frac{1}{c_1}\left(\frac{N}{T}+\theta^2\right)$;

- $E[a'\Sigma a] = \frac{1}{c_1}\left(\frac{\theta^4}{\theta^2+N/T}\right)$;
- while $E[a'\Sigma b]$ is:

$$
\begin{aligned}
E[a'\Sigma b] = E\left[\left(\hat{\Sigma}_{c^*}^{-1}\hat{\mu}\right)'\Sigma\left(\hat{\Sigma}_{c_3}^{-1}\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)^2(T-N-4)^2}{T^2(T-2)^2}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)E\left[\left(\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\Sigma\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)^2(T-N-4)^2}{T^2(T-2)^2}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)E\left[\left(\hat{\mu}'\Sigma^{-1/2}W^{-2}\Sigma^{-1/2}\hat{\mu}\right)\right] = \\
= \frac{(T-N-1)(T-N-4)}{(T-N-2)(T-2)}\left(\frac{\theta^2}{\theta^2+\frac{N}{T}}\right)\left(\frac{N}{T}+\theta^2\right) = \\
E[a'\Sigma b] = \frac{1}{c_1}\theta^2.
\end{aligned}
$$

Hence α^* is:

$$
\begin{aligned}
\alpha^* &= \frac{E[a'\mu] - E[b'\mu] + E[b'\Sigma b] - E[a'\Sigma b]}{E[a'\Sigma a] + E[b'\Sigma b] - 2E[a'\Sigma b]} = \\
\alpha^* &= \frac{\frac{1}{c_1}\left(\frac{\theta^4}{\theta^2+N/T}\right) - \frac{1}{c_1}\theta^2 + \frac{1}{c_1}\left(\frac{N}{T}+\theta^2\right) - \frac{1}{c_1}\theta^2}{\frac{1}{c_1}\left(\frac{N}{T}+\theta^2\right) + \frac{1}{c_1}\left(\frac{N}{T}+\theta^2\right) - 2\frac{1}{c_1}\theta^2} = 1.
\end{aligned}
$$

As we were claiming before.

A.1.6 Percieved expected utilities

Statistical tools that we used to derive results here are the same of Appendix A.1.3. Suppose, first, the case where Σ is known while μ it is not:

$$E[U(\hat{w})|\Sigma,\hat{\mu}] = E\left[\frac{1}{\gamma}\hat{\mu}'\Sigma^{-1}\hat{\mu}\right] - \frac{\gamma}{2}E\left[\frac{1}{\gamma}\hat{\mu}'\Sigma^{-1}\Sigma\frac{1}{\gamma}\Sigma^{-1}\hat{\mu}\right] =$$
$$= \frac{1}{\gamma}E\left[\hat{\mu}'\Sigma^{-1}\hat{\mu}\right] - \frac{\gamma}{2}\frac{1}{\gamma^2}E\left[\hat{\mu}'\Sigma^{-1}\Sigma\Sigma^{-1}\hat{\mu}\right] =$$
$$= \frac{1}{\gamma}E\left[\hat{\mu}'\Sigma^{-1}\hat{\mu}\right] - \frac{1}{2\gamma}E\left[\hat{\mu}'\Sigma^{-1}\hat{\mu}\right] =$$
$$\frac{1}{2\gamma}E\left[\hat{\mu}'\Sigma^{-1}\hat{\mu}\right] = \frac{1}{2\gamma}\left(\frac{N}{T}+\theta^2\right).$$

Then, suppose now the case where Σ is unknown while μ is known:

$$E[U(\hat{w})|\hat{\Sigma}_{ML},\mu] = E\left[\frac{1}{\gamma}\mu'\hat{\Sigma}_{ML}^{-1}\mu\right] - \frac{\gamma}{2}E\left[\frac{1}{\gamma}\mu'\hat{\Sigma}_{ML}^{-1}\hat{\Sigma}_{ML}\frac{1}{\gamma}\hat{\Sigma}_{ML}^{-1}\mu\right] =$$
$$= \frac{1}{\gamma}\mu'E\left[\hat{\Sigma}_{ML}^{-1}\right]\mu - \frac{\gamma}{2}\frac{1}{\gamma^2}\mu'E\left[\hat{\Sigma}_{ML}^{-1}\hat{\Sigma}_{ML}\hat{\Sigma}_{ML}^{-1}\right]\mu =$$
$$= \frac{1}{\gamma}\mu'E\left[\hat{\Sigma}_{ML}^{-1}\right]\mu - \frac{1}{2\gamma}\mu'E\left[\hat{\Sigma}_{ML}^{-1}\right]\mu =$$
$$= \frac{1}{2\gamma}\mu'E\left[\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\right]\mu = \frac{1}{2\gamma}\frac{T-N-2}{T}\theta^2.$$

Suppose, instead, that both are unknown, where $\Sigma = \hat{\Sigma}_{ML}$:

$$E[U(\hat{w}_{ML})|\hat{\Sigma},\hat{\mu}] = E\left[\frac{1}{\gamma}\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] - \frac{\gamma}{2}E\left[\frac{1}{\gamma}\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\Sigma}_{ML}\frac{1}{\gamma}\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] =$$
$$= \frac{1}{\gamma}E\left[\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] - \frac{1}{2\gamma}E\left[\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] = \frac{1}{2\gamma}E\left[\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] =$$
$$= \frac{1}{2\gamma}E\left[\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\hat{\mu}\right] = \frac{1}{2\gamma}\frac{T}{T-N-2}E\left[\hat{\mu}'\Sigma^{-1}\hat{\mu}\right] =$$
$$= \frac{1}{2\gamma}\frac{T}{T-N-2}\left(\frac{N}{T}+\theta^2\right) = \frac{1}{2\gamma}\frac{N}{T-N-2} + \frac{1}{2\gamma}\frac{T}{T-N-2}\theta^2 \implies$$
$$E[U(\hat{w}_{ML})] = \frac{1}{2\gamma}\frac{1}{T-N-2}\left(N+T\theta^2\right)$$

Suppose, then, the case of sample covariance. In this case percieved utility is:

$$E[U(\hat{w}_{SC})|\hat{\Sigma},\hat{\mu}] = E\left[\frac{1}{\gamma}\hat{\mu}'\hat{\Sigma}_{SC}^{-1}\hat{\mu}\right] - \frac{\gamma}{2}E\left[\frac{1}{\gamma}\hat{\mu}'\hat{\Sigma}_{SC}^{-1}\hat{\Sigma}_{SC}\frac{1}{\gamma}\hat{\Sigma}_{SC}^{-1}\hat{\mu}\right] =$$
$$= \frac{1}{\gamma}\frac{T-1}{T}E\left[\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] - \frac{1}{2\gamma}\frac{T-1}{T}E\left[\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] = \frac{1}{2\gamma}\frac{T-1}{T}E\left[\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] =$$
$$= \frac{1}{2\gamma}\frac{T-1}{T}E\left[\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\hat{\mu}\right] = \frac{1}{2\gamma}\frac{T-1}{T}\frac{T}{T-N-2}E\left[\hat{\mu}'\Sigma^{-1}\hat{\mu}\right] =$$
$$E[U(\hat{w}_{SC})] = \frac{1}{2\gamma}\frac{T-1}{T-N-2}\left(\frac{N}{T}+\theta^2\right).$$

Then, we consider the Unbiased Precision (PM) Estimator. We have that:

$$E[U(\hat{w}_{PM})|\hat{\Sigma},\hat{\mu}] = E\left[\frac{1}{\gamma}\hat{\mu}'\hat{\Sigma}_{PM}^{-1}\hat{\mu}\right] - \frac{\gamma}{2}E\left[\frac{1}{\gamma}\hat{\mu}'\hat{\Sigma}_{PM}^{-1}\hat{\Sigma}_{PM}\frac{1}{\gamma}\hat{\Sigma}_{PM}^{-1}\hat{\mu}\right] =$$
$$= \frac{1}{\gamma}\frac{T-N-2}{T}E\left[\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] - \frac{1}{2\gamma}\frac{T-N-2}{T}E\left[\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] = \frac{1}{2\gamma}\frac{T-N-2}{T}E\left[\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] =$$
$$= \frac{1}{2\gamma}\frac{T-N-2}{T}E\left[\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\hat{\mu}\right] = \frac{1}{2\gamma}\frac{T-N-2}{T}\frac{T}{T-N-2}E\left[\hat{\mu}'\Sigma^{-1}\hat{\mu}\right] =$$
$$E[U(\hat{w}_{PM})] = \frac{1}{2\gamma}\left(\frac{N}{T}+\theta^2\right).$$

In the case of equally weighted strategy, where the investor estimate mean and covariance via MLE, perceived utility is:

$$E[U(\hat{w}_{ew})|\hat{\Sigma},\hat{\mu}] = E[w_e'\hat{\mu}] - \frac{\gamma}{2}E\left[w_e'\hat{\Sigma}w_e\right] = w_e'E[\hat{\mu}] - \frac{\gamma}{2}w_e'E[\hat{\Sigma}_{ML}]w_e =$$
$$= w_e'\mu - \frac{\gamma}{2}w_e'E[\Sigma^{1/2}W\Sigma^{1/2}]w_e = w_e'\mu - \frac{\gamma}{2}\frac{T-N-2}{T}w_e'\Sigma w_e.$$

where $w_e = (1/N,\ldots,1/N)$ is a constant vector.

A.1.7 Maximum likelihood is the (perceived) optimal strategy

$$
\begin{aligned}
E\left[\frac{c}{\gamma}\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] - \frac{\gamma}{2}E\left[\frac{c^2}{\gamma^2}\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\Sigma}_{ML}\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] = \\
= \frac{c}{\gamma}E\left[\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] - \frac{c^2}{2\gamma}E\left[\hat{\mu}'\hat{\Sigma}_{ML}^{-1}\hat{\mu}\right] = \\
\frac{c}{\gamma}E\left[\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\hat{\mu}\right] - \frac{c^2}{2\gamma}E\left[\hat{\mu}'\Sigma^{-1/2}W^{-1}\Sigma^{-1/2}\hat{\mu}\right] = \\
\frac{c}{\gamma}\frac{T}{T-N-2}E\left[\hat{\mu}'\Sigma^{-1}\hat{\mu}\right] - \frac{c^2}{2\gamma}\frac{T}{T-N-2}E\left[\hat{\mu}'\Sigma^{-1}\hat{\mu}\right] = \\
\frac{c}{\gamma}\frac{T}{T-N-2}\left(\frac{N}{T}+\theta^2\right) - \frac{c^2}{2\gamma}\frac{T}{T-N-2}\left(\frac{N}{T}+\theta^2\right) = \\
\frac{\partial \frac{c}{\gamma}\frac{T}{T-N-2}\left(\frac{N}{T}+\theta^2\right) - \frac{c^2}{2\gamma}\frac{T}{T-N-2}\left(\frac{N}{T}+\theta^2\right)}{\partial c} = 0 \\
\frac{1}{\gamma}\frac{T}{T-N-2}\left(\frac{N}{T}+\theta^2\right) - \frac{c}{\gamma}\frac{T}{T-N-2}\left(\frac{N}{T}+\theta^2\right) = 0 \\
\frac{1}{\gamma}\frac{T}{T-N-2}\left(\frac{N}{T}+\theta^2\right) = \frac{c}{\gamma}\frac{T}{T-N-2}\left(\frac{N}{T}+\theta^2\right) = \\
c^* = 1. \qquad \text{(A.1)}
\end{aligned}
$$

A.2 Accuracy of the simulations contained in the Chapter 1

Kan & Zhou (2007) studied analytically the estimation error. Estimation error has been defined as the loss in investor's utility due to mistakes in estimating optimal portfolio weights. In other words, remember formula (1.3):

$$\ell(w^*, \hat{w}) = U(w^*) - E[U(\hat{w})]$$

where w^* is the vector of optimal portfolio weights and $\hat{w}$ is the vector of the estimated weights, by plug-in estimators for μ and Σ that we define $\hat{\mu}$ and $\hat{\Sigma}$. Here we have to specify that, while in $U(w^*)$ there is no estimation at all and all quantities are known, for $E[U(\hat{w})]$ parameter estimation is considered **just** within the weights. Indeed, note that:

$$E[U(\hat{w})] = E[\hat{w}'\mu] - \frac{\gamma}{2}E\left[\hat{w}'\Sigma\hat{w}\right]$$

where as we know:

$$\hat{w} = \frac{1}{\gamma}\hat{\Sigma}^{-1}\hat{\mu}$$

Now, as in Kan and Zhou (2007), for $\hat{\mu}$ we always we consider maximum likelihood estimator, while for $\hat{\Sigma}$ we consider several alternative estimators (e.g. maximum likelihood, sample covariance, Unbiased Precision estimator, Identity and later on a new shrinkage). Before to start estimators comparisons we want first to show that our simulations will be correct, in the sense that the expected losses obtained by simulations have to be very accurate approximations of closed formula derived by Kan and Zhou (2007). Indeed, differently from Kan and Zhou (2007), we will not mix closed losses with simulated one but we will instead evaluate all strategies via simulations for a more fair comparison. This is the reason why simulation results must be very good approximations of closed results. In order to do so, we focus on maximum likelihood (therefore, MLE) estimation where both $\hat{\mu}$ and $\hat{\Sigma}$ are estimated via MLE.

Suppose first to consider the case where, in estimating $\hat{w}$, Σ is known and μ is estimated via MLE so we have:

$$\hat{w} = \frac{1}{\gamma}\Sigma^{-1}\hat{\mu}$$

Closed loss, that we call ℓ_1 in this case is:

$$\ell_1 = \frac{N}{2\gamma T} \tag{A.2}$$

Tables A.1 and A.2 report a comparison of ℓ_1 values with the results obtained from M simulations, considering $N = 5$ and $N = 30$ assets with $T = 60$ observations and different values of γ.

From both tables we can recognize that simulation accuracy for this first case is very good, since difference between ℓ_1 and simulated loss is of the order $1/10000$ and in several cases also of $1/1000000$. With known covariance there is not a significant difference in accuracy between $N = 5$ and $N = 30$.

	ℓ_1	Simulations	Difference
$\gamma = 0.3$			
M=1000	0.1388889	0.1449699	-0.006081043
M=10000	0.1388889	0.139362	-0.0004731072
M=100000	0.1388889	0.1390584	-0.0001694949
$\gamma = 1$			
M=1000	0.04166667	0.04126821	0.0003984527
M=10000	0.04166667	0.04173014	-6.347148e^{-05}
M=100000	0.04166667	0.04167376	-7.096134e^{-06}
$\gamma = 3$			
M=1000	0.01388889	0.01403942	-0.0001505312
M=10000	0.01388889	0.01373378	0.0001551044
M=100000	0.01388889	0.0139588	-6.991176e^{-05}

TABLE A.1: Allocation $\hat{w}$ with Σ and $\hat{\mu}$: closed loss vs simulations with $N = 5$

Time series length	ℓ_1	Simulations	Difference
$\gamma = 0.3$			
M=1000	0.8333333	0.8301875	0.003145805
M=10000	0.8333333	0.8324443	0.0008890506
M=100000	0.8333333	0.833451	-0.0001177057
$\gamma = 1$			
M=1000	0.25	0.2488644	0.001135552
M=10000	0.25	0.2487208	0.00127924
M=100000	0.25	0.2504122	-0.000412231
$\gamma = 3$			
M=1000	0.08333333	0.08339203	-5.870147e^{-05}
M=10000	0.08333333	0.08339914	-6.580329e^{-05}
M=100000	0.08333333	0.08330997	2.336758e^{-05}

TABLE A.2: Allocation $\hat{w}$ with Σ and $\hat{\mu}$: closed loss vs simulations with $N = 30$

Then, consider the case where in estimating $\hat{w}$, μ is known and Σ is estimated via MLE so we have:

$$\hat{w} = \frac{1}{\gamma}\hat{\Sigma}^{-1}\mu$$

Closed loss, that we call ℓ_2 in this case is:

$$\ell_2 = (1 - k_1)\frac{\theta^2}{2\gamma} \tag{A.3}$$

where θ^2 is the true squared Sharpe ratio. Tables A.3 and A.4 report a comparison of ℓ_2 values with the results obtained from M simulations.

In this second case simulation accuracy is much higher with $N = 5$ rather than $N = 30$. Indeed, while the approximation order is almost about $1/100000$ in the first case, it is of $1/100$ in the second one. This is probably due to covariance matrix inversion of simulated data that is problematical not only from simulation itself but also by the fact that $T - N$ is lower in the second scenario.

Time series length	ℓ_2	Simulations	Difference
$\gamma = 0.3$			
M=1000	0.01347972	0.01321439	0.0002653283
M=10000	0.01347972	0.01290361	0.0005761089
M=100000	0.01347972	0.0127488	0.0007309113
$\gamma = 1$			
M=1000	0.004043915	0.0039757	$6.82147e^{-05}$
M=10000	0.004043915	0.003750547	0.0002933673
M=100000	0.004043915	0.003829548	0.0002143666
$\gamma = 3$			
M=1000	0.001347972	0.001260536	$8.743548e^{-05}$
M=10000	0.001347972	0.001263504	$8.446758e^{-05}$
M=100000	0.001347972	0.001270044	$7.792721e^{-05}$

TABLE A.3: Allocation $\hat{w}$ with $\hat{\Sigma}$ and μ: closed loss vs simulations with $N = 5$

Time series length	ℓ_2	Simulations	Difference
$\gamma = 0.3$			
M=1000	0.9950749	0.8232067	0.1718681
M=10000	0.9950749	0.8520843	0.1429906
M=100000	0.9950749	0.8534032	0.1416717
$\gamma = 1$			
M=1000	0.2985225	0.2429061	0.05561637
M=10000	0.2985225	0.2568785	0.04164394
M=100000	0.2985225	0.2564265	0.04209601
$\gamma = 3$			
M=1000	0.09950749	0.08117736	0.01833013
M=10000	0.09950749	0.08489173	0.01461576
M=100000	0.09950749	0.08577764	0.01372985

TABLE A.4: Allocation $\hat{w}$ with $\hat{\Sigma}$ and μ: closed loss vs simulations with $N = 30$

Last, we consider the case where within $\hat{w}$ both μ and Σ are estimated via MLE so we have:

$$\hat{w} = \frac{1}{\gamma}\hat{\Sigma}^{-1}\hat{\mu}$$

Closed loss, that we call ℓ_3 in this case is:

$$\ell_3 = (1 - k_1)\frac{\theta^2}{2\gamma} + \frac{NT(T-2)}{2\gamma(T-N-1)(T-N-2)(T-N-4)} \tag{A.4}$$

Tables A.5 and A.6 report a comparison of ℓ_3 values with the results obtained from M simulations.

In this last case simulation accuracy is not so problematic with $N = 5$ assets, even if we need to make a lot of simulations (at least $M = 100000$) to get a difference of the order $1/1000$. Instead, with $N = 30$ assets trough simulations we do not get very good results, in the sense that loss is a bit far from the actual value, even with $M = 100000$. Nevertheless, with increasing M we increase our accuracy. Therefore, in this case we need a lot of simulations (that takes a lot of time to be completed) for

well approximating ℓ_3. In general these results suggests that our simulation scheme is generally correct. Nevertheless, for a fair comparisons do not seems to be the case of comparing closed formulas with simulations as in Kan and Zhou (2007), since in several scenarios simulations under estimate too much the loss with respect to the closed formula. This point is very important and has to be stressed. Since for several strategies do not exist any closed loss, if we systemically under estimate that trough simulation, we can conclude that a certain strategy over performs another one for which we use a closed loss ℓ. Therefore, if our aim is to compare different asset allocation strategies for which we have only partially closed losses, we suggest to use directly simulated losses for all the strategies. In this way we'll never make mistakes.

Time series length	ℓ_3	Simulations	Difference
$\gamma = 0.3$			
M=1000	0.2121616	0.1724365	0.03972505
M=10000	0.2121616	0.1760426	0.03611894
M=100000	0.2121616	0.1768457	0.03531585
$\gamma = 1$			
M=1000	0.06364847	0.0534151	0.01023337
M=10000	0.06364847	0.05411348	0.009534989
M=100000	0.06364847	0.05321563	0.01043284
$\gamma = 3$			
M=1000	0.02121616	0.01728941	0.003926747
M=10000	0.02121616	0.01796729	0.003248865
M=100000	0.02121616	0.01776633	0.003449827

TABLE A.5: Allocation $\hat{w}$ with $\hat{\Sigma}$ and $\hat{\mu}$: closed loss vs simulations with $N = 5$

Time series length	ℓ_3	Simulations	Difference
$\gamma = 0.3$			
M=1000	9.236833	2.892582	6.344251
M=10000	9.236833	2.887576	6.349258
M=100000	9.236833	2.901796	6.335037
$\gamma = 1$			
M=1000	2.77105	0.8585785	1.912471
M=10000	2.77105	0.8667776	1.904272
M=100000	2.77105	0.8718123	1.899238
$\gamma = 3$			
M=1000	0.9236833	0.2886334	0.6350499
M=10000	0.9236833	0.2878283	0.635855
M=100000	0.9236833	0.2897285	0.6339548

TABLE A.6: Allocation $\hat{w}$ with $\hat{\Sigma}$ and $\hat{\mu}$: closed loss vs simulations with $N = 30$

Appendix B

Volatility clustering in financial markets

The following pages are absolutely not intended as a review about the past and recent developments of statistical techniques for dealing with volatility clustering. There are many books as well as special issues in field journals with this aim. However, in what follows I provide some very simple arguments and few evidences to explain why conditional heteroskedasticity models (mainly GARCH and DCC) are used by both academics and financial industry practitioners. In few words,we can argue that "*...the GARCH specification does not arise directly out of any economic theory but, as in the traditional autoregressive moving average time-series analogue, it provides a close and parsimonious approximation to the form of heteroscedasticity typically encountered with economic time-series data*" (Bollerslev, Engle, and Wooldridge, 1988).

B.1 Modelling conditional heteroskedasticity

The idea of modeling volatility as function of time born 40 years ago and is due to a Rob Engle's paper, published in the 1982, in which he introduced the autoregressive conditional heteroskedasticity (ARCH) model (Engle, 1982). Despite the application provided in the paper was related to a macroeconomic time series (i.e. the inflation), is perhaps more common nowadays the application of such methods to financial markets data. Indeed, the most important developments in conditional volatility modelling have been introduced by financial econometricians.

The main idea underlying ARCH modelling is the following. Let consider a random variable Y_t that is drawn from a conditional density function $f(y_t|y_{t-1})$, meaning that the forecast of today's value are based upon the past information, as happen

with standard autoregressive models. Under standard assumptions, the forecast is simply given by its expected value $E[y_t|y_{t-1}]$. What about the variance? Engle (1982) noted that the variance of the simple one-period forecast is given by $V[y_t|y_{t-1}]$, meaning that the conditional forecast variance depends upon past information and may therefore be a random variable as well!

However, the main limitation of the classical autoregressive processes is that the variance of y_t is just given by the White Noise variance σ_ϵ^2. A model that accommodates the dependence between the conditional variance and past realization of the time series is given by:

$$y_t = \epsilon_t y_{t-1}$$

which variance is given by $\sigma_\epsilon^2 y_{t-1}^2$. However, for this model the unconditional variance is either zero or infinity, so a preferable alternative is given by:

$$\begin{aligned} y_t &= \epsilon_t h_t^{1/2}, \\ h_t &= \alpha_0 + \alpha_1 y_{t-1}^2, \end{aligned}$$

that has $\sigma_\epsilon^2 = 1$. This is the simple Autoregressive Conditional Heteroscedasticity (ARCH) model of order one, that can be generalized in order to include $P(p = 1, \ldots, P)$ lags of y_t. A generalization that includes also $Q(q = 1, \ldots, Q)$ moving average components has been introduced by Bollerslev (1986) (Generalized ARCH or GARCH). Note that the squared values of y_t, included in the *variance equation* h_t, is used as predictor. Another simple explanation can be found in the fact that the variance of a random variable Y can be written as $V(Y) = E[Y^2] - E[Y]^2$, but for a zero-mean process it holds that $V(Y) = E[Y^2]$. Hence, it is natural to model the variance of zero-mean processes as linear combination of previous squared realizations. Moreover, it is important to highlight that, if we consider Y_t as the demeaned returns' process, nowadays both squared and absolute returns are used as proxy of volatility in financial markets (Forsberg and Ghysels, 2007). This happen because absolute returns show the same empirical properties, in terms of persistency of the processes, of squared returns.

B.2 Volatility clustering

As briefly stated before, although applied to macroeconomic time series, the ARCH and GARCH models[1] were quickly found to be relevant for the conditional volatility of financial returns. This happen because conditional heteroskedasticity models are well suited in dealing with the volatility clustering phenomenon, that is common for returns of any kind of financial asset. Volatility clustering refers to the evidence that a certain degree of auto-correlation structure is present in volatility dynamics, such that high volatility periods are followed by high volatility periods and low volatility is followed by low volatility.

Actually, volatility clustering is common not only for financial markets but also for macroeconomic time series. Indeed, the first evidence of volatility clustering has been documented by McNees in 1974. The author, in *Forecasting Record for the 1970's*, wrote that "large and small (forecasting) errors tend to cluster together in contiguous time periods". Then, Bollerslev (1987) popularized the use of GARCH for modeling stock returns' volatility clustering. Since then, hundred of papers that extended or used the simple GARCH process of Bollerslev (1986) have been published. Engle (1982) proposed a Lagrange Multiplier (LM) test for the presence of ARCH effects and Engle (1982) demonstrated the presence of volatility clustering in UK inflation time series. The LM test can be easily summarized as follows. Let consider the univariate time series:

$$y_t = \mu_t + \varepsilon_t,$$

where μ_t is the conditional mean of the process, and ε_t is an uncorrelated innovation process with mean zero. Suppose the innovations are generated by:

$$\varepsilon_t = h_t^{1/2} z_t,$$

where z_t is an independent and identically distributed process with mean 0 and variance 1. Let $\mathcal{F}_t$ denote the history of the process available at time t. The conditional

[1] As well as their later extensions, that are not included here because of brevity. A very good reference for an interesting reader is given by the Handbook of Volatility modeling (Bauwens, Hafner, and Laurent, 2012).

variance of y_t is:

$$\text{Var}\left(y_t \mid \mathcal{F}_{t-1}\right) = \text{Var}\left(\varepsilon_t \mid \mathcal{F}_{t-1}\right) = E\left(\varepsilon_t^2 \mid \mathcal{F}_{t-1}\right) = h_t^2$$

Thus, conditional heteroscedasticity in the variance process is equivalent to autocorrelation in the squared innovation process. Define the residual series:

$$e_t = y_t - \hat{\mu}_t$$

If all autocorrelation in the original series, y_t, is accounted for in the conditional mean model, then the residuals are uncorrelated with mean zero. However, the residuals can still be serially dependent. The alternative hypothesis for Engle's ARCH test is autocorrelation in the squared residuals, given by the regression

$$H_a : e_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + \ldots + \alpha_m e_{t-m}^2 + u_t,$$

where u_t is a white noise error process. The null hypothesis is $H_0 : \alpha_0 = \alpha_1 = \ldots = \alpha_m = 0$. Alternatively, it is possible to check for serial dependence (ARCH effects) in a residual series by conducting a Ljung-Box Q-test on some m lags of the squared residual series.

To see what does volatility clustering means in practice, let consider the daily S&P500 Index time series in the time period 1/1/1990-1/1/2021. In the following Fig. B.1 are reported the log-prices, de-meaned returns, squared and absolute de-meaned returns that are used as proxies of volatility.

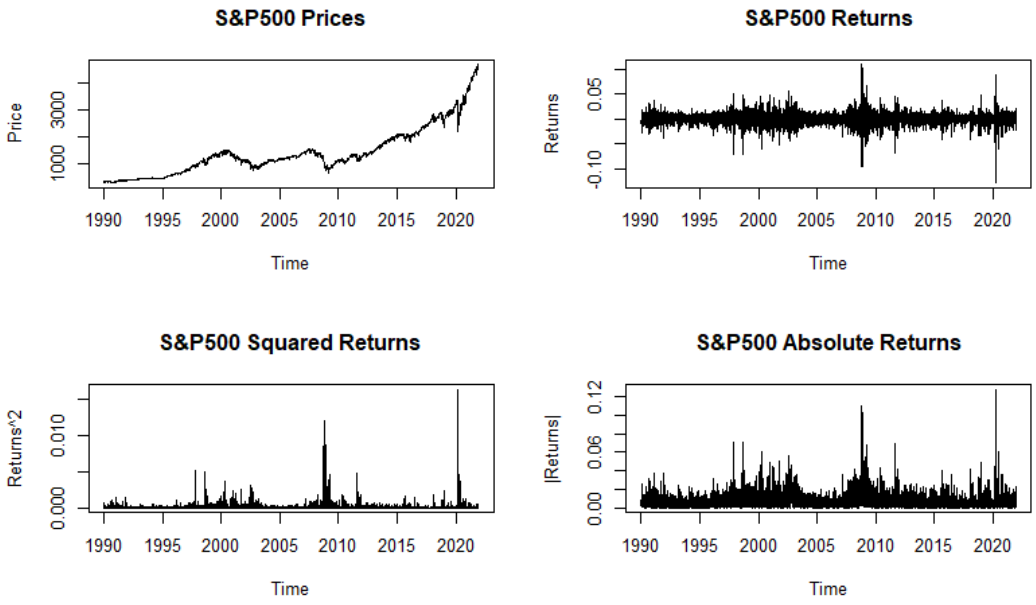
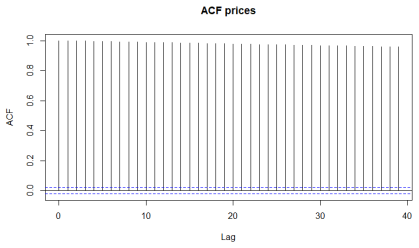
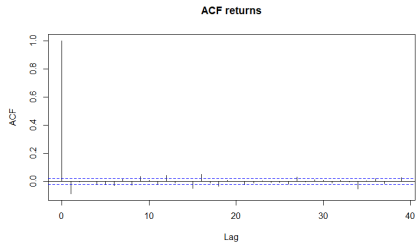


FIGURE B.1: S&P 500 Index: prices and returns

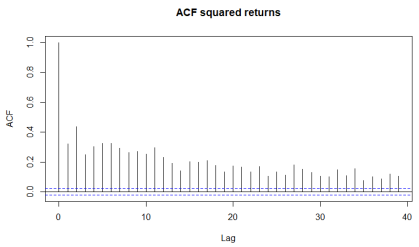
From Fig. B.1 it is evident that volatility clustering phenomenon in both squared and absolute returns. Indeed, it seems that high (low) volatility is followed by high (low) volatility. In other words, the volatility dynamics show an auto-correlation structure. As further argument, Fig. B.2 shows the ACFs for prices and the simple, squared and absolute returns.



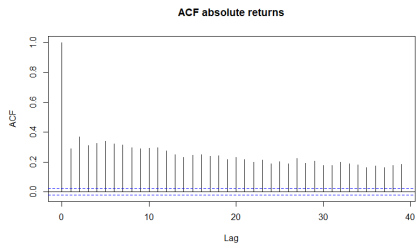
(A) S&P500 prices' ACF



(B) Simple returns' ACF



(C) Squared returns' ACF



(D) Absolute returns' ACF

FIGURE B.2: Auto-correlation functions at different lags for prices and simple, squared and absolute S&P500 returns

From Fig. B.2 we observe a very strong ACF structure for the index prices and a

much weaker for the returns. This is a stylized fact of financial returns (Cont, 2001) since prices are integrated time series. Nevertheless, the most important message from Fig. B.2 is that both the proxy of volatility (i.e. squared and absolute returns) show a strong auto-correlation structure, much greather than simple returns, thus confirming the hypothesis of volatility clustering.

But why does volatility change over time? Schwert (1989) documented the existence of a link between the conditional volatility of macroeconomic variables with the one of stock returns. Moreover, confirming the findings of Christie (1982), he also shown that financial leverage partly explains the phenomenon.

Despite the contribution of Schwert (1989) is old, it remains dateless. Just to mention an example, recent advances in GARCH modeling proposed by Engle and Rangel (2008) and Rangel and Engle (2012) (spline-GARCH models) and Engle, Ghysels, and Sohn (2013) (GARCH-MIDAS), shown that volatility and correlations can be decomposed into short and long-run components and that macroeconomic factors are particularly useful in explaining these quantities in the long-run. Clearly, these studies take inspiration from the Schwert (1989) findings. Overall, it is reasonable to agree that nowadays we have a good understanding about what motivates the dynamic nature of volatility and correlation.

B.3 Modelling conditional variance matrices

If volatility is a time-varying quantity, it is natural to expect that correlation is dynamic as well. Modelling conditional correlations can be useful in many context of finance, such as the portfolio selection, where it is interesting modelling the entire covariance structure. The ARCH and GARCH models, usually be used for modelling volatility of single time series, can be considered as generating processes of the correlation between two time series. Let us consider a simple $N \times N$ covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{2,1} & \cdots & \sigma_{N,1} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{N,2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,N} & \sigma_{2,N} & \cdots & \sigma_N^2 \end{pmatrix}$$

Assuming that both variances and covariances are time-varying, the following dynamic covariance should be considered:

$$\Sigma_t = \begin{pmatrix} \sigma^2_{1,t} & \sigma_{2,1,t} & \cdots & \sigma_{N,1,t} \\ \sigma_{1,2,t} & \sigma^2_{2,t} & \cdots & \sigma_{N,2,t} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,N,t} & \sigma_{2,N,t} & \cdots & \sigma^2_{N,t} \end{pmatrix}$$

An accurate and parsimonious modelling of both variances and covariances that accounts for heteroskedasticity can be achieved by multivariate extension of the GARCH process. Engle, Granger, and Kraft (1984) was the first considering a bivariate ARCH model. In particular, they applied to the forecast errors of two competing models of US inflation, so that their conditional covariance matrix adapts over time. The first financial application is due to Bollerslev, Engle, and Wooldridge (1988), that developed a multivariate GARCH (MGARCH) to model conditional moments rather than unconditional in the CAPM.

The main problem of the MGARCH is that it has too many parameters to be useful for modeling more than two asset returns jointly. Indeed, it is known that practitioners face an asset universe N that is large. Hence, later literature tried to design models that can be estimated for larger dimensions. Important milestones are the Constant Conditional Correlation (CCC) of Bollerslev et al. (1990) and the BEKK of Engle and Kroner (1995).

The CCC model introduced the following decomposition:

$$\Sigma_t = D_t \Gamma D_t$$

with D_t a diagonal matrix with volatilties (i.e. $D_t = H_t^{1/2}$ with H be the matrix containing the variances) and Γ a constant matrix of correlations. Obviously, considering static correlations is counter-intuitive. Indeed, the quantities within Γ should be time varying as D_t varies through time.

The CCC was followed 12 years later by the Dynamic Conditional Correlation model (DCC) of Engle (2002), that is nowadays considered as the benchmark approach for modelling conditional correlations. The DCC model, explained in more details

within the main text of Chapter 2, simply considers the decomposition:

$$\Sigma_t = D_t \Gamma_t D_t$$

with a time-varying correlations Γ_t. Before the introduction of the DCC, a very popular model for covariance matrix was given by the BEKK (Engle and Kroner, 1995):

$$\Sigma_t = CC' + \sum_{i=1}^{p} A_i \left(y_{t-i} y'_{t-i} \right) A'_i + \sum_{j=1}^{q} B_j \Sigma_{t-j} B'_j$$

where $\{A_i\}_{i=1}^{p}$ and $\{B_j\}_{j=1}^{q}$ are non-negative and symmetric matrices. The sandwich products are used to ensure the positive semi definiteness property of the covariance without imposing further constraint. However, despite its merit, the BEKK remains rather complex to handle and computationally more challenging in large dimension than the DCC. This is the reason why DCC is so popular nowadays. Clearly, from the work of Engle (2002) there have been many developments. Discussing or using these models is outside the scope of this appendix and of the whole Chapter 2.

B.4 Forecasting volatility and covariances

Forecasting volatility is a very difficult task because it is a latent concept. Indeed, as already discussed, econometricians use some proxies such as squared or absolute returns as input variables of volatility models.

Adopting this idea, the easiest way to capture volatility clustering is by letting tomorrow's variance be the simple average of the most recent m squared observations:

$$\hat{\sigma}_t^2 = \frac{1}{m} \sum_{\tau=1}^{m} y_{t-\tau}^2 = \sum_{\tau=1}^{m} \frac{1}{m} y_{t-\tau}^2$$

This representation is exactly equivalent of a rolling-window variance estimator for zero-mean processes such as the demeaned returns. This is called a *rolling window variance forecasting model*. However, the fact that the model assigns equal weights on the past observations often yields unwarranted and hard to justify results. This is especially true by considering the sample autocorrelation function of the (absolute) squared returns, that suggest a gradual decline in the effect of past returns on today's

variance (see Fig. B.2). An interesting model that takes this fact into account in forecasting volatility is the JP Morgan's RiskMetrics® that can be defined as follows:

$$\hat{\sigma}_t^2 = (1-\lambda)\sum_{\tau=1}^{\infty}\lambda^{\tau-1}y_{t-\tau}^2 \quad \lambda \in (0,1)$$

clearly, here the weights λ on past squared returns declines exponentially. Because of the mathematical structure, the RiskMetrics® model is also called exponential variance smoother. Exponential smoothers have a long tradition in forecasting literature. Actually, it can be proved that the RiskMetrics® can be written as follows:

$$\hat{\sigma}_t^2 = (1-\lambda)y_{t-1}^2 + \lambda\sigma_{t-1}^2$$

meaning that forecasts of the variance are obtained as a weighted average of static variance and squared return, with weights λ and $1-\lambda$, respectively. A very good property of RiskMetrics® is that it only contains one unknown parameter to estimate λ. Actually, JP Morgan found that the estimates of λ were quite similar across different assets, and therefore suggested to simply set it equal for every asset with daily frequency. In particular they suggest to set $= 0.94$, so no estimation is necessary at all.

RiskMetrics® can be also applied for predicting covariances rather than only variances. The covariance matrix of the multivariate Riskmetrics model is defined as:

$$\mathbf{H}_t = (1-\lambda)\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}_{t-1}' + \lambda\mathbf{H}_{t-1}$$

where $0 < \lambda < 1$ is a scalar, which according to Riskmetrics ® equals to 0.94 for daily data and 0.97 for monthly and quarterly data. Since multivariate Riskmetrics® model is guaranteed to be positive definite and does not require the estimation of any parameters of $\mathbf{H}_t$, it is easy to be used in practice. However, the assumption of imposing the same dynamics on every component in a multivariate ARCH model is difficult to justify.

Therefore, nowadays both univariate and multivariate GARCH-type processes are the most commonly used for predicting volatility in financial markets. This because, RiskMetrics® can be seen as a special case of GARCH with $\omega = 0$ and $\alpha = 1-\beta$,

but with many shortcomings. First of all, RiskMetrics® ignores the fact that a long-run variance exists. Second, while a GARCH with $\alpha + \beta < 1$ is a stationary process, RiskMetrics® is not. For the same property, RiskMetrics® assumes that any shock to current variance is destined to persist forever. Viceversa, the GARCH processes more realistically assume that the variance will revert to its average value.

This is the reason why GARCH is commonly employed not only by the industry practitioners, but also among academics. Hansen and Lunde (2005) represents one of the most famous and cited article about the comparison of statistical models for volatility forecasting. Interestingly, the authors compared only GARCH-type processes for predicting volatility (see. Fig. B.3), confirming the idea that also among scholars the GARCH are the most useful to this aim.

Table I. Specifications for the conditional variance

ARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2$
GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
IGARCH	$\sigma_t^2 = \omega + \varepsilon_{t-1}^2 + \sum_{i=2}^{q} \alpha_i (\varepsilon_{t-i}^2 - \varepsilon_{t-1}^2) + \sum_{j=1}^{p} \beta_j (\sigma_{t-j}^2 - \varepsilon_{t-1}^2)$
Taylor/Schwert:	$\sigma_t = \omega + \sum_{i=1}^{q} \alpha_i \lvert\varepsilon_{t-i}\rvert + \sum_{j=1}^{p} \beta_j \sigma_{t-j}$
A-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} [\alpha_i \varepsilon_{t-i}^2 + \gamma_i \varepsilon_{t-i}] + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
NA-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i (\varepsilon_{t-i} + \gamma_i \sigma_{t-i})^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
V-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i (e_{t-i} + \gamma_i)^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
Thr.-GARCH:	$\sigma_t = \omega + \sum_{i=1}^{q} \alpha_i [(1-\gamma_i)\varepsilon_{t-i}^{+} - (1+\gamma_i)\varepsilon_{t-i}^{-}] + \sum_{j=1}^{p} \beta_j \sigma_{t-j}$
GJR-GARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} [\alpha_i + \gamma_i I_{(\varepsilon_{t-i}>0)}] \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
log-GARCH:	$\log(\sigma_t) = \omega + \sum_{i=1}^{q} \alpha_i \lvert e_{t-i}\rvert + \sum_{j=1}^{p} \beta_j \log(\sigma_{t-j})$
EGARCH:	$\log(\sigma_t^2) = \omega + \sum_{t=1}^{q} [\alpha_i e_{t-i} + \gamma_i (\lvert e_{t-i}\rvert - E\lvert e_{t-i}\rvert)] + \sum_{j=1}^{p} \beta_j \log(\sigma_{t-j}^2)$
NGARCH:[a]	$\sigma_t^\delta = \omega + \sum_{i=1}^{q} \alpha_i \lvert\varepsilon_{t-i}\rvert^\delta + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^\delta$
A-PARCH:	$\sigma^\delta = \omega + \sum_{i=1}^{q} \alpha_i [\lvert\varepsilon_{t-i}\rvert - \gamma_i \varepsilon_{t-i}]^\delta + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^\delta$
GQ-ARCH:	$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i} + \sum_{i=1}^{p} \alpha_{ii} \varepsilon_{t-i}^2 + \sum_{i<j}^{p} \alpha_{ij} \varepsilon_{t-i} \varepsilon_{t-j} + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$
H-GARCH:	$\sigma_t^\delta = \omega + \sum_{i=1}^{q} \alpha_i \delta \sigma_{t-i}^\delta [\lvert e_t - \kappa\rvert - \tau(e_t - \kappa)]^\nu + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^\delta$
Aug-GARCH:[b]	$\sigma_t^2 = \begin{cases} \lvert\delta\phi_t - \delta + 1\rvert^{1/\delta} & \text{if } \delta \neq 0 \\ \exp(\phi_t - 1) & \text{if } \delta = 0 \end{cases}$
	$\phi_t = \omega + \sum_{i=1}^{q} [\alpha_{1i} \lvert\varepsilon_{t-i} - \kappa\rvert^\nu + \alpha_{2i} \max(0, \kappa - \varepsilon_{t-i})^\nu] \phi_{t-j}$
	$+ \sum_{i=1}^{q} [\alpha_{3i} f(\lvert\varepsilon_{t-i} - \kappa\rvert, \nu) + \alpha_{4i} f(\max(0, \kappa - \varepsilon_{t-i}), \nu)] \phi_{t-j}$
	$+ \sum_{j=1}^{p} \beta_j \phi_{t-j}^2$

FIGURE B.3: List of GARCH-type process studied in Hansen and Lunde (2005)

An authoritative and vast overview about forecasting models for volatility is given in Bauwens, Hafner, and Laurent (2012). In general, forecasting models can be classified into four main groups (Poon and Granger, 2003):

- *Historical (HISVOL)*: random walk, historical averages (of squared or absolute returns), moving averages, exponential smoothing (such as RiskMetrics®);
- *GARCH*: ARCH and GARCH processes;
- *Implied (ISD)*: option-based volatility forecasts;
- *Stochastic (SV)*: stochastic volatility models.

According to a survey conducted by Poon and Granger (2003), the overall ranking suggests that Implied volatility models provides the best forecasting with Historical and GARCH almost similarly good. The following Fig. B.4 summarizes the comparisons shown in Poon and Granger (2003).

	Number of Studies	Studies Percentage
HISVOL > GARCH	22	56%
GARCH > HISVOL	17	44%
HISVOL > ISD	8	24%
ISD > HISVOL	26	76%
GARCH > ISD	1	6
ISD > GARCH	17	94
SV > HISVOL	3	
SV > GARCH	3	
GARCH > SV	1	
ISD > SV	1	

FIGURE B.4: Summary of findings shown in Poon and Granger (2003)

The performances of stochastic volatility models were (and it is still nowadays partially true) not very well understood. Indeed, these models are more difficult to estimate and computationally more challenging to be used by not expert users. Moreover, implied volatility forecasts lies on option data that are more difficult to handle and are commonly based on mathematical models (such as the Black & Scholes one) that do not account for many stylized facts of financial returns. On the other side, historical volatility and GARCH models are instead much more simple to understand and computationally easy.

However, we cannot think as the Poon and Granger (2003) findings as conclusive and general. First of all, the contribute of Poon and Granger (2003) is quite old and does not consider returns after the 2008 financial crisis, more advanced GARCH models and much other important aspects. Second, we should have in mind that the results change among different studies also on the basis of how the forecasting accuracy is evaluated. Therefore, even if authoritative, it is clear that a renewed literature review can provide much more accurate results than Poon and Granger (2003). With this respect, the handbook of Bauwens, Hafner, and Laurent (2012) only contains a deep discussion about the statistical models used for forecasting volatility and correlation but not a comparison across the methods. Nevertheless, also the handbook deserves much more space to GARCH models rather than more simple alternatives, documenting the superiority of these class of models with respect of those belonging to the HISVOL one.

On the side of covariance matrices' prediction, two extensive review are given in Caporin and McAleer (2014) and Trucíos et al. (2019). However, these are not the only studies that provided comparisons of multivariate voaltiltiy models. For example Engle and Colacito (2006) showed the superiority of BEKK and DCC-type models over constant correlation models. Moreover Laurent, Rombouts, and Violante (2012), considering the Model Confidence Set (MCS) of Hansen, Lunde, and Nason (2011), showed that over turbulent periods the GO-GARCH and the DCC belonged to the set of superior models while, on the contrary, during calm periods constant conditional correlation model was the best one.

Caporin and McAleer (2014) developed a very deep comparison of alternative forecasting models in terms of predictive accuracy, computed by considering both realized covariances and cross-products as proxies. The authors showed that naive approaches such as the EWMA and RiskMetrics® generally underperform compared with the dynamic models like DCC. Moreover, they also found that during periods of high volatility, most models provided statistically equivalent forecast, even if some preference were given to DCC-type and GoO-GARCH models. However, the same authors concluded assessing that "*..the main message from the empirical analysis ist hat there is no optimal model. The best model must be chosen with respect to a sample period and by using selection criteria that match the purpose of the analysis...*".

More recently, Trucíos et al. (2019) provided another extensive study in terms of forecasting covariance matrices. Differently from Caporin and McAleer (2014), that

compared the models' predictive accuracy, Trucíos et al. (2019) compared the performances of minimum variance portfolios constructed with covariance forecasts. Therefore, the goal of Trucíos et al. (2019) is closer in the spirit to the Chapter 2 of the thesis. However, the authors considered daily data with monthly portfolio rebalancing, hence with estimation window much higher than the number of stocks (i.e. $N = 174$). Therefore, they do not study a real large-dimensional setting. Moreover, the authors did not compare the different models in utility terms and considered only minimum variance portfolios. However, their results, considering the full sample, can summarized in the following Fig. B.5.

Table 1. Annualised performance measures: AV, SD, IR, SR and TO stand for the average, standard deviation, information ratio, Sortino's ratio and turnover of the out-of-sample MVP returns. AV^{net}_{20bp} and AV^{net}_{50bp} stand for the average out-of-sample MVP return net of transaction costs considering 20 and 50 basis-points, respectively. Period January 2004 to November 2017. The shaded cells denote the top five for each criterion. Weights are rebalanced on a daily basis considering short-selling constraints.

	AV	SD	IR	SR	TO	AV^{net}_{20bp}	AV^{net}_{50bp}
1/N	8.302	20.058 (47)	0.414	0.570	-	-	-
CCC	7.706	11.839 (12)	0.651	0.890	0.297	7.509	7.279
CCC LS	7.004	11.881 (14)	0.590	0.807	0.307	6.815	6.578
CCC NLS	7.876	11.932 (17)	0.660	0.905	0.277	7.685	7.470
LS CCC	7.506	11.816 (11)	0.635	0.868	0.302	7.311	7.078
NLS CCC	7.345	11.809 (10)	0.622	0.848	0.298	7.153	6.923
LS CCC LS	6.628	11.918 (16)	0.556	0.759	0.305	6.439	6.205
NLS CCC NLS	7.522	11.910 (15)	0.632	0.865	0.303	7.327	7.091
DCC	7.737	11.613 (2)	0.666	0.908	0.308	7.532	7.296
DCC LS	6.941	11.689 (5)	0.594	0.810	0.314	6.749	6.508
DCC NLS	7.711	11.695 (6)	0.659	0.905	0.285	7.513	7.292
LS DCC	7.707	11.613 (1)	0.664	0.904	0.308	7.502	7.266
NLS DCC	7.629	11.616 (3)	0.657	0.894	0.307	7.424	7.188
LS DCC LS	6.907	11.688 (4)	0.591	0.806	0.314	6.715	6.474
NLS DCC NLS	7.645	11.699 (7)	0.653	0.896	0.283	7.447	7.227
RM2006	8.649	11.809 (9)	0.732	0.995	0.271	8.446	8.234
RM2006 LS	8.746	11.724 (8)	0.746	1.017	0.282	8.564	8.343
RM2006 NLS	8.734	11.865 (13)	0.736	1.011	0.268	8.537	8.327
RM1994	8.502	12.220 (22)	0.696	0.947	0.283	8.289	8.069
RM1994 LS	8.391	12.012 (18)	0.699	0.953	0.277	8.196	7.979
RM1994 NLS	8.763	12.151 (19)	0.721	0.990	0.225	8.581	8.405
DECO	5.980	12.258 (25)	0.488	0.660	0.297	5.797	5.568
DECO NLS	6.103	12.485 (41)	0.489	0.669	0.360	5.884	5.604
LS DECO	5.980	12.257 (24)	0.488	0.660	0.297	5.797	5.568
NLS DECO	5.981	12.257 (23)	0.488	0.660	0.297	5.798	5.569
NLS DECO NLS	6.103	12.485 (42)	0.489	0.669	0.360	5.884	5.604
OGARCH	8.363	12.341 (27)	0.678	0.936	0.095	8.271	8.196
OGARCH LS	7.052	12.544 (43)	0.562	0.773	0.103	6.974	6.893
OGARCH NLS	8.126	12.154 (20)	0.669	0.928	0.072	8.052	7.996
LS OGARCH	7.951	12.477 (39)	0.637	0.877	0.095	7.860	7.786
NLS OGARCH	8.365	12.341 (27)	0.678	0.936	0.095	8.273	8.198
LS OGARCH LS	6.880	12.710 (44)	0.541	0.743	0.101	6.802	6.723
NLS OGARCH NLS	8.126	12.154 (20)	0.669	0.928	0.072	8.051	7.996
GPVC	7.825	12.467 (38)	0.628	0.861	0.132	7.700	7.598
GPVC LS	7.438	12.274 (26)	0.606	0.834	0.106	7.341	7.259
GPVC NLS	6.727	12.369 (31)	0.544	0.749	0.113	6.621	6.533
LS GPVC	7.994	12.452 (36)	0.642	0.891	0.117	7.872	7.781
NLS GPVC	7.672	12.433 (33)	0.617	0.845	0.130	7.547	7.447
LS GPVC LS	7.470	12.429 (32)	0.601	0.826	0.161	7.359	7.238
NLS GPVC NLS	6.725	12.365 (30)	0.544	0.749	0.113	6.619	6.533
RPVC	9.657	12.785 (45)	0.755	1.047	0.222	9.479	9.310
RPCV LS	7.989	12.439 (34)	0.642	0.889	0.180	7.861	7.724
RPVC NLS	9.186	12.485 (40)	0.736	1.026	0.184	9.035	8.893
LS RPVC	8.543	12.347 (29)	0.692	0.953	0.201	8.387	8.235
NLS RPVC	8.064	13.142 (46)	0.614	0.850	0.191	7.904	7.755
LS RPCV LS	7.493	12.439 (35)	0.602	0.828	0.167	7.378	7.252
NLS RPVC NLS	7.658	12.460 (37)	0.615	0.850	0.172	7.509	7.376

FIGURE B.5: Summary of findings shown in Trucíos et al. (2019)

Fig. B.5 shows the very huge comparison in terms of models considered by the authors. In few words, we can argue that DCC and GO-GARCH-type (called O-GARCH, i.e. Oroghonal GARCH in Fig. B.5) of models overperform the others. Indeed, while DCC-type of models allow the construction of portfolios with the lowest variance (see also Engle and Colacito, 2006), the GO-GARCH models allow to reach the lowest turnover possible. Interestingly, also the RiskMetrics® provide good performances in terms of Sharpe ratio, due to the high average returns.

However, even if also in this case there is no evidence in favor of a single model, summarizing the findings of all the previous studies two clear winners arise: the DCC and the GO-GARCH, based on factor structure. For this reason these two models are considered in the applications of the Chapter 2 instead of all the others.

Bibliography

Ahmed, Nesreen K et al. (2010). "An empirical comparison of machine learning models for time series forecasting". In: *Econometric Reviews* 29.5-6, pp. 594–621.

Almadi, Himanshu, David E Rapach, and Anil Suri (2014). "Return predictability and dynamic asset allocation: How often should investors rebalance?" In: *The Journal of Portfolio Management* 40.4, pp. 16–27.

Ang, Andrew and Geert Bekaert (2007). "Stock return predictability: Is it there?" In: *The Review of Financial Studies* 20.3, pp. 651–707.

Arbelaitz, Olatz et al. (2013). "An extensive comparative study of cluster validity indices". In: *Pattern Recognition* 46.1, pp. 243–256.

Arora, Preeti, Shipra Varshney, et al. (2016). "Analysis of k-means and k-medoids algorithm for big data". In: *Procedia Computer Science* 78, pp. 507–512.

Ayebo, Abraham and Tomasz J Kozubowski (2003). "An asymmetric generalization of Gaussian and Laplace laws". In: *Journal of Probability and Statistical Science* 1.2, pp. 187–210.

Bai, Jushan (2003). "Inferential theory for factor models of large dimensions". In: *Econometrica* 71.1, pp. 135–171.

Bai, Jushan and Serena Ng (2013). "Principal components estimation and identification of static factors". In: *Journal of Econometrics* 176.1, pp. 18–29.

Bai, Jushan and Shuzhong Shi (2011). "Estimating High Dimensional Covariance Matrices and its Applications". In: *Annals of Economics and Finance* 12.2, pp. 199–215.

Barry, Christopher B (1974). "Portfolio analysis under uncertain means, variances, and covariances". In: *The Journal of Finance* 29.2, pp. 515–522.

Bastos, João A and Jorge Caiado (2021). "On the classification of financial data with domain agnostic features". In: *International Journal of Approximate Reasoning* 138, pp. 1–11.

Batool, Fatima and Christian Hennig (2021). "Clustering with the average silhouette width". In: *Computational Statistics & Data Analysis* 158, p. 107190.

Bauwens, Luc, Christian M Hafner, and Sébastien Laurent (2012). *Handbook of volatility models and their applications*. Vol. 3. John Wiley & Sons.

Bianchi, Daniele, Matthias Büchner, and Andrea Tamoni (2021). "Bond risk premiums with machine learning". In: *The Review of Financial Studies* 34.2, pp. 1046–1089.

Bollerslev, Tim (1986). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of econometrics* 31.3, pp. 307–327.

— (1987). "A conditionally heteroskedastic time series model for speculative prices and rates of return". In: *The review of economics and statistics*, pp. 542–547.

Bollerslev, Tim, Robert F Engle, and Jeffrey M Wooldridge (1988). "A capital asset pricing model with time-varying covariances". In: *Journal of political Economy* 96.1, pp. 116–131.

Bollerslev, Tim et al. (1990). "Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model". In: *Review of Economics and statistics* 72.3, pp. 498–505.

Boswijk, H Peter and Roy Van der Weide (2011). "Method of moments estimation of go-garch models". In: *Journal of Econometrics* 163.1, pp. 118–126.

Breen, William, Lawrence R Glosten, and Ravi Jagannathan (1989). "Economic significance of predictable variations in stock index returns". In: *The Journal of finance* 44.5, pp. 1177–1189.

Bucci, Andrea (2020a). "Cholesky–ANN models for predicting multivariate realized volatility". In: *Journal of Forecasting* 39.6, pp. 865–876.

— (2020b). "Realized volatility forecasting with neural networks". In: *Journal of Financial Econometrics* 18.3, pp. 502–531.

Caiado, Jorge and Nuno Crato (2010). "Identifying common dynamic features in stock returns". In: *Quantitative Finance* 10.7, pp. 797–807.

Caiado, Jorge, Nuno Crato, and Daniel Peña (2006). "A periodogram-based metric for time series classification". In: *Computational Statistics & Data Analysis* 50.10, pp. 2668–2684.

Caivano, Michele and Andrew Harvey (2014). "Time-series models with an EGB2 conditional distribution". In: *Journal of Time Series Analysis* 35.6, pp. 558–571.

Campbell, John Y and Samuel B Thompson (2008). "Predicting excess stock returns out of sample: Can anything beat the historical average?" In: *The Review of Financial Studies* 21.4, pp. 1509–1531.

Caporin, Massimiliano and Michael McAleer (2014). "Robust ranking of multivariate GARCH models by problem dimension". In: *Computational Statistics & Data Analysis* 76, pp. 172–185.

Cerqueti, Roy, Massimiliano Giacalone, and Raffaele Mattera (2020). "Skewed non-Gaussian GARCH models for cryptocurrencies volatility modelling". In: *Information Sciences* 527, pp. 1–26.

— (2021). "Model-based fuzzy time series clustering of conditional higher moments". In: *International Journal of Approximate Reasoning* 134, pp. 34–52.

Cerqueti, Roy, Massimiliano Giacalone, and Demetrio Panarello (2019). "A Generalized Error Distribution Copula-based method for portfolios risk assessment". In: *Physica A: Statistical Mechanics and its Applications* 524, pp. 687–695.

Chamberlain, Gary (1983). "Funds, factors, and diversification in arbitrage pricing models". In: *Econometrica: Journal of the Econometric Society*, pp. 1305–1323.

Chatfield, Chris (1996). "Model uncertainty and forecast accuracy". In: *Journal of Forecasting* 15.7, pp. 495–508.

Chen, Nai-Fu, Richard Roll, and Stephen A Ross (1986). "Economic forces and the stock market". In: *Journal of Business*, pp. 383–403.

Chopra, Vijay K and William T Ziemba (1993). "The effect of errors in means, variances, and covariances on optimal portfolio choice". In: *Journal of Portfolio Management* 19.2, p. 6.

Christie, Andrew A (1982). "The stochastic behavior of common stock variances: Value, leverage and interest rate effects". In: *Journal of financial Economics* 10.4, pp. 407–432.

Christoffersen, Peter et al. (2010). "Volatility components, affine restrictions, and nonnormal innovations". In: *Journal of Business & Economic Statistics* 28.4, pp. 483–502.

Cont, Rama (2001). "Empirical properties of asset returns: stylized facts and statistical issues". In: *Quantitative Finance* 1, pp. 223–236.

Coppi, Renato and Pierpaolo D'Urso (2006). "Fuzzy unsupervised classification of multivariate time trajectories with the Shannon entropy regularization". In: *Computational statistics & data analysis* 50.6, pp. 1452–1477.

Cox, David R et al. (1981). "Statistical analysis of time series: Some recent developments [with discussion and reply]". In: *Scandinavian Journal of Statistics*, pp. 93–115.

Creal, Drew, Siem Jan Koopman, and André Lucas (2013). "Generalized autoregressive score models with applications". In: *Journal of Applied Econometrics* 28.5, pp. 777–795.

Curto, José Dias, José Castro Pinto, and Gonçalo Nuno Tavares (2009). "Modeling stock markets' volatility using GARCH models with Normal, Student's t and stable Paretian distributions". In: *Statistical Papers* 50.2, p. 311.

Davies, N and P Newbold (1980). "Forecasting with misspecified models". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 29.1, pp. 87–92.

De Miguel, Victor, Lorenzo Garlappi, and Raman Uppal (2007). "Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?" In: *The Review of Financial Studies* 22.5, pp. 1915–1953.

De Nard, Gianluca, Olivier Ledoit, and Michael Wolf (2019). "Factor Models for Portfolio Selection in Large Dimensions: The Good, the Better and the Ugly". In: *Journal of Financial Econometrics*.

DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal (2009). "Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?" In: *The review of Financial studies* 22.5, pp. 1915–1953.

Dey, Dipak K (1987). "Improved estimation of a multinormal precision matrix". In: *Statistics & probability letters* 6.2, pp. 125–128.

Díaz, Sonia Pértega and José A Vilar (2010). "Comparing several parametric and nonparametric approaches to time series clustering: a simulation study". In: *Journal of classification* 27.3, pp. 333–362.

Donaldson, R Glen and Mark Kamstra (1997a). "An artificial neural network-GARCH model for international stock return volatility". In: *Journal of Empirical Finance* 4.1, pp. 17–46.

— (1997b). "An artificial neural network-GARCH model for international stock return volatility". In: *Journal of Empirical Finance* 4.1, pp. 17–46.

Duan, Jin-Chuan (1999). "Conditionally fat-tailed distributions and the volatility smile in options". In: *Rotman School of Management, University of Toronto, Working Paper*.

D'urso, Pierpaolo (2004). "Fuzzy C-means clustering models for multivariate time-varying data: different approaches". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 12.03, pp. 287–326.

D'URSO, PIERPAOLO (2004). "Fuzzy C-means clustering models for multivariate time-varying data: different approaches". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 12.03, pp. 287–326.

D'Urso, Pierpaolo, Livia De Giovanni, and Riccardo Massari (2016). "GARCH-based robust clustering of time series". In: *Fuzzy Sets and Systems* 305, pp. 1–28.

— (2018). "Robust fuzzy clustering of multivariate time trajectories". In: *International Journal of Approximate Reasoning* 99, pp. 12–38.

D'Urso, Pierpaolo, Elizabeth A Maharaj, and Andrés M Alonso (2017). "Fuzzy clustering of time series using extremes". In: *Fuzzy Sets and Systems* 318, pp. 56–79.

D'Urso, Pierpaolo and Elizabeth Ann Maharaj (2012). "Wavelets-based clustering of multivariate time series". In: *Fuzzy Sets and Systems* 193, pp. 33–61.

D'Urso, Pierpaolo and Elizabeth Ann Maharaj (2009). "Autocorrelation-based fuzzy clustering of time series". In: *Fuzzy Sets and Systems* 160.24, pp. 3565–3589.

D'Urso, Pierpaolo et al. (2013). "Clustering of financial time series". In: *Physica A: Statistical Mechanics and its Applications* 392.9, pp. 2114–2129.

D'Urso, Pierpaolo et al. (2019). "Fuzzy clustering with spatial–temporal information". In: *Spatial Statistics* 30, pp. 71–102.

D'Urso, Pierpaolo et al. (2020). "Cepstral-based clustering of financial time series". In: *Expert Systems with Applications* 161, p. 113705.

Efron, Bradley and Carl Morris (1973). "Stein's estimation rule and its competitors—an empirical Bayes approach". In: *Journal of the American Statistical Association* 68.341, pp. 117–130.

Eliasy, Ashkan and Justyna Przychodzen (2020). "The role of AI in capital structure to enhance corporate funding strategies". In: *Array* 6, p. 100017.

Elman, Jeffrey L (1990). "Finding structure in time". In: *Cognitive science* 14.2, pp. 179–211.

Engle, Robert (2002). "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models". In: *Journal of Business & Economic Statistics* 20.3, pp. 339–350.

Engle, Robert and Riccardo Colacito (2006). "Testing and valuing dynamic correlations for asset allocation". In: *Journal of Business & Economic Statistics* 24.2, pp. 238–253.

Engle, Robert F (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". In: *Econometrica: Journal of the Econometric Society*, pp. 987–1007.

Engle, Robert F, Eric Ghysels, and Bumjean Sohn (2013). "Stock market volatility and macroeconomic fundamentals". In: *Review of Economics and Statistics* 95.3, pp. 776–797.

Engle, Robert F, Clive WJ Granger, and Dennis Kraft (1984). "Combining competing forecasts of inflation using a bivariate ARCH model". In: *Journal of economic dynamics and control* 8.2, pp. 151–165.

Engle, Robert F and Kenneth F Kroner (1995). "Multivariate simultaneous generalized ARCH". In: *Econometric theory* 11.1, pp. 122–150.

Engle, Robert F, Olivier Ledoit, and Michael Wolf (2019). "Large dynamic covariance matrices". In: *Journal of Business & Economic Statistics* 37.2, pp. 363–375.

Engle, Robert F and Jose Gonzalo Rangel (2008). "The spline-GARCH model for low-frequency volatility and its global macroeconomic causes". In: *The review of financial studies* 21.3, pp. 1187–1222.

Enke, David and Suraphan Thawornwong (2005). "The use of data mining and neural networks for forecasting stock market returns". In: *Expert Systems with applications* 29.4, pp. 927–940.

Fama, Eugene F and Kenneth R French (1993). "Common risk factors in the returns on stocks and bonds". In: *Journal of Financial Economics* 33.1, pp. 3–56.

— (2015). "A five-factor asset pricing model". In: *Journal of financial economics* 116.1, pp. 1–22.

Fama, Eugene F and G William Schwert (1977). "Asset returns and inflation". In: *Journal of financial economics* 5.2, pp. 115–146.

Fan, Jianqing, Yingying Fan, and Jinchi Lv (2008). "High dimensional covariance matrix estimation using a factor model". In: *Journal of Econometrics* 147.1, pp. 186–197.

Fan, Jianqing, Yuan Liao, and Martina Mincheva (2013). "Large covariance estimation by thresholding principal orthogonal complements". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75.4, pp. 603–680.

Fernandez, Carmen, Jacek Osiewalski, and Mark FJ Steel (1995). "Modeling and inference with v-spherical distributions". In: *Journal of the American Statistical Association* 90.432, pp. 1331–1340.

Fernández, Carmen and Mark FJ Steel (1998). "On Bayesian modeling of fat tails and skewness". In: *Journal of the american statistical association* 93.441, pp. 359–371.

Fleming, Jeff, Chris Kirby, and Barbara Ostdiek (2001). "The economic value of volatility timing". In: *The Journal of Finance* 56.1, pp. 329–352.

— (2003). "The economic value of volatility timing using "realized" volatility". In: *Journal of Financial Economics* 67.3, pp. 473–509.

Forsberg, Lars and Eric Ghysels (2007). "Why do absolute returns predict volatility so well?" In: *Journal of Financial Econometrics* 5.1, pp. 31–67.

Foster, DP and DB Nelson (1996). "Continuous record asymptotics for rolling sample variance estimators". In: *Econometrica* 64.1, pp. 139–174.

Frost, Peter A and James E Savarino (1986a). "An empirical Bayes approach to efficient portfolio selection". In: *Journal of Financial and Quantitative Analysis* 21.3, pp. 293–305.

— (1986b). "Portfolio size and estimation risk". In: *The Journal of Portfolio Management* 12.4, pp. 60–64.

Fulcher, Ben D and Nick S Jones (2014). "Highly comparative feature-based time-series classification". In: *IEEE Transactions on Knowledge and Data Engineering* 26.12, pp. 3026–3037.

Geweke, John and Gianni Amisano (2012). "Prediction with misspecified models". In: *American Economic Review* 102.3, pp. 482–86.

Götze, Tobias, Marc Gürtler, and Eileen Witowski (2020). "Improving CAT bond pricing models via machine learning". In: *Journal of Asset Management* 21.5, pp. 428–446.

Grant, Dwight (1978). "Market timing and portfolio management". In: *The Journal of Finance* 33.4, pp. 1119–1131.

Green, Jeremiah, John RM Hand, and X Frank Zhang (2013). "The supraview of return predictive signals". In: *Review of Accounting Studies* 18.3, pp. 692–730.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020). "Empirical asset pricing via machine learning". In: *The Review of Financial Studies* 33.5, pp. 2223–2273.

Haff, LR (1977). "Minimax estimators for a multinormal precision matrix". In: *Journal of Multivariate Analysis* 7.3, pp. 374–385.

— (1979). "Estimation of the inverse covariance matrix: Random mixtures of the inverse Wishart matrix and the identity". In: *The Annals of Statistics*, pp. 1264–1276.

Hafner, Christian M and Olga Reznikova (2012). "On the estimation of dynamic conditional correlation models". In: *Computational Statistics & Data Analysis* 56.11, pp. 3533–3545.

Hansen, Peter R and Asger Lunde (2005). "A forecast comparison of volatility models: does anything beat a GARCH (1, 1)?" In: *Journal of applied econometrics* 20.7, pp. 873–889.

Hansen, Peter R, Asger Lunde, and James M Nason (2011). "The model confidence set". In: *Econometrica* 79.2, pp. 453–497.

Harvey, Andrew and Rutger-Jan Lange (2017). "Volatility modeling with a generalized t distribution". In: *Journal of Time Series Analysis* 38.2, pp. 175–190.

Harvey, Andrew and Genaro Sucarrat (2014). "EGARCH models with fat tails, skewness and leverage". In: *Computational Statistics & Data Analysis* 76, pp. 320–338.

Harvey, Andrew C (2013). *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*. Vol. 52. Cambridge University Press.

Harvey, Campbell R and Guofu Zhou (1993). "International asset pricing with alternative distributional specifications". In: *Journal of Empirical Finance* 1.1, pp. 107–131.

Hewamalage, Hansika, Christoph Bergmeir, and Kasun Bandara (2021). "Recurrent neural networks for time series forecasting: Current status and future directions". In: *International Journal of Forecasting* 37.1, pp. 388–427.

Hill, Tim, Marcus O'Connor, and William Remus (1996). "Neural network models for time series forecasts". In: *Management science* 42.7, pp. 1082–1092.

Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks". In: *Neural networks* 4.2, pp. 251–257.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5, pp. 359–366.

Hsieh, David A (1989). "Modeling heteroscedasticity in daily foreign-exchange rates". In: *Journal of Business & Economic Statistics* 7.3, pp. 307–317.

Hsu, Ming-Wei et al. (2016). "Bridging the divide in financial market forecasting: machine learners vs. financial economists". In: *Expert Systems with Applications* 61, pp. 215–234.

Hubert, Lawrence and Phipps Arabie (1985). "Comparing partitions". In: *Journal of classification* 2.1, pp. 193–218.

Iorio, Carmela et al. (2016). "Parsimonious time series clustering using p-splines". In: *Expert Systems with Applications* 52, pp. 26–38.

— (2018). "A P-spline based clustering approach for portfolio selection". In: *Expert Systems with Applications* 95, pp. 88–103.

Jagannathan, Ravi and Tongshu Ma (2003). "Risk reduction in large portfolios: Why imposing the wrong constraints helps". In: *The Journal of Finance* 58.4, pp. 1651–1683.

Jarque, Carlos M and Anil K Bera (1987). "A test for normality of observations and regression residuals". In: *International Statistical Review/Revue Internationale de Statistique*, pp. 163–172.

Jing, Liping, Michael K Ng, and Joshua Zhexue Huang (2007). "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data". In: *IEEE Transactions on knowledge and data engineering* 19.8, pp. 1026–1041.

Jobson, J Dave and Bob M Korkie (1981). "Performance hypothesis testing with the Sharpe and Treynor measures". In: *Journal of Finance*, pp. 889–908.

Jobson, J David and Bob Korkie (1980). "Estimation for Markowitz efficient portfolios". In: *Journal of the American Statistical Association* 75.371, pp. 544–554.

Jordan, Michael I (1997). "Serial order: A parallel distributed processing approach". In: *Advances in psychology*. Vol. 121. Elsevier, pp. 471–495.

Jorion, Philippe (1986). "Bayes-Stein estimation for portfolio analysis". In: *Journal of Financial and Quantitative Analysis* 21.3, pp. 279–292.

— (1991). "Bayesian and CAPM estimators of the means: Implications for portfolio selection". In: *Journal of Banking & Finance* 15.3, pp. 717–727.

Kaastra, Iebeling and Milton Boyd (1996). "Designing a neural network for forecasting financial and economic time series". In: *Neurocomputing* 10.3, pp. 215–236.

Kan, Raymond and Guofu Zhou (2007). "Optimal portfolio choice with parameter uncertainty". In: *Journal of Financial and Quantitative Analysis* 42.3, pp. 621–656.

Kim, Ha Young and Chang Hyun Won (2018). "Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models". In: *Expert Systems with Applications* 103, pp. 25–37.

Komunjer, Ivana (2007). "Asymmetric power distribution: Theory and applications to risk measurement". In: *Journal of applied econometrics* 22.5, pp. 891–921.

Kong, Aiguo et al. (2011). "Predicting market components out of sample: asset allocation implications". In: *The Journal of Portfolio Management* 37.4, pp. 29–41.

Kourtis, Apostolos, George Dotsis, and Raphael N Markellos (2012). "Parameter uncertainty in portfolio selection: Shrinking the inverse covariance matrix". In: *Journal of Banking & Finance* 36.9, pp. 2522–2531.

Kristjanpoller, Werner, Anton Fadic, and Marcel C Minutolo (2014). "Volatility forecast using hybrid neural network models". In: *Expert Systems with Applications* 41.5, pp. 2437–2442.

Kristjanpoller, Werner and Marcel C Minutolo (2015). "Gold price volatility: A forecasting approach using the Artificial Neural Network–GARCH model". In: *Expert systems with applications* 42.20, pp. 7245–7251.

— (2016). "Forecasting volatility of oil price using an artificial neural network-GARCH model". In: *Expert Systems with Applications* 65, pp. 233–241.

Kubokawa, Tatsuya and Muni S Srivastava (2008). "Estimation of the precision matrix of a singular Wishart distribution and its application in high-dimensional data". In: *Journal of multivariate Analysis* 99.9, pp. 1906–1928.

Lahmiri, Salim (2016). "Clustering of Casablanca stock market based on hurst exponent estimates". In: *Physica A: Statistical Mechanics and its Applications* 456, pp. 310–318.

Laurent, Sébastien, Jeroen VK Rombouts, and Francesco Violante (2012). "On the forecasting accuracy of multivariate GARCH models". In: *Journal of Applied Econometrics* 27.6, pp. 934–955.

Ledoit, Oliver and Michael Wolf (2008). "Robust performance hypothesis testing with the Sharpe ratio". In: *Journal of Empirical Finance* 15.5, pp. 850–859.

Ledoit, Olivier and Michael Wolf (2003). "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection". In: *Journal of Empirical Finance* 10.5, pp. 603–621.

— (2004a). "A well-conditioned estimator for large-dimensional covariance matrices". In: *Journal of multivariate analysis* 88.2, pp. 365–411.

— (2004b). "Honey, I shrunk the sample covariance matrix". In: *The Journal of Portfolio Management* 30.4, pp. 110–119.

— (2012). "Nonlinear shrinkage estimation of large-dimensional covariance matrices". In: *The Annals of Statistics* 40.2, pp. 1024–1060.

— (2017). "Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks". In: *The Review of Financial Studies* 30.12, pp. 4349–4388.

León, Ángel, Gonzalo Rubio, and Gregorio Serna (2005). "Autoregresive conditional volatility, skewness and kurtosis". In: *The Quarterly Review of Economics and Finance* 45.4-5, pp. 599–618.

Liao, T Warren (2005). "Clustering of time series data—a survey". In: *Pattern recognition* 38.11, pp. 1857–1874.

Liu, Nana et al. (2021). "An Agglomerative Hierarchical Clustering Algorithm for Linear Ordinal Rankings". In: *Information Sciences*.

Lo, Andrew W (1991). "Long-term memory in stock market prices". In: *Econometrica: Journal of the Econometric Society*, pp. 1279–1313.

Maharaj, Elizabeth Ann and Pierpaolo D'Urso (2011). "Fuzzy clustering of time series in the frequency domain". In: *Information Sciences* 181.7, pp. 1187–1211.

Mantegna, Rosario N (1999). "Hierarchical structure in financial markets". In: *The European Physical Journal B-Condensed Matter and Complex Systems* 11.1, pp. 193–197.

Markowitz, Harry (1952). "Portfolio selection". In: *The Journal of Finance* 7.1, pp. 77–91.

Marquering, Wessel and Marno Verbeek (2004). "The economic value of predicting stock index returns and volatility". In: *Journal of Financial and Quantitative Analysis*, pp. 407–429.

Marx, DL and RR Hocking (1977). "Moments of certain functions of elements in the inverse Wishart matrix". In: *Paper presented at the meeting of the American Statistical Association*.

Mattera, Raffaele, Massimiliano Giacalone, and Karina Gibert (2021). "Distribution-Based Entropy Weighting Clustering of Skewed and Heavy Tailed Time Series". In: *Symmetry* 13.6, p. 959.

Merton, Robert C (1973). "An intertemporal capital asset pricing model". In: *Econometrica: Journal of the Econometric Society*, pp. 867–887.

— (1980). "On estimating the expected return on the market: An exploratory investigation". In: *Journal of financial economics* 8.4, pp. 323–361.

Michaud, Richard O (1989). "The Markowitz optimization enigma: Is 'optimized'optimal?" In: *Financial Analysts Journal* 45.1, pp. 31–42.

Mudholkar, Govind S and Alan D Hutson (2000). "The epsilon–skew–normal distribution for analyzing near-normal data". In: *Journal of statistical planning and inference* 83.2, pp. 291–309.

Muirhead, Robb J (1982). "Aspects of multivariate statistical analysis." In: *JOHN WILEY & SONS*.

Mullainathan, Sendhil and Jann Spiess (2017). "Machine learning: an applied econometric approach". In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

Nanda, SR, Biswajit Mahanty, and MK Tiwari (2010). "Clustering Indian stock market data for portfolio management". In: *Expert Systems with Applications* 37.12, pp. 8793–8798.

Nanopoulos, Alex, Rob Alcock, and Yannis Manolopoulos (2001). "Feature-based classification of time-series data". In: *International Journal of Computer Research* 10.3, pp. 49–61.

Nelson, Daniel B (1991). "Conditional heteroskedasticity in asset returns: A new approach". In: *Econometrica: Journal of the Econometric Society*, pp. 347–370.

Okhrin, Yarema and Wolfgang Schmid (2006). "Distributional properties of portfolio weights". In: *Journal of econometrics* 134.1, pp. 235–256.

Otranto, Edoardo (2008). "Clustering heteroskedastic time series by model-based procedures". In: *Computational Statistics & Data Analysis* 52.10, pp. 4685–4698.

Pakel, Cavit et al. (2017). "Fitting vast dimensional time-varying covariance models". In:

Park, Hae-Sang and Chi-Hyuck Jun (2009). "A simple and fast algorithm for K-medoids clustering". In: *Expert systems with applications* 36.2, pp. 3336–3341.

Patton, Andrew J (2020). "Comparing possibly misspecified forecasts". In: *Journal of Business & Economic Statistics* 38.4, pp. 796–809.

Pesaran, M Hashem and Allan Timmermann (1995). "Predictability of stock returns: Robustness and economic significance". In: *The Journal of Finance* 50.4, pp. 1201–1228.

Piccolo, Domenico (1990). "A distance measure for classifying ARIMA models". In: *Journal of Time Series Analysis* 11.2, pp. 153–164.

Poon, Ser-Huang and Clive WJ Granger (2003). "Forecasting volatility in financial markets: A review". In: *Journal of economic literature* 41.2, pp. 478–539.

Raffinot, Thomas (2017). "Hierarchical clustering-based asset allocation". In: *The Journal of Portfolio Management* 44.2, pp. 89–99.

Rand, William M (1971). "Objective criteria for the evaluation of clustering methods". In: *Journal of the American Statistical association* 66.336, pp. 846–850.

Rangel, José Gonzalo and Robert F Engle (2012). "The Factor–Spline–GARCH model for high and low frequency correlations". In: *Journal of Business & Economic Statistics* 30.1, pp. 109–124.

Rapach, David and Guofu Zhou (2013). "Forecasting stock returns". In: *Handbook of economic forecasting*. Vol. 2. Elsevier, pp. 328–383.

Rather, Akhter Mohiuddin, Arun Agarwal, and VN Sastry (2015). "Recurrent neural network and a hybrid model for prediction of stock returns". In: *Expert Systems with Applications* 42.6, pp. 3234–3241.

Ross, Stephen (1976). "The Arbitrage Theory of Capital Asset Pricing". In: *Journal of Economic Theory*, pp. 341–360.

Rousseeuw, Peter J (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20, pp. 53–65.

Schwert, G William (1989). "Why does stock market volatility change over time?" In: *The Journal of Finance* 44.5, pp. 1115–1153.

Sharpe, William F (1963). "A simplified model for portfolio analysis". In: *Management science* 9.2, pp. 277–293.

Stambaugh, Robert F (1999). "Predictive regressions". In: *Journal of Financial Economics* 54.3, pp. 375–421.

Stein, Charles (1956). *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution*. Tech. rep. Stanford University Stanford United States.

Sun, Ruili, Tiefeng Ma, and Shuangzhe Liu (2018). "Portfolio selection: shrinking the time-varying inverse conditional covariance matrix". In: *Statistical Papers*, pp. 1–22.

Theodossiou, Panayiotis (2015). "Skewed generalized error distribution of financial assets and option pricing". In: *Multinational Finance Journal* 19.4, pp. 223–266.

Trucíos, Carlos et al. (2019). "Covariance prediction in large portfolio allocation". In: *Econometrics* 7.2, p. 19.

Tu, Jun and Guofu Zhou (2011). "Markowitz meets Talmud: A combination of sophisticated and naive diversification strategies". In: *Journal of Financial Economics* 99.1, pp. 204–215.

Ushakov, Anton V and Igor Vasilyev (2021). "Near-optimal large-scale k-medoids clustering". In: *Information Sciences* 545, pp. 344–362.

Wang, Jian-Zhou et al. (2011). "Forecasting stock indices with back propagation neural network". In: *Expert Systems with Applications* 38.11, pp. 14346–14355.

Wang, Xiaozhe, Kate Smith, and Rob Hyndman (2006). "Characteristic-based clustering for time series data". In: *Data mining and knowledge Discovery* 13.3, pp. 335–364.

Wang, Yi-Hsien (2009). "Nonlinear neural network forecasting model for stock index option price: Hybrid GJR–GARCH approach". In: *Expert Systems with Applications* 36.1, pp. 564–570.

Weide, Roy Van der (2002). "GO-GARCH: a multivariate generalized orthogonal GARCH model". In: *Journal of Applied Econometrics* 17.5, pp. 549–564.

Wilhelmsson, Anders (2006). "GARCH forecasting performance under different distribution assumptions". In: *Journal of Forecasting* 25.8, pp. 561–578.

Xiang, Cheng, Shenqiang Q Ding, and Tong Heng Lee (2005). "Geometrical interpretation and architecture selection of MLP". In: *IEEE transactions on neural networks* 16.1, pp. 84–96.

Xie, Wen-Bo et al. (2020). "Hierarchical clustering supported by reciprocal nearest neighbors". In: *Information Sciences* 527, pp. 279–292.

Zhang, Guoqiang, B Eddy Patuwo, and Michael Y Hu (1998). "Forecasting with artificial neural networks:: The state of the art". In: *International journal of forecasting* 14.1, pp. 35–62.

Zhu, Dongming and Victoria Zinde-Walsh (2009). "Properties and estimation of asymmetric exponential power distribution". In: *Journal of econometrics* 148.1, pp. 86–99.